EXPLORING THE GENETIC BASIS OF SEED COAT AND NUTRITIONAL QUALITY TRAITS IN COMMON BEAN AND MAIZE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Di Wu

August 2021

EXPLORING THE GENETIC BASIS OF SEED COAT AND NUTRITIONAL QUALITY TRAITS IN COMMON BEAN AND MAIZE

Di Wu, Ph. D.

Cornell University 2021

Common bean (*Phaseolus vulgaris* L.) and maize (*Zea mays* L.) are two crops central to indigenous America and of great global agricultural importance. However, the landraces of common bean are largely underrepresented in genebanks, and despite the importance of elements and tocochromanols to plant function and human health, there are still gaps in the understanding of the transport and accumulation of these nutrients in maize grain. Through the array of research tools offered by the field of population genomics and quantitative genetics, this dissertation works towards addressing such gaps. The genomic characterization of ~ 300 accessions of common bean from Native Seeds/SEARCH collected from southwestern US and northwestern Mexico established it as a unique and underrepresented resource that contained important genetic diversity. Five genes encoding MYB transcription factors proximal to the *C* locus were identified, which is a complex genomic region responsible for the primary control of seed coat patterns. An additional novel association for partial colored seed coats was identified on chromosome 10. Through genome-wide association studies (GWAS) with high density SNP set and the 1500-line Ames panel, I investigated the genetic basis of natural variation for the concentration of 11 elements in grain and identified a total of nine causal genes encoding metal chelator or transporter. Notably, two novel associations were reported between *rte2* and *irt1* with boron and nickel, respectively, and a potential biofortification target, *nas5*, was identified for both zinc and iron.

Similar moderate predictive abilities (0.33–0.53) were obtained for the 11 grain elemental phenotypes with Bayesian Ridge Regression (BRR) and BayesB. However, BayesB, allowing SNPs to have large effects, had a better fit to the genetic architecture of nickel, molybdenum, and copper, thus outperforming BRR by 4-10%. Finally, through GWAS, transcriptome-wide association studies (TWAS) and expression quantitative trait locus (eQTL) mapping, 13 causal genes that were mostly under strong cis-regulatory control were identified to associate with tocochromanol levels in maize grain. Four genes were pinpointed to be associated with tocochromanol concentrations in maize grain, including *vte5*, *dxs1*, *vte7*, and *samt*. Overall, this dissertation demonstrates a multidisciplinary approach to characterize a unique common bean collection and the genetic control of its seed coat pattern, and provides a comprehensive assessment of the genetic basis of nutritional qualities in maize grain.

BIOGRAPHICAL SKETCH

Di Wu was born in 1993 in Hefei, Anhui province in China as the only child of Andong Wu and Ying Zhou. Di became interested in biology and genetics at an early age, and upon high school graduation, she decided to pursue agriculture as her major. Di attended China Agricultural University in Beijing from 2010 to 2014 where she got her B.S. degree in Seed Science and Technology. As an undergraduate research assistant, she worked on rice embryo culturability traits under the direction of Dr. Yongcai Fu and Dr. Chuanqing Sun. In August 2014, Di came to Cornell University and started a M.S. program in Plant Breeding and Genetics under the supervision of Dr. Michael Gore. After a year, she decided to transfer and continue her research in the Ph.D. program. Di is passionate about understanding the genetics of nutrient accumulation in plants and hopes that her Ph.D. research would help alleviate hidden hunger problems around the world.

This dissertation is dedicated to my grandparents, Ben Zhengde and Zhou Yuelan, who never saw this adventure of mine. You will always be in my heart.

这篇论文谨献给我的敬爱的阿公贲正德和阿婆周月兰，

感谢你们对我的爱和鼓励。你们会永远活在我的心里。

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my parents, who have always supported and believed in me throughout my journey. I will never be where I am now without your love and encouragement. I would also like to thank my cousins, Teng Ben and Fang Ben, you have been my role models growing up. And I want to acknowledge all my family members, who provided me great supports along the way.

I want to say a special thank you to my advisor Dr. Michael Gore. I am forever in your debt for the guidance and patience you have given me. You see potential in me and always push me to go with a higher standard. I am thankful to my committee members, Dr. Olena Vatamaniuk and Dr. Li Li, for your intellectual contributions to my development as a scientist.

I would also like to thank all the Gore lab members. As a graduate student, I can ask no more than what I have got here, friendship, teamwork, skills, and communications. I would like to thank Dan Ilut, I have learned a lot working alongside you. I am grateful for Drs. Ryokei Tanaka and Xiaowei Li, it has been a joy working on the same project with you.

I wish to thank National Science Foundation, HarvestPlus, Native Seeds/SEARCH, and Plant Breeding and Genetics Section for supporting my graduate studies and research projects. All the faculty, staffs and graduate students in the Section have been very supportive and helpful for my research.

Last but not least, I am lucky to have wonderful roommates, Yichang Liu, Yifan Yang, and Jing Ning. Special thanks to Meng Lin, who is my lab mate, roommate, and best friend. I feel grateful for all the friends I got to know throughout my stay here. You have taught me the true meaning of friendship and made my Cornell life much more colorful and memorable.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

Agriculture in indigenous America emerged around 7,000 years ago, and with the domestications of maize (*Zea mays* L.), bean (*Phaseolus vulgaris* L.), and squash (*Cucurbita pepo* L.), the Three Sisters garden was created in which the three crops were planted together (Landon, 2008; Terry *et al.*, 2020). This form of sustainable polyculture had spread throughout Mesoamerica by 3,500 years ago, providing a balance among the three crops, both in terms of the agronomic practices and the nutritions they would provide for human subsistence. Agronomically, these three crops can benefit each other when planted together. Maize plants can provide tall stalks for bean plants to climb upon and compete well against weeds, bean functions as a nitrogen fixing legume providing nitrogen required for maize plant growth, and squash plants can grow rapidly along the ground with large leaves to retain soil moisture and suppress weeds for maize and bean (Hart, 2008). These three crops also provide complementary values to diets when consumed together. Maize is high in calories mainly through carbohydrates, but is relatively low in fiber and protein and missing two essential amino acids (lysine and tryptophan). Bean provides a rich source of protein, but lacks the essential amino acid methionine. As a result, the dietary combination of bean and maize can provide a complete essential amino acid profile for human health (Kaplan, 1965; Terry *et al.*, 2020). In addition, squash is high in vitamins and minerals and together with beans and maize, the Three Sisters can provide a more complete array of nutrients for human consumption.

The Three Sisters, particularly bean and maize, are of great agricultural importance.

Common bean is the most widely consumed legume species of the genus *Phaseolus*, especially in Latin American and African countries, and accounts for more than 90% of the total production of the five domesticated *Phaseolus* species (*P. coccineus, P. acutifolius, P. lunatus, P. dumosus,* and *P. vulgaris*) (FAO, 2019). Similarly, maize is currently the crop species with the highest production worldwide (FAO, 2019), serving as food crop for humans, as well as animal feeds and for ethanol fuel production, among other commodities.

This widespread success of these two crops can be attributed to breeding, along with the mechanization of production and commercial use of fertilizers and pesticides. However, loss of genetic diversity in populations of these species can be catastrophic for breeding, resulting in increases of disease and pest susceptibility and lower responsiveness to changing climates (Plucknett *et al.,* 1983). Genebanks can offer a great solution for counteracting such limitations, as the conserved landraces and wild relatives can hold important genes and alleles that were otherwise lost or overlooked during domestication and improvement processes (Plucknett *et al.,* 1983; Plucknett & Smith, 2014). In addition to breeding, such germplasm collections are also valuable resources for genetic studies, providing genetic diversity and mapping resolution that will help with the identification of favorable alleles and detection of causal variants for traits of interest (Diepenbrock & Gore, 2015; Plucknett *et al.,* 1983). Apart from genebanks, a large amount of genetic diversity still exists within landraces in the hands of small-holder farmers and also in wild relatives or progenitors that are yet to be explored and incorporated into breeding. For example, wild common bean possesses a wide range of adaptive traits (Acosta-Gallegos *et al.,* 2007) and improved nutritional value (Guzmán-Maldonado *et al.,* 2000), but are underrepresented in the genebanks of many countries

(Dennis *et al.,* 2014; Escribano *et al.,* 1998; Espinosa-Alonso *et al.,* 2006; Leitão *et al.,* 2017; Okii *et al.,* 2014).

To utilize this extant genetic diversity in an effort to understand the genetic basis of phenotypic variation, quantitative genetics offer a variety of tools for such efforts. In particular, genome-wide association studies (GWAS) proves to be one approach that exploits historical recombination events within a population to find associations between genetic markers and traits of interest (Korte & Farlow, 2013; Nordborg & Weigel, 2008). The patterns of non-random association between alleles at two or more loci, or linkage disequilibrium (LD), is the result of recombination and other evolutionary processes. The size of LD blocks, which are comprised of genetic markers in strong LD with each other, impact the resolution of GWAS and are generally smaller for out-crossing species such as maize relative to self-pollinated species such as common bean (Remington *et al.,* 2001). As such, natural diversity panels with large sample sizes, high allelic diversity, and rapid LD decay are ideal for GWAS (Korte & Farlow, 2013). In addition to GWAS, genome-wide transcription profiles from tissues and developmental time points of interest can be a complementary source of information for causal gene identification that is generally independent of LD. The integrated approach of transcriptome-wide association studies (TWAS), which tests for associations between transcript abundance and traits of interest, and GWAS, has proven to increase statistical power and prioritize candidate gene selection in maize and sorghum (Ferguson *et al.,* 2020; Kremling *et al.,* 2019; Pignon *et al.,* 2021).

In this dissertation, Chapter 2 focuses on the genomic characterization of a novel and underrepresented common bean genetic resource from Native Seeds/SEARCH collection and

establishes this collection as a unique source of traditional landraces. Additionally, GWAS was conducted to identify putative causal genes responsible for variation in seed coat phenotypes. Chapters 3 and 4 are devoted to understanding the genetic basis of nutritional qualities in maize grain, another crop of the Three Sisters. The maize Ames diversity panel consisting of ~1,500 inbred lines was used for both projects, and GWAS was conducted to investigate the associations between high-density SNP sets and the natural variation of elemental and tocochromanol (vitamin E and antioxidants) levels in grain. In Chapter 3, apart from GWAS, two whole-genome prediction models were utilized to evaluate the predictive abilities of the elemental grain phenotypes. In Chapter 4, the addition of transcriptome data from developing kernels was shown to enhance our current understanding of grain tocochromanol accumulation and regulatory control of causal genes. Furthermore, Chapter 5 describes a unique experiment that explores the involvement of chlorophyll and two previously identified chlorophyll pathway genes, *protochlorophyllide reductases* (*por1* and *por2*) (Diepenbrock *et al.,* 2017) in tocopherol biosynthesis in maize grain. Overall, this dissertation utilized population genomics and quantitative genetics tools to characterize a unique genetic resource, as well as to further our knowledge of the genetic control of the common bean seed coat pattern and nutritional qualities in the maize grain.

REFERENCES

Acosta-Gallegos, J. A., Kelly, J. D., & Gepts, P. (2007). Prebreeding in common bean and use of genetic diversity from wild germplasm. Crop Science, 47, S – 44.

Dennis, O., Phinehas, T., James, K., Annet, N., Pamela, P., Michael, U., & Paul, G. (2014). The genetic diversity and population structure of common bean (*Phaseolus vulgaris* L) germplasm in Uganda. African Journal of Biotechnology, 13(29), 2935–2949.

Diepenbrock, C. H., & Gore, M. A. (2015). Closing the divide between human nutrition and plant breeding. Crop Science, 55(4), 1437–1448.

Diepenbrock, C. H., Kandianis, C. B., Lipka, A. E., Magallanes-Lundback, M., Vaillancourt, B., Góngora-Castillo, E., Wallace, J. G., Cepela, J., Mesberg, A., Bradbury, P. J., Ilut, D. C., Mateos-Hernandez, M., Hamilton, J., Owens, B. F., Tiede, T., Buckler, E. S., Rocheford, T., Buell, C. R., Gore, M. A., & DellaPenna, D. (2017). Novel loci underlie natural variation in vitamin E levels in maize grain. Plant Cell, 29(10), 2374–2392.

Escribano, M. R., Santalla, M., Casquero, P. A., & de Ron, A. M. (1998). Patterns of genetic diversity in landraces of common bean (*Phaseolus vulgaris* L.) from Galicia. Plant Breeding, 117(1), 49–56.

Espinosa-Alonso, L. G., Lygin, A., Widholm, J. M., Valverde, M. E., & Paredes-Lopez, O. (2006). Polyphenols in wild and weedy Mexican common beans (*Phaseolus vulgaris* L.). Journal of Agricultural and Food Chemistry, 54(12), 4436–4444.

FAO. (2019). FAOSTAT. In: FAOSTAT. http://www.fao.org/faostat/en/#data/CC/visualize. Accessed Jun 14, 2021.

Ferguson, J., Fernandes, S., Monier, B., Miller, N. D., & Allan, D. (2020). Machine learning enabled phenotyping for GWAS and TWAS of WUE traits in 869 field-grown sorghum accessions. bioRxiv. https://www.biorxiv.org/content/biorxiv/early/2020/11/03/2020.11.02.365213.full.pdf

Guzmán-Maldonado, S. H., Acosta-Gallegos, J., & Paredes-López, O. (2000). Protein and mineral content of a novel collection of wild and weedy common bean (*Phaseolus vulgaris* L). Journal of the Science of Food and Agriculture, 80(13), 1874–1881.

Hart, J. P. (2008). Evolving the three sisters: The changing histories of maize, bean, and squash in New York and the greater Northeast. Current Northeast Paleoethnobotany II, 87–99.

Kaplan, L. (1965). Archeology and domestication in American Phaseolus (beans). Economic Botany, 19(4), 358–368.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. Plant Methods, 9(1), 1–9.

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, 9(9), 3023–3033.

Landon, A. J. (2008). The “how” of the three sisters: The origins of agriculture in Mesoamerica and the human niche. https://digitalcommons.unl.edu/nebanthro/40/

Leitão, S. T., Dinis, M., Veloso, M. M., Šatović, Z., & Vaz Patto, M. C. (2017). Establishing

the bases for introducing the unexplored Portuguese common bean germplasm into the breeding world. Frontiers in Plant Science, 8, 1296.

Nordborg, M., & Weigel, D. (2008). Next-generation genetics in plants. Nature, 456(7223), 720–723.

Okii, D., Tukamuhabwa, P., Odong, T., Namayanja, A., Mukabaranga, J., Paparu, P., & Gepts, P. (2014). Morphological diversity of tropical common bean germplasm. African Crop Science Journal, 22(1), 59–68.

Pignon, C. P., Fernandes, S. B., Valluru, R., Bandillo, N., Lozano, R., Buckler, E., Gore, M. A., Long, S. P., Brown, P. J., & Leakey, A. D. B. (2021). Phenotyping stomatal closure by thermal imaging for GWAS and TWAS of water use efficiency-related genes. In bioRxiv (p. 2021.05.06.442962). https://doi.org/10.1101/2021.05.06.442962

Plucknett, D. L., & Smith, N. J. H. (2014). Gene Banks and the World's Food. Princeton University Press.

Plucknett, D. L., Smith, N. J., Williams, J. T., & Anishetty, N. M. (1983). Crop germplasm conservation and developing countries. Science, 220(4593), 163–169.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M., & Buckler, E. S., 4th. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. U. S. A., 98(20), 11479–11484.

Terry, P. A., Pearson, D., & Holder, G. (2020). Sustainable agriculture: nutrition of indigenous American 3 Sisters garden compared to monoculture corn production and a cool old squash. Journal of Scientific Research and Reports, 99–108.

Chapter 2 Genomic characterization of the native seeds/search common bean (*Phaseolus vulgaris* L.) collection and its seed coat patterns[1]

ABSTRACT

Common bean (*Phaseolus vulgaris* L.) is one of the most important legume crops for human consumption. The Native Seeds/SEARCH common bean collection consists of locally-adapted accessions collected from the southwestern US and northwestern Mexico. In this study, a representative panel of nearly 300 accessions from this collection was genotyped with more than 10,000 high-quality SNP markers and phenotyped for seed coat patterns. The collection consists primarily of accessions from the Mesoamerican gene pool, and they separate into three distinct subpopulations, with strong population differentiation ($F_{ST} > 0.4$) observed between them. Through a genome-wide association study with the Mesoamerican accessions, we identified several SNPs on chromosome 8 that are associated with seed coat pattern traits and reside proximal to the putative location of the *C* locus, a locus previously shown to control the pattern of the seed coat. Five *myb* transcription factors linked to these SNPs were identified as candidate causal genes for seed coat patterns controlled by the *C* locus. Furthermore, we identified a potentially novel locus on chromosome 10 that appears to control the Anasazi seed coat phenotype. Our work is the first to characterize the genetic

---

[1]Wu, D., Hought, J., Baseggio, M., Hart, J. P., Gore, M. A., & Ilut, D. C. (2019). Genomic characterization of the native seeds/search common bean (*Phaseolus vulgaris* L.) Collection and its seed coat patterns. *Genetic Resources and Crop Evolution*, *66*(7), 1469-1482.

diversity of the Native Seeds/SEARCH common bean collection, providing valuable genetic information for germplasm conservation efforts.

Abbreviations: BIC, Bayesian information criterion; FDR, False discovery rate; F_{IT}, Index of panmixia; F_{ST}, Fixation index; GBS, Genotyping-by-sequencing; GWAS, Genome-wide association study; IBS, Identity-by-state; LD, Linkage disequilibrium; MLMM, Multi-locus mixed model; NPGS, National plant germplasm system; NS/S, Native Seeds/SEARCH; PA, Proanthocyanidins; PCA, Principal component analysis; SFS, Site frequency spectrum; SNP, Single-nucleotide polymorphism.

INTRODUCTION

Common bean (*Phaseolus vulgaris* L.) serves as a nutrient-rich staple food and major source of protein in many areas of the world, and it exhibits adaptation to a wide range of environments and cropping systems (Jones 1999). Common bean production has continued to increase worldwide over the past 20 years, reaching a total production of over 26 million tons in 2016, with more than half of the common bean production taking place in Latin America and Africa (FAO 2018). With the exception of Argentina where beans are mostly produced by large modern farms, common bean production is generally the result of smallholder farming (Wortmann 1998; Broughton *et al.* 2003). Yearly consumption of common bean has also increased in Latin America over the past 20 years, with the majority of production destined for local consumption (FAO 2018). Smallholder farming systems are expected to be particularly sensitive to the effects of climate change in the coming decades (Eitzinger *et al.* 2017), and it is likely that new varieties will be needed to address such challenges.

Currently, international germplasm repositories such as the U.S. National Plant Germplasm System (NPGS) represent the primary sources of material for the development of new varieties. However, many native landraces are not necessarily well represented in such repositories. For this study, we had the opportunity to genotype for the first time accessions of common bean landraces in the Native Seeds/SEARCH (NS/S; https://www.nativeseeds.org) collection. NS/S is a nonprofit organization based in Tucson, Arizona, dedicated to the conservation of the genetic diversity of endangered traditional seeds from the southwestern US and northwestern Mexico (Burgess 1994). The NS/S seed bank houses over 1,900 locally-adapted accessions of traditional crops utilized as food, fiber, and dye from local tribes and agriculturists (van Schoonhoven 1991). Over half of the accessions preserved by NS/S are the three sisters, *i.e.* maize, bean and squash, as well as over 100 additional species of crops and crop wild relatives. The common bean collection in the NS/S seed bank is comprised of over 300 accessions, collected mostly from Arizona and New Mexico (US), and Sonora and Chihuahua (Mexico).

The collection range for these NS/S common bean accessions overlaps considerably with the center of origin for common bean (Bitocchi *et al.* 2012b). Two major gene pools generally corresponding to distinct geographical regions have been described by several research groups using biochemical and molecular markers (Singh *et al.* 1991; Blair *et al.* 2006; Bassett 2007; Kwak and Gepts 2009; Schmutz *et al.* 2014): the Mesoamerican gene pool, representing varieties grown in Colombia, Central America, and Mexico, and the Andean gene pool, representing varieties grown in Peru, Bolivia, and Argentina (Bitocchi *et al.* 2012b). The wild progenitors of the two gene pools were derived from a common ancestral

wild population in Mesoamerica (Bitocchi *et al.* 2012b), followed by parallel domestications in Central and South America which generated the Mesoamerica and Andean gene pools respectively (Bitocchi *et al.* 2012a; Mamidi *et al.* 2013; Schmutz *et al.* 2014). Higher genetic diversity has been observed in the Mesoamerican gene pool as compared to the Andean gene pool (Kwak and Gepts 2009; Bitocchi *et al.* 2012b; Schmutz *et al.* 2014; Ariani *et al.* 2018), supporting the Mesoamerican origin of the common bean. Microsatellite data have identified five common bean groups in the Mesoamerican gene pool and four groups in the Andean gene pool (Kwak and Gepts 2009). Significant gene pool differentiation as well as racial differentiation within gene pools is observed, and the domesticated populations generally possess lower genetic diversity and higher F_{ST} compared with wild populations. While domesticated common bean accessions have been sampled across the vast majority of Mesoamerican and Andean locations, few studies have focused on the landraces and varieties cultivated and/or preserved from northern Mexico and the United States, and the NS/S common bean collection is a rich resource to bring under the lens of genetic characterization.

Being the primary agricultural product of the plant, the presentation of the seed is an important factor in consumer preference, with local, cultural, and aesthetic factors working together to form hyper-local preferences for seed colors and patterns. The distinct cultural and commercial market classes of both dry and immature podded beans have been defined based on their seed type and deployed extensively in agriculture, and the current commercial market classes of common bean in North America largely continue the gene pool, race structure, and seed-type distinctions selected by early agriculturists (van Schoonhoven 1991). Therefore, in addition to suitability to a specific local environment, the uptake of candidate varieties for

cultivar replacement is likely dependent on consumer preference as well. As reviewed in Bassett (2007), seed coat colors and patterns are controlled by different Mendelian loci with complex epistatic interactions, with seed coat patterning phenotypes primarily controlled by *C* and *T*. The *C* locus appears to be a complex of closely linked interacting genes controlling sharply defined patternings, characterized by a contrast of darker pattern color and lighter background color (Prakken 1974). Recessive alleles at the *T* locus are required for partial coloring, which is characterized by a white background and a colored zone on the seed coat. Other modifying genes, such as those at the *J* (=*L*) and *Bip* (=*Ana*) loci, have epistatic interactions with *C* and *T* for specific patterns. To our knowledge, the causal genes underlying these loci have not been conclusively identified.

The overall goal of our study was to complete a genetic characterization of the NS/S common bean collection and to identify candidate genomic regions likely involved in seed coat appearance. Our specific objectives were to: (1) investigate the levels and patterns of genetic diversity in the NS/S common bean collection, and (2) identify candidate genes responsible for the major differentiation features of seed coat patterns.

MATERIALS AND METHODS

Plant materials and growth conditions

The common bean panel analyzed here was assembled with 324 representative accessions from Native Seeds/SEARCH (NS/S) (Supplemental Table S2.1), all of which are directly available to the public from NS/S. The collection is comprised of accessions that capture a wide range of variation for traits such as seed size, shape, color, and plant growth habit. In

June 2015, two seeds of each accession were planted in a single 5.7L (1.5 gallons) pot that contained Cornell soil mix (Boodley and Sheldrake 1972) and thinned to one seedling per pot after seedling establishment. In addition, we planted seeds of accession G19833 to generate eight pots (biological replicates) with a single plant each. Accession G19833 in the NPGS collection was originally sourced from the International Center for Tropical Agriculture (CIAT) in Cali, Colombia, and these seeds were derived from a subset of G19833 seeds that had been previously purified via inbreeding for the purpose of reference genome sequencing (Schmutz *et al.* 2014) and are therefore expected to be more genetically similar to each other than a similar size sample from the G19833 accession in the NPGS collection. Plants were grown under natural light in a greenhouse at the Guterman Bioclimatic Lab (Ithaca, NY, USA). All plants were fertilized once a day with irrigation using 21-5-20 All Purpose LX nutrient solution (JR Peters Inc, Allentown, PA, USA), and a trellis was set up before flowering to provide a stable growing environment for climbing beans.

In order to provide evolutionary and genetic context for this common bean panel, we selected an additional 46 accessions from the National Plant Germplasm System (NPGS) maintained by the Western Regional Plant Introduction System at Pullman, Washington, with 23 accessions representing each of the Mesoamerican and Andean gene pools (Supplemental Table S2.1). These accessions were selected to include representatives of the specific trait complexes within each gene pool, as well as representatives of most of the commercial seed types important in North America (pinto, great northern, black, navy, small red, pink, white kidney, red kidney, yellow, cranberry, and snap beans). The race of each accession was inferred using the previously established seed type and plant morphology descriptions (Singh

et al. 1991). One seed each of these accessions was planted in 15-cell trays on greenhouse benches for leaf tissue harvest. Only genotype data were collected for these plants.

Seed phenotyping

Mature pods were harvested from each of the 324 plants, with each plant representing a distinct accession derived from the NS/S collection. For each plant, threshed seeds were bulked and dried at 37 ºC for 3 days. On average, a subsample of eight seeds randomly selected from each bulk was visually scored for one of the following seven seed coat pattern categories: striped, netted, mottled, a combined pattern which included two of the three aforementioned patterns, Anasazi, Little Appaloosa, and solid (single color, no pattern) (Supplemental Table S2.2). Representative images of these patterns are presented in Supplemental Figures S2.1 and S2.2. Each seed subsample was homogeneous for a single pattern category. These pattern classifications were further reduced to three classes that are expected to be controlled by distinct genetic loci, following the treatment of Bassett (2007): solid color, partial color (containing the "Anasazi" pattern), and patterned (containing striped, netted, mottled, combined and Little Appaloosa patterns). In addition, for each of the three pattern classes, we created presence/absence (binary) classes that were used to investigate the association of genetic loci with specific pattern classes.

Tissue collection, genotyping, and initial SNP filtering

We harvested the first emerged trifoliolate leaves of each plant from 3-week-old seedlings and stored them at -80°C for at least 1 h before lyophilizing for 72 h. For each plant, including NS/S accessions, G19833 and USDA NPGS accessions, 20 mg lyophilized leaf tissue samples

were ground individually using a SPEX SamplePrep Geno/Grinder (SPEX SamplePrep, Metuchen, NJ, USA) and total genomic DNA was isolated with the Qiagen DNeasy Plant Mini Kit (Qiagen Inc., Valencia, CA, USA).

Sequencing libraries for 378 samples were generated using the standard genotyping-by-sequencing (GBS) protocol (Elshire *et al.* 2011) with restriction enzyme ApeKI at the Cornell Biotechnology Resource Center Genomics Facility (Cornell University, Ithaca, NY, USA). The libraries were sequenced on an Illumina sequencer HiSeq 2500, with 96 samples multiplexed in each lane (four lanes total). Sequence data from this study have been deposited at the National Center of Biotechnology Information Sequence Read Archive under BioProject accession number PRJNA542132.

The single-end 100 bp reads were aligned to the non-masked reference genome *P.vulgaris* v2.0 (downloaded from Phytozome website https://phytozome.jgi.doe.gov/pz/portal.html, accessed 2018-05-08) with Bowtie (version 2; Langmead and Salzberg 2012). Through the implementation of the TASSEL (version 5.2.4; Bradbury *et al.* 2007) GBS analysis pipeline v2 (Glaubitz *et al.* 2014), we identified 181,137 putative SNPs, which were further filtered using VCFtools (version 0.1.13; Danecek *et al.* 2011) and custom scripts as described below.

First, using the initial 181,137 putative SNPs, we set all genotype calls with a genotype quality less than 30 to missing, and removed clustered SNPs that had physical positions of 1 bp apart from each other. Subsequently, we evaluated allele counts and distributions in the 378 samples, removing SNPs with more than two alleles and biallelic SNPs with minor alleles observed in only one sample (singletons and doubletons).

Next, for each sample, we inspected the read alignment depth at each genotype call and calculated an upper threshold of read depth defined as the lowest value expected for the 5% high tail of a Poisson distribution with λ equal to the mean observed read depth across SNPs. Furthermore, for heterozygous genotype calls, we calculated an allele balance score by dividing the lowest allele read depth by the total read depth. We filtered out genotype calls with a read depth smaller than 2 or larger than the sample's upper threshold, as well as heterozygous genotype calls with an allele balance score less than 0.3, by setting the genotype call as missing. With the remaining 127,409 SNPs, we removed three samples (15DW257, 15DW338, 15DW388) with a genotype call rate of less than 0.6, resulting in a total of 375 samples retained for further filtering.

SNP data subsetting and subpopulation assignment

Starting with the 375 samples and 127,409 SNPs retained after the initial filtering, we applied further quality control filters to generate four different sets of SNPs (SNP sets 0, I, II, and III) targeted toward different analyses (Supplemental Figure S2.3, Supplemental Table S2.3). SNP Set 0 was used to identify unintended sample replicates that are likely to result from independent sampling of widely distributed bean varieties with distinct local names, and replace each group of such replicates with a representative sample for that group. SNP Set I was used to classify NS/S accessions as from the Andean or Mesoamerican gene pool. SNP sets II and III were limited to Mesoamerican NS/S accessions with unintended sample replicates removed, and were used for the genome-wide association study (GWAS) and population structure analysis, respectively. They differed primarily by whether missing data were imputed (SNP Set II) or not (SNP Set III).

For SNP filtering purposes, at each SNP over all 375 samples we calculated the index of panmixia (F_{IT}; Romay *et al.* 2013), and used Fisher's exact test to determine if the two alleles were independent in our population, calculating the level of significance ($\alpha = 0.001$) at which the null hypothesis of independence was rejected. We generated SNP Set 0 by removing SNPs matching any of the following criteria: 1) $F_{IT} < 0.9$; 2) proportion of missing genotype calls $\geqslant 0.3$; 3) minor allele frequency < 0.05; 4) Fisher's exact test $\alpha = 0.001$; or 5) a proportion of heterozygous genotype calls > 0.1. A total of 11,472 SNPs were retained after this filtering step, and they were used to calculate the centered identity-by-state (IBS) value (Endelman and Jannink 2012) using TASSEL. The IBS threshold for identifying unintended sample replicates was set to 99.94%, the minimum IBS value among all pairs of the eight known biological replicates of G19833. Using this threshold, a total of 51 samples were assigned to 16 distinct groups such that the IBS value between any members of a group was above the threshold identified above, and the sample with the highest genotype call rate within each group was selected as a representative sample. As a result, a total of 340 accessions, which consisted of 294 NS/S accessions and 46 NPGS accessions, were retained for further filtering.

For SNP Set I, we used all 340 accessions and removed SNPs with a $F_{IT} < 0.9$ and a proportion of missing genotype calls $\geqslant 0.8$. The missing genotype calls for the remaining 22,064 SNPs were imputed using TASSEL FILLIN (Swarts *et al.* 2014), and the post-imputation SNP set was further filtered by removing SNPs that matched any of the following criteria: 1) proportion of missing genotype calls $\geqslant 0.3$; 2) minor allele frequency < 0.05; 3) Fisher's exact test $\alpha = 0.001$; or 4) a proportion of heterozygous genotype calls > 0.1. After

filtering, SNP Set I contained 340 accessions and 13,846 SNPs.

We used SNP Set I to perform a principal component analysis (PCA) in R (R Core Team 2019) using the probabilistic approach for missing value estimation implemented in the package 'pcaMethods' (version 1.60.0, Stacklies *et al.* 2007). We used fastSTRUCTURE (Raj *et al.* 2014) to determine the most likely number of subpopulations using the simple prior approach, followed by using the logistic prior approach to determine population composition for each of the 340 samples using the previously determined K value (K = 2). We used a population assignment value of $Q \geqslant 0.5$ to assign samples to each of the two subpopulations, and the labeling of the subpopulations as Andean or Mesoamerican was determined by the membership of the 46 known accessions from those gene pools obtained from the USDA NPGS (Supplemental Figure S2.4).

To generate SNP Set II, out of the 294 NS/S accessions remaining after removal of unintended sample replicates we selected only the 281 NS/S accessions assigned to the Mesoamerican gene pool by the SNP Set I analysis, removed any SNPs that were monomorphic in this subset, and applied the same SNP filtering and imputation criteria as for SNP Set I. The remaining 5,766 SNPs were used to identify Mesoamerican subpopulations and classify samples based on subpopulation membership. Specifically, we combined the results of simple prior fastSTRUCTURE, PCA (Supplemental Figure S2.5), and a phylogenetic analysis (Supplemental Figure S2.6) to select K = 3 as the number of NS/S Mesoamerican subpopulations, and used the logistic prior approach in fastSTRUCTURE to assign samples with a subpopulation-specific value of $Q \geqslant 0.8$ to the respective subpopulation. Samples with $Q < 0.8$ for all three population were labeled as admixed and

excluded from population genetics analysis. After subpopulation classification, we further filtered SNPs to remove those that have fewer than 10 genotype calls in each subpopulation, resulting in a total of 281 accessions and 5,732 SNPs for SNP Set II. For F_{ST} comparison purposes (Supplemental Table S2.4), we also assigned samples to subpopulations with a more relaxed cutoff of $Q \geqslant 0.5$ using the same procedure, but this classification was not used in any other analysis.

Finally, SNP Set III, which was used for our population genetic analysis of the Mesoamerican gene pool, was generated in a manner similar to SNP Set II. We used the same 281 Mesoamerican accessions, but did not perform imputation and applied a more stringent filtering threshold for the proportion of missing genotype call (0.2) and a more stringent percent heterozygosity criterion (5%). Filters for F_{IT} (0.9), Fisher's exact test ($\alpha = 0.001$), and minimum genotype call counts in subpopulations (10) were as before. SNP Set III contains a total of 281 accessions and 4,872 SNPs.

Population genetic analysis

The 281 accessions of Mesoamerica origin and 4,872 SNPs from SNP Set III were used for all further population genetic analysis. Nucleotide diversity (Hudson *et al.* 1992), Tajima's D (Tajima 1989), and all pairwise fixation index (F_{ST}) values among the three subpopulations were calculated and averaged on 100-kb genomic bins containing at least three variants using VCFtools. In addition to the full set of SNPs, we generated two subsets of SNPs that address ascertainment bias per pairwise F_{ST}: one subset excluded SNPs that were not polymorphic in the first subpopulation, and the other excluded SNPs that were not polymorphic in the second subpopulation. Total number of SNPs that were polymorphic in each of the three

subpopulations are 2,701, 1,994 and 1,814, respectively.

Genome-wide association and linkage disequilibrium estimation

Using SNP Set II, consisting of 281 accessions and 5,732 partially imputed SNPs, we conducted a genome-wide association study for the seed coat pattern traits. Principal components (Price *et al.* 2006) and a kinship matrix based on VanRaden's method 1 (VanRaden 2008) were calculated based on SNP Set II and included in the multi-locus mixed model (MLMM) to control for population structure and unequal relatedness. The Bayesian information criterion (BIC; Schwarz 1978) was used to determine the optimal number of principal components included as covariates in the mixed linear model for each trait. Any remaining missing genotypes were conservatively imputed as heterozygous using the Genome Association and Prediction Integrated Tool (GAPIT, version 2017.08.18; Lipka *et al.* 2012) in R with the 'middle' option. We employed MLMM to account for several loci with large effects (Segura *et al.* 2012), given that this was the expected genetic architecture for seed coat pattern traits in common bean (Bassett 2007). The MLMM uses a stepwise mixed-model regression with forward inclusion and backward elimination, while re-estimating the genetic and error variances at each step. The optimal model was selected using the extended Bayesian information criterion (eBIC; Chen and Chen 2008).

Linkage disequilibrium (LD) between all pairs of SNPs on each chromosome was estimated in TASSEL using the squared allele-frequency correlation (r^2) method of Hill and Weir (Hill and Weir 1988). SNP Set II with 5,732 SNPs was used to estimate LD and approximate physical distance of where median LD decayed to a genome-wide background level of $r^2 = 0.1$.

We used the primer sequences for sequence-tagged site (STS) markers developed by McClean *et al.* (2002) to identify the most likely genomic location for several genetic loci previously determined to be related to our seed coat traits of interest. Specifically, we used NCBI BLAST (Camacho *et al.* 2009) to align the primer sequences to the *P. vulgaris* reference genome (version 2.1) and selected likely locations for these markers on the reference genome (Supplemental Table S2.5). First, we identified locations where at least one of the forward or reverse primer sequences had a unique best (lowest E-value) match on the chromosome indicated by the genetic map associated with the marker. Second, for markers that had multiple likely matches for the second primer sequence, we selected the closest location to the unique best match identified in the first step. We used the genomic location of these primer sequences as a proxy for the likely genomic location of the corresponding STS marker.

RESULTS

Population structure and genetic diversity of the NS/S common bean collection

Principal component analysis of SNP Set I revealed two distinct subgroups along the first principal component (PC1, explaining 41.7% of the variation; Supplemental Figure S2.4), separating the Andean and Mesoamerican NPGS accessions as expected. Using fastSTRUCTURE, we identified K = 2 as the most likely number of subpopulations and partitioned the 294 NS/S accessions in SNP Set I into two distinct groups: 281 accessions clustering with NPGS accessions from the Mesoamerican gene pool, and 13 accessions clustering with NPGS accessions from the Andean gene pool. The NS/S accessions were

predominantly assigned to the Mesoamerican gene pool, which is consistent with their collection locales. Therefore, we performed the remaining analysis exclusively on the 281 NS/S accessions from the Mesoamerican gene pool, which were further divided into three subpopulations (Figure 2.1, Supplemental Table S2.1, Supplemental Figures S2.5 and S2.6). If the probability of an accession being derived from a given subpopulation was at least 80% (as determined by fastSTRUCTURE analysis), the accession was assigned to that subpopulation, otherwise it was classified as admixed. The three subpopulations (SP1, SP2, and SP3) have population sizes of 139, 29 and 55, respectively, and a total of 58 samples were assigned as admixed samples. All the accessions originally collected in the US were in SP3, and SP1 and SP2 consisted of accessions exclusively collected from Mexico.

In order to better understand the differences between these three subpopulations, we used SNP Set III to quantify the genetic differentiation using measurements of fixation index (F_{ST}), nucleotide diversity (π), and Tajima's D, correcting for differences in sample sizes via either an estimator that is independent of population size (Hudson F_{ST}), or different ascertainment schemes (Table 2.1, Supplemental Table S2.4). Nucleotide diversity estimates (π) at the genome-wide level for the three subpopulations (SP1, SP2 and SP3) were similar to each other (range: 8.08e–06 to 9.09e–06), but lower than the population of 281 Mesoamerican NS/S accessions (1.14e–05). Genome-wide Tajima's D values were all positive for the three subpopulations, with a slightly higher D value for the overall population. Without ascertainment bias correction, the weighted average genome-wide F_{ST} ranged from 0.45 to 0.64 for all three pairs when using $Q \geqslant 0.8$ as the population assignment filter to define the three subpopulations. When using $Q \geqslant 0.5$ as the population assignment filter, we assigned

170, 36 and 68 samples to SP1, SP2 and SP3 respectively, and the weighted average genome-wide F_{ST} (without ascertainment bias correction) was approximately 15% lower than using the 0.8 threshold, ranging from 0.37 to 0.56 (Supplemental Table S2.4).

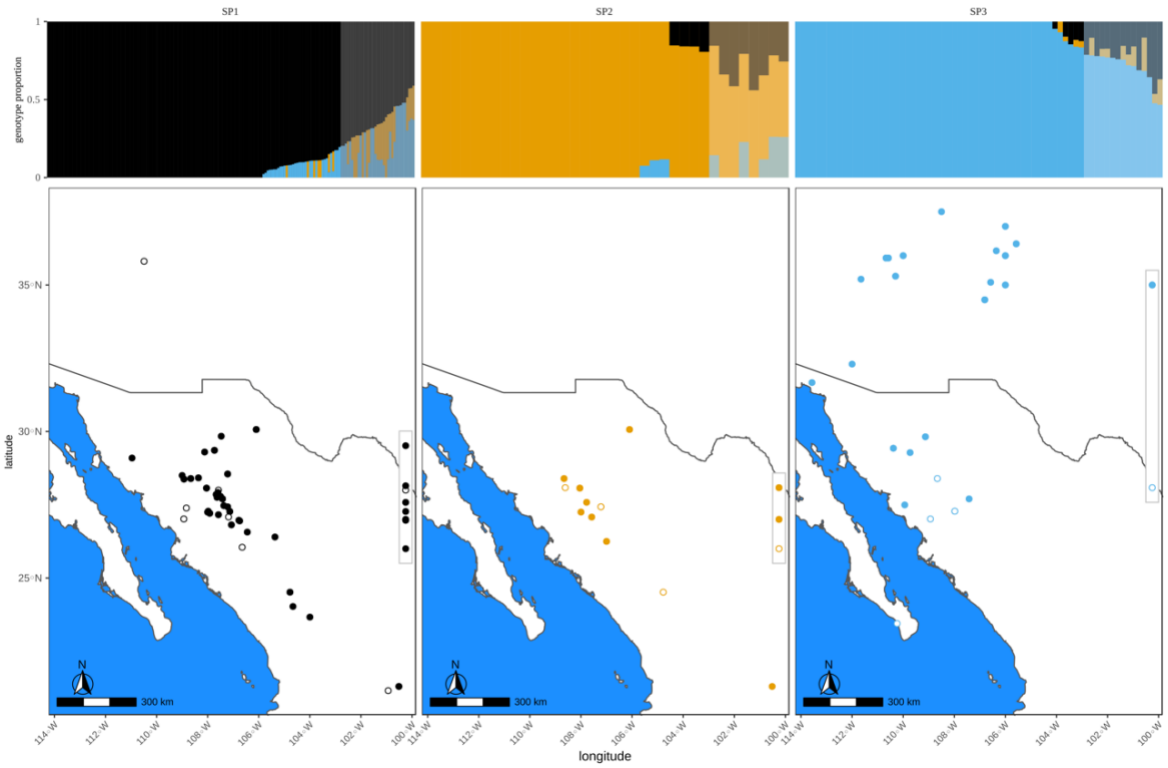

Figure 2.1. Subpopulation structure of the NS/S Mesoamerican population. Subpopulation contribution (top) and geographics distribution (bottom) is shown for SP1 (black), SP2 (orange), and SP3 (blue), with subpopulation membership for each sample assigned using a Q $\geqslant$ 0.5 cutoff. Accessions categorized as admixed at a Q $\geqslant$ 0.8 cutoff are represented by lighter color bars (top) and open circles (bottom). The right edge strip on each map contains accessions for which latitude information was inferred, and no longitude information was available.

Table 2.1. Population genetic statistics for the NS/S Mesoamerican population and its three subpopulations, calculated using SNP Set III.

Subpopulation	π	D	F_{ST} SP1	F_{ST} SP2	F_{ST} SP3
All	1.14E–05	1.21	–	–	–
SP1	8.22E–06	0.58	–	0.50 (0.42)	0.43 (0.36)
SP2	9.09E–06	0.67	0.55 (0.45)	–	0.62 (0.56)
SP3	8.08E–06	0.47	0.46 (0.38)	0.55 (0.54)	–

F_{ST} values are the result of calculations without ascertainment bias. F_{ST} values are reported for subpopulation assignment thresholds of $Q \geqslant 0.8$ (primary number) and $Q \geqslant 0.5$ (alternative, in parenthesis). All, all 281 NS/S Mesoamerican accessions; π, nucleotide diversity; D, Tajima's D statistics; F_{ST}, weighted fixation index

Genome-wide association study of seed coat traits

The genetic control of seed coat patterning in the NS/S common bean population of 281 accessions of Mesoamerican origin was dissected via GWAS using 5,732 genome-wide SNP markers from SNP Set II. The MLMM analysis selected three SNPs on chromosome 8 (S8_3258880, S8_3229903 and S8_3258963) and one SNP (S10_36816230) on chromosome 10 (Figure 2.2, Supplemental Figure S2.7, Supplemental Table S2.6) that were associated with the seed coat pattern phenotype (solid color, partially colored, and patterned categories) observed in this panel.

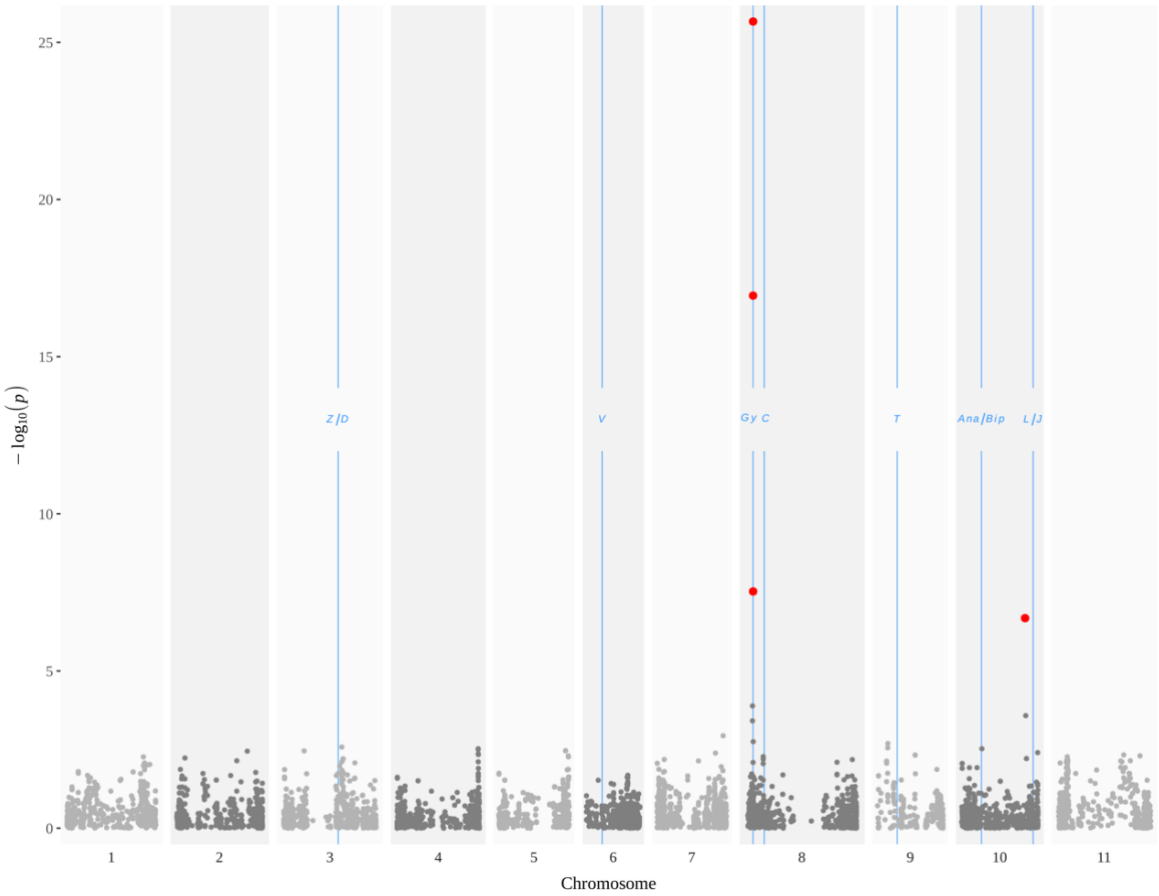

Figure 2.2. GWAS results for the three-class seed coat pattern trait. Genetic loci controlling seed coat color and pattern identified in previous work are indicated with blue vertical lines corresponding to their putative genomic location, and the lines are labelled with the corresponding gene names. SNPs selected by the MLMM as significantly associated with the three-class seed coat pattern trait are shown in red, while the remaining unselected SNPs are shown in grey.

In order to better assign the association signals to the three different categories that defined the seed coat pattern phenotype, we performed MLMM analyses on the three binary seed coat pattern classifications (Supplemental Table S2.7). With the “Anasazi binary” trait, six SNPs were found to be associated with this trait: the previously identified SNP on

chromosome 10 (S10_36816230), two SNPs on chromosome 6, two on chromosome 7, and one on chromosome 5. For "pattern binary" and "color binary" traits, MLMM results were the same, identifying five SNPs on chromosome 8 (S8_3258880, S8_3229903, S8_3258963, S8_2905184, S8_2907111), the first three of which were the same as identified by MLMM with the three class seed coat pattern trait.

Within this group of 281 Mesoamerican samples, the genome-wide median LD (50th percentile) decayed to background levels ($r^2 < 0.1$) by ~128 Kb, (Supplemental Figure S2.8) and, as such, the candidate gene search space was limited to ± 128 Kb of the GWAS-detected SNP markers. The region on chromosome 8 contains, within 28 Kb of the most significant SNP (S8_3258880), the putative location of the *Gy*-STS (OW17) marker previously genetically mapped proximal (2.7 cM) to the *C* locus (McClean *et al.* 2002). Previous work (Prakken 1974) identified this as a complex locus consisting of many closely linked genes that are responsible for various seed coat patterns including striped, mottled and netted seed coats. When searching for candidate genes in the ± 128 Kb regions flanking the three MLMM-selected SNPs on chromosome 8, five *myb* transcription factors (Phvul.008G038000, Phvul.008G038200, Phvul.008G038400, Phvul.008G038500, and Phvul.008G038600) were identified within this region (Supplemental Table S2.7).

A single SNP on chromosome 10 (S10_36816230) was associated with seed coat pattern traits by both the two-class and three-class MLMM analysis. Two loci associated with the partial color phenotype have been previously mapped to this chromosome: *J* and *Bip*. The putative physical position of the *J* locus, estimated using the STS marker $OL4S_{500}$ 1.2 cM away from the *J* locus (McClean *et al.* 2002) as a proxy, is ~ 4 Mb away from this SNP. The

putative physical position of the *Ana* locus, estimated using the STS marker OM9S$_{200}$ 5.4 cM away from the *Ana* locus (McClean *et al.* 2002) as a proxy, is ~ 35 Mb away from this SNP. The *J* locus had been previously shown (Bassett *et al.* 2002b; McClean *et al.* 2002; Bassett 2007) to be associated with the expression of seed coat color (*J*) and expression of pattern restriction (*L*) in partly colored seed coat, while the *Bip* locus has been associated with the Anasazi (*Ana*) and bipunctata (*Bip*) seed coat patterns (Bassett *et al.* 2000). No clear candidate genes were found proximal to this SNP on chromosome 10.

DISCUSSION

The NS/S common bean collection represents a unique sampling of landraces throughout the northern native range of Mesoamerican common bean. Only a small subset of this collection, primarily from the US, is available through the USDA NPGS, with most of the accessions originating in Mexico available only from the NS/S collection. Our study is the first high-density genomic characterization of the genetic diversity and population structure within this collection, as well as the first identification and genetic diversity comparison among these three subpopulations of the Mesoamerican common bean gene pool. Furthermore, the wide variation of seed coat phenotypes among these accessions enabled us to provide higher resolution mapping and identify several candidate genes at the *C* locus, as well as identify a novel locus potentially associated with the Anasazi seed coat pattern phenotype.

Population structure

Due to the paucity of Andean genotypes in our panel, we focused genetic analyses on the Mesoamerican population. Previous molecular-based studies, whether using GBS (Ariani *et*

al. 2018), select nuclear genes (Bitocchi *et al.* 2012b), or microsatellite markers (Kwak and Gepts 2009), have consistently recovered strong population structure within Mesoamerican common bean, identifying between three and four subpopulations, including sympatric subpopulations. Consistent with those findings, our study identified three subpopulations that are strongly differentiated (pairwise weighted $F_{ST} > 0.4$), with two of them having completely overlapping geographical ranges in sampling.

Given that the vast majority of the NS/S accessions in our study originate further North than previous studies, with little to no overlap in sampling locale (Supplemental Figure S2.9), and most of our accessions represent unique collections of landraces cultivated by indigenous groups, a straightforward comparison of the subpopulations in our study and those from previous work is not possible. Only one accession genotyped in our study, PI 615391, was previously genotyped, and it was assigned to the K9 (Races Jalisco and Durango) subpopulation of Kwak and Gepts (2009). This accession is genetically similar to SP3 (Supplemental Figure S2.6), a subpopulation almost exclusively found in the southwest US in our study, and this is consistent with the relationships of other NPGS accessions labelled as "Jalisco" or "Durango". It is therefore plausible that SP3 corresponds to subpopulations labelled as "Jalisco & Durango" in previous studies, containing the "Jalisco" and "Durango" Mesoamerican common bean races, but our study suggests this subpopulation is endemic to southwest US rather than Mexico. The other Mesoamerican common bean race, "Mesoamerica," is represented in our study by five NPGS accessions (Supplemental Table S2.1), and in our phylogenetic analysis they are contained within a clade primarily consisting of admixed samples (Supplemental Figure S2.6), distinct from both the SP1 and SP2 clades.

Therefore, both SP1 and SP2 appear to be novel subpopulations that do not correspond to previously described Mesoamerican common bean races. Further joint analysis of samples from our and previous work would be needed to draw confident parallels between Mesoamerican subpopulations identified here and previously. However, due to differences in technology used (select genes, microsatellite markers) or enzyme choice (GBS), this is not possible with the existing data, and new genotyping on a common platform will be necessary for this work.

Differentiation among Mesoamerican subpopulations in this study is relatively high, with pairwise weighted F_{ST} values, regardless of the ascertainment scheme of population assignment threshold, higher than those observed between Mesoamerican and Andean gene pools. At the same time, although the overall diversity within our Mesoamerican gene pool (π = 1.14e–05) is in line with previously reported values (π = 1.42e–05; Ariani *et al.* 2018), diversity within subpopulations is lower (π = 8.08e–06 to 9.09e–06). This suggests that the genetic diversity within the Mesoamerican gene pool is partitioned among subpopulations. Given that the accessions in this study consisted exclusively of domesticated beans, this allelic segregation, as well as the various degrees of admixture observed among these subpopulations, are likely the result of not only population genetic factors but also ethnographic and ethnobotanical factors such as historical inter-locality trade flow and local preferences for common bean landraces. Future interdisciplinary work would be needed to elucidate the relative contributions of these factors and paint a fuller picture of population differentiation within this Mesoamerican germplasm pool.

Seed coat trait-associated genomic loci

The genetic control of seed coat patterning and coloring has been of long standing interest to the common bean community, with research into the genetic control of these traits spanning decades (Bassett 2007 and references within). Complementing the designs of previous studies, which followed pre-determined genetic crosses, GWAS allowed us to leverage the phenotypic and genotypic diversity of our mapping population and identify SNP markers associated with seed coat traits. Given the complex nature of genetic control for color (Bassett 2007) and the difficulty in reliably phenotyping color variation in the presence of various coat patterns, we chose to focus on the three major classes of patterning, following Bassett (2007), as traits of interest: solid, patterned, and partial coloring. Previous work has shown a complex architecture controlling seed coat patterning, consisting of two primary genes (*C* and *T*) and a larger number of modifier genes such as *Z, J*, *Bip*, etc (McClean *et al.* 2002; Bassett 2007). A gene within the *C* locus is expected to control patterned seed coat, whereas the *T* locus is expected to control partial coloring. Alternate alleles at these loci result in the default "solid" pattern. Using the primer sequences for STS markers linked to these loci published in McClean *et al.* (2002), we were able to identify the likely physical location of several of these markers on the *P. vulgaris* genome. This allowed us to compare the SNP positions identified by GWAS with the expected location of relevant loci, such as the *C* associated marker $OAP2S_{650}$ at approximately 9.7 Mb on chromosome 8 and the *T* associated marker $OM19S_{350}$ at approximately 11.7 Mb on chromosome 9.

We did indeed find three independent SNPs on chromosome 8 associated with seed coat pattern, consistent with the expectation that the *C* locus is comprised of a gene complex.

Moreover, these SNPs were shown to be significantly associated both in the three-class GWAS (solid, patterned, partial color) and the two-class GWAS involving solid or patterned phenotypes, but not in the two-class GWAS involving the partial color phenotype. This again is consistent with the expectation that the *C* locus controls the patterned phenotype, but is independent of the partial color phenotype controlled by *T*. Although these SNPs are approximately 6.5 Mb away from the likely location of the *C* associated STS marker $OAP2S_{650}$, they span a region that overlaps the likely location of the *Gy* associated STS marker $OW17S_{600}$ previously genetically mapped proximal (2.7 cM) to the *C* locus (Bassett *et al.* 2002a; McClean *et al.* 2002). Together, this evidence strongly suggests that the SNPs found by GWAS to be associated with the patterned traits are in strong LD with causal genes at the (potentially expansive) *C* locus.

The *C* locus controls common bean seed coat patterns by regulating the differential expression of flavonoids via epistatic interactions with *J* and *V* loci (Feenstra 1960). Within the family of flavonoids that accumulate in the seed coat of common bean, proanthocyanidins (PA), or condensed tannin, has been shown to be partially controlled by the *C* locus in certain genetic backgrounds (Caldas and Blair 2009). Our search for plausible causal genes in the vicinity of the seed coat pattern associated SNPs on chromosome 8 resulted in the identification of five *myb* genes. Multiple lines of evidence implicate the role of *myb* genes, the most abundant transcription factor family in common bean (Kalavacharla *et al.* 2011), in the PA biosynthesis pathway in the seed coat of plant species (Nesi *et al.* 2001; Aharoni *et al.* 2001; Yang *et al.* 2010; Zabala and Vodkin 2014; Liu *et al.* 2014; Hong *et al.* 2017). Furthermore, in addition to evidence for involvement in PA accumulation, differential

expression analysis in soybean implicated the involvement of a *myb* gene in disruptive cell wall formation and net pattern of seed coat (Kour *et al.* 2014). Therefore, among the genes proximal to the MLMM-selected SNPs on chromosome 8, the MYB transcription factors identified are the most biologically plausible candidate causal genes for the patterning of common bean seed coat. The presence of five *myb* genes within this 80-Kb genomic region is consistent with the hypothesis of Prakken (1974) that the *C* locus is comprised of a set of tightly linked genes.

In contrast, we did not identify any SNPs on chromosome 9 (the expected location of the *T* locus) significantly associated with the partial color phenotype. Given that only 2% (6 out of 281) Mesoamerican accessions exhibited this phenotype, and the SNP set used in GWAS was filtered to remove SNPs with minor allele frequency below 5%, we did not expect to find any SNPs in perfect LD with the causal gene for this phenotype, but rather proximal SNPs above that threshold that would be in partial LD with the locus. However, in our GWAS SNP set, the closest SNP to the expected location of the *T* associated marker $OM19S_{350}$ was approximately 600 Kb away, almost 5 times further away than the genome-wide average LD decay distance. It is therefore not surprising that we were unable to detect SNPs associated with the *T* locus, because we did not evaluate any SNPs that were likely to be in LD with it. We did, however, identify a SNP on chromosome 10 that was associated with seed coat pattern, the same chromosome where the *Bip* and *J* loci, two loci modulating partial coloring, are located. This SNP was significantly associated with the trait both in the three-class trait GWAS, as well as in the two-class trait GWAS involving the partial color phenotype, but was not associated with the trait in the two-class trait GWAS involving solid

or patterned phenotype, indicating that it is associated with the partial color phenotype.

A closer inspection of the phenotypes classified as partial color (Supplemental Figure S2.2) shows that they appear much more similar to the Anasazi phenotype (Bassett *et al.* 2000, Fig. 1) rather than the various phenotypes controlled by the *J* locus and its interactions (Bassett 2007 Fig. 8.3). The Anasazi pattern is expected to be controlled by the *Bip* locus (Bassett *et al.* 2000). However, our associated SNP is more than 25 Mb downstream of the likely location of the *Ana* associated STS marker $OM9S_{200}$ of McClean *et al.* (2002), on a different chromosome arm. There are several plausible explanations for this incongruency. Although the distance is too large to be explained alone by low mapping resolution from only six accessions with the Anasazi phenotype, the detected SNP could be indirectly associated with the causal gene of this rare phenotype due to cryptic population structure. Accessions exhibiting the Anasazi phenotype are likely derived from ancient beans cultivated in the four corners region of the SW US (Bassett *et al.* 2000) and are thus expected to have very similar genetic backgrounds. Three of these six accessions are indeed from that region, whereas the other three are from the SP1 subpopulation that is limited to Mexico. However, all three of the SP1 accessions were classified as admixed. The accessions possess a mixture of SP1 and SP3 (primarily of US origin) alleles, with more than 20% of alleles derived from SP3 (Supplemental Figure S2.10). Although the MLMM is expected to control for population structure and unequal relatedness through PCs and a kinship matrix, it may not have been able to adequately control for such partial population structure where only a subset of loci is introgressed into an otherwise distinct genetic background. Therefore, it is possible that the SNP detected is associated with a larger *Ana*-containing introgression, or a group of

introgressed loci, rather than the causal locus itself. Indeed, the multiple MLMM selected SNPs in the Anasazi-binary GWAS (Supplemental Figure S2.7), at four distinct loci across the genome, would be consistent with a population structure signal representing a group of introgressed loci. Moreover, unlike the *T* locus, our dataset did include proximal SNPs, as close as 18 Kb, that are expected to be in LD with the likely locus for the *Ana*-linked marker, but no significant association was found for these SNPs. Overall, under the assumption that *Ana* is the controlling gene for the observed phenotypes, this suggests that our analysis did not recover a signal for the *Bip* locus, but rather a cryptic population structure signal. However, it is also possible that the SNP is tagging a novel allele of the *J* locus whose resultant phenotype strongly resembles the Anasazi phenotype, as several seed coat partial color patterns controlled by *J*, such as "expansa" or "white ends" (Bassett 2007, Fig. 8.3 C and E respectively) partially resemble Anasazi. Further work with an expanded panel of Anasazi beans will be necessary to elucidate this issue and more accurately locate the locus or loci responsible for the Anasazi phenotype.

Finally, it should be noted that common bean is one of several Phaseolus species with interesting variation in seed coat patterning and coloring. Previous work in lima bean (Phaseolus lunatus L.) suggests a similar genetic control mechanism for seed coat color in that species (Allard 1953, Bemis 1957), and future comparative genomics work between common bean and lima bean might be able to shed light on the evolution of these loci. Seed coat color diversity appears to be maintained in scarlet runner bean (*Phaseolus coccineus* L.) landraces as well (Acampora *et al.* 2007). Given their overlapping centers of origin and domestication, it would be instructive for future comparative genomics work to include landraces of all three

of these species.

CONCLUSIONS

Foremost, our study represents the genomic characterization and population genetic analysis of a unique set of common bean accessions from previously underrepresented geographic and ethnographic sources. Using this novel genetic resource, we identified both allopatric and sympatric population structure, suggesting that, in addition to natural segregation, human selection factors have shaped the population structure of domesticated Mesoamerican common beans. This highlights the importance of the NS/S collection for understanding the various ethnographic factors that have shaped the history of domestication and cultivation of Mesoamerican common beans, and it establishes the NS/S collection as a unique resource of traditional landraces that contains important genetic diversity and likely possesses important alleles for biotic and abiotic stress tolerance that have been selected by smallholder farmers over thousands of years. With respect to overall common bean genetic architecture, we were able to identify putative causal genes for the various seed coat pattern phenotypes controlled by the *C* locus and identified a potential novel locus controlling a subset of partial coloring phenotypes. This brings us closer to understanding the complex genetic control of the extremely varied patterning and coloring of common bean seed coats.

ACKNOWLEDGEMENTS

This research was supported by Cornell University startup funds (M.A.G). We thank the Department of Energy Joint Genome Institute and collaborators (Scott Jackson, Phil McClean, and Jeremy Schmutz), for pre-publication access to release v2.1 of the Phaseolus vulgaris

(common bean) genome and annotation. We thank Dr. Phil McClean for seeds of G19833. We are grateful for the help from Nicholas Kaczmar for planting and harvesting, as well as staff of Cornell's Guterman Bioclimatic Laboratory for greenhouse management.

SUPPLEMENTAL INFORMATION

Supplemental Figure S2.1 Visual examples of the seed coat pattern classification used in this study.

Supplemental Figure S2.2 Visual examples of the seed coat pattern for all six accessions classified as "Anasazi".

Supplemental Figure S2.3 Schematic overview of the SNP filtering pipeline. SNP Set 0 was used to identify unintended sample duplicates. SNP Set I was used to separate accessions into Andean and Mesoamerican groups. SNP Set II was used for GWAS. SNP Set III was used for population genetics analysis.

Supplemental Figure S2.4 Principal component analysis of SNP Set I. Mesoamerican NPGS and Andean NPGS classifications are based on a priori annotation of NPGS accessions. Mesoamerican NSS and Andean NSS classifications are based on K-means clustering (K = 2) along PC1 with the respective group of NPGS accessions.

Supplemental Figure S2.5 Principal component analysis of SNP Set II. Subpopulation labels are based on fastSTRUCTURE analysis of SNP Set III, with a subpopulation assignment criterion of $Q \geqslant 0.8$.

Supplemental Figure S2.6 Neighbour joining phylogenetic tree of the NS/S (SP1, SP2, SP3, SP mixed) and NPGS (NPGS Durango, NPGS Jalisco, NPGS Mesoamerica)

Mesoamerican accessions, rooted using the reference genome accession (G19833, 8 samples, Andean genotype). NPGS accession PI 615391, used in both our study and Kwak and Gepts (2009), is indicated by the dotted line and label. Subpopulation labels are based on fastSTRUCTURE analysis of SNP Set III, with a subpopulation assignment criterion of $Q \geqslant 0.8$.

Supplemental Figure S2.7 GWAS results of three-class and binary coding of seed coat pattern traits. SNPs selected by the MLMM as significantly associated with the trait are shown in red, while the remaining SNPs are shown in black.

Supplemental Figure S2.8 Linkage disequilibrium (LD) estimates between SNPs in the NS/S Mesoamerican population of 281 accessions. The distribution of SNPs at different percentile cutoffs are indicated by the labeled lines. Median LD is indicated by the solid line labelled 50%, which decays to background levels ($r^2 < 0.1$) at a physical distance beyond 128 Kb.

Supplemental Figure S2.9 Histograms of latitude distributions for accessions used in this study and three previous studies. The red color indicates accessions collected from Mexico.

Supplemental Figure S2.10 Subpopulation composition of the six accessions with Anasazi phenotypes using fastSTRUCTURE results. Three accessions (15DW320, 15DW324, 15DW306) are assigned to SP1 at $Q \geqslant 0.5$, but classified as mixed at $Q \geqslant 0.8$.

Supplemental Table S2.1 The complete list of 375 accessions included in this study and relevant metadata. The “Group ID” column indicates whether or not an accession was part of an unintended sample duplicate group, and the “Choice” column indicates whether or

not an accession was selected as a representative for its unintended sample duplicate group. The NS/S-assigned accession number, catalog number, common name, and phenotype description are reported for accessions sourced from NS/S. For NPGS sourced accessions, the data refers to annotation from the GRIN database. The race of each NPGS accession was inferred using the documented seed type and morphology description with criteria from Singh *et al.* (1991).

Supplemental Table S2.2 Accession geographical provenance and seed coat pattern trait encoding for the NS/S accessions of Mesoamerican origin. The last column indicates the corresponding seed coat pattern and color descriptors according to IBPGR (1982) and Kornerup (1967).

Supplemental Table S2.3 Summary of the four SNP sets used in this study.

Supplemental Table S2.4 F_{ST} results of all pairwise comparisons between the three subpopulations within the NS/S Mesoamerican population.

Supplemental Table S2.5 BLAST results for STS marker sequences for *a priori* seed coat color and pattern genes. Putative genomic locations of seven out of the 11 STS markers described in Table 2 of McClean *et al.* (2002) are presented.

Supplemental Table S2.6 Seed coat pattern trait GWAS results and genomic location information for MLMM selected SNPs.

Supplemental Table S2.7 All annotated genes within 128 Kb of the MLMM selected SNPs reported in Supplemental Table S2.6.

REFERENCES

Acampora A, Ciaffi M, De Pace C, *et al.* (2007) Pattern of variation for seed size traits and molecular markers in Italian germplasm of *Phaseolus coccineus* L. Euphytica 157:69–82

Aharoni A, De Vos CH, Wein M, *et al.* (2001) The strawberry FaMYB1 transcription factor suppresses anthocyanin and flavonol accumulation in transgenic tobacco. Plant J 28:319–332

Allard RW (1953) Inheritance of some seed-coat colors and patterns in lima beans. Hilgardia 22:167–177

Ariani A, Berny Mier Y Teran JC, Gepts P (2018) Spatial and temporal scales of range expansion in wild *Phaseolus vulgaris*. Mol Biol Evol 35:119–131

Bassett MJ (2007) Genetics of seed coat color and pattern in common bean. In: Plant Breeding Reviews. pp 239–315

Bassett MJ, Hartel K, McClean P (2000) Inheritance of the Anasazi pattern of partly colored seedcoats in common bean. Journal of the American Society for Horticultural Science 125:340–343

Bassett MJ, Lee R, Otto C, McClean PE (2002a) Classical and molecular genetic studies of the strong greenish yellow seedcoat color in `Wagenaar' and `Enola' common bean. J Am Soc Hortic Sci 127:50–55

Bassett MJ, Lee R, Symanietz T, McClean PE (2002b) Inheritance of reverse margo seedcoat pattern and allelism between the genes J for seedcoat color and L for partly colored seedcoat pattern in common bean. J Am Soc Hortic Sci 127:56–61

Bemis WP (1957) Inheritance of a base seed-coat color factor in lima beans. Journal of Heredity 48:124–127

Bitocchi E, Bellucci E, Giardini A, *et al.* (2012a) Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. New Phytol 197:300–313

Bitocchi E, Nanni L, Bellucci E, *et al.* (2012b) Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. Proc Natl Acad Sci USA 109:

Blair MW, Giraldo MC, Buendía HF, *et al.* (2006) Microsatellite marker diversity in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 113:100–109

Boodley JW, Sheldrake R (1972) Cornell peat-lite mixes for commercial growing. Inf Bull 43, Cornell Univ Ithaca, NY

Bradbury PJ, Zhang Z, Kroon DE, *et al.* (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635

Broughton WJ, Hernández G, Blair M, *et al.* (2003) Beans (Phaseolus spp.) – model food legumes. Plant Soil 252:55–128

Burgess MA (1994) Cultural responsibility in the preservation of local economic plant resources. Biodivers Conserv 3:126–136

Caldas GV, Blair MW (2009) Inheritance of seed condensed tannins and their relationship with seed-coat color and pattern genes in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet

119:131–142

Camacho C, Coulouris G, Avagyan V, *et al.* (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421

Chen J, Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95:759–771

Common bean descriptors. https://www.bioversityinternational.org/fileadmin/user_upload/online_library/publications/pdfs/160.pdf. Accessed 30 Jul 2019

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

Eitzinger A, Läderach P, Rodriguez B, *et al.* (2017) Assessing high-impact spots of climate change: spatial yield simulations with Decision Support System for Agrotechnology Transfer (DSSAT) model. Mitig Adapt Strateg Glob Chang 22:743–760

Elshire RJ, Glaubitz JC, Sun Q, *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379

Endelman JB, Jannink J-L (2012) Shrinkage estimation of the realized relationship matrix. G3 2:1405–1413

FAO (2018) FAOSTAT. In: FAOSTAT. http://www.fao.org/faostat/en/#data/CC/visualize. Accessed 4 Nov 2018

Feenstra WJ (1960) Biochemical aspects of seedcoat colour inheritance in *Phaseolus vulgaris* L. Veenman

Glaubitz JC, Casstevens TM, Lu F, *et al.* (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9:e90346

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33:54–78

Hong M, Hu K, Tian T, *et al.* (2017) Transcriptomic analysis of seed coats in yellow-seeded *Brassica napus* reveals novel genes that influence proanthocyanidin biosynthesis. Front Plant Sci 8.: doi: 10.3389/fpls.2017.01674

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132:583–589

Jones AL (1999) *Phaseolus* bean post-harvest operations

Kalavacharla V, Liu Z, Meyers BC, *et al.* (2011) Identification and analysis of common bean (*Phaseolus vulgaris* L.) transcriptomes by massively parallel pyrosequencing. BMC Plant Biol 11:135

Kornerup A (1967) Methuen handbook of colour. Hastings House Pub

Kour A, Boone AM, Vodkin LO (2014) RNA-Seq profiling of a defective seed coat mutation in *Glycine max* reveals differential expression of proline-rich and other cell wall protein transcripts. PLoS One 9:e96342

Kwak M, Gepts P (2009) Structure of genetic diversity in the two major gene pools of

common bean (*Phaseolus vulgaris* L., Fabaceae). Theor Appl Genet 118:979–992

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Lipka AE, Tian F, Wang Q, *et al.* (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399

Liu C, Jun JH, Dixon RA (2014) MYB5 and MYB14 play pivotal roles in seed coat polymer biosynthesis in *Medicago truncatula*. Plant Physiol 165:1424–1439

Mamidi S, Rossi M, Moghaddam SM, *et al.* (2013) Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L. Heredity 110:267–276

McClean PE, Lee RK, Otto C, *et al.* (2002) Molecular and phenotypic mapping of genes controlling seed coat pattern and color in common bean (*Phaseolus vulgaris* L.). J Hered 93:148–152

Nesi N, Jond C, Debeaujon I, *et al.* (2001) The Arabidopsis TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. Plant Cell 13:2099–2114

Prakken R (1974) Inheritance of colours in *Phaseolus vulgaris* L. IV Recombination within the "Complex Locus C". *Meded Landbouwhogesch Wageningen*, *24*, 1-36.

Price AL, Patterson NJ, Plenge RM, *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

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

Romay MC, Millard MJ, Glaubitz JC, *et al.* (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14:R55

Schmutz J, McClean PE, Mamidi S, *et al.* (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Segura V, Vilhjálmsson BJ, Platt A, *et al.* (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44:825–830

Singh SP, Gepts P, Debouck DG (1991) Races of common bean (*Phaseolus vulgaris*, Fabaceae). Econ Bot 45:379–396

Stacklies W, Redestig H, Scholz M, *et al.* (2007) pcaMethods--a bioconductor package providing PCA methods for incomplete data. Bioinformatics 23:1164–1167

Swarts K, Li HH, Navarro JAR, *et al.* (2014) Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7.: doi: 10.3835/plantgenome2014.05.0023

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

van Schoonhoven A (1991) Common beans: research for crop improvement. CAB International, Centro Internacional de Agricultura Tropical (CIAT), Cali

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

van Schoonhoven A (1991) Common beans: research for crop improvement.CAB International; Cali, CO: Centro Internacional de Agricultura Tropical (CIAT), 980 p.

Wortmann CS (1998) Atlas of common bean (*Phaseolus vulgaris* L.) production in Africa.Centro Internacional de Agricultura Tropical (CIAT), Cali, CO. 131 p. (CIAT publication no. 297)

Yang K, Jeong N, Moon J-K, *et al.* (2010) Genetic analysis of genes controlling natural variation of seed coat and flower colors in soybean. J Hered 101:757–768

Zabala G, Vodkin LO (2014) Methylation affects transposition and splicing of a large CACTA transposon from a MYB transcription factor regulating anthocyanin synthase genes in soybean seed coats. PLoS One 9:e111959

Chapter 3 High-resolution genome-wide association study pinpoints metal transporter and chelator genes involved in the genetic control of element levels in maize grain[1]

ABSTRACT

Despite its importance to plant function and human health, the genetics underpinning element levels in maize grain remain largely unknown. Through a genome-wide association study in the maize Ames panel of nearly 2,000 inbred lines that was imputed with ~7.7 million SNP markers, we investigated the genetic basis of natural variation for the concentration of 11 elements in grain. Novel associations were detected for the metal transporter genes *rte2* (*rotten ear2*) and *irt1* (*iron-regulated transporter1*) with boron and nickel, respectively. We also further resolved loci that were previously found to be associated with one or more of five elements (copper, iron, manganese, molybdenum, and/or zinc), with two metal chelator and five metal transporter candidate causal genes identified. The *nas5* (*nicotianamine synthase5*) gene involved in the synthesis of nicotianamine, a metal chelator, was found associated with both zinc and iron and suggests a common genetic basis controlling the accumulation of these two metals in the grain. Furthermore, moderate predictive abilities were obtained for the 11 elemental grain phenotypes with two whole-genome prediction models: Bayesian Ridge Regression (0.33-0.51) and BayesB (0.33-0.53). Of the two models, BayesB, with its greater

[1] Wu, D., Tanaka, R., Li, X., Ramstein, G. P., Cu, S., Hamilton, J. P., ... & Gore, M. A. (2021). High-resolution genome-wide association study pinpoints metal transporter and chelator genes involved in the genetic control of element levels in maize grain. G3, 11(4), jkab059.

emphasis on large-effect loci, showed ~4-10% higher predictive abilities for nickel, molybdenum, and copper. Altogether, our findings contribute to an improved genotype-phenotype map for grain element accumulation in maize.

INTRODUCTION

Elements are important in every aspect of organismal development. In higher plants, at least 20 elements are involved in key biological functions (Mengel and Kirkby 2012). To maintain elemental homeostasis, plants require the activities of metal transporters, chelators, and signaling pathways for the regulation of optimal uptake, transport, and storage of metal ions (Clemens 2001). The complex network responsible for elemental accumulation in various plant organs and tissues such as physiologically mature seed is coordinated at the genetic level, but it can be perturbed by alterations in the soil chemical environment, plant architecture, physiology, and metabolism (Baxter 2009). However, the genetic underpinnings of the biological processes that regulate elemental uptake, transport, and storage have yet to be fully elucidated in model plants and crop species.

Maize (*Zea mays* L.) is a globally important staple crop, serving as a critical source of calories in Sub-Saharan Africa and Latin America (FAOSTAT 2018). However, the declining soil fertility of farming systems contributes in part to the unrealized potential yield of maize in these geographies (Dixon *et al.* 2001). Not only does the deficiency or excess of one or more key elements in the soil limit maize plant productivity (ten Berge *et al.* 2019), but it also has implications for human nutrition if this causes an unfavorable elemental profile in the maize grain (Graham and Welch 1996; Welch 2002; Welch and Graham 2004). This could pose

serious malnutrition-related health problems in a maize-based diet because such a diet may not provide the recommended dietary allowances of micronutrients such as iron (Fe) and zinc (Zn) (Welch and Graham 2002). The development of crop varieties with improved nutritional quality through plant breeding, a strategy known as "biofortification," has the potential to sustainably address micronutrient deficiencies in developing nations (Diepenbrock and Gore 2015; Bouis and Saltzman 2017).

Depending on the element and plant species, elements accumulated in seed could originate from direct root uptake or remobilization from senescing tissues through the involvement of transporters and chelators (Waters and Sankaran 2011). Several metal transporter and chelator protein families, such as METAL TOLERANCE PROTEIN (MTP), NATURAL RESISTANCE-ASSOCIATED MACROPHAGE PROTEIN (NRAMP), NICOTIANAMINE SYNTHASE (NAS), YELLOW STRIPE-LIKE (YSL), and ZINC-REGULATED TRANSPORTER (ZRT)/IRON-REGULATED TRANSPORTER (IRT)-LIKE PROTEIN (ZIP), have been bioinformatically identified in the genomes of Arabidopsis [*Arabidopsis thaliana* (L.) Heynh.], rice (*Oryza sativa* L.), maize and other plant species, yet only a subset from each family have been functionally characterized (Whitt *et al.* 2020). Many metal transporters and chelators have broad substrate specificity (Axelsen and Palmgren 2001), making it difficult to infer their primary roles with homology-based approaches. In maize, only a few metal transporter genes have been functionally studied including *rotten ear1* (*rte1*), *rte2*, and *tassel-less1* (*tls1*) for boron (B) (Chatterjee *et al.* 2014, 2017; Durbak *et al.* 2014), *yellow stripe1* (*ys1*) and *ys3* for Fe (Von Wiren *et al.* 1994; Chan-Rodriguez and Walker 2018), and *ysl2* (Zang *et al.* 2020) and *zip5* (Li *et al.* 2019) for Fe and

Zn. Of these, transgenic maize overexpressing *zip5* with an endosperm-specific promoter was shown to accumulate higher levels of Fe and Zn in grain (Li *et al.* 2019).

Genetic mapping approaches offer another opportunity to identify the largely unknown genes responsible for elemental concentration in maize grain. There have been a number of linkage analysis studies that have used biparental populations to identify quantitative trait loci (QTL) associated with an elemental concentration in maize grain, especially for Fe and Zn (Lung'aho *et al.* 2011; Šimić *et al.* 2011; Qin *et al.* 2012; Baxter *et al.* 2013; Jin *et al.* 2013; Gu *et al.* 2015; Asaro *et al.* 2016; H. Zhang *et al.* 2017; Ziegler *et al.* 2017; Fikas *et al.* 2019). However, the biparental populations used in these studies did not provide gene-level mapping resolution due to the limited number of recent recombination events (Zhu *et al.* 2008; Myles *et al.* 2009). Thus, the causal genes presumably residing in the large QTL intervals with low resolution could not be conclusively identified.

Genome-wide association studies (GWAS) that exploit the extensive phenotypic variation and ancient recombination of many individuals comprising a diversity population (association panel) often offer higher mapping resolution to dissect complex traits than biparental mapping populations (Myles *et al.* 2009; Lipka *et al.* 2015). A total of 46 marker-trait associations for the concentration of Zn and Fe in grain were identified in a tropical maize association panel (Hindu *et al.* 2018). Some of these associations were independently supported by separate QTL analyses in biparental populations. However, the association signals were not definitively resolved to causal genes. Through joint-linkage (JL) analysis and GWAS in the US maize nested association (NAM) panel, Ziegler *et al.* (2017) identified six high confidence candidate genes underlying association signals for four elements (manganese,

Mn; molybdenum, Mo; phosphorus, P; and rubidium, Rb), but not all signals for these and other elements could be unambiguously mapped to single genes. Overall, the promise of GWAS for identifying the causal genes responsible for elemental accumulation in maize grain has yet to be fully realized, but efforts could be improved with the use of larger, more diverse association panels which have been densely genotyped.

When GWAS is employed to elucidate the molecular genetic basis of phenotypes, the significantly associated markers tend to be those in strong linkage-disequilibrium (LD) with causal loci of large effect (Myles *et al.* 2009). Therefore, if a phenotype is genetically controlled by mostly small-effect loci, the heritable fraction of a phenotype may not be completely explained by GWAS-detected loci alone. If this occurs, genomic prediction models that employ all available genome-wide markers to account for a range of small to large marker effects across the entire genome (*i.e.*, whole-genome prediction, WGP) could be used to improve trait prediction accuracy (Meuwissen *et al.* 2001; Gianola *et al.* 2009; de Los Campos *et al.* 2013). Furthermore, trained WGP models are used in genomic selection to increase genetic gain per unit of time when breeding for phenotypes having polygenic inheritance, as marker-assisted selection is better suited for Mendelian and oligogenic traits (Lorenz *et al.* 2011; Desta and Ortiz 2014; Owens *et al.* 2014). To date, WGP models have only been evaluated on elemental grain phenotypes of wheat (*Triticum aestivum* L.) (Velu *et al.* 2016; Manickavelu *et al.* 2017; Alomari *et al.* 2018) and only Zn for maize (Guo *et al.* 2020; Mageto *et al.* 2020).

In our study, a maize inbred association panel consisting of 1,813 individuals imputed with ~7.7 million SNP markers was used for the genetic dissection and prediction of natural

variation for elemental concentration in grain. The objectives of our study were to (i) assess the extent of phenotypic variation and heritability of 11 elemental grain phenotypes, (ii) conduct a GWAS to identify candidate causal genes controlling variation for 11 elemental phenotypes in maize grain, (iii) compare detected GWAS signals with genetic mapping results from the U.S. maize NAM panel, and (iv) evaluate the predictive abilities of two WGP models having different assumptions of the underlying genetic architecture for the elemental grain phenotypes.

MATERIALS AND METHODS

Plant materials and experimental design

We evaluated more than 2,400 maize inbred lines from the North Central Regional Plant Introduction Station association panel (hereafter, Ames panel) (Romay *et al.* 2013) at Purdue University's Agronomy Center for Research and Education in West Lafayette, IN, on Raub silt loam (fine-silty, mixed, superactive, and mesic Aquic Argiudolls) and Chalmers silty clay loam (fine-silty, mixed, superactive, and mesic Typic Endoaquolls) soils in 2 consecutive years (2012-2013). A single replicate of the entire experiment was grown in each of the two years following a design that has been previously described in Owens *et al.* (2019). Briefly, the maize inbred lines were partitioned into six sets according to their flowering time, with each set arranged as a 20 × 24 incomplete block design. Within a set, each incomplete block was augmented with the random positioning of a B73 plot (experiment-wide check) and two plots of a set-specific check. Experimental units were one-row plots that had a length of 3.81m, with ~15 plants per plot. The physiologically mature grain from the hand-harvested,

dried, and shelled self-pollinated ears (at most six) of each harvestable plot were bulked to generate a representative, composite sample for element analysis.

Phenotypic data analysis

We ground 4,406 grain samples weighing 10 g each from 2,177 inbred lines and a separate set of 11 repeated check lines with a Retsch ZM200 mill (Retsch, Haan, Germany). For inductively coupled plasma mass spectrometry (ICP-MS) analysis, ~0.3 g of each ground sample, which had been oven dried at 80°C for 4 h to remove remaining moisture, was acid-digested in a closed tube as described in Wheal *et al.* (2011). Elemental concentrations of samples were measured using ICP-MS (7500x; Agilent, Santa Clara, CA) according to the method of Palmer *et al.* (2014). The 18 quantified elements were aluminum (Al; for only monitoring contamination with soil), arsenic (As), boron (B), calcium (Ca), cadmium (Cd), cobalt (Co), copper (Cu), Fe, potassium (K), magnesium (Mg), Mn, Mo, sodium (Na), nickel (Ni), P, lead (Pb), selenium (Se), and Zn in $\mu g\ g^{-1}$ on a dry weight basis. In each of 10 digestion batches, a blank and a certified reference material (CRM; NIST 8433 corn bran) were added for quality assurance. Additionally, 6 to 7 experimental samples were replicated twice within each batch, allowing the assessment of technical (measurement) error. Technical replicate sample pairs with a relative standard deviation > 10% were removed, which resulted in the removal of three inbred lines. Samples (0.8%) with Al present at $> 5\ \mu g\ g^{-1}$ were considered to have unacceptable levels of purported soil contamination (Yasmin *et al.* 2014), thus resulting in the removal of an additional eight inbred lines from the dataset.

To improve the quality of the resultant dataset of 4,351 samples from the remaining 2,166 inbred lines and separate set of 11 repeated check lines, we assessed phenotypes for

missing values due to the limit of detection (LOD) for ICP-MS. The levels of Ca and Ni were below the LOD for 1.98% and 18.30% of samples, respectively. Separately for each of these two elements, a $\mu g\ g^{-1}$ value was approximated for the missing value of each of these samples by imputing a uniform random variable ranging from 0 to the minimum ICP-MS detection value for the given element within each year (Lubin *et al.* 2004; Lipka *et al.* 2013). Given the potential for biased results (Lubin *et al.* 2004), we excluded six elements (As, Cd, Co, Na, Pb, and Se) that had more than 70% of samples with a concentration below the LOD for ICP-MS.

We screened the generated dataset of 11 elemental phenotypes without missing values from the 2,166 inbred lines and separate set of 11 repeated check lines for significant outliers according to the procedure of Owens *et al.* (2019). Briefly, the full mixed linear model (equation 1) of Owens *et al.* (2019) was fitted for each elemental phenotype in ASReml-R version 3.0 (Gilmour *et al.* 2009). The model terms included check as a fixed effect and genotype (noncheck line), year, genotype-by-year (G × Y) interaction, set within year, plot grid row within year, incomplete block within set within year, and ICP-MS batch as random effects. Studentized deleted residuals (Neter *et al.* 1996) produced from these mixed linear models were examined to remove significant outlier observations for each phenotype after a Bonferroni correction for $\alpha = 0.05$. The variance component estimates generated by refitting the full model (Figure 3.1) for each outlier screened phenotype were used to calculate heritability on a line-mean basis (Holland *et al.* 2003; Hung *et al.* 2012), with the delta method (Lynch and Walsh, 1998; Holland *et al.* 2003) used to calculate their standard errors.

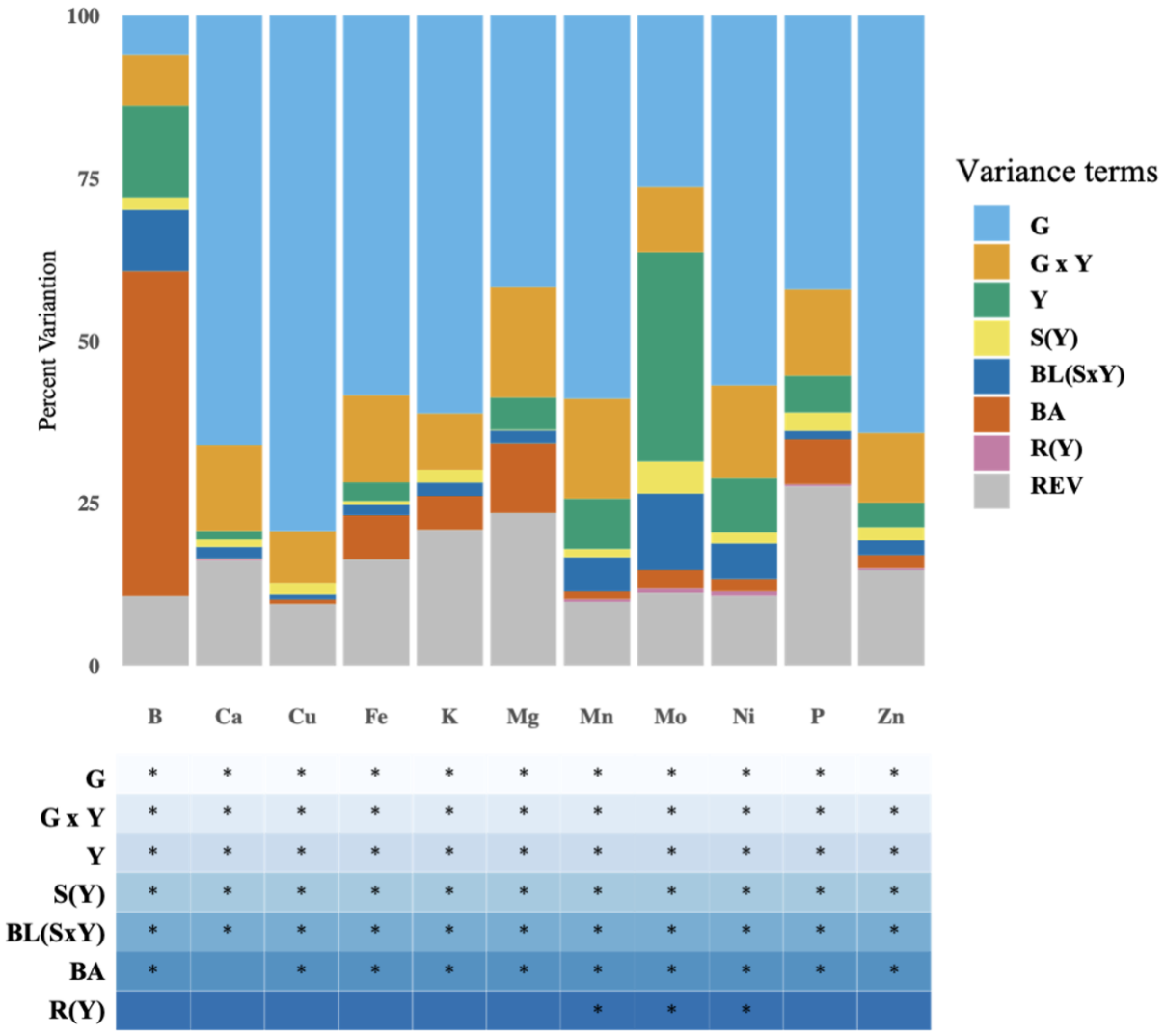

Figure 3.1. Sources of variation for 11 elemental grain phenotypes in the Ames panel. The phenotypic variance was statistically partitioned into the following components: genotype (G), genotype-by-year interaction (G×Y), year (Y), set within year [S(Y)], block within set within year [BL(S×Y)], inductively coupled plasma mass spectrometry (ICP-MS) batch (BA), row within year [R(Y)], and residual error variance (REV). Variance component estimates were calculated for all random effects from the full model equation 1 of Owens *et al.* (2019). The table below indicates which random effects were significant (*) according to a likelihood ratio test ($\alpha = 0.05$).

To generate best linear unbiased predictor (BLUP) values, a best fit model was selected for each outlier-screened phenotype through iteratively fitting the above full mixed linear model in ASReml-R version 3.0 and retention of only random effect terms found to be significant ($\alpha = 0.05$) in a likelihood ratio test (Littell *et al.* 2006). The final best fitted model was used to obtain a BLUP for each elemental phenotype for each inbred line (Supplemental Table S3.1). Given that elements are not always distributed evenly among seed tissues (*e.g.*, pericarp, endosperm, and embryo) and extreme grain phenotypes could have substantially altered elemental composition (Lombi *et al.* 2009, 2011; Pongrac *et al.* 2013; Baxter *et al.* 2014), 247 inbred lines classified according to Romay *et al.* (2013) and Germplasm Resources Information Network (GRIN; https://www.ars-grin.gov/) as sweet corn, popcorn, or with an endosperm mutation were conservatively removed from the dataset. All of the remaining 1,919 inbred lines had BLUP values for 10 or more of the 11 elemental phenotypes.

Genotype data processing and imputation

We used target and reference SNP genotype sets in B73 RefGen_v4 coordinates (B73 v4) to increase the marker density of the Ames panel with an approach similar to Ramstein *et al.* (2020). In brief, the raw genotypes of genotyping-by-sequencing (GBS) SNPs (ZeaGBSv27_publicSamples_raw_AGPv4-181023.h5, available on CyVerse at http://datacommons.cyverse.org/browse/iplant/home/shared/panzea/genotypes/GBS/v27) scored at 943,455 loci were obtained for the Ames panel from Romay *et al.* (2013), providing a total of 2,172 GBS samples with a call rate ≥ 20% from 1,839 of the 1,919 inbred lines with phenotypic data for constructing the target set. We initially used a stringent filtered dataset of 35,082 SNPs [call rate ≥ 50%, % heterozygosity ≤ 10%, index of panmixia $F_{IT} \geq 0.8$, and

linkage disequilibrium (LD) $r^2 \leq 0.2$] derived from the Romay *et al.* (2013) dataset to calculate pairwise identity by state (IBS) between multiple samples of the same "accession number" for each of 260 lines in PLINK version 1.9 (Purcell *et al.* 2007). This analysis resulted in the detection and removal of all samples of 23 inbred lines that had a mean IBS value < 0.95 for all within-line sample comparisons, producing a concordance-enhanced dataset of 2,098 GBS samples from 1,816 inbred lines that segregated for biallelic SNPs at 477,155 of the 943,455 SNP loci. To merge two or more GBS samples from the same line, SNP genotype calls with ≥ 50% occurrence were selected as the consensus genotype, whereas calls with < 50% occurrence were set to missing. After consensus-calling, 1,813 lines with a call rate ≥ 0.2, % heterozygosity ≤ 10%, and inbreeding coefficient (F) ≥ 0.8 were retained, which comprised the final set of lines used for downstream genetic analyses. Finally, heterozygous genotype calls were set to missing given their potential to be the result of paralogous alignments.

To construct the reference SNP genotype set, we used the maize HapMap 3.2.1 unimputed datasets (hmp321_agpv4_chrx.vcf.gz, where x is 1 to 10, available on CyVerse at https://datacommons.cyverse.org/browse/iplant/home/shared/panzea/hapmap3/hmp321/unimputed/uplifted_APGv4/) consisting of ~83 million variants identified from more than 1,200 lines (Bukowski *et al.* 2018) that included variants called from the higher coverage sequencing (average of ~7x) of the maize 282 (Goodman-Buckler) panel of Flint-Garcia *et al.* (2005). This dataset was processed in the following manner: selection of 14,613,169 SNPs [biallelic, call rate ≥ 50%, minor allele frequency (MAF) ≥ 1%, local LD flag present, and NI5 flag absent], heterozygous genotype calls set to missing, and imputation of all missing

SNP genotype data. With the resultant dataset serving as the reference panel, SNP genotypes at the 14,613,169 loci were imputed based on GBS SNPs (target set) in the final set of 1,813 lines from the Ames panel with BLUP values (Supplemental Table S3.1). All imputation was conducted in BEAGLE v5.0 (Browning *et al.* 2018) with 10 iterations for initial burn-in, 15 sampling interactions, an effective population size of 50,000 (Ross-Ibarra *et al.* 2009), and the U.S. maize NAM genetic linkage map (McMullen *et al.* 2009) (https://www.maizegdb.org/data_center/map) to provide further information on the recombination landscape. The quality of the imputed genotypes was further enhanced by retaining only biallelic SNPs with MAF ≥ 1% and predicted dosage r^2 (DR2) ≥ 0.80, resulting in 12,211,420 SNPs. In PLINK version 1.9 (Purcell *et al.* 2007), this SNP dataset was LD pruned with a sliding window of 100 SNPs and step size of 25 SNPs to construct datasets for the 1,813 lines that included only those SNPs with pairwise $r^2 < 0.99$ (7,719,799 SNPs for marker-trait association tests) or $r^2 < 0.10$ (361,302 SNPs for population structure and relatedness estimation).

Genome-wide association study

We conducted marker-trait associations at the genome-wide level as previously described in Owens *et al.* (2019). Briefly, to reduce heteroscedasticity and nonnormality of the residuals, the Box-Cox power transformation procedure (Box and Cox 1964) was invoked for each phenotype with an intercept-only model through the 'boxcox' function in MASS package version 7.3-50 in R version 3.5.1 (R Core Team 2018) that chose the optimal convenient lambda (Supplemental Table S3.2) for transformation of BLUP values (Supplemental Table S3.3). With a mixed linear model that used the population parameters previously determined

approximation (Zhang *et al.* 2010), each of the 7,719,799 SNP markers was tested for an association with transformed BLUP values of each phenotype from the 1,813 lines in the R package GAPIT version 2018.08.18 (Lipka *et al.* 2012). The fitted mixed linear models included principal components (PCs) and a genomic relationship matrix (kinship) to control for population structure and relatedness. In GAPIT, the 1,813 line × 361,302 SNP genotype matrix was used to calculate the kinship matrix with VanRaden's method 1 (VanRaden 2008) and PCs with the prcomp function from the R base package. The Bayesian information criterion (Schwarz 1978) was used to determine the optimal number of PCs for model inclusion. The amount of phenotypic variation explained by a SNP was approximated as the difference between the likelihood-ratio-based R^2 statistic (R^2_{LR}) (Sun *et al.* 2010) of a mixed linear model with or without a given SNP included. The Benjamini–Hochberg procedure (Benjamini and Hochberg 1995) was used to control the false discovery rate (FDR) at the 5% level for each phenotype.

To better clarify complex association signals, the multi-locus mixed-model (MLMM) approach (Segura *et al.* 2012) that employs forward-backward stepwise regression to sequentially add significant markers as fixed effects (covariates) was used to control for major-effect loci in genome-wide scans for marker-trait associations. The multiple-Bonferroni criterion (mBonf) was used to choose the optimal model. The statistical control of major-effect loci was further evaluated by reconducting GWAS with the MLMM-selected SNPs included as covariates in the mixed linear models fitted in GAPIT.

Population structure analysis

We classified the inbred lines of the Ames panel to better understand the allele frequency

patterns of associated SNPs across subpopulations. The 1,813 line × 361,302 SNP genotype matrix, which had been also used for a principal component analysis (PCA; see Genome-wide association study section of Materials and Methods), served as the input dataset for the estimation of population structure with fastSTRUCTURE (Raj *et al.* 2014). The number of ancestral populations (K) were varied from 1 to 10 with the simple prior when conducting the fastSTRUCTURE analysis. We selected K=3 as the number of subpopulations based on the collective evaluation of the fastSTRUCTURE and PCA results in combination with earlier findings on patterns of population structure in the Ames panel (Romay *et al.* 2013). The 1,813 lines were assigned to one of three subpopulations (SP1, SP2, or SP3) if they had an assignment value of $Q \geq 0.7$. If lines had assignment values of $Q < 0.7$ for all three subpopulations, they were considered to be admixed (Supplemental Table S3.4). The SP1, SP2, and SP3 subpopulations predominantly consisted of lines classified as nonstiff stalk (NSS), tropical, and stiff stalk (SS), respectively.

Candidate gene identification

To identify candidate genes, we first constructed a set of distinct loci significantly associated with the elemental phenotypes. A locus was defined as an association signal composed of at least two SNPs significant at 5% FDR within 100 kb from one another, with the most significant SNP designated as the peak marker at a locus. Estimates of pairwise LD (r^2) between a peak SNP and all SNPs within ± 5 Mb were calculated in TASSEL v5.2.49 (Bradbury *et al.* 2007). If two or more peak SNPs occurred within ± 5 Mb of each other, a locus was declared distinct if its peak SNP had an r^2 value < 0.2 with all other designated peak SNPs. The genomic search space to identify candidate genes was limited to within ± 100

kb of each peak SNP, given the rapid LD decay in this maize association panel (Romay *et al.* 2013). In addition, the candidate gene search process was partly informed by a curated list of genes involved in the accumulation of elements in plants (Whitt *et al.* 2020). The top three unique best hits of the nine most plausible candidate genes in Arabidopsis (Columbia-0 ecotype) and rice (*O. sativa* L. ssp. Japonica cv. "Nipponbare") with E-values < 1 were identified by BLASTP with default parameters at TAIR (https://www.arabidopsis.org) and RAP-DB (https://rapdb.dna.affrc.go.jp) databases, respectively (Supplemental Table S3.5).

Integration of genetic mapping results

The genetic mapping results from joint linkage (JL) analysis and GWAS of grain elemental phenotypes in the U.S. nested association mapping (NAM) panel (Ziegler *et al.* 2017) were joined with those generated from our GWAS in the Ames panel (Supplemental Tables S3.6 and S3.7). Given that the four field sites (New York, Florida, North Carolina, and Puerto Rico) included in the study of Ziegler *et al.* (2017) had climates and soil types different from those of the Indiana field site, we focused the comparative on NAM genetic mapping results based on BLUP phenotypes generated from a combined analysis of all four locations (All Locs). To accomplish this, first the markers used for JL analysis (SNPs) and GWAS (SNPs and small indels) in the NAM panel were uplifted from B73 RefGen_AGPv2 (B73 v2) to B73 RefGen_AGPv4 (B73 v4). To uplift markers, 50 nt flanking sequences (101 nt total) were clipped from each side of the marker position in the B73 v2 assembly, followed by alignment of the flanking sequences to the B73 v4 assembly through the use of Vmatch v2.3.0 (Kurtz, 2010) with the following options: -d -p -complete -h1. Alignments to B73 v4 were filtered to retain the highest scoring and unique alignment for each individual marker. Markers not

having a high confidence, unique alignment were discarded from the uplifted results. If markers defining the upper or lower bounds of a QTL support interval could not be uplifted to B73 v4, then the next closest outer SNP marker that could be uplifted was used so as not to compromise the calculated 95% support interval.

Whole-genome prediction

We evaluated two WGP models, Bayesian ridge regression (BRR) and BayesB (Habier *et al* 2011; Pérez and de los Campos 2014). The BGLR package version 1.0.8 (Pérez and de los Campos 2014) was used to implement the two WGP models for the transformed BLUP values of each phenotype from the 1,813 lines with a Markov chain Monte Carlo process as follows: 12,000 iterations with a burn-in of 4,000 and a thinning of 5. As a compromise between model run time and performance, the LD-pruned ($r^2 < 0.10$) dataset of 361,302 genome-wide SNPs was used for both computationally intensive WGP models with an expected minimal loss of information. A stratified five-fold cross-validation scheme that accounted for population structure was conducted 10 times for each of the 11 phenotypes, with predictive ability calculated as the mean Pearson's correlation of transformed BLUP values with genomic estimated breeding values across folds. Both models used the same fold assignments, and each fold had the same subpopulation (SP) proportion (SP1, SP2, SP3, and admixed) as calculated for the entire panel (Supplemental Table S3.4).

RESULTS

Phenotypic variation

On average, K, P, and Mg were the most abundant (> 1,000 μg g^{-1}) elements in grain from the Ames panel, followed by Ca at an almost two orders of magnitude lower average concentration (Table 3.1). For the other elements, the average concentrations of Zn and Fe were closest to Ca, whereas Mn, Cu, B, Mo, and Ni had average concentrations < 7 μg g^{-1}. The calculated correlations between the 11 elements ranged from essentially nonexistent (r < 0.01) between Mo and Ni to very strong (r = 0.70) between P and Mg (Supplemental Figure S3.1). Interestingly, we detected a strong correlation (r = 0.55) between Fe and Zn, which suggests that these two elements could have a partially shared genetic architecture. All 11 elements showed significant genotype-by-year interaction, but which accounted for only a small percentage of the total phenotypic variance (Figure 3.1). The 11 phenotypes had an average heritability of 0.70, with a range of 0.33 for B to 0.87 for Cu (Table 3.1). Even though these phenotypes were influenced by the environment, our results indicate that the exhibited phenotypic variation was mostly attributable to genetic variation among inbred lines.

Genome-wide association study

A GWAS was conducted for the 11 elemental phenotypes with 1,813 lines of the Ames panel imputed with ~7.7 million SNPs. Collectively, 1,917 significant marker-trait associations were detected for B, Cu, Mn, Mo, Ni, and Zn, but none were found for Ca, Fe, Mg, K, and P at a genome-wide FDR of 5% (Figure 3.2, Table 3.2, and Supplemental Figure S3.2).

Examination of local LD patterns resolved the 1,917 marker-trait associations into a robust set of 33 loci (Supplemental Table S3.8). The search space for candidate genes was defined as ± 100 kb of the most significant SNP (*i.e.*, peak SNP) at each of the 33 loci, a window size considerate of high marker density, wide variance in rapid rate of LD decay (mean r^2 of 0.2 within ~1-10 kb) in the panel (Romay *et al.* 2013), and distant cis-regulatory variants (Salvi *et al.* 2007; Studer *et al.* 2011; Wallace *et al.* 2014; Rodgers-Melnick *et al.* 2016; Huang *et al.* 2018; Ricci *et al.* 2019).

Table 3.1. Means, ranges, and standard deviations (Std. Dev.) of untransformed BLUP values (in μg g^{-1}) for 11 elemental grain phenotypes evaluated in the Ames panel and estimated heritability on a line-mean basis and their standard errors (Std. Err.) across two years.

Phenotype	Number of lines	BLUPs			Heritabilities	
		Mean	Range	Std. Dev.	Estimate	Std. Err.
B	1812	2.19	1.59 - 3.09	0.21	0.33	0.03
Ca	1813	39.72	8.60 - 121.08	12.7	0.77	0.01
Cu	1812	2.32	0.91 - 5.75	0.68	0.87	0.01
Fe	1810	23.59	14.62 - 36.33	3.29	0.75	0.01
K	1813	4435.72	2944.20 - 6671.02	431.62	0.76	0.01
Mg	1813	1334.16	955.20 - 1814.08	115.97	0.61	0.02
Mn	1812	6.12	2.38 - 11.69	1.44	0.78	0.01
Mo	1812	0.49	0.29 - 0.85	0.07	0.65	0.02
Ni	1809	0.23	-0.04 - 1.12	0.14	0.77	0.01
P	1813	3298.76	2453.00 - 4341.12	277.24	0.61	0.02
Zn	1813	30.68	12.59 - 52.32	4.36	0.79	0.01

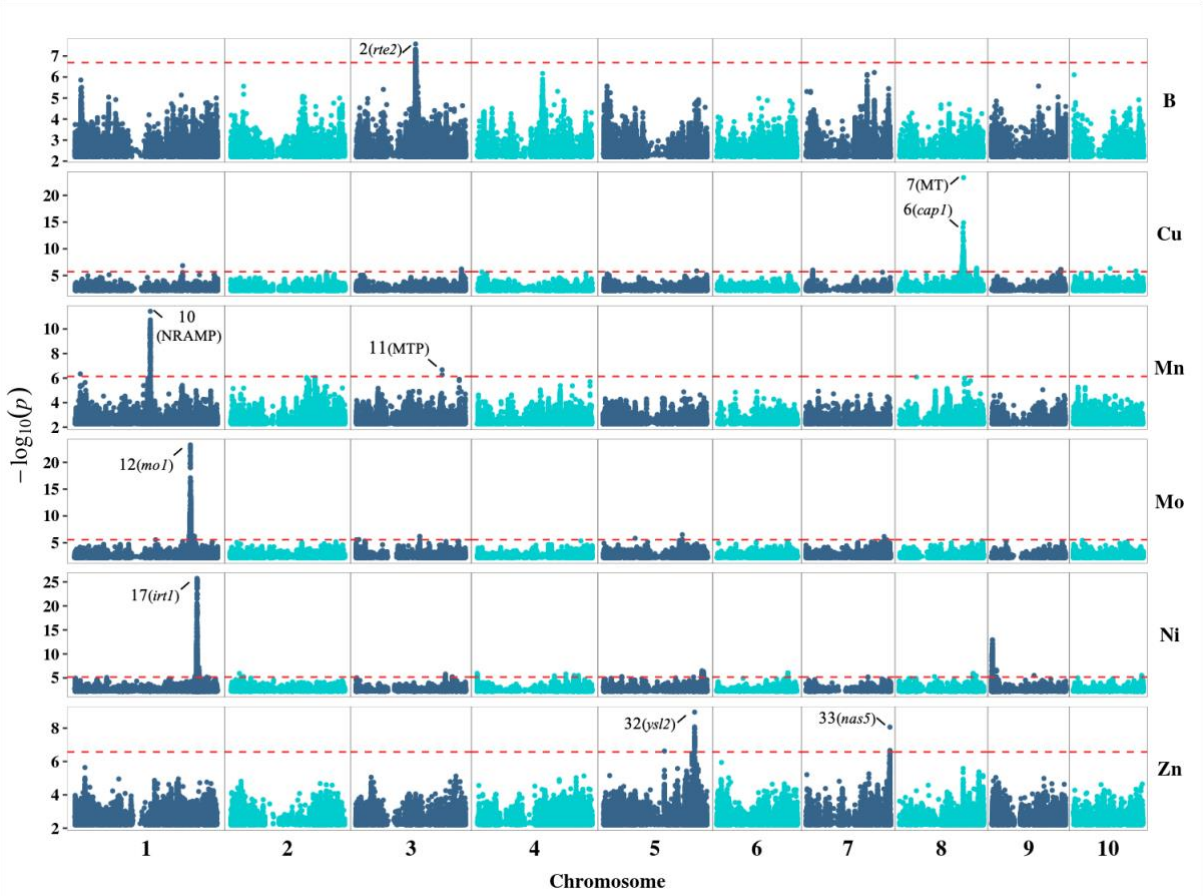

Figure 3.2. Manhattan plot of results from a genome-wide association study of the six elemental grain phenotypes with significant associations at the 5% false discovery rate (FDR) level in the Ames panel. Each point represents a SNP with its -log$_{10}$ *P*-value (y-axis) from a mixed linear model analysis plotted as a function of physical position (B73 RefGen_v4) across the 10 chromosomes of maize (x-axis). The red horizontal dashed line indicates the -log$_{10}$ *P*-value of the least statistically significant SNP at 5% FDR. The most probable candidate genes within ± 100 kb of the most significant SNP (*i.e.*, peak SNP) of each numbered locus are labeled above their corresponding association signals.

Table 3.2. Most plausible candidate genes identified through a genome-wide association study of 11 elemental phenotypes in grain from the Ames panel.

Phenotype	Locus number	SNP ID[a]	*P*-value	FDR-adjusted *P*-value	SNP R^{2b}	Gene ID	Annotated gene function
B	2	3-128693026	6.47E–08	4.59E–02	0.01	Zm00001d041590	B transporter (*rte2*)
Cu	6	8-136857539	9.10E–15	2.34E–08	0.03	Zm00001d011013	Ca transporter (*cap1*)
Cu	7	8-137939692	4.69E–24	3.62E–17	0.04	Zm00001d011063	Metal chelator (MT)
Mn	10	1-162962818	3.61E–12	2.79E–05	0.02	Zm00001d030846	Metal transporter (NRAMP)
Mn	11	3-184559931	2.11E–07	1.71E–02	0.01	Zm00001d042939	Metal transporter (MTP)
Mo	12	1-248672716	5.58E–24	4.31E–17	0.04	Zm00001d033053	Mo transporter (MOT)
Ni	17	1-262893725	1.98E–26	6.75E–20	0.05	Zm00001d033446	Metal transporter (ZIP; *irt1*)
Zn	32	5-195765640	1.10E–09	8.51E–03	0.02	Zm00001d017427	Metal-NA transporter (YSL; *ysl2*)
Zn	33	7-179962589	8.67E–09	1.75E–02	0.01	Zm00001d022557	Metal chelator (NAS; *nas5*)
Fe		7-180077496	1.06E–07	1.53E–01	0.01	Zm00001d022557	Metal chelator (NAS; *nas5*)

[a]SNP ID nomenclature consists of chromosome number, followed by physical position (bp) in B73 RefGen_v4 coordinates
[b]SNP R^2 is calculated as follows: R^2 likelihood ratio of model with SNP minus R^2 likelihood ratio of model without SNP (Supplemental Table S3.8)

The two loci significantly associated with B comprised a mildly complex association signal spanning from 127.4 to 128.7 Mb on chromosome 3 (Figures 3.2 and 3.3; Supplemental Table S3.8). The peak SNP of each locus (locus 1: 3-127841465, *P*-value 2.68E–08; locus 2: 3-128693026, *P*-value 6.47E–08) was separated by a physical distance of ~851 kb, with virtually no LD (r^2 = 0.03) between them. The peak SNP of the second locus, 3-128693026, was located ~59 kb from the open reading frame (ORF) of the *rotten ear2* (*rte2*) gene (Zm00001d041590) encoding a B efflux transporter (Chatterjee *et al.* 2017).

The peak SNPs for the strongest two of five association signals for Cu on chromosome 8 (Figure 3.2) were separated by a physical distance of ~1.1 Mb and in weak LD (r^2 = 0.15) with each other. Of the two, the more significant peak SNP (8-137939692; *P*-value 4.69E–24) was located within a gene (Zm00001d011063) (Supplemental Figure S3.3) coding for a protein 43-60% identical at the amino acid sequence level to three type 2 metallothioneins (MTs) in rice (Supplemental Table S3.5) (Zhou *et al.* 2006; Kumar *et al.* 2012). Members of the plant MT family are low-molecular weight, cysteine-rich metal-binding proteins and of which some have been shown to bind Cu and other metal ions (Guo *et al.* 2008; Benatti *et al.* 2014).

The least significant of the two peak SNPs (8-136857539; *P*-value 9.10E–15) resided within the *calcium pump1* (*cap1*) gene (Zm00001d011013) (Supplemental Figure S3.4) encoding a calmodulin-regulated P-type Ca^{2+}-ATPase that had been shown to have slightly enhanced mRNA expression in maize roots under anoxic conditions (Subbaiah and Sachs 2000). Although plausible, to our knowledge it had never been reported to transport Cu. The peak SNPs for the other three loci (3-5) on chromosome 8, as well as the two loci (8-9) on

chromosomes 3 and 7 were within ± 100 kb of candidate genes (Supplemental Table S3.9) less likely to be involved in Cu chelation or transport.

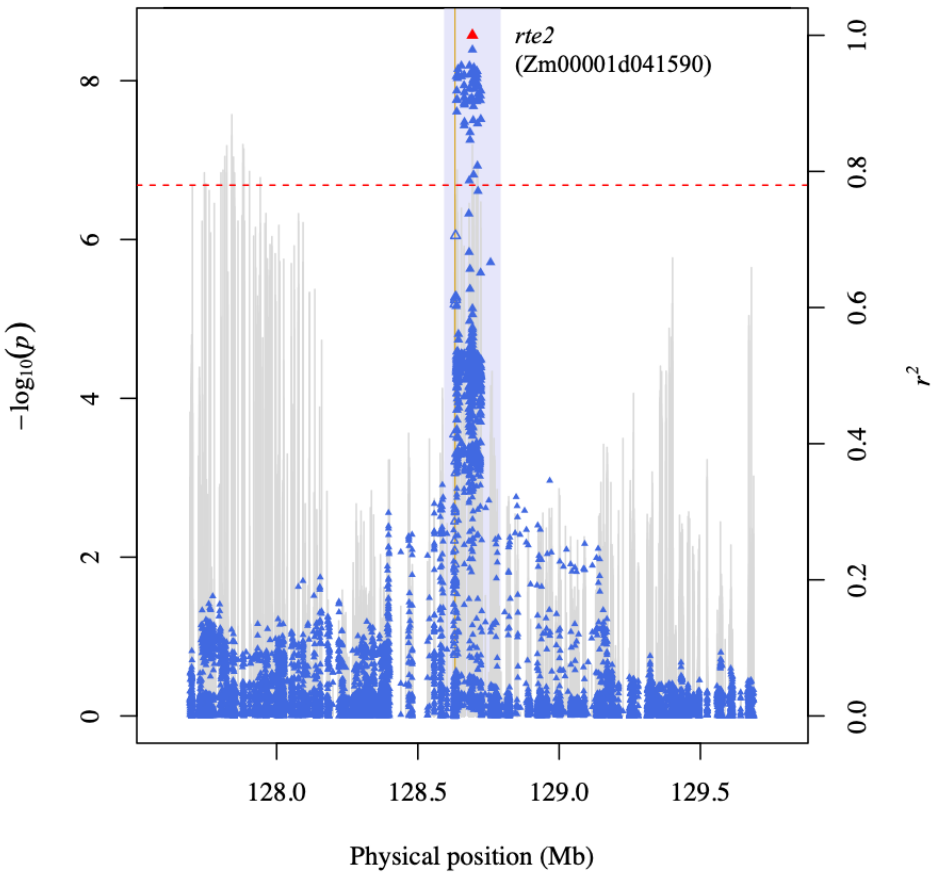

Figure 3.3. A regional Manhattan plot of locus 2. Scatter plot of association results from a mixed model analysis of B grain concentration and linkage disequilibrium (LD) estimates (r^2) for a genomic region that contains the peak SNP (3-128693026) at locus 2. Each vertical line represents the -$\log_{10}$ *P*-value of a SNP. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 128,693,026 bp (B73 RefGen_v4) on chromosome 3. The red horizontal dashed line indicates the -$\log_{10}$ *P*-value of the least statistically significant SNP at a genome-wide false discovery rate of 5%. The yellow vertical line indicates the genomic position of the *rotten ear2* (*rte2*) gene Zm00001d041590. The open triangles indicate SNPs that are within the candidate gene. The light blue rectangle demarcates the ± 100 kb candidate gene search space surrounding the peak SNP.

Of the two loci associated with Mn (Figure 3.2), the strongest signal was located 162.9 to 163.2Mb on chromosome 1 (Supplemental Figure S3.5). The peak SNP (1-162962818, *P*-value 3.61E–12) of this locus resided about 2.2kb from a gene (Zm00001d030846) encoding a protein with 74% and 72% sequence identity to NRAMP3 and NRAMP4 of Arabidopsis (Supplemental Table S3.5) that in addition to Fe, export Mn from vacuoles to chloroplasts in leaf mesophyll cells (Lanquar *et al.* 2005, 2010). An additional four SNPs within this gene were significantly associated (*P*-values 7.48E–11 to 3.18E–10) with Mn and in very strong LD ($r^2 > 0.90$) with the peak SNP.

The weaker effect locus at ~184.6Mb on chromosome 3 (Supplemental Figure S3.6) for Mn was defined by two significant SNPs. Both SNPs were in very strong LD ($r^2 = 0.79$) with one another. The peak (3-184559931; *P*-value 2.11E–07) and second SNPs (3-184590243; *P*-value 5.46E–07) were approximately 29 and 0.78kb, respectively, from a gene (Zm00001d042939) that codes for a protein with 80% sequence identity to METAL TOLERANCE PROTEIN 11 (MTP11) of Arabidopsis (Supplemental Table S3.5) that transports Mn^{2+} (Delhaize *et al.* 2007).

The strongest signal for Mo spanned from 246.5 to 250.3Mb on chromosome 1 (Supplemental Figure S3.7), with the peak SNP (1-248672716; *P*-value 5.58E–24) ~71 kb from the *molybdate transporter1* (*mo1*) gene (Zm00001d033053) that codes for a protein having 69% sequence identity to the mitochondrial-localized MOLYBDATE TRANSPORTER 1 (MOT1) from Arabidopsis (Supplemental Table S3.5) that specifically transports Mo (Tomatsu *et al.* 2007; Baxter *et al.* 2008). Furthermore, the peak SNP was in very strong LD ($r^2 = 0.95$) with a highly significant SNP (*P*-value 1.03E–21) located within

the gene. The other four loci (13-16) collectively consisted of 10 significant SNPs across three chromosomes but were within ± 100 kb of less probable candidate genes (Supplemental Table S3.9).

Of the 15 loci associated with Ni, the strongest signal mapped from 261.8 to 263.3 Mb on chromosome 1 (Figure 3.2). The peak SNP (1-262893725, *P*-value 1.98E–26) at this locus (Figure 3.4) was located ~82 kb from the *iron-regulated transporter1* (*irt1*) gene (Zm00001d033446) (Mondal *et al.* 2014), which encodes a protein sharing amino acid sequence similarity (55-57% identical) with members of the ZIP transporter family in Arabidopsis that transport a variety of divalent metal ions including Ni^{2+} (Vert *et al.* 2009; Nishida *et al.* 2011; Li *et al.* 2019). In addition, 15 significant SNPs within *irt1* were associated with Ni and, on average, were in strong LD (mean r^2 of 0.52) with the peak SNP. The peak SNPs for the other 14 Ni-associated loci (18-31), however, were within ± 100 kb of candidate genes (Supplemental Table S3.9) with more speculative involvement in Ni accumulation.

Two significant SNPs comprised the locus associated with Zn at ~179.9Mb on chromosome 7 (Supplemental Table S3.8). Also, these two SNPs were in moderately strong LD (r^2 = 0.69) with each other. Of these two SNPs, the peak SNP (7-179962589; *P*-value 8.67E–09) was nearest (~1.9 kb) to the *nicotianamine synthase5* (*nas5*) gene (Zm00001d022557) that codes for a class II NAS purportedly involved in synthesizing the metal ion chelator nicotianamine (Zhou *et al.* 2013). Notably, a weaker association signal significant at 15% FDR was identified for Fe with a peak SNP (7-180077496; *P*-value 1.06E–07) at a distance of ~112 kb from *nas5* (Supplemental Figure S3.8). The minor allele of each

peak SNP, which was associated with a higher mean concentration of either Zn or Fe, occurred at very low frequencies in the tropical (~5%) and Stiff Stalk (~1%) subpopulations (Supplemental Table S3.10).

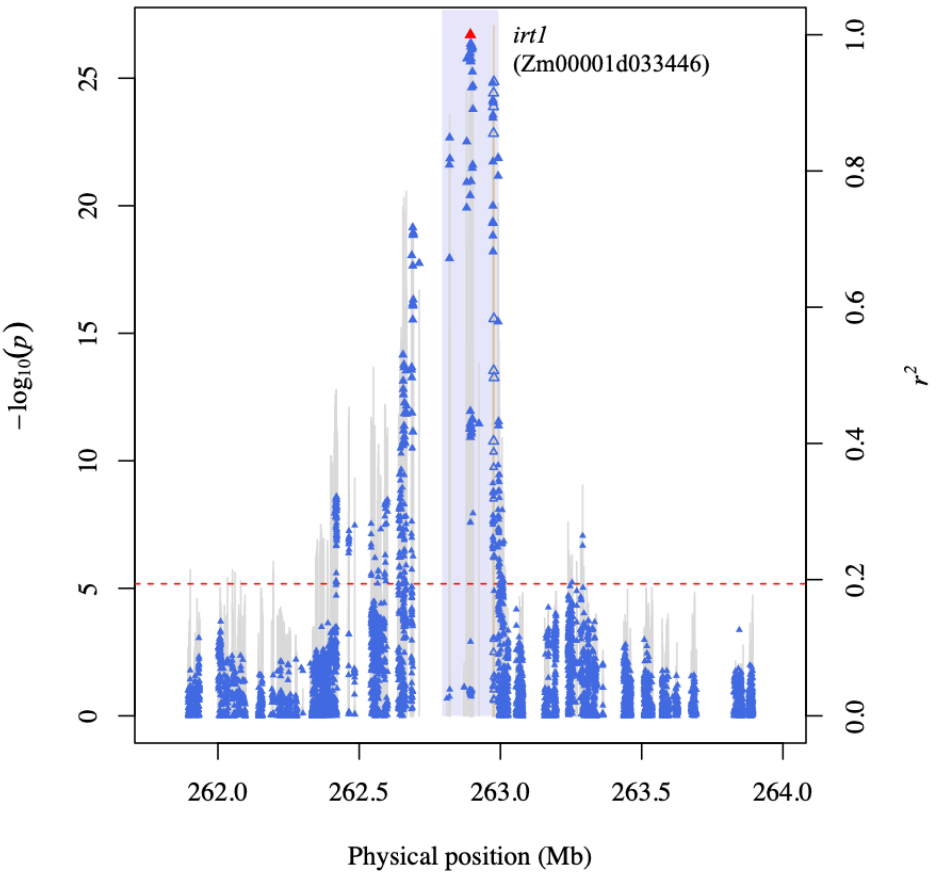

Figure 3.4. A regional Manhattan plot of locus 17. Scatter plot of association results from a mixed model analysis of Ni grain concentration and linkage disequilibrium (LD) estimates (r^2) for a genomic region that contains the peak SNP (1-262893725) at locus 17. Each vertical line represents the -$\log_{10}$ *P*-value of a SNP. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 262,893,725 bp (B73 RefGen_v4) on chromosome 1. The red horizontal dashed line indicates the -$\log_{10}$ *P*-value of the least statistically significant SNP at a genome-wide false discovery rate of 5%. The yellow vertical line indicates the genomic position of the *iron-regulated transporter1* (*irt1*) gene Zm00001d033446. The open triangles indicate SNPs that are within the candidate gene. The light blue rectangle demarcates the ± 100 kb candidate gene search space surrounding the peak SNP.

On chromosome 5, the association signal for Zn ranged from 195.6 to 195.8Mb (Supplemental Figure S3.9), with the peak SNP (5-195765640; *P*-value 1.10E–09) only 1.2 kb from the *yellow stripe-like2* (*ysl2*) gene (Zm00001d017427). The protein encoded by *ysl2* has 58% and 63% sequence identity with AtYSL1 and AtYSL3 (Yordem *et al.* 2011) that transport metal-nicotianamine complexes to various Arabidopsis plant tissues (Waters *et al.* 2006). Also, three significant SNPs (*P*-values 7.72E–08 to 1.38E–07) within this gene were in moderately strong LD (mean r^2 of 0.39) with the peak SNP.

Clarification of association signals to identify the largest-effect loci

The MLMM approach, which helps resolve complex association signals, selected one or more SNPs for Cu, Mn, Mo, Ni, and Zn, but none for the presumably weaker signals of the other six elements (Supplemental Table S3.11). The top one or two most significant peak SNP markers that had been detected with the mixed linear model in GAPIT were selected by the MLMM for Cu (locus 7; MT), Mn (locus 10; NRAMP), Mo (locus 12; *mo1*), Ni (locus 17; *irt1*), and Zn (locus 32, *ysl2*; locus 33, *nas5*). Furthermore, the MLMM had selected an additional two SNPs, 9-1875785 (locus 29) and 9-745061 (locus 28), for Ni. When a conditional mixed linear model analysis was conducted separately for Cu, Mn, Mo, Ni, and Zn with their respective MLMM-selected SNPs as covariates, there were no remaining significant SNPs at a genome-wide 5% FDR (Supplemental Figure S3.10). This suggests the MLMM had identified and conditional models had accounted for the major loci contributing to phenotypic variation.

Comparison of genetic mapping results

We assessed the findings of our study through a comparison involving the JL-QTL analysis and GWAS results of these elemental grain phenotypes in the U.S. maize NAM panel (Supplemental Tables S3.6 and S3.7) (Ziegler *et al.* 2017). Of the identified candidate genes, *rte2* and *irt1* were novel associations, whereas the other seven candidate genes were coincident with NAM JL-QTL (Ziegler *et al.* 2017). However, *cap1*, *mo1*, Zm00001d011063 (MT), Zm00001d030846 (NRAMP), *nas5* for Zn, and *ysl2* were distant from their respective NAM GWAS peak variants within JL-QTL support intervals (average: ~1.68Mb; range: 0.356 - 4.67Mb), whereas *nas5* for Fe and Zm00001d042939 (MTP) were 126.9 and 0.203kb, respectively, from their closest peak NAM GWAS variant. The joint consideration of GWAS results suggests that the large-effect loci associated with natural variation for the six grain elements in the NAM panel were all resolved to the level of highly probable causal genes in the Ames panel.

Whole-genome prediction

We evaluated the predictive ability of WGP for the 11 elemental phenotypes with two models that have different assumptions about the distribution of underlying genetic effects, BRR and BayesB (Habier *et al.* 2011; Pérez and de los Campos 2014). On average, BRR and BayesB had nearly identical prediction abilities of 0.45 and 0.46, respectively, across the 11 phenotypes (Table 3.3). As expected, given the results of Combs and Bernardo (2013), the predictive abilities of both WGP models were strongly correlated with the heritabilities of all phenotypes (BRR, $r = 0.66$, P-value < 0.05; BayesB, $r = 0.65$, P-value < 0.05). While the

predictive abilities from both models were essentially equivalent for most phenotypes, the predictive abilities of Ni, Mo, and Cu increased by 10.42%, 4.00%, and 3.92%, respectively, with BayesB relative to BRR.

Table 3.3. Predictive abilities of 11 elemental grain phenotypes of the Ames panel from Bayesian ridge regression (BRR) and BayesB models.

Phenotype	BRR		BayesB	
	Predictive ability	Std. Dev.	Predictive ability	Std. Dev.
B	0.33	0.01	0.33	0.01
Ca	0.47	0.01	0.47	0.01
Cu	0.51	0.01	0.53	0.01
Fe	0.46	0.01	0.46	0.01
K	0.34	0.01	0.34	0.01
Mg	0.45	0.01	0.45	0.01
Mn	0.50	0.01	0.50	0.01
Mo	0.50	0.01	0.52	0.01
Ni	0.48	0.01	0.53	0.01
P	0.40	0.01	0.40	0.01
Zn	0.50	0.01	0.50	0.01

DISCUSSION

Elemental homeostasis is critically important, with prolonged deficiencies or toxicities in essential elements negatively affecting plants (Marschner 2011). To date, the identification of causal genes via GWAS has mostly centered on elemental levels in roots and shoots for model and crop plant species (Huang and Salt 2016; Yang *et al.* 2018), thus the prioritization of candidate genes contributing to elemental accumulation in grain of staple crops remains an important pursuit. To this end, we conducted GWAS on the concentrations of 11 elements in grain from ~2,000 lines of the maize Ames panel imputed with ~7.7 million SNP markers. By

leveraging the tremendous genetic diversity and rapid intragenic LD decay of this powerful genetic resource, we identified nine candidate genes encoding proteins with functions relevant to the accumulation of elements in maize grain. We also demonstrated moderate prediction abilities for the 11 elements with two different WGP models, which is especially relevant for Fe and Zn biofortification of maize grain (Welch and Graham 2002).

Novel loci associate with B and Ni

We detected novel associations of *rte2* and *irt1* with levels of B and Ni in maize grain, respectively. The *rte2* gene, coding for a B efflux transporter, is one of six members of a small gene family (*rte1-6*) (Chatterjee *et al.* 2017). Even though the duplicated *rte1* and *rte2* genes were reported to have contrasting tissue-specific expression patterns across maize reproductive and root tissues, it was also shown that they work in concert to provide B for maize plants growing in B-depleted soils (Chatterjee *et al.* 2017). It is possible that *rte1* (maize1 subgenome) and *rte2* (maize2 subgenome) functionally diverged following the most recent tetraploidization event for the maize genome (Schnable *et al.* 2011), potentially explaining why not even a very weak association signal was detected with B at *rte1*. Given that *rte2* has high sequence similarity to the class I B transporters of Arabidopsis and rice (Miwa *et al.* 2006; Nakagawa *et al.* 2007; Miwa and Fujiwara 2010; Chatterjee *et al.* 2014, 2017), we hypothesize in our study that *rte2* had an indirect involvement in the accumulation of B in grain by playing a role in xylem loading of B.

The *irt1* gene, which underpinned an association signal for Ni on chromosome 1, is in the maize gene family with sequence similarity to the ZIP family of transporters (Mondal *et*

al. 2014) that transport Fe, Zn, and other divalent metal ions in other plants (Eide *et al.* 1996; Grotz *et al.* 1998; Korshunova *et al.* 1999; Li *et al.* 2019). AtIRT1, which is 55% identical in amino acid sequence to ZmIRT1 (Mondal *et al.* 2014), is a plasma membrane protein demonstrated to be critical for Fe^{2+} uptake inside Arabidopsis root epidermal cells (Vert *et al.* 2002), but also showed to have enhanced activity as a transporter of Ni^{2+} in Arabidopsis roots under Ni excess conditions (Nishida *et al.* 2011, 2015). Li *et al.* (2015) reported that overexpression of *Zmirt1* in Arabidopsis produced higher concentrations of Fe and Zn in roots and seeds. Therefore, we speculate that *irt1* contributed to Ni accumulation in maize grain as a metal transporter with a yet to be characterized broader range of specificity that includes Ni^{2+}.

Higher mapping resolution afforded by the Ames panel

The other seven identified candidate genes co-localized with NAM JL-QTL and GWAS signals. With the exception of Zm00001d042939 (MTP), they were more finely mapped in the Ames panel than in the U.S. NAM panel. The proteins with the highest identity (80% and 93%) to Zm00001d042939 in Arabidopsis (AtMTP11) and rice (OsMTP11) (Supplemental Table S3.5) are Golgi-localized Mn transporters involved in conferral of Mn tolerance by a mechanism hypothesized to involve one or both of vesicular transport to the vacuole or extracellular secretion (Delhaize *et al.* 2007; Peiter *et al.* 2007; Farthing *et al.* 2017; Zhang and Liu 2017; Tsunemitsu *et al.* 2018). Notably, through a GWAS in a sorghum association panel, a syntenic ortholog (Sobic.003G349200) of Zm00001d042939 (Y. Zhang *et al.* 2017) was implicated in the genetic control of Mn grain levels (Shakoor *et al.* 2016). Although

different in cellular function and localization compared to MTP11, Zm00001d030846 (NRAMP), a member of a largely uncharacterized maize gene family (Jin *et al.* 2015), encodes a protein closely related to the multispecific metal transporters AtNRAMP3 and AtNRAMP4 in Arabidopsis. In addition to their roles as vacuolar iron effluxers, these two NRAMP proteins were shown by Lanquar *et al.* (2005, 2010) to be functionally redundant vacuolar membrane-localized transporters involved in the export of Mn to the cytosol from the vacuole of mature leaf mesophyll cells in Arabidopsis.

The *mo1* gene, inferred to be orthologous to MOT1 proteins in Arabidopsis and rice (Supplemental Table S3.5), underlied the Mo association signal on chromosome 1. *AtMOT1* was the first cloned and characterized Mo-specific transporter in plants (Tomatsu *et al.* 2007; Baxter *et al.* 2008) and hypothesized to regulate Mo content (Baxter *et al.* 2008). Complementation studies with Arabidopsis ecotypes also showed that natural allelic variants of *AtMOT1* altered shoot Mo content (Baxter *et al.* 2008). Comparatively, a QTL identified for the genetic control of shoot and grain Mo concentration in a rice mapping population was fine mapped to a molybdate transporter (*OsMOT1;1*), with the Mo transport activity of this causal gene confirmed via genetic and transgenic complementation (Huang *et al.* 2019). Furthermore, Huang *et al.* (2019) showed that a knockout of *OsMOT1;1*, a gene shown to have strong root expression, produced lower levels of Mo in grain, resulting likely from lower root-to-shoot translocation of Mo.

The stronger of two association signals for Cu on chromosome 8 was underlain by the candidate Zm00001d011063, which encodes an uncharacterized protein possessing weak amino acid sequence similarity to Arabidopsis MTs (Supplemental Table S3.5) that are

involved in homeostasis and remobilization of Cu (Benatti *et al.* 2014). Although not yet implicated in Cu accumulation, the rice protein with the highest sequence identity to Zm00001d011063, *OSMT2b* (also named as *OsMT-I-2c*) (Supplemental Table S3.5), had altered transcript abundance in rice shoot and root seedling tissues after Cu treatment (Yuan *et al.* 2008). The second genetically distinct signal coincided with *cap1*, a gene that codes for a calmodulin regulated P-type Ca^{2+}-ATPase (Subbaiah and Sachs 2000). The CAP1 protein is 80% identical in sequence to ECA1 in Arabidopsis (Supplemental Table S3.5), which is an ER-localized P_{2A}-type Ca^{2+}-ATPase reported to transport Ca^{2+}, Mn^{2+}, and potentially other divalent cations in root cells (Wu *et al.* 2002). This is a somewhat unexpected but still plausible finding, given that heavy metal P_{1B}-type ATPase subfamily members from Arabidopsis and rice have demonstrated Cu transport activity (Hirayama *et al.* 1999; Andrés-Colás *et al.* 2006; Kobayashi *et al.* 2008; Deng *et al.* 2013; Huang *et al.* 2016).

A key step toward the biofortification of maize grain

Suggestive of a pleiotropic locus for two correlated phenotypes, *nas5* underpinned the coincident association signals for Fe and Zn on chromosome 7. This gene family member encodes a class II NAS putatively responsible for synthesizing nicotianamine–a divalent metal chelator responsible for the internal transport of trace metals including Fe and Zn (reviewed in Curie *et al.* 2009; Schuler *et al.* 2012). Nicotianamine is also a precursor for producing mugineic acid family phytosiderophores exuded by roots of graminaceous plants to facilitate Fe uptake (reviewed in Curie *et al.* 2009; Swamy *et al.* 2016). In particular, activation tagging of *OsNAS3*, the rice protein with the highest sequence identity to *nas5* (Zhou *et al.* 2013),

resulted in higher nicotianamine levels that led to increased Fe and Zn in rice grain (Lee *et al.* 2009). In maize, *nas5* was found to be strongly expressed in stems and induced under excessive Fe and Zn conditions, suggesting its more localized involvement in homeostasis and transport, but this has yet to be extensively characterized (Zhou *et al.* 2013). Nonetheless, the identification of SNPs tagging the low frequency *nas5* alleles associated with increasing Fe or Zn grain levels is a key step towards facilitating biofortification of tropical maize. Many people with deficiencies for both of these elements subsist predominantly on maize grain in developing nations (Welch 2002; Welch and Graham 2004).

The *ysl2* gene associated with Zn on chromosome 5 encodes a protein with amino acid sequence similarity to the YSL family of transporters that uptake metal-phytosiderophores or metal-nicotianamine complexes (reviewed in Curie *et al.* 2009). Of the three Arabidopsis proteins (AtYSL1-3) with high sequence identity to *ysl2*, AtYSL1 and AtYSL3 (Yordem *et al.* 2011) were both implicated in the remobilization of Zn from senescing leaves as a complex with nicotianamine to developing seeds (Waters *et al.* 2006). Recently, Zang *et al.* (2020) showed that ZmYSL2 is a metal-nicotianamine transporter involved in the transport of Fe from the endosperm to embryo in the developing maize grain, but importantly they also implicated ZmYSL2 in the transport of Zn. Interestingly, the *ys1* gene that encodes a Fe(III)-PS transporter (Curie *et al.* 2001), the gene family member most closely related to *ysl2* (Yordem *et al.* 2011), was ~68 kb from the peak SNP for Zn on maize chromosome 5, but has contradictory support as a key contributor for Zn uptake or allocation (Wiren *et al.* 1996; Roberts *et al.* 2004; Schaaf *et al.* 2004; Chan-Rodriguez and Walker 2018). Therefore, *ys1* and *ysl2* merit joint consideration in future fine mapping and mutagenesis studies to more

conclusively determine their independent or collective contribution to Zn accumulation in grain.

Generalizability of genetic mapping results

Importantly, our GWAS findings for all 11 elemental traits may not be generalizable beyond the Ames panel itself or where it was grown. As an example, the number of JL-QTL detected by Ziegler *et al.* (2017) for each of the 11 elemental grain phenotypes ranged from 3 (B) to 17 (Mn) with varied effect sizes (R^2 = 0.8 to 37.6%) in the U.S. NAM panel that affords higher statistical power (Yu *et al.* 2008). The findings of Ziegler *et al.* (2017) include a total of 11, 5, 12, and 11 JL-QTL identified for Ca, K, Mg, and P, respectively, which are the four elements that lacked significant associations in the Ames panel (Supplemental Table S3.6). Although these detected genetic differences could be attributed to environmental factors that influence elemental accumulation in grain (Ziegler *et al.* 2017), it is also possible that in the Ames panel the genetic architecture for each of these high concentration macroelements is predominated by rare variants of weak to modest effect, thus limiting their detectability even with a high density of SNP markers used in GWAS. Despite the genetic differences between association panels for macronutrients, seven of the nine candidate causal genes identified for micronutrients in the Ames panel co-localized with NAM JL-QTL and GWAS signals, thus these genetic signals are more likely to be reproduced in further independent genetic mapping panels and environments.

Informing whole-genome prediction with genetic mapping results

We conducted WGP of grain elemental phenotypes in maize, resulting in, on average,

moderate predictive abilities from BRR (0.45) and BayesB (0.46) across all phenotypes that are comparable to those obtained for elemental phenotypes in wheat grain (Velu *et al.* 2016; Manickavelu *et al.* 2017; Alomari *et al.* 2018) and for Zn in maize grain (Guo *et al.* 2020; Mageto *et al.* 2020). For Ni, Mo, and Cu, the BayesB model that allows for a few of many genome-wide markers to have large effects (Meuwissen *et al.* 2001; Gianola *et al.* 2009; de Los Campos *et al.* 2013) modestly outperformed (3.92-10.42%; Table 3.3) the BRR model with homogeneous shrinkage across all markers (Gianola 2013; de Los Campos *et al.* 2013). These three elements had the highest number of associated loci (5 to 15) and largest amount of phenotypic variation explained by peak SNPs (3 to 5%) tagging a candidate causal gene (Figure 3.2; Table 3.3), implying that BayesB could better fit the genetic architecture of Ni, Mo, and Cu (de Los Campos *et al.* 2013). Taken together, our GWAS-informed WGP results provide a foundational framework for exploring the additional modeling of identified large-effect loci when conducting genomic selection of elemental grain phenotypes in maize breeding populations (Bernardo 2014).

CONCLUSIONS

We found 11 elemental grain phenotypes to be moderately heritable in the maize Ames panel, with minor but significant genotype-by-year interaction. The novel associations of *rte2* and *irt1* with B and Ni, respectively, in combination with enhanced pinpointing of seven candidate causal genes for Cu, Fe, Mn, Mo, and/or Zn illustrate the high level of statistical power and mapping resolution conferred by the Ames panel for genetically dissecting complex trait variation in maize. However, not all detected GWAS signals were resolved down to an

individual gene with a definitive role in metal transport or chelation, thus potentially revealing novel candidate genes that could be further assessed for function in mutagenesis experiments. Additionally, we identified two loci (*nas5* and *ysl2*) that could be leveraged with marker-based breeding approaches to increase Zn levels in maize grain. Notably, the *nas5* gene also associated with the concentration of Fe in grain, thus helping to enable multi-trait selection (Jia and Jannink 2012) for developing biofortified maize varieties to help combat dietary Fe and Zn deficiencies that collectively affect more than 2 billion people worldwide (Viteri 1998; Prasad 2014). Furthermore, the moderate WGP prediction accuracies for Zn and Fe concentrations imply that both grain phenotypes should respond favorably to genomic selection approaches. Overall, our work has provided new insights into the genetic architecture of elemental accumulation in maize grain and strengthened the knowledge base needed to accelerate genomics-assisted breeding efforts for increased grain concentrations of Zn and Fe in maize breeding populations.

DATA AVAILABILITY

The raw genotypes of GBS SNPs Samples_raw_AGPv4-181023.h5) are available on CyVerse (at http://datacommons.cyverse.org/browse/iplant/home/shared/ panzea/genotypes/GBS/v27). The maize HapMap 3.2.1 unim- puted datasets (hmp321_agpv4_chrx.vcf.gz, where x is 1 to 10) are available on CyVerse (at https://datacommons.cyverse.org/ browse/iplant/home/shared/panzea/hapmap3/hmp321/unim puted/uplifted_APGv4/). The untransformed and transformed BLUP values of the phenotypes are provided in Supplemental Tables S3.1 and S3.3, respectively.

ACKNOWLEDGEMENTS

We would like to thank Flinders Analytical for elemental analy- sis. We thank James Schnable for the maize, sorghum, and rice ortholog list. We are grateful to Olena Vatamaniuk, Meng Lin, Jenna Hershberger, and Matheus Baseggio for providing valuable feedback on the manuscript, and we appreciate the insights of Ivan Baxter on candidate genes and maize NAM genetic mapping results. We also thank Jeff Doyle and Daniel Ilut for the discus- sion of gene family evolution and homology.

AUTHOR CONTRIBUTIONS

M.A.G., J.S. and T.R.R. designed this project and supervised the research. D.W. and M.A.G. wrote the manuscript, and all co-authors were involved in editing the manuscript. T.R.R. conducted the field experiment. S.C. and J.S. performed elemental analysis. J.P.H., C.R.B. and D.W. uplifted JL-QTL and GWAS results in the maize NAM panel. D.W., X.L. and G.P.R. conducted genotype processing and imputation. D.W. conducted phenotype processing and GWAS. R.T. conducted WGP.

FUNDING

This research was supported by the National Institute of Food and Agriculture; the USDA Hatch under accession numbers 1013637 (M.A.G), 1013641 (M.A.G), and 1007766 (T.R.R.), HarvestPlus (T.R.R. and M.A.G.), National Science Foundation (IOS-1546657 to M.A.G. and C.R.B.), Cornell University startup funds (M.A.G.) and Patterson Chair Funds (T.R.R.).

SUPPLEMENTAL INFORMATION

Supplemental Figure S3.1. Sources of variation for 11 elemental grain phenotypes in the Ames panel. The phenotypic variance was statistically partitioned into the following components: genotype (G), genotype-by-year interaction (G×Y), year (Y), set within year [S(Y)], block within set within year [BL(S×Y)], inductively coupled plasma mass spectrometry (ICP-MS) batch (BA), row within year [R(Y)], and residual error variance (REV). Variance component estimates were calculated for all random effects from the full model equation 1 of Owens *et al.* (2019). The table below indicates which random effects were significant (*) according to a likelihood ratio test ($\alpha = 0.05$).

Supplemental Figure S3.2. Correlation matrix for untransformed BLUP values for 11 elemental grain phenotypes in the Ames panel. Pearson's correlation coefficients (*r*) calculated with the function 'cor' in R are presented in the upper triangle, whereas the corresponding *P*-values for the significance of correlations ($\alpha = 0.05$) are displayed below the diagonal. The untransformed BLUP values were used to represent the true directionality of the relationship between phenotypes.

Supplemental Figure S3.3. Manhattan plot of results from a genome-wide association study of the five elemental grain phenotypes without significant associations at the 5% false discovery rate level in the Ames panel. Each point represents a SNP with its $-\log_{10}$ *P*-value (y-axis) from a mixed linear model analysis plotted as a function of physical position (B73 RefGen_v4) across the 10 chromosomes of maize (x-axis).

Supplemental Figure S3.4. A regional Manhattan plot of locus 6. Scatter plot of

association results from a mixed model analysis of Cu grain concentration and linkage disequilibrium (LD) estimates (r^2) for a genomic region that contains the peak SNP (8-137939692) at locus 6. Each vertical line represents the -$\log_{10}$ *P*-value of a SNP. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 137,939,692 bp (B73 RefGen_v4) on chromosome 8. The red horizontal dashed line indicates the -$\log_{10}$ *P*-value of the least statistically significant SNP at a genome-wide false discovery rate of 5%. The yellow vertical line indicates the genomic position of the gene Zm00001d011063 with homology to metallothioneins (MTs) in rice and Arabidopsis (Supplemental Table S3.5). The open triangles indicate SNPs that are within the candidate gene. The light blue rectangle demarcates the ± 100 kb candidate gene search space surrounding the peak SNP.

Supplemental Figure S3.5. A regional Manhattan plot of locus 7. Scatter plot of association results from a mixed model analysis of Cu concentration and linkage disequilibrium (LD) estimates (r^2) for a genomic region that contains the peak SNP (8-136857539) at locus 7. Each vertical line represents the -$\log_{10}$ *P*-value of a SNP. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 136,857,539 bp (B73 RefGen_v4) on chromosome 8. The red horizontal dashed line indicates the -$\log_{10}$ *P*-value of the least statistically significant SNP at a genome-wide false discovery rate of 5%. The yellow vertical line indicates the genomic position of the *calcium pump1* (*cap1*) gene Zm00001d011013. The open triangles indicate SNPs that are within the candidate gene. The light blue rectangle demarcates the ± 100 kb candidate gene search space surrounding the peak SNP.

Supplemental Figure S3.6. A regional Manhattan plot of locus 10. Scatter plot of

association results from a mixed model analysis of Mn grain concentration and linkage disequilibrium (LD) estimates (r^2) for a genomic region that contains the peak SNP (1-162962818) at locus 10. Each vertical line represents the -$\log_{10}$ *P*-value of a SNP. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 162,962,818 bp (B73 RefGen_v4) on chromosome 1. The red horizontal dashed line indicates the -$\log_{10}$ *P*-value of the least statistically significant SNP at a genome-wide false discovery rate of 5%. The yellow vertical line indicates the genomic position of the gene Zm00001d030846 with homology to NATURAL RESISTANCE-ASSOCIATED MACROPHAGE PROTEINs (NRAMPs) in rice and Arabidopsis (Supplemental Table S3.5). The open triangles indicate SNPs that are within the candidate gene. The light blue rectangle demarcates the ± 100 kb candidate gene search space surrounding the peak SNP.

Supplemental Figure S3.7. A regional Manhattan plot of locus 11. Scatter plot of association results from a mixed model analysis of Mn grain concentration and linkage disequilibrium (LD) estimates (r^2) for a genomic region that contains the peak SNP (3-184559931) at locus 11. Each vertical line represents the -$\log_{10}$ *P*-value of a SNP. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 184,559,931 bp (B73 RefGen_v4) on chromosome 3. The red horizontal dashed line indicates the -$\log_{10}$ *P*-value of the least statistically significant SNP at a genome-wide false discovery rate of 5%. The yellow vertical line indicates the genomic position of the gene Zm00001d042939 with homology to METAL TOLERANCE PROTEINs (MTPs) in rice and Arabidopsis (Supplemental Table S3.5). The open triangles indicate SNPs that are within the candidate gene. The light blue rectangle demarcates the ± 100 kb candidate gene search space

surrounding the peak SNP.

Supplemental Figure S3.8. A regional Manhattan plot of locus 12. Scatter plot of association results from a mixed model analysis of Mo grain concentration and linkage disequilibrium (LD) estimates (r^2) for a genomic region that contains the peak SNP (1-248672716) at locus 12. Each vertical line represents the -$\log_{10}$ *P*-value of a SNP. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 248,672,716 bp (B73 RefGen_v4) on chromosome 1. The red horizontal dashed line indicates the -$\log_{10}$ *P*-value of the least statistically significant SNP at a genome-wide false discovery rate of 5%. The yellow vertical line indicates the genomic position of the gene Zm00001d033053 with homology to MOLYBDATE TRANSPORTER (MOT) in rice and Arabidopsis (Supplemental Table S3.5). The open triangles indicate SNPs that are within the candidate gene. The light blue rectangle demarcates the ± 100 kb candidate gene search space surrounding the peak SNP.

Supplemental Figure S3.9. A regional Manhattan plot of the *nas5* gene. Scatter plot of association results from a mixed model analysis of Zn (a) and Fe (b) grain concentration and linkage disequilibrium (LD) estimates (r^2) for a genomic region that contains the peak SNPs (Zn: 7-179962589, locus 33; Fe: 7-180077496) at the *nas5* gene. Each vertical line represents the -$\log_{10}$ *P*-value of a SNP. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 179,962,589 bp and 180,077,496 bp (B73 RefGen_v4) on chromosome 7 for Zn and Fe, respectively. The yellow triangle indicates the peak SNP for Fe in the Zn Manhattan plot, and vice versa. The red and green horizontal dashed lines indicate the -$\log_{10}$ *P*-value of the least statistically significant SNP at a genome-wide false discovery

rate of 5% (Zn) and 20% (Fe), respectively. The yellow vertical line indicates the genomic position of the *nicotianamine synthase5* (*nas5*) gene (Zm00001d022557). The open triangles indicate SNPs that are within the candidate gene. The light blue rectangle demarcates the ± 100 kb candidate gene search space surrounding the peak SNP.

Supplemental Figure S3.10. A regional Manhattan plot of locus 32. Scatter plot of association results from a mixed model analysis of Zn grain concentration and linkage disequilibrium (LD) estimates (r^2) for a genomic region that contains the peak SNP (5-195765640) at locus 32. Each vertical line represents the -$\log_{10}$ *P*-value of a SNP. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 195,765,640 bp (B73 RefGen_v4) on chromosome 5. The red horizontal dashed line indicates the -$\log_{10}$ *P*-value of the least statistically significant SNP at a genome-wide false discovery rate of 5%. The yellow vertical line indicates the genomic position of the *yellow stripe-like2* (*ysl2*) gene (Zm00001d017427). The open triangles indicate SNPs that are within the candidate gene. The light blue rectangle demarcates the ± 100 kb candidate gene search space surrounding the peak SNP.

Supplemental Figure S3.11. Manhattan plot of results from a conditional genome-wide association study (GWAS) of five elemental grain phenotypes in the Ames panel. For each phenotype, the SNPs selected by the optimal multi-locus mixed model were included as covariates in the mixed linear model to control for large-effect loci. Each point represents a SNP with its -$\log_{10}$ *P*-value (y-axis) from a mixed linear model analysis plotted as a function of physical position (B73 RefGen_v4) across the 10 chromosomes of maize (x-axis). None of the tested SNPs in the conditional GWAS were significant at a genome-wide false-discovery

rate of 5%.

Supplemental Table S3.1. Untransformed best linear unbiased predictors of 11 elemental grain phenotypes used for the genome-wide association study and whole-genome prediction in the Ames panel.

Supplemental Table S3.2. Lambda values used in Box-Cox transformation of 11 elemental grain phenotypes in the Ames panel.

Supplemental Table S3.3. Transformed best linear unbiased predictors of 11 elemental grain phenotypes used for the genome-wide association study and whole-genome prediction in the Ames panel.

Supplemental Table S3.4. Population structure analysis of the Ames panel. Included in the table are the first two principal components (PCs) calculated from the SNP genotype matrix (PC1 and PC2), assignment value (Q) for each subpopulation (subpopulations 1 to 3), group classification of lines following assignments of Romay *et al.* (2013), largest Q assignment value, and the subpopulation (Subpopulation 1, 2, 3, or Admixed) to which each line was assigned (Category) in our study.

Supplemental Table S3.5. Rice and Arabidopsis homologs of the nine candidate genes identified for elemental grain phenotypes via a genome-wide association study in the maize Ames panel.

Supplemental Table S3.6. Joint-linkage QTL support intervals (SI) of 11 elemental grain phenotypes analyzed in the maize NAM panel (Ziegler *et al.*, 2017) uplifted from RefGen_v2 to v4 and their relationship to peak SNPs within ± 100 kb of candidate genes identified from a genome-wide association study in the Ames panel (Table 3.2 and

Supplemental Table S3.8).

Supplemental Table S3.7. Genome-wide association study results of 11 elemental grain phenotypes analyzed in the maize NAM panel (Ziegler *et al.* 2017) uplifted from RefGen_v2 to v4. Only NAM marker variants with resample model inclusion probability (RMIP) ≥ 5 are shown and those that reside within joint-linkage QTL support intervals (Supplemental Table S3.6) are demarcated in the "NAM QTL number" column. The relationship of NAM marker variants to the candidate genes identified in genome-wide association study in the Ames panel (Table 3.2) are also presented.

Supplemental Table S3.8. Statistically significant results from a genome-wide association study of 11 elemental grain phenotypes in the Ames panel.

Supplemental Table S3.9. Genomic information (RefGen_v4) for the candidate genes within ± 100 kb of the peak SNPs identified in the genome-wide association study.

Supplemental Table S3.10. Frequencies, mean concentrations (in μg g^{-1}), and standard deviations (Std. Dev.) for peak SNPs within ± 100 kb of candidate genes identified in the genome-wide association study.

Supplemental Table S3.11. Statistically significant results from a multi-locus mixed-model analysis of 11 elemental grain phenotypes in the Ames panel.

REFERENCES

Alomari, D. Z., K. Eggert, N. von Wirén, A. Polley, J. Plieske *et al.*, 2018 Whole-genome association mapping and genomic prediction for iron concentration in wheat grains. Int. J. Mol. Sci. 20: 76.

Andrés-Colás, N., V. Sancenón, S. Rodríguez-Navarro, S. Mayo, D. J. Thiele *et al.*, 2006 The Arabidopsis heavy metal P-type ATPase HMA5 interacts with metallochaperones and functions in copper detoxification of roots. Plant J. 45: 225–236.

Asaro, A., G. Ziegler, C. Ziyomo, O. A. Hoekenga, B. P. Dilkes *et al.*, 2016 The interaction of genotype and environment determines variation in the maize kernel ionome. G3 6: 4175–4183.

Axelsen, K. B., and M. G. Palmgren, 2001 Inventory of the superfamily of P-type ion pumps in Arabidopsis. Plant Physiol. 126: 696–706.

Baxter, I., 2009 Ionomics: studying the social network of mineral nutrients. Curr. Opin. Plant Biol. 12: 381–386.

Baxter, I., B. Muthukumar, H. C. Park, P. Buchner, B. Lahner *et al.*, 2008 Variation in molybdenum content across broadly distributed populations of *Arabidopsis thaliana* is controlled by a mitochondrial molybdenum transporter (MOT1). PLoS Genet. 4: e1000004.

Baxter, I.R., Gustin, J.L., Settles, A.M., and Hoekenga, O.A. (2013) Ionomic characterization of maize kernels in the intermated B73× Mo17 population. Crop Sci., **53**, 208–220.

Baxter, I. R., G. Ziegler, B. Lahner, M. V. Mickelbart, R. Foley *et al.*, 2014 Single-kernel ionomic profiles are highly heritable indicators of genetic and environmental influences on elemental accumulation in maize grain (*Zea mays*). PLoS ONE 9: e87628.

Benatti, M., N. Yookongkaew, M. Meetam, W.-J. Guo, N. Punyasuk *et al.*, 2014 Metallothionein deficiency impacts copper accumulation and redistribution in leaves and seeds of Arabidopsis. New Phytol. 202: 940–951.

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

ten Berge, H. F. M., R. Hijbeek, M. P. van Loon, J. Rurinda, K. Tesfaye *et al.*, 2019 Maize crop nutrient input requirements for food security in sub-Saharan Africa. Glob. Food Sec. 23: 9–21.

Bernardo, R., 2014 Genomewide selection when major genes are known. Crop Sci. 54: 68–75.

Bouis, H. E., and A. Saltzman, 2017 Improving nutrition through biofortification: a review of evidence from HarvestPlus, 2003 through 2016. Glob. Food Sec. 6: 49–58.

Box, G. E. P., and D. R. Cox, 1964 An analysis of transformations. J. R. Stat. Soc. Series B Stat. Methodol. 26: 211–243.

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

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

Bukowski, R., X. Guo, Y. Lu, C. Zou, B. He *et al.*, 2018 Construction of the third-generation

Zea mays haplotype map. Gigascience 7: 1–12.

Chan-Rodriguez, D., and E. L. Walker, 2018 Analysis of yellow striped mutants of *Zea mays* reveals novel loci contributing to iron deficiency chlorosis. Front. Plant Sci. 9: 157.

Chatterjee, M., Q. Liu, C. Menello, M. Galli, and A. Gallavotti, 2017 The combined action of duplicated boron transporters is required for maize growth in boron-deficient conditions. Genetics 206: 2041–2051.

Chatterjee, M., Z. Tabi, M. Galli, S. Malcomber, A. Buck *et al.*, 2014 The boron efflux transporter ROTTEN EAR is required for maize inflorescence development and fertility. Plant Cell 26: 2962–2977.

Clemens, S., 2001 Molecular mechanisms of plant metal tolerance and homeostasis. Planta 212: 475–486.

Combs, E., and R. Bernardo, 2013 Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6: 1–7.

Curie, C., G. Cassin, D. Couch, F. Divol, K. Higuchi *et al.*, 2009 Metal movement within the plant: contribution of nicotianamine and yellow stripe 1-like transporters. Ann. Bot. 103: 1–11.

Curie, C., Z. Panaviene, C. Loulergue, S. L. Dellaporta, J. F. Briat *et al.*, 2001 Maize *yellow stripe1* encodes a membrane protein directly involved in Fe(III) uptake. Nature 409: 346–349.

Delhaize, E., B. D. Gruber, J. K. Pittman, R. G. White, H. Leung *et al.*, 2007 A role for the *AtMTP11* gene of Arabidopsis in manganese transport and tolerance. Plant J. 51: 198–210.

Deng, F., N. Yamaji, J. Xia, and J. F. Ma, 2013 A member of the heavy metal P-type ATPase OsHMA5 is involved in xylem loading of copper in rice. Plant Physiol. 163: 1353–1362.

Desta, Z. A., and R. Ortiz, 2014 Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 19: 592–601.

Diepenbrock, C. H., and M. A. Gore, 2015 Closing the divide between human nutrition and plant breeding. Crop Sci. 55: 1437–1448.

Dixon, J. A., A. Gulliver, and D. P. Gibbon, 2001 Farming systems and poverty: Improving farmers' livelihoods in a changing world. Rome, Italy: Food and Agriculture Organization of the United Nations (FAO); Washington, DC: World Bank.

Durbak, A. R., K. A. Phillips, S. Pike, M. A. O'Neill, J. Mares *et al.*, 2014 Transport of boron by the *tassel-less1* aquaporin is critical for vegetative and reproductive development in maize. Plant Cell 26: 2978–2995.

Eide, D., M. Broderius, J. Fett, and M. L. Guerinot, 1996 A novel iron-regulated metal transporter from plants identified by functional expression in yeast. Proc. Natl. Acad. Sci. U. S. A. 93: 5624–5628.

FAOSTAT, 2018 FAO Statistical databases. http://www.fao.org/faostat/en/#data/CC/visualize. Accessed January 17, 2021.

Farthing, E. C., P. K. Menguer, J. P. Fett, and L. E. Williams, 2017 OsMTP11 is localised at the Golgi and contributes to Mn tolerance. Sci. Rep. 7: 1–13.

Fikas, A. A., B. P. Dilkes, and I. Baxter, 2019 Multivariate analysis reveals environmental and

genetic determinants of element covariation in the maize grain ionome. Plant Direct 3: e00139.

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

Gianola, D., 2013 Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194: 573–596.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. Genetics 183: 347–363.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, R. and Thompson, R., 2009 ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Graham, R. D., and R. M. Welch, 1996 Breeding for staple food crops with high micronutrient density, pp. 1–72 in *Agricultural Strategies for Micronutrients Working Paper No. 3*, International Food Policy Research Institute, Washington, DC.

Grotz, N., T. Fox, E. Connolly, W. Park, M. L. Guerinot *et al.*, 1998 Identification of a family of zinc transporter genes from Arabidopsis that respond to zinc deficiency. Proc. Natl. Acad. Sci. U. S. A. 95: 7220–7224.

Gu, R., F. Chen, B. Liu, X. Wang, J. Liu *et al.*, 2015 Comprehensive phenotypic analysis and quantitative trait locus identification for grain mineral concentration, content, and yield in maize (*Zea mays* L.). Theor. Appl. Genet. 128: 1777–1789.

Guo, R., T. Dhliwayo, E. K. Mageto, N. Palacios-Rojas, M. Lee *et al.*, 2020 Genomic prediction of kernel zinc concentration in multiple maize populations using genotyping-by-sequencing and repeat amplification sequencing markers. Front. Plant Sci. 11.

Guo, W.-J., M. Meetam, and P. B. Goldsbrough, 2008 Examining the specific contributions of individual Arabidopsis metallothioneins to copper distribution and metal tolerance. Plant Physiol. 146: 1697–1706.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Hindu, V., N. Palacios-Rojas, R. Babu, W. B. Suwarno, Z. Rashid *et al.*, 2018 Identification and validation of genomic regions influencing kernel zinc and iron in maize. Theor. Appl. Genet. 131: 1443–1457.

Hirayama, T., J. J. Kieber, N. Hirayama, M. Kogan, P. Guzman *et al.*, 1999 RESPONSIVE-TO-ANTAGONIST1, a Menkes/Wilson disease–related copper transporter, is required for ethylene signaling in Arabidopsis. Cell 97: 383–393.

Holland, J. B., W. E. Nyquist, and C. T. Cervantes-Martínez, 2003 Estimating and interpreting heritability for plant breeding: an update. Plant Breed. Rev. 22: 9–112.

Huang, C., H. Sun, D. Xu, Q. Chen, Y. Liang *et al.*, 2018 *ZmCCT9* enhances maize adaptation to higher latitudes. Proc. Natl. Acad. Sci. U. S. A. 115: E334–E341.

Huang, X.-Y., F. Deng, N. Yamaji, S. R. M. Pinson, M. Fujii-Kashino *et al.*, 2016 A heavy metal P-type ATPase OsHMA4 prevents copper accumulation in rice grain. Nat. Commun. 7: 12138.

Huang, X.-Y., H. Liu, Y.-F. Zhu, S. R. M. Pinson, H.-X. Lin *et al.*, 2019 Natural variation in a molybdate transporter controls grain molybdenum concentration in rice. New Phytol. 221: 1983–1997.

Huang, X.-Y., and D. E. Salt, 2016 Plant ionomics: from elemental profiling to environmental adaptation. Mol. Plant 9: 787–797.

Hung, H.-Y., C. Browne, K. Guill, N. Coles, M. Eller *et al.*, 2012 The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. Heredity 108: 490–499.

Jia, Y., and J.-L. Jannink, 2012 Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192: 1513–1522.

Jin, T., J. Chen, L. Zhu, Y. Zhao, J. Guo *et al.*, 2015 Comparative mapping combined with homology-based cloning of the rice genome reveals candidate genes for grain zinc and iron concentration in maize. BMC Genet. 16: 17.

Jin, T., J. Zhou, J. Chen, L. Zhu, Y. Zhao *et al.*, 2013 The genetic architecture of zinc and iron content in maize grains as revealed by QTL mapping and meta-analysis. Breed. Sci. 63: 317–324.

Kobayashi, Y., K. Kuroda, K. Kimura, J. L. Southron-Francis, A. Furuzawa *et al.*, 2008 Amino acid polymorphisms in strictly conserved domains of a P-Type ATPase HMA5 are involved in the mechanism of copper tolerance variation in Arabidopsis. Plant Physiol. 148: 969–980.

Korshunova, Y. O., D. Eide, W. G. Clark, M. L. Guerinot, and H. B. Pakrasi, 1999 The IRT1 protein from *Arabidopsis thaliana* is a metal transporter with a broad substrate range. Plant Mol. Biol. 40: 37–44.

Kumar, G., H. Kushwaha, V. Panjabi-Sabharwal, S. Kumari, R. Joshi *et al.*, 2012 Clustered metallothionein genes are co-regulated in rice and ectopic expression of OsMT1e-P confers multiple abiotic stress tolerance in tobacco via ROS scavenging. BMC Plant Biology 12: 107.

Kurtz, S., 2010 The Vmatch large scale sequence analysis software - a manual. http://www.vmatch.de/virtman.pdf.

Lanquar, V., F. Lelièvre, S. Bolte, C. Hamès, C. Alcon *et al.*, 2005 Mobilization of vacuolar iron by AtNRAMP3 and AtNRAMP4 is essential for seed germination on low iron. EMBO J. 24: 4041–4051.

Lanquar, V., M. S. Ramos, F. Lelièvre, H. Barbier-Brygoo, A. Krieger-Liszkay *et al.*, 2010 Export of vacuolar manganese by AtNRAMP3 and AtNRAMP4 is required for optimal photosynthesis and growth under manganese deficiency. Plant Physiol. 152: 1986–1999.

Lee, S., U. S. Jeon, S. J. Lee, Y.-K. Kim, D. P. Persson *et al.*, 2009 Iron fortification of rice seeds through activation of the nicotianamine synthase gene. Proc. Natl. Acad. Sci. U. S. A. 106: 22014–22019.

Lipka, A. E., M. A. Gore, M. Magallanes-Lundback, A. Mesberg, H. Lin *et al.*, 2013 Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. G3 3: 1287–1299.

Lipka, A. E., C. B. Kandianis, M. E. Hudson, J. Yu, J. Drnevich *et al.*, 2015 From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr. Opin. Plant Biol. 24: 110–118.

Lipka, A. E., F. Tian, Q. Wang, J. Peiffer, M. Li *et al.*, 2012 GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397–2399.

Li, S., X. Liu, X. Zhou, Y. Li, W. Yang *et al.*, 2019 Improving zinc and iron accumulation in maize grains using the zinc and iron transporter ZmZIP5. Plant Cell Physiol. 60: 2077–2085.

Li, S., X. Zhou, H. Li, Y. Liu, L. Zhu *et al.*, 2015 Overexpression of *ZmIRT1* and *ZmZIP3* enhances iron and zinc accumulation in transgenic Arabidopsis. PLoS ONE 10: e0136647.

Littell, R. C., G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger, 2006 Appendix 1: Linear mixed model theory. SAS for mixed models. SAS Institute Inc., Cary, NC 733–756.

Lombi, E., K. G. Scheckel, J. Pallon, A. M. Carey, Y. G. Zhu *et al.*, 2009 Speciation and distribution of arsenic and localization of nutrients in rice grains. New Phytol. 184: 193–201.

Lombi, E., E. Smith, T. H. Hansen, D. Paterson, M. D. de Jonge *et al.*, 2011 Megapixel imaging of (micro)nutrients in mature barley grains. J. Exp. Bot. 62: 273–282.

Lorenz, A. J., S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi *et al.*, 2011 Genomic selection in plant breeding: knowledge and prospects, pp. 77–123 in *Advances in agronomy*, Elsevier.

de Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–345.

Lubin, J. H., J. S. Colt, D. Camann, S. Davis, J. R. Cerhan *et al.*, 2004 Epidemiologic evaluation of measurement data in the presence of detection limits. Environ. Health Perspect. 112: 1691–1696.

Lung'aho, M. G., A. M. Mwaniki, S. J. Szalma, J. J. Hart, M. A. Rutzke *et al.*, 2011 Genetic and physiological analysis of iron biofortification in maize kernels. PLoS ONE 6: e20429.

Lynch, M., and B. Walsh, 1998 *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA.

Mageto, E. K., J. Crossa, P. Pérez-Rodríguez, T. Dhliwayo, N. Palacios-Rojas *et al.*, 2020 Genomic prediction with genotype by environment interaction analysis for kernel zinc concentration in tropical maize germplasm. G3 8: 2629–2639.

Manickavelu, A., T. Hattori, S. Yamaoka, K. Yoshimura, Y. Kondou *et al.*, 2017 Genetic nature of elemental contents in wheat grains and its genomic prediction: Toward the effective use of wheat landraces from Afghanistan. PLoS ONE 12: e0169416.

Marschner, P., 2011 *Marschner's Mineral Nutrition of Higher Plants*. 3rd edition. Edited by P. Marschner. Amsterdam, Netherlands: Elsevier/Academic Press, pp. 684, ISBN 978-0-12-384905-2. Elsvier.

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

Mengel, K., and E. A. Kirkby, 2012 *Principles of Plant Nutrition*. Springer Science & Business Media.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value

using genome-wide dense marker maps. Genetics 157: 1819–1829.

Miwa, K., and T. Fujiwara, 2010 Boron transport in plants: co-ordinated regulation of transporters. Ann. Bot. 105: 1103–1108.

Miwa, K., J. Takano, and T. Fujiwara, 2006 Improvement of seed yields under boron-limiting conditions through overexpression of BOR1, a boron transporter for xylem loading, in *Arabidopsis thaliana*. Plant J. 46: 1084–1091.

Mondal, T. K., S. A. Ganie, M. K. Rana, and T. R. Sharma, 2014 Genome-wide analysis of zinc transporter genes of maize (*Zea mays*). Plant Mol. Biol. Rep. 32: 605–616.

Myles, S., J. Peiffer, P. J. Brown, E. S. Ersoz, Z. Zhang *et al.*, 2009 Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21: 2194–2202.

Nakagawa, Y., H. Hanaoka, M. Kobayashi, K. Miyoshi, K. Miwa *et al.*, 2007 Cell-type specificity of the expression of *OsBOR1*, a rice efflux boron transporter gene, is regulated in response to boron availability for efficient boron uptake and xylem loading. Plant Cell 19: 2624–2635.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, 1996 *Applied linear statistical models*. Irwin Chicago.

Nishida, S., A. Kato, C. Tsuzuki, J. Yoshida, and T. Mizuno, 2015 Induction of nickel accumulation in response to zinc deficiency in *Arabidopsis thaliana*. Int. J. Mol. Sci. 16: 9420–9430.

Nishida, S., C. Tsuzuki, A. Kato, A. Aisu, J. Yoshida *et al.*, 2011 AtIRT1, the primary iron uptake transporter in the root, mediates excess nickel accumulation in *Arabidopsis thaliana*. Plant Cell Physiol. 52: 1433–1442.

Owens, B. F., A. E. Lipka, M. Magallanes-Lundback, T. Tiede, C. H. Diepenbrock *et al.*, 2014 A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. Genetics 198: 1699–1716.

Owens, B. F., D. Mathew, C. H. Diepenbrock, T. Tiede, D. Wu *et al.*, 2019 Genome-wide association study and pathway-level analysis of kernel color in maize. G3 9: 1945–1955.

Palmer, L. J., L. T. Palmer, M. A. Rutzke, R. D. Graham, and J. C. R. Stangoulis, 2014 Nutrient variability in phloem: examining changes in K, Mg, Zn and Fe concentration during grain loading in common wheat (*Triticum aestivum*). Physiol. Plant. 152: 729–737.

Peiter, E., B. Montanini, A. Gobert, P. Pedas, S. Husted *et al.*, 2007 A secretory pathway-localized cation diffusion facilitator confers plant manganese tolerance. Proc. Natl. Acad. Sci. U. S. A. 104: 8532–8537.

Pérez, P., and G. de los Campos, 2014 Genome-wide regression and prediction with the BGLR statistical package. Genetics 198: 483–495.

Pongrac, P., K. Vogel-Mikuš, L. Jeromel, P. Vavpetič, P. Pelicon *et al.*, 2013 Spatially resolved distributions of the mineral elements in the grain of tartary buckwheat (*Fagopyrum tataricum*). Food Res. Int. 54: 125–131.

Prasad, A. S., 2014 Impact of the discovery of human zinc deficiency on health. J. Trace Elem. Med. Biol. 28: 357–363.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a

tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

Qin, H., Y. Cai, Z. Liu, G. Wang, J. Wang *et al.*, 2012 Identification of QTL for zinc and iron concentration in maize kernel and cob. Euphytica 187: 345–358.

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

Ramstein, G. P., S. J. Larsson, J. P. Cook, J. W. Edwards, E. S. Ersoz *et al.*, 2020 Dominance effects and functional enrichments improve prediction of agronomic traits in hybrid maize. Genetics 215: 215–230.

R Core Team, 2018 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ricci, W. A., Z. Lu, L. Ji, A. P. Marand, C. L. Ethridge *et al.*, 2019 Widespread long-range cis-regulatory elements in the maize genome. Nat. Plants 6: 328.

Roberts, L. A., A. J. Pierson, Z. Panaviene, and E. L. Walker, 2004 Yellow stripe1. Expanded roles for the maize iron-phytosiderophore transporter. Plant Physiol. 135: 112–120.

Rodgers-Melnick, E., D. L. Vera, H. W. Bass, and E. S. Buckler, 2016 Open chromatin reveals the functional maize genome. Proc. Natl. Acad. Sci. U. S. A. 113: E3177–3184.

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

Ross-Ibarra, J., M. Tenaillon, and B. S. Gaut, 2009 Historical divergence and gene flow in the genus Zea. Genetics 181: 1399–1413.

Salvi, S., G. Sponza, M. Morgante, D. Tomes, X. Niu *et al.*, 2007 Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. U. S. A. 104: 11376–11381.

Schaaf, G., U. Ludewig, B. E. Erenoglu, S. Mori, T. Kitahara *et al.*, 2004 ZmYS1 functions as a proton-coupled symporter for phytosiderophore- and nicotianamine-chelated metals. J. Biol. Chem. 279: 9091–9096.

Schnable, J. C., N. M. Springer, and M. Freeling, 2011 Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. U. S. A. 108: 4069–4074.

Schuler, M., R. Rellán-Álvarez, C. Fink-Straube, J. Abadía, and P. Bauer, 2012 Nicotianamine functions in the phloem-based transport of iron to sink organs, in pollen development and pollen tube growth in *Arabidopsis*. Plant Cell 24: 2380–2400.

Schwarz, G., 1978 Estimating the dimension of a model. Ann. Stat. 6: 461–464.

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

Shakoor, N., G. Ziegler, B. P. Dilkes, Z. Brenton, R. Boyles *et al.*, 2016 Integration of experiments across diverse environments identifies the genetic determinants of variation in *Sorghum*

bicolor seed element composition. Plant Physiol. 170: 1989–1998.

Šimić, D., S. Mladenović Drinić, and Z. Zdunić, 2011 Quantitative trait loci for biofortification traits in maize grain. J. Hered., **103**, 47–54.

Studer, A., Q. Zhao, J. Ross-Ibarra, and J. Doebley, 2011 Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nat. Genet. 43: 1160–1163.

Subbaiah, C. C., and M. M. Sachs, 2000 Maize *cap1* encodes a novel SERCA-type calcium-ATPase with a calmodulin-binding domain. J. Biol. Chem. 275: 21678–21687.

Sun, G., C. Zhu, M. H. Kramer, S.-S. Yang, W. Song *et al.*, 2010 Variation explained in mixed-model association mapping. Heredity 105: 333–340.

Swamy, B. P. M., M. A. Rahman, M. A. Inabangan-Asilo, A. Amparado, C. Manito *et al.*, 2016 Advances in breeding for high grain zinc in rice. Rice 9: 49.

Tomatsu, H., J. Takano, H. Takahashi, A. Watanabe-Takahashi, N. Shibagaki *et al.*, 2007 An *Arabidopsis thaliana* high-affinity molybdate transporter required for efficient uptake of molybdate from soil. Proc. Natl. Acad. Sci. U. S. A. 104: 18807–18812.

Tsunemitsu, Y., M. Genga, T. Okada, N. Yamaji, J. F. Ma *et al.*, 2018 A member of cation diffusion facilitator family, MTP11, is required for manganese tolerance and high fertility in rice. Planta 248: 231–241.

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

Velu, G., J. Crossa, R. P. Singh, Y. Hao, S. Dreisigacker *et al.*, 2016 Genomic prediction for grain zinc and iron concentrations in spring wheat. Theor. Appl. Genet. 129: 1595–1605.

Vert, G., M. Barberon, E. Zelazny, M. Séguéla, J.-F. Briat *et al.*, 2009 Arabidopsis IRT2 cooperates with the high-affinity iron uptake system to maintain iron homeostasis in root epidermal cells. Planta 229: 1171–1179.

Vert, G., N. Grotz, F. Dédaldéchamp, F. Gaymard, M. L. Guerinot *et al.*, 2002 IRT1, an Arabidopsis transporter essential for iron uptake from the soil and for plant growth. Plant Cell 14: 1223–1233.

Viteri, F. E., 1998 A new concept in the control of iron deficiency: community-based preventive supplementation of at-risk groups by the weekly intake of iron supplements. Biomed. Environ. Sci. 11: 46–60.

Von Wiren, N., S. Mori, H. Marschner, and V. Romheld, 1994 Iron inefficiency in maize mutant *ys1* (*Zea mays* L. cv Yellow-Stripe) is caused by a defect in uptake of iron phytosiderophores. Plant Physiol. 106: 71–77.

Wallace, J. G., P. J. Bradbury, N. Zhang, Y. Gibon, M. Stitt *et al.*, 2014 Association mapping across numerous traits reveals patterns of functional variation in maize. PLoS Genetics 10: e1004845.

Waters, B. M., H.-H. Chu, R. J. Didonato, L. A. Roberts, R. B. Eisley *et al.*, 2006 Mutations in Arabidopsis *yellow stripe-like1* and *yellow stripe-like3* reveal their roles in metal ion homeostasis and loading of metal ions in seeds. Plant Physiol. 141: 1446–1458.

Waters, B. M., and R. P. Sankaran, 2011 Moving micronutrients from the soil to the seeds:

genes and physiological processes from a biofortification perspective. Plant Sci. 180: 562–574.

Welch, R. M., 2002 Breeding strategies for biofortified staple plant foods to reduce micronutrient malnutrition globally. J. Nutr. 132: 495S–499S.

Welch, R. M., and R. D. Graham, 2002 Breeding crops for enhanced micronutrient content, pp. 267–276 in *Food security in nutrient-stressed environments: Exploiting plants' genetic capabilities*, edited by J. J. Adu-Gyamfi. Springer Netherlands, Dordrecht.

Welch, R. M., and R. D. Graham, 2004 Breeding for micronutrients in staple food crops from a human nutrition perspective. J. Exp. Bot. 55: 353–364.

Wheal, M. S., T. O. Fowles, and L. T. Palmer, 2011 A cost-effective acid digestion method using closed polypropylene tubes for inductively coupled plasma optical emission spectrometry (ICP-OES) analysis of plant essential elements. Anal. Methods 3: 2854–2863.

Whitt, L., F. K. Ricachenevsky, G. Z. Ziegler, S. Clemens, E. Walker *et al.*, 2020 A curated list of genes that affect the plant ionome. Plant Direct 4: e00272.

Wiren, N. von, N. von Wiren, H. Marschner, and V. Romheld, 1996 Roots of iron-efficient maize also absorb phytosiderophore-chelated zinc. Plant Physiology 111: 1119–1125.

Wu, Z., F. Liang, B. Hong, J. C. Young, M. R. Sussman *et al.*, 2002 An endoplasmic reticulum-bound Ca^{2+}/Mn^{2+} pump, ECA1, supports plant growth and confers tolerance to Mn^{2+} stress. Plant Physiol. 130: 128–137.

Yang, M., K. Lu, F.-J. Zhao, W. Xie, P. Ramakrishna *et al.*, 2018 Genome-wide association studies reveal the genetic basis of ionomic variation in rice. Plant Cell 30: 2720–2740.

Yasmin, Z., N. Paltridge, R. Graham, B.-L. Huynh, and J. Stangoulis, 2014 Measuring genotypic variation in wheat seed iron first requires stringent protocols to minimize soil iron contamination. Crop Sci. 54: 255–264.

Yordem, B. K., S. S. Conte, J. F. Ma, K. Yokosho, K. A. Vasques *et al.*, 2011 *Brachypodium distachyon* as a new model system for understanding iron homeostasis in grasses: phylogenetic and expression analysis of Yellow Stripe-Like (YSL) transporters. Annals of Botany 108: 821–833.

Yuan, J., D. Chen, Y. Ren, X. Zhang, and J. Zhao, 2008 Characteristic and expression analysis of a metallothionein gene, *OsMT2b*, down-regulated by cytokinin suggests functions in root development and seed embryo germination of rice. Plant Physiol. 146: 1637–1650.

Yu, J., J. B. Holland, M. D. McMullen, and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics 178: 539–551.

Zang, J., Y. Huo, J. Liu, H. Zhang, J. Liu *et al.*, 2020 Maize YSL2 is required for iron distribution and development in kernels. J. Exp. Bot. 71: 5896–5910.

Zhang, H., J. Liu, T. Jin, Y. Huang, J. Chen *et al.*, 2017 Identification of quantitative trait locus and prediction of candidate genes for grain mineral concentration in maize across multiple environments. Euphytica 213: 90.

Zhang, M., and B. Liu, 2017 Identification of a rice metal tolerance protein OsMTP11 as a manganese transporter. PLoS ONE 12: e0174987.

Zhang, Y., D. W. Ngu, D. Carvalho, Z. Liang, Y. Qiu *et al.*, 2017 Differentially regulated

orthologs in sorghum and the subgenomes of maize. Plant Cell 29: 1938–1951.

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari *et al.*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nature Genetics 42: 355–360.

Zhou, G., Y. Xu, J. Li, L. Yang, and J.-Y. Liu, 2006 Molecular analyses of the metallothionein gene family in rice (*Oryza sativa* L.). J. Biochem. Mol. Biol. 39: 595–606.

Zhou, X., S. Li, Q. Zhao, X. Liu, S. Zhang *et al.*, 2013 Genome-wide identification, classification and expression profiling of nicotianamine synthase (NAS) gene family in maize. BMC Genomics 14: 238.

Zhu, C., M. Gore, E. S. Buckler, and J. Yu, 2008 Status and Prospects of Association Mapping in Plants. Plant Genome 1: 5–20.

Ziegler, G., P. J. Kear, D. Wu, C. Ziyomo, A. E. Lipka *et al*, 2017 Elemental accumulation in kernels of the maize nested association mapping panel reveals signals of gene by environment interactions. bioRxiv 164962.

Chapter 4 Integrating GWAS and TWAS to identify causal genes for tocochromanol levels in maize grain

ABSTRACT

Tocochromanols (tocopherols and tocotrienols) are lipid-soluble antioxidants that are important for human health and plant fitness. Tocochromanols from plant-derived products, notably from seed oils, are the main sources of vitamin E in the human diet. The tocochromanol biosynthesis pathway has been elucidated in *Arabidopsis thaliana*, but the complete role of pathway genes and their regulators in the genetic control of natural variation of tocochromanol levels in seed of cereal crops largely remains to be determined. To more fully catalogue the causal genes involved in tocochromanol accumulation in maize grain, we utilized the high mapping resolution of the maize Ames panel of ~1,500 inbred lines scored with 12.2 million SNPs and generated metabolomic (mature grain) and transcriptomic (developing grain) datasets for genetic mapping. Through integrating results from genome- and transcriptome-wide association studies, we identified a total of 13 causal candidate genes and of which four were associated with maize grain tocochromanols for the first time. These four include *vte7* (*alpha/beta hydrolase*), *samt1* (*S-adenosylmethionine transporter 1*), *vte5* (*phytol kinase*), and *dxs1* (*1-deoxy-D-xylulose-5-phosphate synthase*). Expression quantitative trait locus (eQTL) mapping was conducted for the identified 13 genes, revealing they were predominantly regulated by cis-eQTL. *phytoene synthase 1*, which encodes an enzyme that catalyzes the first committed step in carotenoid synthesis, was identified to underlie a trans-eQTL that was the strongest signal for *dxs2*. Through these integrated -omics analyses, we

have performed the most detailed genetic dissection of tocochromanol accumulation in maize grain to date and provide insights for biofortification breeding efforts in maize and other cereals.

INTRODUCTION

Tocochromanols, which include the biosynthetically related tocopherols and tocotrienols, are a group of plant-synthesized lipid-soluble antioxidants that have a chromanol ring head derived from homogentisic acid (HGA) and hydrophobic polyprenyl side chain. Tocopherols have a saturated side chain derived from phytol-diphosphate (PDP), whereas the side chain of tocotrienols derives from geranylgeranyl diphosphate (GGDP) and has three double bonds. Within each tocochromanol class, there are four isoforms (α, β, δ, and γ) that vary in their degree and position of methyl groups on the chromanol ring head. Among the tocochromanols, α-tocopherol has the highest vitamin E activity (DellaPenna & Mène-Saffrané, 2011), while tocotrienols tend to have greater antioxidant activity (Sen *et al.,* 2006). Although severe vitamin E deficiency leading to ataxia and myopathy is rare in human populations (Traber, 2012), less than optimal dietary intake of vitamin E exists in certain population segments (Ford *et al.,* 2006; McBurney *et al.,* 2015) and has been linked to an elevated risk of cardiovascular diseases (Knekt *et al.,* 1994; Kushi *et al.,* 1996). Tocochromanols are found at high levels in plant seeds where they confer protection against lipid peroxidation during seed storage and germination (Collakova & DellaPenna, 2003; Liu *et al.,* 2008; Sattler *et al.,* 2004); however, α-tocopherol is not the major tocochromanol in cereal seed and their extracted oil, thus limiting vitamin E in the diet of both humans and

animals (DellaPenna & Mène-Saffrané, 2011).

Tocochromanols are only synthesized by photosynthetic organisms, with the tocochromanol biosynthetic pathway deciphered and highly conserved in the plant kingdom (DellaPenna & Mène-Saffrané, 2011). In the committed step of tocopherol synthesis (Figure 4.1), a homogentisate phytyltransferase (VTE2) condenses PDP and HGA from the shikimate pathway to produce MPBQ (2-methyl-6-phytyl-1,4-benzoquinol) (Sattler *et al.,* 2004). In the monocot lineage, HGA is condensed with geranylgeranyl diphosphate (GGDP) from the methylerythritol 4-phosphate (MEP) pathway by homogentisate geranylgeranyltransferase (HGGT1) to generate MGGBQ (2-methyl-6-geranylgeranyl-1,4-benzoquinol) in the committed step in tocotrienol synthesis. The committed precursors for tocopherols (MPBQ) and tocotrienols (MGGBQ) serve as the starting substrate for a sequence of cyclization (VTE1, tocopherol cyclase) and methylation (VTE3, MPBQ/MGGBQ methyltransferase; and VTE4, γ-tocopherol methyltransferase) reactions to generate the α, β, δ, and γ isoforms of tocopherols and tocotrienols (Cheng *et al.,* 2003; Porfirova *et al.,* 2002; Sattler *et al.,* 2004; Shintani & DellaPenna, 1998; Van Eenennaam *et al.,* 2003). The methylation reactions catalyzed by VTE3 and VTE4 transfer a methyl group from S-adenosyl-L-methionine (SAM). The generation of PDP as a precursor in the committed step of tocopherol synthesis can directly result from the enzymatic reduction of GGDP, or indirectly through the sequential phosphorylation of chlorophyll-derived phytol by VTE5 (phytol kinase) and VTE6 (phytol phosphate kinase) (Valentin *et al.*, 2006; Vom Dorp *et al.,* 2015). The major source of phytol for tocopherol synthesis is provided directly from chlorophyll biosynthesis through the hydrolase activity of VTE7 (alpha/beta hydrolase) in seeds (Arabidopsis and maize) and

leaves (maize) (Albert *et al.* in review).

In the past decade, a number of loci associated with natural variation for the content and composition of tocochromanols in maize grain were identified via genome-wide association studies (GWAS) in mapping panels. Several studies have reported strong associations between *vte4* and α-tocopherol concentration in maize grain (Baseggio *et al.*, 2019; Lipka *et al.*, 2013; Q. Li *et al.*, 2012; Wang *et al.*, 2018), with relatively weaker associations detected for *vte1*, *hggt1*, and an arogenate/prephenate dehydratase with maize grain tocotrienol levels (Baseggio *et al.*, 2019; Lipka *et al.*, 2013). Suggesting the importance of genes outside of the core tocochromanol biosynthesis pathway, Wang *et al.* (2018) implicated genes responsible for fatty acid biosynthesis, chlorophyll metabolism, and chloroplast function as having involvement in the genetic control of grain tocopherol levels. Despite the progress made towards better understanding the genetic basis of tocochromanol levels in maize grain, these studies were limited by panel sizes and marker density.

Through a joint-linkage (JL) analysis and GWAS in the 5,000-line U.S. maize nested association mapping (NAM) panel with superior statistical power, Diepenbrock *et al.* (2017) identified 50 unique QTL for tocopherol and tocotrienol grain traits. Of these, 13 QTL were resolved to seven *a priori* pathway genes (*arodeH2*, *dxs2*, *hggt1*, *sds*, *vte3*, *hppd1*, and *vte4*) and six novel genes (*por1*, *por2*, *snare*, *ltp*, *phd*, and *fbn*) encoding a predicted function not known to affect tocochromanol traits. Even though maize grain is a non-green, non-photosynthetic tissue, the two protochlorophyllide reductases (*por1* and *por2*) found to be major loci for controlling total tocopherols were hypothesized to be part of a cycle that provides chlorophyll-derived phytol for tocopherol synthesis in the maize embryo. The

involvement of *por2* in the accumulation of tocopherols in maize grain has since been transgenically confirmed by Zhan *et al.* (2019). Although the U.S. NAM panel provided tremendous mapping resolution, not all of the identified unique QTL could be resolved to the gene level, thus other mapping approaches combined with more diverse mapping panels are needed to better pinpoint the underlying candidate causal genes.

For terminal phenotypes such as grain tocochromanols, intermediate phenotypes or endophenotypes can offer orthogonal genetic information to better connect genotype to phenotype. In a statistical approach that generates insight on biological processes, transcriptome-wide association studies (TWAS) correlates mRNA abundance with complex trait variation, allowing for the linkage of an intermediate phenotype to a terminal phenotype. In an assessment of TWAS for the genetic dissection of tocochromanol and carotenoid grain traits in the Goodman-Buckler association panel, Kremling *et al.* (2019) showed that the statistical power to detect causal genes could be increased through an ensemble approach combining GWAS and TWAS results with the Fisher's combined test (FCT). Additionally, this approach was used to identify plausible causal genes associated with natural variation for water use efficiency traits in sorghum (Ferguson *et al.*, 2020; Pignon *et al.*, 2021). The genetic markers used in GWAS could also be linked to mRNA abundance via expression quantitative trait locus (eQTL) mapping (reviewed in Majewski & Pastinen, 2011), enabling the regulatory landscape of gene expression to be better explored as it relates to a terminal phenotype as has been conducted for oil content in maize grain (Hui Li *et al.*, 2013).

In this study, we conducted a comprehensive genetic dissection of tocochromanol grain phenotypes in a large maize association panel that leveraged the scoring of ~12.2

million SNP markers and transcript abundances from developing grain as an endophenotype. Our integrated GWAS and TWAS approach combined with eQTL mapping offered increased statistical power and mapping resolution to pinpoint multiple candidate causal genes and uncover their regulatory control.

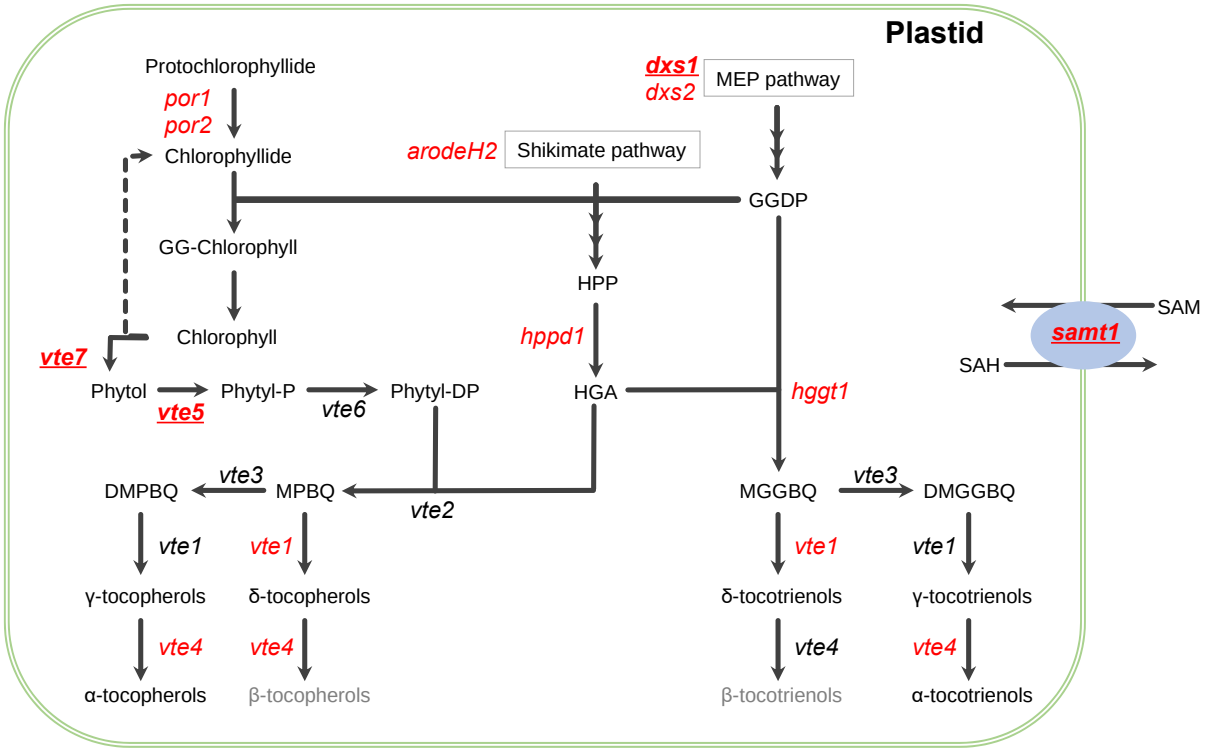

Figure 4.1. Tocochromanol biosynthetic pathways in maize. Precursor pathways are summarized in gray boxes. Key *a priori* genes are in italicized text at the pathway step(s) executed by their encoded enzyme, with the 13 causal genes identified in this study highlighted in red. Four genes that had not been previously shown to be associated with natural variations of tocochromanols in maize grains are bolded and marked with underscores. Compound abbreviations: DMGGBQ, 2,3-dimethyl-5-geranylgeranyl-1,4-benzoquinol; DMPBQ, 2,3-dimethyl-6-phytyl-1,4-benzoquinol; GGDP, geranylgeranyl diphosphate; GG-Chlorophyll, geranylgeranyl-chlorophyll a; HGA, homogentisic acid; HPP, p-hydroxyphenylpyruvate; MGGBQ, 2-methyl-6-geranylgeranyl-1,4-benzoquinol; MPBQ, 2-

methyl-6-phytyl-1,4-benzoquinol; Phytyl-DP, phytyl diphosphate; Phytyl-P, phytyl monophosphate; SAM, S-adenosyl-L-methionine; SAH, S-adenosyl-L-homocysteine. Gene abbreviations: 1-deoxy-D-xylulose-5-phosphate synthase (*dxs1* and *2*); α-/β-hydrolase family protein (*vte7*); arogenate/prephenate dehydrogenase family protein (*arodeH2*); phytol kinase (*vte5*); phytol phosphate kinase (*vte6*); *p*-hydroxyphenylpyruvate dioxygenase (*hppd1*); protochlorophyllide reductase (*por1* and *2*); homogentisate geranylgeranyltransferase (*hggt1*); homogentisate phytyltransferase (*vte2*); MPBQ/MGGBQ methyltransferase (*vte3*); γ-tocopherol methyltransferase (*vte4*); S-adenosylmethionine transporter 1 (*samt1*); tocopherol cyclase (*vte1*).

MATERIALS AND METHODS

Experimental design for genetic mapping

A total of 1,815 maize inbred lines from the North Central Regional Plant Introduction Station association panel (hereafter Ames panel) (Romay *et al.*, 2013) were grown as a single replicate at Iowa State University's Agricultural Engineering and Agronomy Research Farm in Ames, IA, in 2015 and 2017. The Ames panel was arranged in an augmented complete block design with B73 as a repeated check. For each year, two blocking directions were assigned: each range block consisted of three adjacent rows of plots, and each pass block consisted of eight adjacent columns of plots. At least one B73 check plot was planted within each pass block and range block. The inbred lines were grouped into two and three tiers for 2015 and 2017, respectively, according to their days to silk (flowering time) recorded in Romay *et al.* (2013) for the 2015 experimental design and the recorded pollination date in 2015 for the 2017 experimental design. Experimental units were one-row plots that had a length of 3.05 m, with 0.76 m inter-row spacing. There was a 0.76 m alley at the end of each plot. Each plot consisted of ~18 maize plants. An average of six plants (at most 11) per plot were self-pollinated, with pollination dates recorded. Ears that had been self-pollinated were

hand-harvested at physiological maturity. The kernels from all dried and shelled ears of each harvestable plot were bulked to comprise a representative sample for quantification of tocochromanols.

The extraction of tocochromanols from 15-20 mg of ground kernels and their quantification by high-performance liquid chromatography (HPLC) and fluorometry were as previously described (Lipka *et al.*, 2013). Briefly, two types of tocochromanols (tocopherols and tocotrienols) were assessed on 3,539 grain samples from 1,762 inbred lines and the repeated B73 check plots. The nine evaluated tocopherol and tocotrienol phenotypes in $\mu g\ g^{-1}$ dry seed were as follows: α-tocopherol (αT), δ-tocopherol (δT), γ-tocopherol (γT), α-tocotrienol (αT3), δ-tocotrienol (δT3), γ-tocotrienol (γT3), total tocopherols (ΣT, αT + δT + γT), total tocotrienols (ΣT3, αT3 + δT3 + γT3), and total tocochromanols (ΣTT3, total tocopherols + total tocotrienols).

Phenotypic data analysis

To screen the raw HPLC data for significant outliers, we fitted a mixed linear model for each tocochromanol phenotype in ASReml-R version 3.0 (Gilmour *et al.*, 2009). The full model (Equation 4.1) fitted to the data was as follows:

$$Y_{ijklmn} = \mu + check_i + genotype_i + year_j + year \times grp_{ij} + genotype \times year_{ij} + tier(year)_{jk} + pass(tier \times year)_{jkl} + range(tier \times year)_{jkm} + plate(year)_{jn} + \varepsilon_{ijklmn} \quad \text{(Equation 4.1)}$$

in which Y_{ijklmn} is an individual phenotypic observation; μ is the grand mean; $check_i$ is the fixed effect for the ith check, where it is set to 0 if the genotype is a noncheck line; $genotype_i$ is the fixed effect of the ith genotype (noncheck line), where it is set as 0 and omitted if the ith

observation is of a check line; $year_j$ is the effect of the jth year; year × grp_{ij} is the interaction term between the jth year and ith grp, where grp is an indicator variable with two levels that indicates whether the ith observation is from a check or noncheck line; genotype × $year_{ij}$ is the effect of the interaction between the ith genotype and jth year; $tier(year)_{jk}$ is the effect of the kth tier within the jth year; $pass(tier \times year)_{jkl}$ is the effect of the lth pass within the kth tier within the jth year; $range(tier \times year)_{jkm}$ is the effect of the mth range within the kth tier within the jth year; $plate_n$ is the effect of the nth HPLC autosampler plate; and ε_{ijklmn} is the residual error effect assumed to be independently and identically distributed (i.i.d.) according to a normal distribution with mean zero and variance σ_ε^2, that is ~iid $N(0, \sigma_\varepsilon^2)$. Of these terms, μ, check, and genotype were modeled as fixed effects, while all other terms were modeled as random effects. Degrees of freedom were calculated with the Kenward-Roger approximation (Kenward & Roger, 1997). Studentized deleted residuals (Neter *et al.,* 1996) generated from Equation 4.1 were examined to remove outlier observations based on the Bonferroni correction at the $\alpha = 0.05$ level of significance. The removal of 147 significant outliers produced a dataset consisting of raw HPLC data for one or more tocochromanol phenotypes from 1,762 inbred lines and the repeated B73 check plots.

With the outlier-screened phenotypic dataset, we generated best linear unbiased estimator (BLUE) values for the 1,762 inbred lines across years by fitting the full model (Equation 4.1) in ASReml-R (Gilmour *et al.,* 2009). The full model was refitted with genotype as a random effect to generate variance component estimates for the calculation of heritability on a line-mean basis (Holland *et al.,* 2003; Hung *et al.,* 2012). Standard errors of the heritability estimates were calculated with the delta method (Holland *et al.,* 2003; Lynch

& Walsh, 1998). Considering that tocopherols and tocotrienols are unevenly distributed between the embryo and endosperm (Grams *et al.*, 1970; Weber, 1987), kernel starch synthesis genes associate with tocotrienols in sweet corn (Baseggio *et al.*, 2019), and morphologically extreme grain types can potentially have inflated tocochromanol concentrations based on a dry sample weight basis, we conservatively excluded 265 inbred lines that had been classified by Romay *et al.* (2013) and Germplasm Resources Information Network (GRIN; https://www.ars-grin.gov/) as sweet corn, popcorn, or having an endosperm mutation. The remaining set of 1,497 lines had BLUE values for one or more of the nine tocochromanol phenotypes.

Genotype data processing and imputation

The genetic imputation approach implemented in Wu *et al.* (2021) was used to generate a high-density single-nucleotide polymorphism (SNP) marker set in B73 RefGen_v4 coordinates for the Ames panel. To construct the target SNP genotype set, unimputed genotyping-by-sequencing (GBS) SNP genotypes scored at 943,455 loci in the Ames panel by Romay *et al.* (2013) were downloaded from CyVerse (ZeaGBSv27_publicSamples_raw_AGPv4-181023.vcf.gz, available at http://datacommons.cyverse.org/browse/iplant/home/shared/panzea/genotypes/GBS/v27), which provided 1,779 GBS samples for 1,493 of the 1,497 lines that had BLUE values for tocochromanol phenotypes. Given that there were 220 lines with more than one corresponding GBS sample having a call rate ≥ 20%, we followed the approach of Wu *et al.* (2021) to merge two or more GBS samples from the same line. Briefly, a stringently filtered SNP set [call rate

≥ 50%, % heterozygosity ≤ 10%, index of panmixia F_{IT} ≥ 0.8, MAF ≥ 0.01 and linkage disequilibrium (LD) r^2 ≤ 0.2] of 32,267 SNPs derived from the Romay *et al.* (2013) unimputed marker dataset was used to calculate average pairwise identity-by-state (IBS) between multiple samples of the same line using PLINK version 1.9 (Purcell *et al.,* 2007). A total of 19 lines with a mean IBS value < 0.95 for all within-line sample comparisons were removed from the analysis, followed by consensus genotype calling for the remaining 201 lines. Collectively, the final target dataset consisted of a retained 1,462 lines with a call rate ≥ 0.2%, heterozygosity ≤ 10%, and inbreeding coefficient (F) ≥ 0.8. All heterozygous genotype calls were set to missing prior to imputation.

The reference SNP genotype set, which was identical to that constructed in Wu *et al.* (2021), consisted of 14,613,169 SNPs derived from maize HapMap 3.2.1 (Bukowski *et al.,* 2018). In BEAGLE v5.0 (Browning *et al.,* 2018) with parameters as previously specified in Wu *et al.* (2021), the genotypes at the 14,613,169 SNP loci were imputed based on 443,419 GBS SNPs (target set) in the 1,462 Ames panel lines. The resultant imputed dataset was further filtered for SNP quality, resulting in 12,184,805 biallelic SNPs with MAF ≥ 1% and predicted dosage r^2 (DR2) ≥ 0.80 for conducting marker-trait association tests with a mixed linear model. In PLINK version 1.9 with a sliding window of 100 kb and step size of 25 SNPs, the complete set of 12,184,805 SNPs was LD pruned to construct two reduced marker sets: 1) 7,319,895 SNPs with pairwise r^2 < 0.99 for performing marker-trait association tests with a multi-locus mixed model (MLMM), and 2) 344,469 SNPs with pairwise r^2 < 0.10 for estimation of population structure and relatedness.

Genome-wide association study

We conducted GWAS of the tocochromanol grain phenotypes measured on the 1,462 lines with two complementary mixed linear model approaches as previously described (Wu *et al.,* 2021). In brief, to correct for heteroscedasticity and non-normality of error terms, the Box-Cox power transformation procedure (Box & Cox, 1964) was implemented through an intercept-only model with the MASS package version 7.3-50 in R version 3.5.1 (R Core Team, 2018) to select an optimal value (highest log-likelihood in a 95% confidence interval) of convenient lambda (−2 to +2, 0.5 increments) for transforming the non-negative BLUE values of each phenotype (Supplemental Table S4.1). Given that several negative BLUE values were generated in the model fitting process, we added a constant that made all values positive and no less than 1E−09 for αT, δT, γT, and γT3 before applying the transformation (Supplemental Table S4.1). The untransformed and transformed BLUE values of the nine tocochromanol phenotypes for the 1,462 lines are provided in Supplemental Tables S4.2 and S4.3, respectively. Each of the 12,184,805 SNPs from the complete marker set was tested for an association with transformed BLUE values from the 1,462 lines using a mixed linear model (Yu *et al.,* 2006) that employed the population parameters previously determined approximation (Zhang *et al.,* 2010) in the R package GAPIT version 2018.08.18 (Lipka *et al.,* 2012). The fitted mixed linear models controlled for population stratification and familial relatedness through the inclusion of principal components (PCs) and a genomic relationship matrix (kinship matrix). In GAPIT, the reduced set of 344,469 SNPs was used to calculate PCs and the kinship matrix with VanRaden method I (VanRaden, 2008). The optimal number of PCs to include in the mixed linear model fitted for each phenotype was determined by the

Baysian information criterion (BIC) (Schwarz, 1978). The amount of phenotypic variation explained by a SNP was approximated as the difference between the likelihood-ratio-based R^2 statistic (R^2_{LR}) (Sun *et al.,* 2010) of a mixed linear model with or without the SNP. A Pearson's correlation coefficient (*r*) was calculated between the untransformed BLUEs of each pair of phenotypes from the 1,462 lines with the function '*cor*' in R version 3.5.1 (R Core Team, 2018).

To control for the statistical influence of large-effect loci, we used the MLMM approach of Segura *et al.* (2012) to conduct a GWAS of each tocochromanol phenotype with the reduced set of 7,319,895 SNPs that alleviated model constraints by removing perfectly correlated SNPs. Briefly, a stepwise regression was performed with forward selection and backward elimination of significant markers as covariates, with the multiple-Bonferroni criterion (mBonf) used to select the optimal model. In each fitted model for a phenotype, we included the same kinship matrix and BIC-determined optimal number of PCs that had been used for GWAS with the mixed linear model. To explore the complexity of large-effect association signals, we reconducted GWAS in GAPIT using the complete marker set of 12,184,805 SNPs, with the MLMM-selected SNPs included as covariates in the refitted mixed linear models. The "*p.adjust*" function in base R version 3.5.1 (R Core Team, 2018) was used to apply the Benjamini-Hochberg multiple test correction procedure (Benjamini & Hochberg, 1995) to the *P*-values of tested SNPs for each phenotype in each refitted models to control the false-discovery rate (FDR) at 5%.

Experimental design for transcriptomic profiling of developing kernels

In 2018, 1,022 of the 1,815 maize inbred lines from the Ames panel evaluated for grain tocochromanols plus five founders of the U.S. maize NAM panel (McMullen *et al.*, 2009; Yu *et al.*, 2008) not included within our earlier genetic mapping field trials were grown as a single replicate at Iowa State University's Agricultural Engineering and Agronomy Research Farm in Ames, IA. This germplasm set was initially constructed by including 256 lines that met at least one of the following criteria: 1) extreme for a grain metabolite phenotype in the 2015 field trial; 2) founder of the U.S. NAM panel; or 3) available genome assembly. The additionally selected 771 lines were included to maximize genetic diversity and increase the sample size. The 1,027 noncheck lines were partitioned and randomized into 24 blocks based on pollination dates recorded in the 2015 and 2017 field trials and divided across two tiers. To control for spatial variation across the field, the incomplete block design was augmented by planting a B73 check plot within each block. Additionally, we selected two local checks to be planted at random positions in each block to account for temporal variation across fresh harvest dates that spanned more than a month. Within each block, the latest-pollinated line was selected to serve as one of the two local checks; ties were broken by choosing the line with the highest sample call rate based on a filtered partially imputed GBS dataset (https://de.cyverse.org/data/ds/iplant/home/shared/GoreLab/dataFromPubs/Wu_AmesTocochromonols_2021). Each selected local check was also planted in their adjacent later-flowering block, so that two local check lines were present in blocks 2-24. An additional early-flowering line (S 117) was identified as a local check and planted in block 1, ensuring that two local check lines were planted in block 1 as well as in all other blocks. In addition, 25 local checks

(S 117, C38, A508, A641 Goodman-Buckler, C31, 807, LH202, 764, PHG71, PHB47, SD101, A680, B93, NC292, NC280, LH208, H100, NC252, NC314, LH51, CI 187-2, NC324, Mo11, NC334, and M37W) were also planted in a separate third tier to account for field effects on these lines. Experimental units were plots consisting of one row having a 3.05 m length, with a 0.76 m alley length and 0.76 m inter-row spacing. Of the ~18 maize plants per plot, an average of six plants were self-pollinated and pollination dates recorded. A single self-pollinated ear was hand-harvested from each plot at ~23 days after pollination (DAP), followed by immediately freezing the dehusked ear in liquid N and covering it in dry ice until shelling. To control for temporal effects, a self-pollinated ear of a local check from the third tier was hand-harvested at 23 DAP on each day of fresh harvest, with all harvested ears identically processed with liquid N and dry ice prior to shelling. The mid-section of each frozen ear was individually shelled on dry ice and its kernels stored at −80°C until RNA extraction. In total, 1,012 and 107 samples were collected from non-check and check lines, respectively.

RNA isolation and 3′ mRNA Sequencing

Frozen kernels (8-10) were ground using liquid N cooled grinding cups in an IKA Tube Mill Control (IKA-Werke, Staufen, Germany) and approximately 100 mg of ground tissue was used for RNA isolation using a modified hot borate method (Wan & Wilkins, 1994). RNA was DNase treated with the Ambion Turbo DNA-free Kit (Thermo Fisher Scientific, Waltham, MA) and checked for quality using a combination of NanoDrop spectrophotometer (Thermo Fisher Scientific, Wilmington, DE) readings and agarose gel electrophoresis.

Isolated RNA was randomized into 96-well plates and submitted for 3′ mRNA sequencing at the Genomics Facility of the Cornell Institute of Biotechnology. Included in each plate submission were positive controls consisting of the same pool of B73 control RNA aliquoted into four wells in each plate, as well as four negative controls per plate consisting of water. This resulted in two positive and negative controls per lane of sequencing. Libraries were constructed using the Lexogen QuantSeq 3′ mRNA-Seq Library Kit FWD (Lexogen, Greenland, NH) and sequenced on an Illumina NextSeq 500 (Illumina, San Diego, CA) with each plate being split in half and each half being sequenced on a single lane to achieve desirable coverage.

Expression abundance determination

The 3′ QuantSeq reads were cleaned using two rounds of Cutadapt version 2.3 (Martin, 2011) in accordance with Lexogen recommendations (https://www.lexogen.com/wp-content/uploads/2020/04/015UG009V0252_QuantSeq_Illumina_2020-04-03.pdf). In round one, Illumina adapters were trimmed from the reads and in round two the first 12 bases were clipped and polyA tails trimmed to finish cleaning the reads. Reads were then aligned to the B73 RefGen_v4 reference genome (Jiao *et al.,* 2017) using HISAT2 version 2.1.0 (Kim *et al.,* 2019) with the following parameters: --min-intronlen 20, --max-intronlen 60000, --dta-cufflinks, and --rna-strandness F. The resultant alignments were then sorted using SAMTools version 1.9 (Heng Li *et al.,* 2009). Counts were then generated using the htseq-count function within HTSeq version 0.11.2 (Anders *et al.,* 2015) using the B73 version 4.59 annotation with the following parameters: --format=bam, --order=pos, --stranded=yes, --minaqual=10, --

idattr=ID, --type=gene, and --mode=union. The DESeq2 rlog function (Love *et al.,* 2014) was used to normalize the count data. All genes with a normalized count of less than or equal to zero in all samples were removed from the final count matrix. The normalization of the count data using the rlog function of DESeq2 was performed on a set of 1,171 samples that included 1,012 noncheck, 107 check, and 52 B73 positive control samples.

Expression data set quality control

To verify the quality and integrity of the samples, SNPs were called using the 3′ QuantSeq read alignments and compared to SNP calls from a 942 maize line RNA-Seq dataset (WiDiv-942 panel) (Gage *et al.,* 2019). In total, 375 lines overlapped with the WiDiv-942 panel, for a total of 430 3′ QuantSeq samples and 54 positive controls. First, 3′ QuantSeq reads were mapped to the B73 RefGen_v4 assembly (Jiao *et al.,* 2017) following the HISAT2 mapping protocol indicated above. Duplicate reads were identified and marked using Picard tools MarkDuplicates version 2.20.8 (https://broadinstitute.github.io/picard/). Output was sorted using SAMTools sort version 1.9 and a pileup file created using SAMTools mpileup with BAQ computation disabled (-B) and alignments with a mapQ less than 60 were omitted (-q 60), allowing for only unique alignments to be processed. Only positions with a base quality of greater than or equal to 20 were included and all insertions and deletions were discarded. Genotype calls were made at a position in an individual if the coverage was at least five reads, but not greater than 500 reads, and the allele made up greater than 3% of the calls at that position in the individual. If more than two alleles passed the coverage and frequency cutoff, the position was scored as heterozygous and set to missing data only when calculating percent

identity. After removing positions from the WiDiv-942 SNP matrix that were not called in the 3′ QuantSeq dataset, there were 919,074 remaining positions. Percent identity between the same line in the two datasets was calculated by taking the number of positions that had the same genotype call at a position divided by the total number of positions excluding missing data positions in either dataset.

Stringent filtering was employed to curate the final expression dataset and ensure it contained high quality data. Samples were filtered out based on the following criteria: sampling concerns (12 samples removed), number of cleaned reads were below 5 million (1 sample removed), a HISAT2 alignment rate of less than or equal to 65% (17 samples removed), a Pearson's correlation value (r) less than 0.90 with 40 or more samples (3 samples removed), samples that had less than 95% identity when compared to their high confidence WiDiv-942 panel counterpart during genotype confirmation assessment (15 samples removed), and finally removal of samples that had an heterozygosity greater than or equal to 10% (339 samples removed). This final heterozygosity filter was employed to remove samples that were contaminated by spillover during library construction at the Cornell Institute of Biotechnology's Genomics Facility. This stringent heterozygosity filtering was employed to ensure the final dataset was free of contaminating reads that may have impacted downstream analysis. The final data set of 784 high confidence, high quality samples included 43 B73 positive control samples and 741 collected field samples of check and noncheck lines. The B73 positive controls were used during data processing for quality control. Only the 741 check and noncheck samples were used for downstream analysis.

Expression data analysis

The B73 expression data sets consisting of 665 samples for 664 noncheck lines and 76 samples for 25 check lines were separately further stringently filtered at the gene level to ensure high quality data for statistical analysis (Supplemental Table S4.4). All genes that were expressed in less than half of the 741 samples were filtered out, resulting in the removal of 2,195 genes. The rlog-transformed values of the 22,141 retained genes were screened for extreme outliers that could contribute to lack of model convergence, with the method of Davies and Gather (1993) used to remove values that exceeded a conservative threshold of 100 median absolute deviations from the median for a given gene. This resulted in the removal of 2,008 extreme rlog-transformed values. The number of removed outliers exceeded 10% for five genes, thus these genes were filtered out. Of the 22,136 retained genes, only 132 had extreme outliers. The removed outliers were replaced with the median rlog-transformed value for a given gene because the downstream applied PEER (probabilistic estimation of expression residuals) approach (Stegle *et al.,* 2012) did not allow for missing data. With the outlier-screened expression data set, we fit a mixed linear model that allows for modelling of genetic and field and laboratory non-genetic effects. To account for the effect of temperature differences given that not all ears could be harvested exactly at 23 DAP and differential rates of grain development among lines, growing degree days (GDD) was also included as a model term. The number of days from pollination date to fresh-harvest date were converted to GDD according to equation 1 from Bollero *et al.* (1996). For each gene, BLUE values were generated for each of the 664 noncheck lines in ASReml-R version 3.0 (Gilmour *et al.,* 2009) as follows:

$$Y_{ijklmn} = \mu + check_i + genotype_i + \alpha \times GDD_j + tier_k + block(tier)_{kl} + plate_m + lane(plate)_{mn} + \varepsilon_{ijklmn} \quad \text{(Equation 4.2)}$$

in which Y_{ijklmn} is an individual rlog-transformed value; μ is the grand mean; $check_i$ is the fixed effect for the ith check, where it is set to 0 if the genotype is a noncheck line; $genotype_i$ is the fixed effect of the ith genotype (noncheck line), where it is set as 0 and omitted if the ith observation is of a check line; α is a scalar regression coefficient for GDD_j for plants harvested on the jth day; $tier_k$ is the kth tier; $block(tier)_{kl}$ is the lth block in the kth tier; $plate_m$ is the mth RNAseq plate; $lane(plate)_{mn}$ is the nth lane (minimum unit of the RNAseq run) in the mth plate; and ε_{ijklmn} is the residual error effect assumed to be ~iid $N(0, \sigma_\varepsilon^2)$. With the exception of the grand mean, check, genotype and GDD, all terms were fitted as random effects. Of the 664 lines, we excluded 104 classified as sweet corn, popcorn, or with an endosperm mutation and an additional 15 lines not analyzed in GWAS. The final data set contained BLUE expression values of 22,136 genes across 545 lines.

To account for inferred confounders that influence expression variation, the PEER approach (Stegle *et al.*, 2012) was separately applied to the 545 line × 22,136 gene matrix of BLUE expression values. The number of PEER hidden factors was determined to be eight based on the "elbow criterion" in a diagnostic plot of the factor relevance, with the number of hidden factors ranging from 1 to 25. The contribution of these factors was subtracted to generate a residual data set of the BLUE expression values (hereafter, PEER values).

To screen the PEER values for significant outliers, we fitted a simple linear model that had only the grand mean of each gene in ASReml-R version 3.0 (Gilmour *et al.*, 2009), with the obtained Studentized deleted residuals (Neter *et al.*, 1996) examined to remove outliers for

each gene at a Bonferroni adjusted significance threshold of $\alpha = 0.05$.

Given that the *vte7* locus consists of tandemly duplicated genes (Zm00001d006778 and Zm00001d006779) with high pairwise nucleotide sequence identity (> 99%) in the B73 RefGen_v4 assembly (Jiao *et al.*, 2017), the reads pertaining to these two genes were not uniquely mappable with the above described bioinformatic pipeline. Therefore, to calculate the transcript abundances at the *vte7* locus, the number of read alignments to the two gene models using multi-mapping reads were summed and normalized to counts per million alignments (CPMA) as (total count within both loci/total reads aligned)*1,000,000. Next, the CPMA values were fitted with the Equation 4.2 model to generate BLUE values, which were then screened for outliers as described for the PEER values.

Transcriptome-wide association study

To associate gene expression values with tocochromanol phenotypes, we conducted a transcriptome-wide association study (TWAS) on the 545 lines and 22,136 genes with a mixed linear model approach that controlled for population stratification and familial relatedness (Yu *et al.*, 2006; Zhang *et al.*, 2010). Briefly, a mixed linear model was fit for the combination of each tocochromanol phenotype (transformed BLUE values, response variable) and expressed gene (outlier-screened PEER values, explanatory variable) similar to that of Kremling *et al.* (2019). To construct the SNP marker set for the 545 lines, 12,018,644 biallelic SNPs (DR2 $\geq$ 0.80; MAF $\geq$ 1%) were subsetted from the full set of 14,613,169 SNP loci and pruned down to 328,892 SNPs with pairwise $r^2 < 0.10$ in PLINK version 1.9 as described in the *Genotype data processing and imputation* section. The PCs and the kinship matrix were

generated from the 328,892 SNPs in the R package GAPIT version 2018.08.18 (Lipka *et al.,* 2012) as described in the *Genome-wide association study* section. The mixed linear model fitted for each tocochromanol phenotype*gene expression combination was implemented with the "*GWAS*" function, which allows for continuous explanatory variables, and setting the "P3D" option to FALSE in the R package rrBLUP version 4.6 (Endelman, 2011). The optimal models for all tocochromanol phenotypes included kinship and no PCs, as determined by the BIC (Schwarz, 1978). BIC values were based on log-likelihoods calculated with the method of Kang *et al.* (2008) using estimated variance components and effects of PCs generated from the "*mixed.solve*" function in rrBLUP. The association analysis with transcript abundances from the *vte7* locus was conducted separately from the other genes, given that reads for the *vte7* locus were uniquely processed. However, the *P*-value of the association test was compared to those from TWAS.

FCT of GWAS and TWAS

The top 10% of the most associated SNPs (1,218,480 SNPs) from GWAS with the mixed linear model were selected to perform FCT following Kremling *et al.* (2019). In brief, the GWAS *P*-value of each top 10% SNP was assigned to its nearest gene (gene direction was not considered) based on the B73 RefGen_v4 assembly and then paired with the TWAS *P*-value for that gene. We used the B73 version 4.59 GFF file to identify the physical position of genes, as this same GFF file was used for 3′ QuantSeq read alignments. The GFF file was initially filtered to remove the following genes: 1) 855 genes that were mapped to contigs or scaffolds, and 2) 627 non-protein-coding genes that were duplicated in annotation. This

resulted in a set of 44,918 genes available for the FCT. When two or more gene regions overlapped (4,464 of the 44,918 genes), we merged them into a single gene region based on the union of their physical positions. In these instances, the nearest SNPs were paired with each of the merged genes. For genes that were not scored by 3′ QuantSeq or filtered out in previous quality control steps, their TWAS *P*-values were set to 1 before combining with GWAS *P*-values. FCT was conducted with the "*sumlog*" function implemented in the R package metap version 1.1 (Dewey, 2019).

Candidate gene identification

Given that GWAS, TWAS, and FCT differ in their statistical power and structure, we did not directly compare *P*-values thresholds across methods but instead used the rankings of *P*-values according to Kremling *et al.* (2019). The top 0.02% of SNPs were selected according to their *P*-value from GWAS results for each phenotype, with selection of the percentage threshold guided by the oligogenic genetic architecture of these phenotypes in the U.S. maize NAM panel (Diepenbrock *et al.*, 2014). To allow for a separate assessment of GWAS results, the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) was used to control for the multiple testing problem at 5% FDR as described in the 'Genome-wide association study' section. Given the complex association signals detected via GWAS with the mixed linear model, a set of associated loci was constructed from the selected SNPs separately for each tocochromanol phenotype following the approach of Wu *et al.* (2021) with minor modification. Briefly, each declared locus consisted of at least two SNPs within 100 kb of one another, with the peak SNP (*i.e.*, SNP with smallest *P*-value) of the locus having an estimated

pairwise r^2 value < 0.05 with all other peak SNPs at loci on the same chromosome as calculated in TASSEL v5.2.49 (Bradbury *et al.,* 2007). Considering the rapid LD decay in the Ames panel (Romay *et al.,* 2013), the search interval for candidate genes was restricted to ± 100 kb of each peak SNP. The top 0.5% of genes according to their *P*-value were selected from TWAS and FCT results for each phenotype, resulting in a total number of unique genes identified across phenotypes by each method comparable to that of GWAS. The identification of plausible causal genes was assisted by a list of 125 *a priori* candidate genes involved in aromatic head group synthesis, prenyl group synthesis, chlorophyll synthesis, and from the core tocochromanol pathway (Supplemental Table S4.5). The physical positions of 50 unique JL-QTL common support intervals (CSIs) and GWAS markers associated with these tocochromanol phenotypes in the U.S. NAM panel (Diepenbrock *et al.,* 2017) were uplifted via Vmatch version 2.3.0 (Kurtz, 2010) to B73 RefGen_v4 coordinates (Supplemental Tables S4.6 and S4.7) following the approach described in Wu *et al.* (2021). A BLASTP with default parameters was conducted to identify the top three non-redundant hits (E-values < 1) of undescribed candidate causal genes in Arabidopsis (Columbia-0 ecotype) and rice (*Oryza sativa* L. ssp. Japonica cv. 'Nipponbare') at TAIR (https://www.arabidopsis.org) and RAP-DB (https://rapdb.dna.affrc.go.jp) databases, respectively (Supplemental Table S4.8).

eQTL mapping

We performed expression QTL (eQTL) mapping to better understand the genetic regulatory landscape of gene expression variation for the identified candidate causal genes in developing grain. To conduct eQTL mapping, the 12,018,644 SNPs used in the TWAS approach were

individually tested for association with PEER values of each candidate causal gene using a mixed linear model implemented in GAPIT version 2018.08.18 (Lipka *et al.,* 2012) in R 3.5.1 (R Core Team, 2018). The calculated PCs and kinship matrix used in TWAS were also used in eQTL mapping to control for population structure and familial relatedness, with the optimal number of PCs determined by BIC. To have a stringent control of the Type I error rate in the presence of complex LD patterns and strong association signals, we accounted for multiple testing with a 5% Bonferroni significance threshold (*P*-value ≤ 4.16E−09). Significant loci were declared as described for GWAS signals in the *Candidate gene identification* section.

To integrate the GWAS and TWAS findings with eQTL analysis, the candidate causal genes were separated into two groups based on r^2 values between cis-eQTL and GWAS peak SNP signals (< 0.05: independent; > 0.05: correlated). The differences in three genomic features between the two groups were examined: 1) the distance of a GWAS peak SNP signal to the gene, 2) the distance of a cis-eQTL peak SNP signal to the gene, and 3) the TWAS ranking of each gene. For each genomic feature, a one-tailed *t*-test assuming equal variances as determined by the Levene's test in 'car' package version 3.0-10 was used to test for differences between means of the two groups in R version 3.5.1 (R Core Team, 2018).

RESULTS

Phenotypic variation

We assessed the extent of quantitative variation for tocochromanol concentrations in physiologically mature grain samples harvested from two outgrowths of the maize Ames panel. The measurement of six tocochromanol compounds by HPLC showed that γT (~55%)

and γT3 (~23%) collectively accounted for nearly 80% of ΣTT3, whereas the α- and δ-species for both tocopherols and tocotrienols individually represented approximately 1 (δT3) to 10% (αT3) of ΣTT3 (Table 4.1). The tocochromanol compound with the highest vitamin E activity, αT, had the third lowest mean concentration (5.83 μg g^{-1} dry seed) and accounted for only ~8% of ΣTT3. Indicative of common genetic control, pairwise correlations were strongest within a compound class between the δ- and γ-species for tocopherols (r = 0.67) and tocotrienols (r = 0.62) and between compound classes for αT with αT3 (r = 0.45). However, only relatively weaker correlations (−0.15 to 0.19) were found between all other compounds despite having a shared biosynthetic pathway (Supplemental Figure S4.1). As inferred from the high estimates of heritability on a line-mean basis (0.77 to 0.94), the majority of variation for each of the six tocochromanol compounds and three sum phenotypes was attributable to genetic variation in the full Ames panel (Table 4.1 and Supplemental Figure S4.2).

Genetic analysis of grain tocochromanol levels

We integrated GWAS and TWAS results through FCT, an ensemble approach shown to have enhanced statistical power over either GWAS or TWAS alone for the detection of causal genes associated with natural variation for tocochromanol grain phenotypes in maize (Kremling *et al.*, 2019). The findings from FCT (top 0.5%), GWAS (top 0.02%), and TWAS (top 0.5%) for each phenotype were integrated with the genetic mapping results of the same grain phenotypes in the U.S. NAM panel (Table 4.2), with the intent to further resolve loci previously found in the NAM panel to the level of causal genes (Figure 4.2, Supplemental Figures S4.3 - S4.10). A total of 720, 676, and 918 unique genes were identified across the

nine phenotypes in GWAS, TWAS, and FCT, respectively (Supplemental Table S4.9, S4.10, and S4.11). Of these, 330 (GWAS), 299 (TWAS), and 646 (FCT) genes were located within NAM JL-QTL CSIs for the nine tocochromanol phenotypes (Diepenbrock *et al.,* 2017).

Table 4.1. Means, ranges, and standard deviations (Std. Dev.) of untransformed BLUE values (in μg g^{-1}) for nine tocochromanol grain phenotypes evaluated in the Ames panel and estimated heritability on a line-mean basis and genomic heritability and their standard errors (Std. Err.) across two years.

Phenotype	Number of lines	BLUEs			Heritabilities	
		Mean	Range	Std. Dev.	Estimate	Std. Err.
αT	1452	5.83	–1.79 - 41.36	4.59	0.87	0.006
δT	1456	1.74	–0.33 - 14.32	1.62	0.85	0.007
γT	1458	42.19	–1.32 - 158.91	21.34	0.86	0.006
ΣT	1460	49.95	1.79 - 174.84	22.65	0.85	0.007
αT3	1456	7.87	0.89 - 23.39	3.21	0.77	0.010
δT3	1454	0.93	0.01 - 17.05	0.97	0.94	0.003
γT3	1458	17.60	–1.79 - 90.39	11.71	0.93	0.003
ΣT3	1458	26.55	2.62 - 111.01	13.32	0.91	0.004
ΣTT3	1460	77.04	18.13 - 205.36	28.08	0.87	0.006

Of the 14 genes identified to associate with grain tocochromanols in the U.S. NAM panel by Diepenbrock *et al.* (2017), five (*por1*, *por2*, *vte4*, *hggt1*, and *hppd1*), which tended to be large-effect loci in the NAM panel, were detected by FCT for one or more phenotypes in the Ames panel (Table 4.2). Of the five genes, *por1*, *por2*, *vte4*, and *hggt1* were also identified by both GWAS and TWAS, whereas *hppd1* was only detected by GWAS. Two copies of *arodeH2* (Zm00001d014734 and Zm00001d014737) were within 100 kb of GWAS peak SNPs for γT3 and ΣT3, with the Zm00001d014734 gene having been previously implicated by Diepenbrock *et al.* (2017) in the genetic control of αT3 and ΣT3. Interestingly,

in the Ames panel, Zm00001d014737 was detected by both FCT and GWAS, whereas Zm00001d014734 was detected by GWAS only. An additional one of the 14 genes identified by Diepenbrock *et al.* (2017), *dxs2*, was detected by TWAS only.

The detection of these seven *a priori* pathway genes illustrated the gene-level resolution of our integrated genetic mapping approach, thus it was applied to better resolve NAM JL-QTL CSIs and detect loci novel to the Ames panel. In total, four NAM JL-QTL CSIs were further dissected, resulting in novel associations with three genes (*samt1*, *vte7*, and *dxs1*) and more precise mapping of a fourth gene (*vte1*) not fully resolved in the US NAM panel. A gene encoding a SAM transporter (*samt1*, Zm00001d017937) was detected by FCT, GWAS, and TWAS. This gene encodes a predicted protein that has 77% identity at the amino acid sequence level to SAMT1/SAMC1 (At4g39460) in Arabidopsis (Supplemental Table S4.8), which transports SAM, a tocopherol cosubstrate, through plastid envelopes and affects leaf tocopherol levels when silenced or knocked out (Bouvier *et al.*, 2006; Palmieri *et al.*, 2006). The *vte7* locus was found to be ~64 kb from a single SNP associated with δT in GWAS at 0.02% (Figure 4.2). Providing stronger evidence for the detection of *vte7* in the Ames panel, this same SNP served as the peak of a declared δT-associated locus consisting of 45 significant SNPs at an FDR of 5% (Supplemental Table S4.12). Two additional *a priori* pathway genes, *dxs1* (TWAS) and *vte1* (FCT and GWAS), were associated with several tocochromanol phenotypes. Indicating the value of the Ames panel beyond genetically dissecting unresolved NAM QTL, we detected an association of *vte5* with ΣTT3 by GWAS

alone, which is the first report of this locus to be associated via GWAS with any tocochromanol grain trait in maize.

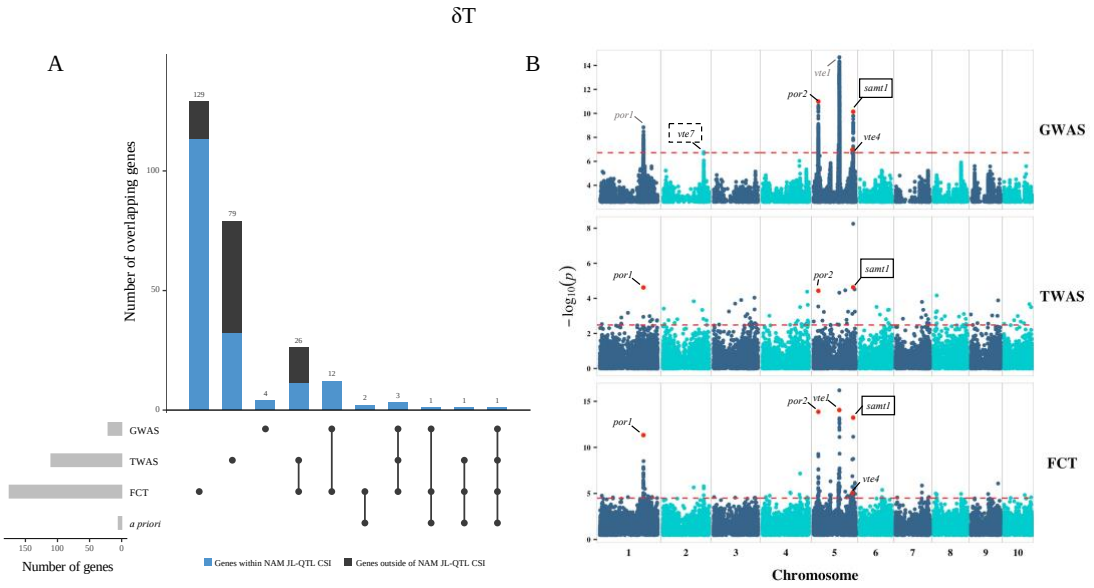

Figure 4.2. GWAS, TWAS and FCT results for δT. A: Upset plot showing the number of overlapping genes between GWAS, TWAS, FCT, and *a priori* pathway genes involved in the biosynthesis of chlorophylls and tocochromanols (Supplemental Table S4.5). The number of genes that are within the NAM JL-QTL CSI for δT is highlighted in blue in the bar plots. B: Manhattan plots of GWAS, TWAS, and FCT results. Red horizontal dashed lines indicate the thresholds of top 0.02%, top 0.5% and top 0.5% for GWAS, TWAS, and FCT, respectively. Causal genes (Table 4.2) that are within 100 kb of a top 0.02% GWAS peak SNP or ranked top 0.5% in TWAS or FCT are highlighted with red dots and labeled in black in the Manhattan plots. Causal genes that are within 1 Mb of a top 0.02% GWAS peak SNP are labeled in gray. Novel genes are marked with a black rectangle. Novel gene that passed 5% FDR in GWAS are marked with a black rectangle with dashed line.

We further supported our findings through conducting a GWAS with the MLMM approach, allowing us to control for large-effect loci. Of the eleven genes detected by GWAS with the MLM, eight genes (*por2*, *vte1*, both *arodeH2* copies, *hppd1*, *vte4*, *samt1*, and *hggt1*)

were located within 100 kb of at least one MLMM-selected SNP for one or more of the nine tocochromanol phenotypes (Supplemental Table S4.13). Although at slightly lower mapping resolution, the *por1* gene was located ~162 kb from one of the multiple MLMM-selected SNPs for tocopherol phenotypes. Of the MLMM-selected SNPs within 100 kb of *vte4*, 2-4 SNPs were selected for αT, αT3, and γT, whereas only a single SNP each was selected for δT and γT3. Comparably, 2-3 SNPs from a ~1.2 Mb genomic region that included *hggt1* were selected by the MLMM for δT3, γT3, and ΣT3; however, only two of the MLMM-selected SNPs were located within 100 kb of *hggt1*. The selection of multiple independent SNPs by the MLMM implies that multiple alleles (*i.e.*, allelic heterogeneity) exist at the *vte4* and *hggt1* loci in the Ames panel.

eQTL mapping

To gain insights into the regulatory patterns of the causal genes identified through GWAS, TWAS, and FCT in the Ames panel (Table 4.2), eQTL mapping was conducted for each of the 13 identified candidate causal genes (Figure 4.3, Supplemental Figure S4.11). Of the 13 genes, cis-eQTL (peak SNP within 1 Mb of gene) were identified for all but one gene (*arodeH2*, Zm00001d014737), whereas five trans-eQTL were identified for four genes (*vte5*, *por2*, *dxs1*, and *dxs2*) (Supplemental Table S4.14). The peak SNPs for cis-eQTL were within 100 kb of their respective gene, with the exception of *arodeH2* (Zm00001d014734, 227 kb), *dxs2* (808 kb), and *vte1* (220 kb) (Supplemental Table S4.14). In general, cis-eQTL were more statistically significant than trans-eQTL, but the trans-eQTL for *dxs2* was more significant than its cis-eQTL. This trans-eQTL for *dxs2* was located on chromosome 6, having

a peak SNP 1.5 kb from *phytoene synthase1* (*psy1*, Zm00001d036345)—a gene encoding an enzyme involved in the first committed step of carotenoid biosynthesis (Hirschberg, 2001).

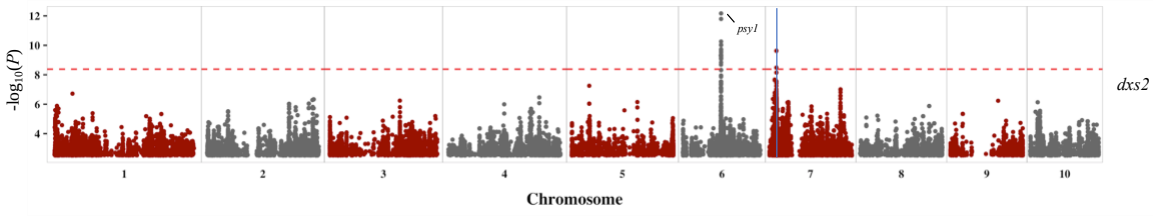

Figure 4.3. Manhattan plot of eQTL mapping results of *dxs2*. Each point represents a SNP with its -$\log_{10}$ *P*-value (y-axis) from a mixed linear model analysis plotted as a function of physical position (B73 RefGen_v4) across the 10 chromosomes of maize (x-axis). The red horizontal dashed line indicates the significant threshold after Bonferroni correction (α = 0.05). Genomic position of *dxs2* is marked by the blue vertical lines in the respective plots. Causal candidate gene within 100 kb of the eQTL peak SNP are labeled in the Manhattan plot.

The extent of LD between cis-eQTL and GWAS signals at casual genes and its relationship to the rankings of these genes in TWAS were investigated to better understand the impact of regulatory variation on the grain tocochromanol phenotypes (Figure 4.4). Our tested hypothesis was that the causal genes with GWAS signals that co-locate with cis-eQTL (*i.e.*, regulatory variants are tagged by SNPs in GWAS) near their ORF will have higher ranking in TWAS. When focusing on the 10 causal genes identified by GWAS that had detected cis-eQTL, we found that r^2 values between cis-eQTL and GWAS peak SNPs for six gene*phenotype combinations were less than 0.05 (independent), whereas the r^2 values exceeded 0.05 (correlated) for the other 16 gene*phenotype combinations. The correlated group included six genes (*por2*, *hggt1*, *vte4*, *samt1*, *vte1*, and *hppd1*), while the independent

group included five genes [*vte5*, *vte7*, *arodeH2* (Zm00001d014734), *vte4*, and *por1*]. The *vte4* gene was included in both groups, as its r^2 values were > 0.05 for αT, γT, αT3, γT3, and ΣTT3, but < 0.05 for δT (0.03). Significant differences ($P < 0.05$) were detected between the means of the correlated and independent groups for the distance of cis-eQTL (26.0 vs. 98.2 kb) and GWAS signal (15.0 vs. 53.5 kb) peak SNPs from their respective causal genes, with even greater separation between groups (correlated 0.7 and 1.1 kb vs independent 49.4 and 59.1 kb) observed for the median distance values of both cis-eQTL and GWAS signals. The gene*phenotype combinations in the correlated group (mean 8.06%) ranked significantly higher ($P < 0.05$) in TWAS relative to the independent group (mean 40.88%), but the median of rankings (correlated 0.03% vs. independent 35.62%) revealed a greater distinction between groups. Collectively, our findings support our hypothesis that expression-level variation most strongly correlates with grain tocochromanol levels in TWAS when cis-eQTL and GWAS signals co-locate in genomic regions where promoters and short-range regulatory elements are expected to reside.

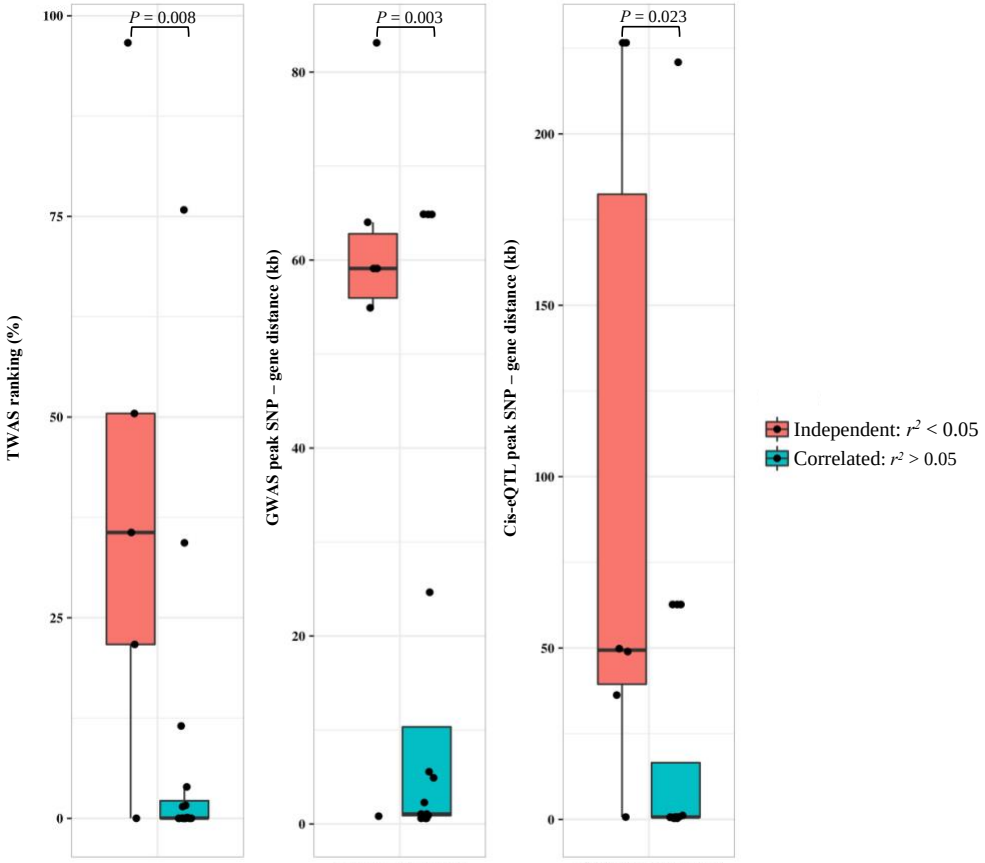

Figure 4.4. Box plots showing the GWAS, TWAS and eQTL results of the causal genes. Results were grouped according to the LD (r^2) between GWAS and cis-eQTL peak SNPs: Independent: $r^2 < 0.05$; Correlated $r^2 > 0.05$. For GWAS and eQTL, results of peak SNPs were used.

.

Table 4.2. GWAS, TWAS, FCT results of the nine tocochromanol phenotypes in the Ames panel.

Gene ID	Gene	Chr	Gene start	Gene end	GWAS	TWAS	FCT	NAM JL-QTL CSI ID[a]
Zm00001d032576	*por1*	1	231,120,510	231,123,615	ΣT	δT, γT, ΣT, ΣTT3	δT, γT, ΣT, ΣTT3	NAM_JL_5
Zm00001d001896	*vte5*	2	2,509,567	2,511,414	ΣTT3			
Zm00001d006778	*vte7*	2	216,443,683	216,448,352	γT			NAM_JL_12
Zm00001d013937	*por2*	5	25,431,430	25,434,346	δT, γT, ΣT, ΣTT3	αT, δT, γT, ΣT, ΣTT3	αT, δT, γT, ΣT, ΣTT3	NAM_JL_24
Zm00001d014734	*arodeH2*	5	61,099,110	61,100,192	γT3, ΣT3			NAM_JL_25
Zm00001d014737	*arodeH2*	5	61,117,986	61,119,432	γT3, ΣT3		δT3, γT3, ΣT3	NAM_JL_25
Zm00001d015356	*hppd1*	5	86,084,655	86,086,755	γT3, ΣT3		δT3, γT3, ΣT3, ΣTT3	NAM_JL_26
Zm00001d015985	*vte1*	5	136,805,708	136,822,194	δT3		δT, δT3	NAM_JL_26
Zm00001d017746	*vte4*	5	205,825,586	205,829,216	αT, αT3, δT, γT, γT3, ΣTT3	αT, αT3, γT3	αT, αT3, δT, γT, γT3	NAM_JL_28
Zm00001d017937	*samt1*	5	210,385,310	210,401,948	αT3, δT	δT	δT, δT3	NAM_JL_29
Zm00001d038170	*dxs1*	6	150,418,144	150,422,431		γT3		NAM_JL_32
Zm00001d019060	*dxs2*	7	14,494,700	14,497,925		δT3, γT3, ΣT3		NAM_JL_35
Zm00001d046558	*hggt1*	9	95,895,575	95,899,061	αT3, δT3, γT3	δT3, γT3, ΣT3	δT3, δT3, γT3, ΣT3, ΣTT3	NAM_JL_45

[a]Common support intervals (CSI) from joint-linkage QTL (JL-QTL) results of nine tocochromanol grain phenotypes analyzed in the maize NAM panel (Diepenbrock *et al.*, 2017) that contain the open reading frame of the gene (Supplemental Table S4.6)

DISCUSSION

In this study, we provided the most comprehensive study to date on the genetic basis of tocochromanol levels in maize grains by integration of GWAS, TWAS, FCT, and eQTL analyses. By utilizing the Ames diversity panel of over one thousand diverse inbred lines scored with 12 million SNP markers and the expression profiles of 22,136 genes at 23 DAP developing grains, we were able to identify 13 causal genes controlling the natural variation of tocochromanol levels in maize grain. Of the 13 genes identified, two novel genes (*samt1, vte7*) were first implicated as large-effect loci, and two *a priori* genes (*dxs1, vte5*) identified had not been previously shown to be associated with natural variation of tocochromanols in maize grain (Li *et al.*, 2012; Lipka *et al.*, 2013; Diepenbrock *et al.*, 2017; Wang *et al.*, 2018).

A total of 11 *a priori* pathway genes were identified in this study through the combination of GWAS, TWAS and FCT. These 11 genes are involved in the head group biosynthesis (two *arodeH2* copies, *hppd1*), tail group biosynthesis (*dxs1*, *dxs2*, *vte5*, *por1*, and *por2*), and core tocochromanol pathway (*hggt1*, *vte1*, and *vte4*), which has been consistent with their known biological activities (Shintani and DellaPenna, 1998; Cahoon *et al.*, 2003; Cheng *et al.*, 2003; Collakova and DellaPenna, 2003; Karunanandaa *et al.*, 2005; Valentin *et al.*, 2006; Hunter and Cahoon, 2007; DellaPenna and Mène-Saffrané, 2011). As a result of the high resolution of this Ames panel, *vte1* and *hppd1*, located in the pericentromeric region of chromosome 5 and previously unresolvable in the U.S. NAM and the Goodman-Buckler association panel (Lipka *et al.*, 2013; Diepenbrock *et al.*, 2017), can now be distinctly resolved to gene levels in our study (Table 4.2, Supplemental Figures S4.3 - S4.10,

Supplemental Table S4.8). Notably, the two *por* genes, encoding chlorophyll biosynthesis pathway enzymes and novel genes firstly implicated in Diepenbrock *et al.* (2017) to control the level of tocopherols in non-green, non-photosynthetic maize grains, were also top-ranked genes in GWAS, TWAS and FCT for tocopherols in this study.

On chromosome 5, two *arodeH2* copies (Zm00001d014734 and Zm00001d014737) were identified in GWAS of γT3 and ΣT3 (Table 4.2). These two genes were ~18 kb apart, with Zm00001d014737 closer to the peak GWAS SNP (5_61159296). These two genes were both within NAM JL-QTL CSI 25 of Diepenbrock *et al.* (2017), and only Zm00001d014734 was identified as the causal gene. Interestingly, although Zm00001d014734 was not a ceeQTL in U.S. NAM (defined as 'significant correlations between expression values and JL-QTL allelic effect estimates at more than two time points for at least one trait'), the expression of Zm00001d014737 was significantly correlated with JL-QTL allelic effect estimates of ΣT3 at 12 and 16 DAP developing grain, making it a ceeQTL. Therefore, it is possible that Zm00001d014737 is the actual causal gene within the genomic region, or both copies are involved in the head group biosynthesis. However, we do not have enough evidence to more strongly support one hypothesis over the other, and further studies would be needed to illustrate their respective functions and interactions.

A pair of homologs encoding 1-deoxy-D-xylulose-5-phosphate synthases were identified in TWAS for δT3 (*dxs2*), γT3 (*dxs1* and *dxs2*), and ΣT3 (*dxs2*). Of the two genes, *dxs2* has been previously reported to be associated with tocotrienols (Diepenbrock *et al.*, 2017), while to the best of our knowledge, this is the first time that the expression of *dxs1* has been reported to affect the tocotrienol levels in mature maize grain. In the U.S. NAM panel,

dxs1 resides within QTL32 (Supplemental Table S4.6), which spanned ~26 Mb on chromosome 6 for αT, γT3, ΣT3 and ΣTT3; however, the significant GWAS SNPs were all more than 2 Mb from *dxs1* (Supplemental Table S4.7). In agreement with Diepenbrock *et al.* (2017), the expression of *dxs2* was strongly associated with solely tocotrienol phenotypes, although *dxs2* was not identified in GWAS for any tocotrienol phenotypes in this study. The cis-eQTL of *dxs2* was the second highest peak in eQTL and its peak SNP was located 800 kb from the ORF of *dxs2*, potentially explaining the lack of GWAS detection of *dxs2* in this panel. The most significant eQTL for *dxs2* was located on chromosome 6 and 1.5 kb from Zm00001d036345 (*phytoene synthase 1*; *psy1*) (Supplemental Table S4.13), which encodes the enzyme in the first committed step of carotenoid biosynthesis (Hirschberg, 2001). The epistatic interaction of *psy1* and *dxs2* had been reported before in tomato (Lois *et al.*, 2000; Kachanovsky *et al.*, 2012), but this is the first time that this regulatory role of *psy1* to *dxs2* has been observed in maize grain.

A novel gene (*samt1*) was identified for δT on chromosome 5 through both GWAS and TWAS, which encodes a S-adenosylmethionine carrier 1 that transports SAM through plastid envelopes (Bouvier *et al.*, 2006; Palmieri *et al.*, 2006). In the U.S. NAM panel, this gene was located within JL-QTL CSI 29 for δT, δT3, and ΣTT3; however, as the GWAS signal was more than 100 Kb away from the gene this interval was not resolved to gene level (Diepenbrock *et al.*, 2017). SAM is an important substrate for both chlorophyll and tocochromanol biosynthesis, with Arabidopsis knockout mutants showing significantly lowered chlorophyll and αT levels and slightly increased γT level in Arabidopsis (Bouvier *et al.*, 2006). This same pattern of chlorophyll and αT level decreases and γT level increase was

also observed in the independent experiment with the *SAMT1*–silenced *N. benthamiana* plants (Bouvier *et al.,* 2006). It is possible that this SAM transporter can affect the level of δT both directly in the tocochromanol biosynthesis pathway as the cosubstrate of *vte3* and *vte4* (Lipka *et al.*, 2013) and also indirectly through chlorophyll biosynthesis pathway, given the large effect sizes of *por1* and *por2* suggesting vital role of chlorophyll in tocopherol biosynthesis (Diepenbrock *et al.*, 2017). In the U.S. NAM panel, *samt1* was a ceeQTL, its expression in 16 and 20 DAP developing grains having significant and negative correlations ($r < -0.5$) with the allelic effect estimates of δT. Given that chlorophyll levels should be positively correlated with tocopherol levels as the direct and major donor of phytol group for tocopherol biosynthesis (Diepenbrock *et al.*, 2017), these negative correlations would suggest that *samt1* is more likely to affect δT level directly as cosubstrate of *vte4*, as increased SAM could lead to more δT being catalyzed by *vte4.* Interestingly, we did not find strong evidence of *samt1* affecting the levels of the γ- and α- branch of tocopherol. It is possible that *vte3* and *vte4* have different affinity towards SAM and as MAF of eight SNPs within *vte3* are all less than 0.1 in the Ames panel, statistical power was not enough to differentiate the effect of *samt1* on the γ- and α- branch of tocopherol biosynthesis. Further study would be needed to fully illustrate the functionality of *samt1* in tocochromanol biosynthesis.

Another novel locus identified in GWAS was *vte7* on chromosome 2 for δT, which consists of tandemly duplicated genes that encode an alpha/beta-hydrolase (Albert *et al.* in prep). This gene was recently characterized to be a plastid-localized hydrolase that links the chlorophyll and tocopherol biosynthesis pathways and regulates the tocopherol levels in both Arabidopsis and maize. Separate TWAS was conducted for *vte7* due to the tandem

duplication (Zm00001d006778 and Zm00001d006779) in the B73 RefGen_v4 genome, and compared with TWAS of genome-wide genes. None of the rankings were within top 0.5% in TWAS, with the highest ranking being top 0.6% for δT. This gene was located within the JL-QTL CSI 12 in the U.S. NAM panel for δT, αT, γT, ΣT and ΣTT3 (Diepenbrock *et al.*, 2017). However, the most significant GWAS marker was ~500 Kb from the gene and therefore this JL-QTL was not resolved in that study. Notably, only one copy of *vte7* was observed in the *de novo* assemblies of all other NAM parents (Hufford, Seetharam and Woodhouse, 2021). We also checked for tandem duplication in the *de novo* assemblies of five stiff stalks (B84, LH145, NKH8431, PHB47, PHJ40) (available at http://maize.plantbiology.msu.edu), and only NKH8431, which has a B73 background, has the tandem duplication within the genome. In the U.S. NAM panel, as all of the other parents have a single copy of *vte7*, we would expect this QTL to be detected in all families if copy number variation at this locus is causal (Diepenbrock *et al.*, 2017). In fact, this QTL was detected in 18, 17, and 16 families out of the 25 NAM families for δT, γT, ΣT, respectively. Therefore, we lack enough evidence to determine whether causality is attributed to structural (tandem gene duplication) and/or non-structural variants (SNP, indel) at this locus. In addition, suggestive of an association between the expression level of the *vte7* locus and δT accumulation, we detected a weak TWAS signal for this locus with merged read counts, ranking at 0.6% for δT.

Through the integration of GWAS, TWAS and eQTL results, we observed that the distance of GWAS to eQTL signal peak SNPs could be a strong indicator for our ability to detect the causal genes in TWAS. Although we only had a total of 13 and nine gene*phenotype combinations within the proximal group and distal group, respectively, we

still observed a weakly significant difference of the means of TWAS rankings between the two groups. Interestingly, in a study with 41 diverse maize phenotypes, Wallace *et al.* (2014) observed a main peak at ~25 kb between genome-wide SNPs to their nearest genes, as well as a secondary peak for only GWAS-hit SNPs at ~1-5 kb. This is consistent with our observation that the proximal group has a median of 1.0 and 0.7 kb for the distances between the gene and its GWAS and eQTL peak SNPs, respectively, coinciding with the expected location of promoters and short-range regulatory elements. For the proximal group, where the GWAS signal co-locate with the cis-eQTL signal, the causal variants that regulate gene expression and tocochromanol levels are likely to be the same or in strong LD. This in turn could result in a strong association between the transcript abundance and tocochromanol level, and a high ranking of the gene in TWAS. Contrastingly, the cis-eQTL and GWAS peak SNPs are not tagging the same genomic region in the distal group. As we only examined the peak SNPs from each analysis, there could be a less significant GWAS SNP signal in the vicinity of the gene that co-localize with eQTL peak SNP, and vice versa. Therefore, the regulatory region for gene expression may or may not be strongly associated with tocochromanol levels, and as a result, we observed a wide range of TWAS ranking in the distal group (0.009% - 96.634%).

CONCLUSIONS

We identified a total of 13 causal genes controlling the accumulation of nine tocochromanols in the grain of maize Ames panel through an integrated approach that combined GWAS and TWAS results. The identification of the novel genes *samt1* and *vte7*, together with the novel association of two *a priori* pathway genes *vte5* and *dxs1* with tocochromanol accumulation,

demonstrates the superior statistical power and mapping resolution achieved by the Ames panel and the combined GWAS/TWAS approach to study the genetic basis for complex traits. The novel gene *vte7* is hypothesized to be the missing gene that connects the chlorophyll and tocopherol biosynthesis pathways by providing phytol from a chlorophyll-based cycle proposed in Diepenbrock *et al.* (2017). And the identification of *samt1* implies that, together with *vte4* and *vte3*, the composition of tocochromanol isoforms could be more accurately modified through breeding. However, additional studies are needed to further validate the functions of the novel genes through gene editing or mutagenesis experiments. Through eQTL mapping, we determined that the 13 causal genes were regulated primarily by cis-eQTL, with only five trans-eQTL identified for four genes. A carotenoid pathway gene, *psy1*, which had a known epistatic interaction with *dxs2*, was first proposed to play a regulatory role in *dxs2* expression in maize grain. In summary, our work presented the most comprehensive analysis that utilized genomic and transcriptomic data on the large-scale to genetically dissect natural variations of tocochromanols in maize grain, and provided insights into the new breeding targets that could be tailored to provide an ideal tocochromanol profile for human nutritional intake.

SUPPLEMENTAL INFORMATION

Supplemental Figure S4.1. Correlation matrix for untransformed BLUE values for nine tocochromanol grain phenotypes in the Ames panel. Pearson's correlation coefficients (*r*) calculated with the function 'cor' in R are presented in the upper triangle, whereas the corresponding *P*-values for the significance of correlations ($\alpha = 0.05$) are displayed below the

diagonal. The untransformed BLUE values were used to represent the true directionality of the relationship between phenotypes.

Supplemental Figure S4.2. Sources of variation for nine tocochromanol grain phenotypes in the Ames panel. The phenotypic variance was statistically partitioned into the following components: G: genotype; G×Y: genotype × year; Y: year; Field: tier(year), pass(tier × year), range(tier × year); Plate: plate(year); and REV: residual error variance. Variance components were estimated by refitting the full model from Equation 4.1 with genotype as a random effect.

Supplemental Figure S4.3. GWAS, TWAS and FCT results for αT. A: Upset plot showing the number of overlapping genes between GWAS, TWAS, FCT, and *a priori* pathway genes involved in the biosynthesis of chlorophylls and tocochromanols (Supplemental Table S4.5). The number of genes that are within the NAM JL-QTL CSI for αT3 is highlighted in blue in the bar plots. B: Manhattan plots of GWAS, TWAS, and FCT results. Red horizontal dashed lines indicate the thresholds of top 0.02%, top 0.5% and top 0.5% for GWAS, TWAS, and FCT, respectively. Causal genes (Table 4.2) that are within 100 kb of a top 0.02% GWAS peak SNP or ranked top 0.5% in TWAS or FCT are highlighted with red dots and labeled in black in the Manhattan plots. Causal genes that are within 1 Mb of a top 0.02% GWAS peak SNP are labeled in gray. Novel genes are marked with a black rectangle.

Supplemental Figure S4.4. GWAS, TWAS and FCT results for αT3. A: Upset plot showing the number of overlapping genes between GWAS, TWAS, FCT, and *a priori* pathway genes involved in the biosynthesis of chlorophylls and tocochromanols

(Supplemental Table S4.5). The number of genes that are within the NAM JL-QTL CSI for αT3 is highlighted in blue in the bar plots. B: Manhattan plots of GWAS, TWAS, and FCT results. Red horizontal dashed lines indicate the thresholds of top 0.02%, top 0.5% and top 0.5% for GWAS, TWAS, and FCT, respectively. Causal genes (Table 4.2) that are within 100 kb of a top 0.02% GWAS peak SNP or ranked top 0.5% in TWAS or FCT are highlighted with red dots and labeled in black in the Manhattan plots. Causal genes that are within 1 Mb of a top 0.02% GWAS peak SNP are labeled in gray. Novel genes are marked with a black rectangle.

Supplemental Figure S4.5. GWAS, TWAS and FCT results for δT3. A: Upset plot showing the number of overlapping genes between GWAS, TWAS, FCT, and *a priori* pathway genes involved in the biosynthesis of chlorophylls and tocochromanols (Supplemental Table S4.5). The number of genes that are within the NAM JL-QTL CSI for αT3 is highlighted in blue in the bar plots. B: Manhattan plots of GWAS, TWAS, and FCT results. Red horizontal dashed lines indicate the thresholds of top 0.02%, top 0.5% and top 0.5% for GWAS, TWAS, and FCT, respectively. Causal genes (Table 4.2) that are within 100 kb of a top 0.02% GWAS peak SNP or ranked top 0.5% in TWAS or FCT are highlighted with red dots and labeled in black in the Manhattan plots. Causal genes that are within 1 Mb of a top 0.02% GWAS peak SNP are labeled in gray. Novel genes are marked with a black rectangle.

Supplemental Figure S4.6. GWAS, TWAS and FCT results for γT. A: Upset plot showing the number of overlapping genes between GWAS, TWAS, FCT, and *a priori* pathway genes involved in the biosynthesis of chlorophylls and tocochromanols

(Supplemental Table S4.5). The number of genes that are within the NAM JL-QTL CSI for αT3 is highlighted in blue in the bar plots. B: Manhattan plots of GWAS, TWAS, and FCT results. Red horizontal dashed lines indicate the thresholds of top 0.02%, top 0.5% and top 0.5% for GWAS, TWAS, and FCT, respectively. Causal genes (Table 4.2) that are within 100 kb of a top 0.02% GWAS peak SNP or ranked top 0.5% in TWAS or FCT are highlighted with red dots and labeled in black in the Manhattan plots. Causal genes that are within 1 Mb of a top 0.02% GWAS peak SNP are labeled in gray. Novel genes are marked with a black rectangle.

Supplemental Figure S4.7. GWAS, TWAS and FCT results for γT3. A: Upset plot showing the number of overlapping genes between GWAS, TWAS, FCT, and *a priori* pathway genes involved in the biosynthesis of chlorophylls and tocochromanols (Supplemental Table S4.5). The number of genes that are within the NAM JL-QTL CSI for αT3 is highlighted in blue in the bar plots. B: Manhattan plots of GWAS, TWAS, and FCT results. Red horizontal dashed lines indicate the thresholds of top 0.02%, top 0.5% and top 0.5% for GWAS, TWAS, and FCT, respectively. Causal genes (Table 4.2) that are within 100 kb of a top 0.02% GWAS peak SNP or ranked top 0.5% in TWAS or FCT are highlighted with red dots and labeled in black in the Manhattan plots. Causal genes that are within 1 Mb of a top 0.02% GWAS peak SNP are labeled in gray. Novel genes are marked with a black rectangle.

Supplemental Figure S4.8. GWAS, TWAS and FCT results for ΣT. A: Upset plot showing the number of overlapping genes between GWAS, TWAS, FCT, and *a priori* pathway genes involved in the biosynthesis of chlorophylls and tocochromanols

(Supplemental Table S4.5). The number of genes that are within the NAM JL-QTL CSI for αT3 is highlighted in blue in the bar plots. B: Manhattan plots of GWAS, TWAS, and FCT results. Red horizontal dashed lines indicate the thresholds of top 0.02%, top 0.5% and top 0.5% for GWAS, TWAS, and FCT, respectively. Causal genes (Table 4.2) that are within 100 kb of a top 0.02% GWAS peak SNP or ranked top 0.5% in TWAS or FCT are highlighted with red dots and labeled in black in the Manhattan plots. Causal genes that are within 1 Mb of a top 0.02% GWAS peak SNP are labeled in gray. Novel genes are marked with a black rectangle.

Supplemental Figure S4.9. GWAS, TWAS and FCT results for ΣT3. A: Upset plot showing the number of overlapping genes between GWAS, TWAS, FCT, and *a priori* pathway genes involved in the biosynthesis of chlorophylls and tocochromanols (Supplemental Table S4.5). The number of genes that are within the NAM JL-QTL CSI for αT3 is highlighted in blue in the bar plots. B: Manhattan plots of GWAS, TWAS, and FCT results. Red horizontal dashed lines indicate the thresholds of top 0.02%, top 0.5% and top 0.5% for GWAS, TWAS, and FCT, respectively. Causal genes (Table 4.2) that are within 100 kb of a top 0.02% GWAS peak SNP or ranked top 0.5% in TWAS or FCT are highlighted with red dots and labeled in black in the Manhattan plots. Causal genes that are within 1 Mb of a top 0.02% GWAS peak SNP are labeled in gray. Novel genes are marked with a black rectangle.

Supplemental Figure S4.10. GWAS, TWAS and FCT results for ΣTT3. A: Upset plot showing the number of overlapping genes between GWAS, TWAS, FCT, and *a priori* pathway genes involved in the biosynthesis of chlorophylls and tocochromanols

(Supplemental Table S4.5). The number of genes that are within the NAM JL-QTL CSI for αT3 is highlighted in blue in the bar plots. B: Manhattan plots of GWAS, TWAS, and FCT results. Red horizontal dashed lines indicate the thresholds of top 0.02%, top 0.5% and top 0.5% for GWAS, TWAS, and FCT, respectively. Causal genes (Table 4.2) that are within 100 kb of a top 0.02% GWAS peak SNP or ranked top 0.5% in TWAS or FCT are highlighted with red dots and labeled in black in the Manhattan plots. Causal genes that are within 1 Mb of a top 0.02% GWAS peak SNP are labeled in gray. Novel genes are marked with a black rectangle.

Supplemental Figure S4.11. Manhattan plots of eQTL mapping results of the causal genes identified for the tocochromanol grain phenotypes (Table 4.2) in the Ames panel. Each point represents a SNP with its $-\log_{10}$ *P*-value (y-axis) from a mixed linear model analysis plotted as a function of physical position (B73 RefGen_v4) across the 10 chromosomes of maize (x-axis). The red horizontal dashed line indicates the significant threshold after Bonferroni correction ($\alpha = 0.05$).

Supplemental Figure S4.12. Box plots showing the GWAS, TWAS and eQTL results of the causal genes. Results were grouped according to the physical distance between GWAS and cis-eQTL peak SNPs. For GWAS and eQTL, results of peak SNPs were used.

Supplemental Table S4.1. Lambda values used in Box-Cox transformation of nine tocochromanol grain phenotypes in the Ames panel.

Supplemental Table S4.2. Untransformed best linear unbiased estimators ($\mu g\ g^{-1}$) of nine tocochromanol grain phenotypes in the Ames panel.

Supplemental Table S4.3. Transformed best linear unbiased estimators ($\mu g\ g^{-1}$) of

nine tocochromanol grain phenotypes in the Ames panel.

Supplemental Table S4.4. Number of samples and genes in the TWAS pipeline.

Supplemental Table S4.5. Genomic information for the 125 *a priori* candidate genes in the tocochromanol and chlorophyll biosynthesis pathways. Causal genes that were identified in GWAS, TWAS and FCT in this Ames panel are bolded (Supports Table 4.2).

Supplemental Table S4.6. Joint-linkage QTL results of nine tocochromanol grain phenotypes analyzed in the maize NAM panel (Diepenbrock *et al.*, 2017) uplifted from RefGen_v2 to v4.

Supplemental Table S4.7. Genome-wide association study results of nine tocochromanol grain phenotypes analyzed in the maize NAM panel (Diepenbrock *et al.*, 2017) uplifted from RefGen_v2 to v4. Only NAM marker variants with resample model inclusion probability (RMIP) ≥ 0.05 are shown and those that reside within joint-linkage QTL support intervals (Supplemental Table S4.6) are demarcated in the "NAM JL-QTL CSI ID" column.

Supplemental Table S4.8. Rice and Arabidopsis homologs of *samt1*.

Supplemental Table S4.9. Genomic information (RefGen_v4) for the candidate genes within ± 100 kb of the peak SNPs identified in the genome-wide association study.

Supplemental Table S4.10. Genomic information (B73 RefGen_v4) for the genes in top 0.5% for the nine tocochromanol grain phenotypes in the transcriptome-wide association studies.

Supplemental Table S4.11. Genomic information (RefGen_v4) for the SNP-gene pairs that were in top 0.5% for the nine tocochromanol grain phenotypes in the Fisher's

combined test.

Supplemental Table S4.12. Significant results from a genome-wide association study of nine tocochromanol grain phenotypes in the Ames panel.

Supplemental Table S4.13. Significant results from a multi-locus mixed-model analysis of nine tocochromanol grain phenotypes in the Ames panel.

Supplemental Table S4.14. Cis- and trans-eQTL of the 13 causal genes (Table 4.2) identified for tocochromanol phenotypes in the Ames panel. Genomic information (RefGen_v4) for the candidate genes within ± 100 kb of the trans-eQTL peak SNPs were presented. Two candidate causal genes underlying trans-eQTL of *dxs1* and *dxs2* are bolded (supports Figure 4.3).

Supplemental Table S4.15. Directionality of association between gene expression (PEER) and nine tocochromanol phenotypes. Gene-trait pairs that were significant in TWAS (Table 4.2) are bolded.

REFERENCE

Albert, E., Kim, S., Deason, N., Bao, Y., Magallanes-Lundback, M., Danilo, B., Wu, D., Li, X., Gore, M. A., Wood, J. C., Buell, R. C., DellaPenna, D. (2021). Genome-wide association identifies a missing hydrolase for tocopherol biosynthesis in plants. *Proc. Natl. Acad. Sci. U. S. A.*. In review.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2), 166–169.

Baseggio, M., Murray, M., Magallanes-Lundback, M., Kaczmar, N., Chamness, J., Buckler, E. S., Smith, M. E., DellaPenna, D., Tracy, W. F., & Gore, M. A. (2019). Genome-wide association and genomic prediction models of tocochromanols in fresh sweet corn kernels. *The Plant Genome*, *12*(1). https://doi.org/10.3835/plantgenome2018.06.0038

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, *57*(1), 289–300.

Bollero, G. A., Bullock, D. G., & Hollinger, S. E. (1996). Soil temperature and planting date effects on corn yield, leaf area, and plant development. *Agron. J.*, *88*(3), 385–390.

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

Bouvier, F., Linka, N., Isner, J.-C., Mutterer, J., Weber, A. P. M., & Camara, B. (2006). Arabidopsis SAMT1 defines a plastid transporter regulating plastid biogenesis and plant development. *Plant Cell*, *18*(11), 3088–3105.

Bouvier, F., Rahier, A., & Camara, B. (2005). Biogenesis, molecular regulation and function of plant isoprenoids. *Prog. Lipid Res.*, *44*(6), 357–429.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. Series B Stat. Methodol.*, *26*(2), 211–243.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635.

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.*, *103*(3), 338–348.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C., Fan, L., Gao, S., Xu, X., Zhang, G., Li, Y., Jiao, Y., Doebley, J. F., Ross-Ibarra, J., Lorant, A., … Xu, Y. (2018). Construction of the third-generation *Zea mays* haplotype map. *GigaScience*, *7*(4), 1–12.

Cahoon, E. B., Hall, S. E., Ripp, K. G., Ganzke, T. S., Hitz, W. D., & Coughlan, S. J. (2003). Metabolic redesign of vitamin E biosynthesis in plants for tocotrienol production and increased antioxidant content. *Nat. Biotechnol.*, *21*(9), 1082–1087.

Cheng, Z., Sattler, S., Maeda, H., Sakuragi, Y., Bryant, D. A., & DellaPenna, D. (2003). Highly

divergent methyltransferases catalyze a conserved reaction in tocopherol and plastoquinone synthesis in cyanobacteria and photosynthetic eukaryotes. *Plant Cell*, *15*(10), 2343–2356.

Collakova, E., & DellaPenna, D. (2003). The role of homogentisate phytyltransferase and other tocopherol pathway enzymes in the regulation of tocopherol synthesis during abiotic stress. *Plant Physiol.*, *133*(2), 930–940.

Davies, L., & Gather, U. (1993). The identification of multiple outliers. *J. Am. Stat. Assoc.*, *88*(423), 782–792.

DellaPenna, D. (2005). A decade of progress in understanding vitamin E synthesis in plants. *Plant Physiol.*, *162*(7), 729–737.

DellaPenna, D., & Mène-Saffrané, L. (2011). Vitamin E. In F. Rébeillé & R. Douce (Eds.), *Advances in Botanical Research* (Vol. 59, pp. 179–227). Elsevier Ltd., Amsterdam, The Netherlands.

Dewey, M. (2019). *metap: Meta-analysis of significance values. R package version 1.1*.

Diepenbrock, C. H., Kandianis, C. B., Lipka, A. E., Magallanes-Lundback, M., Vaillancourt, B., Góngora-Castillo, E., Wallace, J. G., Cepela, J., Mesberg, A., Bradbury, P. J., Ilut, D. C., Mateos-Hernandez, M., Hamilton, J., Owens, B. F., Tiede, T., Buckler, E. S., Rocheford, T., Buell, C. R., Gore, M. A., & DellaPenna, D. (2017). Novel loci underlie natural variation in vitamin E levels in maize grain. *Plant Cell*, *29*(10), 2374–2392.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, *4*(3), 250.

Ferguson, J., Fernandes, S., Monier, B., Miller, N. D., & Allan, D. (2020). Machine learning enabled phenotyping for GWAS and TWAS of WUE traits in 869 field-grown sorghum accessions. *bioRxiv*. https://www.biorxiv.org/content/biorxiv/early/2020/11/03/2020.11.02.365213.full.pdf

Ford, E. S., Schleicher, R. L., Mokdad, A. H., Ajani, U. A., & Liu, S. (2006). Distribution of serum concentrations of α-tocopherol and γ-tocopherol in the US population. *Am. J. Clin. Nutr.*, *84*(2), 375–383.

Gage, J. L., Vaillancourt, B., Hamilton, J. P., Manrique-Carpintero, N. C., Gustafson, T. J., Barry, K., Lipzen, A., Tracy, W. F., Mikel, M. A., Kaeppler, S. M., Buell, C. R., & de Leon, N. (2019). Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. *The Plant Genome*, *12*(2). https://doi.org/10.3835/plantgenome2018.09.0069

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Thompson, R., Butler, D., & Others. (2009). ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

Grams, G. W., Blessin, C. W., & Inglett, G. E. (1970). Distribution of tocopherols within the corn kernel. *J. Am. Oil Chem. Soc.*, *47*(9), 337–339.

Hirschberg, J. (2001). Carotenoid biosynthesis in flowering plants. *Curr. Opin. Plant Biol.*, *4*(3), 210–218.

Holland, J. B., Nyquist, W. E., & Cervantes-Martínez, C. T. (2003). Estimating and interpreting heritability for plant breeding: an update. *Plant Breed. Rev*, *22*, 9–112.

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

https://www.biorxiv.org/content/biorxiv/early/2021/01/16/2021.01.14.426684.full.pdf

Hung, H.-Y., Browne, C., Guill, K., Coles, N., Eller, M., Garcia, A., Lepak, N., Melia-Hancock, S., Oropeza-Rosas, M., Salvo, S., Upadyayula, N., Buckler, E. S., Flint-Garcia, S., McMullen, M. D., Rocheford, T. R., & Holland, J. B. (2012). The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity*, *108*(5), 490–499.

Hunter, S. C., & Cahoon, E. B. (2007). Enhancing vitamin E in oilseeds: unraveling tocopherol and tocotrienol biosynthesis. *Lipids*, *42*(2), 97–108.

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.

Kachanovsky, D. E., Filler, S., Isaacson, T., & Hirschberg, J. (2012). Epistasis in tomato color mutations involves regulation of *phytoene synthase 1* expression by *cis*-carotenoids. *Proc. Natl. Acad. Sci. U. S. A.*, *109*(46), 19021–19026.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*(3), 1709–1723.

Karunanandaa, B., Qi, Q., Hao, M., Baszis, S. R., Jensen, P. K., Wong, Y.-H. H., Jiang, J., Venkatramesh, M., Gruys, K. J., Moshiri, F., Post-Beittenmiller, D., Weiss, J. D., & Valentin, H. E. (2005). Metabolically engineered oilseed crops with enhanced seed tocopherol. *Metab. Eng.*, *7*(5-6), 384–400.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997.

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, *37*(8), 907–915.

Knekt, P., Reunanen, A., Järvinen, R., Seppänen, R., Heliövaara, M., & Aromaa, A. (1994). Antioxidant vitamin intake and coronary mortality in a longitudinal population study. *Am. J. Epidemiol.*, *139*(12), 1180–1189.

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*, *9*(9), 3023–3033.

Kurtz, S. (2010). *The Vmatch large scale sequence analysis software - a manual. http://www.vmatch.de/virtman.pdf*.

Kushi, L. H., Folsom, A. R., Prineas, R. J., Mink, P. J., Wu, Y., & Bostick, R. M. (1996). Dietary antioxidant vitamins and death from coronary heart disease in postmenopausal women. *N. Engl. J. Med.*, *334*(18), 1156–1162.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., Liu, J., Warburton, M. L., Cheng, Y., Hao, X., Zhang, P., Zhao, J., Liu, Y., Wang, G., Li, J., & Yan, J. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.*, *45*(1), 43–50.

Lipka, A. E., Gore, M. A., Magallanes-Lundback, M., Mesberg, A., Lin, H., Tiede, T., Chen, C., Buell, C. R., Buckler, E. S., Rocheford, T., & DellaPenna, D. (2013). Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. *G3*, *3*(8), 1287–1299.

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

Li, Q., Yang, X., Xu, S., Cai, Y., Zhang, D., Han, Y., Li, L., Zhang, Z., Gao, S., Li, J., & Yan, J. (2012). Genome-wide association studies identified three independent polymorphisms associated with α-tocopherol content in maize kernels. *PloS One*, *7*(5), e36807.

Liu, X., Hua, X., Guo, J., Qi, D., Wang, L., Liu, Z., Jin, Z., Chen, S., & Liu, G. (2008). Enhanced tolerance to drought stress in transgenic tobacco plants overexpressing *VTE1* for increased tocopherol production from *Arabidopsis thaliana*. *Biotechnol. Lett.*, *30*(7), 1275–1280.

Lois, L. M., Rodriguez-Concepcion, M., Gallego, F., Campos, N., & Boronat, A. (2000). Carotenoid biosynthesis during tomato fruit development: regulatory role of 1-deoxy-D-xylulose 5-phosphate synthase. *The Plant Journal*, *22*(6), 503–513.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, *15*(12), 550.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits* (Vol. 1). Sinauer Sunderland, MA.

Majewski, J., & Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.*, *27*(2), 72–79.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12.

McBurney, M. I., Yu, E. A., Ciappio, E. D., Bird, J. K., Eggersdorfer, M., & Mehta, S. (2015). Suboptimal serum α-tocopherol concentrations observed among younger adults and those depending exclusively upon food sources, NHANES 2003-2006. *PloS One*, *10*(8), e0135510.

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., … Buckler, E. S. (2009). Genetic properties of the maize nested association mapping population. *Science*, *325*(5941), 737–740.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (Vol. 4). Irwin Chicago.

Norris, S.R., Shen, X., and DellaPenna, D. (1998). Complementation of the Arabidopsis *pds1* mutation with the gene encoding *p*-hydroxyphenylpyruvate dioxygenase. *Plant Physiol*. 117: 1317–1323.

Palmieri, L., Arrigoni, R., Blanco, E., Carrari, F., Zanor, M. I., Studart-Guimaraes, C., Fernie, A. R., & Palmieri, F. (2006). Molecular identification of an arabidopsis S-adenosylmethionine

transporter: Analysis of organ distribution, bacterial expression, reconstitution into liposomes, and functional characterization. *Plant Physiol.*, *142*(3), 855–865.

Pignon, C. P., Fernandes, S. B., Valluru, R., Bandillo, N., Lozano, R., Buckler, E., Gore, M. A., Long, S. P., Brown, P. J., & Leakey, A. D. B. (2021). Phenotyping stomatal closure by thermal imaging for GWAS and TWAS of water use efficiency-related genes. In *bioRxiv* (p. 2021.05.06.442962). https://doi.org/10.1101/2021.05.06.442962

Porfirova, S., Bergmuller, E., Tropf, S., Lemke, R., & Dormann, P. (2002). Isolation of an *Arabidopsis* mutant lacking vitamin E and identification of a cyclase essential for all tocopherol biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.*, *99*(19), 12495–12500.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, *81*(3), 559–575.

R Core Team. (2018). *An Introduction to R*. Samurai Media Limited.

Rippert, P., Scimemi, C., Dubald, M., and Matringe, M. (2004). Engineering plant shikimate pathway for production of tocotrienol and improving herbicide resistance. *Plant Physiol*. 134: 92–100.

Rodríguez-Villalón, A., Gas, E., & Rodríguez-Concepción, M. (2009). Phytoene synthase activity controls the biosynthesis of carotenoids and the supply of their metabolic precursors in dark-grown Arabidopsis seedlings. *The Plant Journal: For Cell and Molecular Biology*, *60*(3), 424–435.

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., Elshire, R. J., Acharya, C. B., Mitchell, S. E., Flint-Garcia, S. A., McMullen, M. D., Holland, J. B., Buckler, E. S., & Gardner, C. A. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.*, *14*(6), R55.

Sattler, S. E., Gilliland, L. U., Magallanes-Lundback, M., Pollard, M., & DellaPenna, D. (2004). Vitamin E is essential for seed longevity and for preventing lipid peroxidation during germination. *Plant Cell*, *16*(6), 1419–1432.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.*, *44*(7), 825–830.

Sen, C. K., Khanna, S., & Roy, S. (2006). Tocotrienols: Vitamin E beyond tocopherols. *Life Sciences*, *78*(18), 2088–2098.

Shintani, D., & DellaPenna, D. (1998). Elevating the vitamin E content of plants through metabolic engineering. *Science*, *282*(5396), 2098–2100.

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. *Nat. Protoc.*, *7*(3), 500–507.

Sun, G., Zhu, C., Kramer, M. H., Yang, S.-S., Song, W., Piepho, H.-P., & Yu, J. (2010). Variation explained in mixed-model association mapping. *Heredity*, *105*(4), 333–340.

Traber, M. G. (2012). Vitamin E. In J. W. Erdman Jr, I. A. MacDonald, & S. H. Zeisel (Eds.), *Present*

Knowledge in Nutrition. John Wiley & Sons.

Valentin, H. E., Lincoln, K., Moshiri, F., Jensen, P. K., Qi, Q., Venkatesh, T. V., Karunanandaa, B., Baszis, S. R., Norris, S. R., Savidge, B., Gruys, K. J., & Last, R. L. (2006). The *Arabidopsis* vitamin E pathway *gene5-1* mutant reveals a critical role for phytol kinase in seed tocopherol biosynthesis. *Plant Cell*, *18*(1), 212–224.

Van Eenennaam, A. L., Lincoln, K., Durrett, T. P., Valentin, H. E., Shewmaker, C. K., Thorne, G. M., Jiang, J., Baszis, S. R., Levering, C. K., Aasen, E. D., Hao, M., Stein, J. C., Norris, S. R., & Last, R. L. (2003). Engineering vitamin E content: From Arabidopsis mutant to soy oil. *Plant Cell*, *15*(12), 3007–3019.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.*, *91*(11), 4414–4423.

Vom Dorp, K., Hölzl, G., Plohmann, C., Eisenhut, M., Abraham, M., Weber, A. P. M., Hanson, A. D., & Dörmann, P. (2015). Remobilization of phytol from chlorophyll degradation is essential for tocopherol synthesis and growth of Arabidopsis. *Plant Cell*, *27*(10), 2846–2859.

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.

Wan, C. Y., & Wilkins, T. A. (1994). A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (*Gossypium hirsutum* L.). *Anal. Biochem.*, *223*(1), 7–12.

Wang, H., Xu, S., Fan, Y., Liu, N., Zhan, W., Liu, H., Xiao, Y., Li, K., Pan, Q., Li, W., Deng, M., Liu, J., Jin, M., Yang, X., Li, J., Li, Q., & Yan, J. (2018). Beyond pathways: genetic dissection of tocopherol content in maize kernels by combining linkage and association analyses. *Plant Biotechnol. J.*, *16*(8), 1464–1475.

Weber, E. J. (1987). Carotenoids and tocols of corn grain determined by HPLC. *J. Am. Oil Chem. Soc.*, *64*(8), 1129–1134.

Wu, D., Tanaka, R., Li, X., Ramstein, G. P., Cu, S., Hamilton, J. P., Buell, C. R., Stangoulis, J., Rocheford, T., & Gore, M. A. (2021). High-resolution genome-wide association study pinpoints metal transporter and chelator genes involved in the genetic control of element levels in maize grain. *G3*. https://doi.org/10.1093/g3journal/jkab059

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.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, *38*(2), 203–208.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.*, *42*(4), 355–360.

Zhan, W., Liu, J., Pan, Q., Wang, H., Yan, S., Li, K., Deng, M., Li, W., Liu, N., Kong, Q., Fernie, A. R., & Yan, J. (2019). An allele of *ZmPORB2* encoding a protochlorophyllide oxidoreductase promotes tocopherol accumulation in both leaves and kernels of maize. *Plant J.*, *100*(1), 114–

127.

Chapter 5 Chlorophyll dephytylation is the main phytol provider for tocopherol synthesis in maize grain

ABSTRACT

Tocopherol is a class of tocochromanols exhibiting high vitamin E activity. However, the mechanism of phytol synthesis, one of the key precursors of tocopherols, is still largely unknown in cereal seeds, a non-photosynthetic tissue. Recently, two protochlorophyllide reductases (*por1* and *por2*) involved in chlorophyll synthesis were identified as large-effect loci controlling the concentration of tocopherols in maize grain in the U.S. nested association mapping (NAM) panel. In this study, we further investigated the involvement of *por1* and *por2* in a hypothesized chlorophyll-based cycle in the provision of phytol for tocopherol synthesis in maize grain. Three differential light treatments (high-light, light-deprived, and control) were applied to developing kernels on self-pollinated ears of several parents of the U.S. NAM panel with contrasting *por1* and *por2* effects. Significant genotype, treatment, and genotype × treatment effects were observed for all tocopherol phenotypes (α-, δ-, γ-, and total tocopherols) measured on mature grain and developing embryo samples. In contrast to tocotrienol phenotypes (α-, δ-, γ-, and total tocotrienols), near-zero and significantly lower tocopherol levels were observed in grain samples from light-deprived ears relative to high-light and control ears, suggesting the involvement of chlorophyll dephytylation in tocopherol synthesis. Expression profiling of developing maize embryos revealed limited light-mediated control of gene expression in the isopentenyl pyrophosphate pathway and prenyl group synthesis relative to the chlorophyll pathway, implying that phytyl diphosphate for producing

tocopherols is mostly synthesized from chlorophyll-derived phytol. Overall, our study provides evidence that strongly supports the hypothesis that the predominant generation of phytol for tocopherol synthesis results from a chlorophyll-based cycle that occurs in the embryo of maize grain.

INTRODUCTION

Vitamin E is an essential micronutrient for the human diet. Substandard dietary intake of vitamin E can result in a higher risk of cardiovascular disease and a weakened immune system (Wolf, 2005). The lipid-soluble compounds synthesized in plants with a range of vitamin E and antioxidant activities are collectively known as tocochromanols, which are structurally separated into two classes, tocotrienols and tocopherols. Each class has four isoforms: α, β, δ, and γ. Of the two tocochromanol classes, tocopherols generally have higher vitamin E activity, with α-tocopherol conferring the highest vitamin E activity on a molar basis (Leth & Søndergaard, 1977). In plants, tocochromanols are important antioxidants, providing protection to lipids by quenching reactive oxygen species in photosynthetic tissues and lipid peroxy radicals resulting from peroxidation of polyunsaturated fatty acids in the seed (Collakova & DellaPenna, 2003; X. Liu *et al.*, 2008; Sattler *et al.*, 2004).

The core tocopherol biosynthesis pathway has been well-characterized in *Arabidopsis thaliana* (DellaPenna & Mène-Saffrané, 2011). The committed step of tocopherol synthesis is catalyzed by *vte2* (homogentisate phytyltransferase), which condenses homogentisic acid (HGA) and phytyl diphosphate (PDP) to produce 2-methyl-6-phytyl-1,4-benzoquinol (MPBQ), the immediate precursor of the four tocopherol isoforms. This condensation reaction

is a rate-limiting step for overall tocopherol biosynthesis (Collakova & DellaPenna, 2003). Also, the low availability of tocopherol precursors, especially HGA and phytol, could significantly limit tocopherol synthesis (reviewed in Mène-Saffrané & Pellaud, 2017). PDP could originate from isoprenoid synthesis via the reduction of geranylgeranyl-diphosphate (GGDP) by geranylgeranyl reductase (GGR) (Keller *et al.,* 1998), or synthesized via the two-step phosphorylation of free phytol by phytol kinase (*vte5*) and phytol phosphate kinase (*vte6*) (Valentin *et al.,* 2006; Vom Dorp *et al.,* 2015). In photosynthetic tissues, the supply of phytol for tocopherol synthesis is majorly dependent on chlorophyll breakdown (Vom Dorp *et al.,* 2015). In addition to chlorophyll breakdown, the chlorophyll-salvage cycle is primarily involved in the release of chlorophyllide a and phytol in the dephytylation of chlorophyll a (Lin *et al.,* 2014, 2016; P. Wang & Grimm, 2021). Despite the identification of enzymes involved in the dephytylation of chlorophyll, the singular or combined role of these enzymes in providing phytol for tocopherol biosynthesis in seed is not well understood (reviewed in Lin & Charng, 2021).

Recently, two genes encoding protochlorophyllide reductases (*por1* and *por2*) were implicated as loci responsible for the genetic control of mature grain tocopherol concentrations in the U.S. nested association mapping (NAM) panel (Diepenbrock *et al.,* 2017). These two homologs encode key enzymes involved in the light-dependent synthesis of chlorophyll (Buhr *et al.,* 2017) and underlie the largest effect loci associated with total tocopherols despite that maize grain is a non-green, non-photosynthetic tissue. Diepenbrock *et al.* (2017) hypothesized that a chlorophyll-based cycle instead of chlorophyll degradation is responsible for supplying the phytol needed for tocopherol synthesis in the embryo. In further

support of this hypothesis, a plastid-localized alpha/beta hydrolase (*vte7*) was identified and proposed to be the primary hydrolase that provides the majority of phytol from chlorophyll, thus regulating tocopherol levels in monocot and dicot seeds (Albert *et al.* in prep). The involvement of *por2* in tocopherol biosynthesis was additionally confirmed by Zhan *et al.* (2019), as the transgenic overexpression of *por2* increased tocopherol levels in maize leaf and grain tissues. Additionally, Zhan *et al.* (2019) observed the concomitant dramatic increase in total tocopherol and *por2* expression levels several days after pollination in the flag leaves of non-transgenic, near-isogenic lines for *por2*. Given these findings, the authors proposed that the large demand of precursors for tocopherol biosynthesis in the grain is met by chlorophyll turnover in maternal leaves through a yet to be described mechanism.

In this study, we conducted a field experiment that utilized several parents of the U.S. NAM panel that had contrasting allelic effects at *por1* and *por2* to investigate the role of these two loci in tocopherol biosynthesis in maize grain. The two main objectives of our study were to (i) confirm the involvement of a chlorophyll-based cycle in synthesis of tocopherol in grain, and (ii) determine which tissue is primarily responsible for the provision of phytol for the synthesis of grain tocopherols.

MATERIALS AND METHODS

Experimental design

To evaluate the involvement of chlorophyll in the synthesis of tocopherols, a field experiment was conducted in 2018 that included seven parents of the U.S. NAM panel with contrasting allelic effects at *por1* and *por2* (Diepenbrock *et al.,* 2017). Included in the experiment were

B97 and M37W with large negative *por1* effects, NC358 and Ki11 with large positive *por2* effects, and MS71 and OH7B with large effects for both *por1* (negative) and *por2* (positive) compared to B73. We also included B73 as a check line, given that it is the common parent for the U.S. NAM panel. The seven NAM parents were planted in a split-plot design with two replications at Cornell University's Musgrave Research Farm in Aurora, NY. Each plot consisted of three subplots, with high-light, light-deprivation, and control treatments randomly assigned to each subplot. Each experimental unit was a one-row subplot with a length of 5.33 m, an inter-row spacing of 0.76 m, and a 0.76 m alley at the end of each subplot. For each subplot, a total of 20 seeds were planted, and the primary ears of at least four plants were self-pollinated by hand to ensure sufficient materials for the experiment. At 12 days after pollination (DAP), pollination bags of the high-light ears were removed, and the outer husk leaves were peeled until only three to four layers of husk leaves remained. This allowed high light exposure (30–50 μmol/m^2/s) while still retaining sufficient moisture for the ear over grain development. For the light-deprived ears, constructed foil bags were used to enshroud pollination bags, thus essentially preventing exposure of ears to light (0.00–0.01 μmol/m^2/s). To represent typical self-pollinated ear conditions, the ears of the control treatment remained in medium water-proof pollination bags (MIDCO Global, St. Louis, MO) that allowed ~50 μmol/m^2/s light to pass through them. Given that there were multiple layers of husk leaves that could additionally reduce light penetration, the light levels reaching the kernels were at even lower levels (< 50 μmol/m^2/s) for the control treatment.

At 24 DAP, three ears per subplot were collected for providing developing kernels for dissection. Harvested ears were dehusked, inserted into a pollination bag, and immediately

placed on wet ice in a cooler (IGLOO, Katy, TX) with a closed lid for transport to the field house. For light-deprived ears, foil bags were placed back on pollination bags for reduced light exposure to ears prior to placing them in the cooler. For each subplot, 20 kernels from each of the three ears were hand-dissected into embryo and endosperm tissues. The kernels were dissected in a dark room with each dissector wearing a green headlamp to limit the initiation of light-regulated reactions. Next, five embryos from each ear were placed in one of four 2 mL centrifuge tubes, resulting in four tubes that each had 15 embryos (*i.e.*, a bulk of embryos, with each tube having five embryos from each of the three ears). These tubes were immediately frozen in liquid nitrogen before temporarily storing on dry ice for transport to the laboratory and storage in a −80 °C freezer prior to RNA extraction and metabolite quantification.

The remaining ears (3–13 individual ears) in each subplot were harvested at physiological maturity. Harvested ears were individually shelled and maintained separately. Tocochromanols were extracted from ground mature grain (~30 mg) and 24 DAP embryo (a ~15 mg subsample from 30 ground embryos) samples and quantified by high-performance liquid chromatography (HPLC) following the method of Lipka *et al.* (2013). A total of nine tocochromanol phenotypes were evaluated in μg g^{-1}: α-tocopherol (αT), δ-tocopherol (δT), γ-tocopherol (γT), α-tocotrienol (αT3), δ-tocotrienol (δT3), γ-tocotrienol (γT3), total tocopherols (ΣT; αT + δT + γT), total tocotrienols (ΣT3; αT3 + δT3 + γT3), and total tocochromanols (ΣTT3; total tocopherols + total tocotrienols). These phenotypes were scored on 343 mature grain and 42 embryo samples from 42 subplots (Supplemental Tables S5.1 and S5.2). In addition to tocochromanols, the level of chlorophyll a was measured on the 42

pooled embryo samples as described in Diepenbrock *et al.* (2017).

In 2019, the experiment was repeated with only four NAM parents, B73, B97, Ki11, and OH7B. These four parents are a subset of the seven NAM parents selected in 2018, and apart from B73, each had large *por1* and/or *por2* effects. Consistent with 2018, a split-plot design with two replications was used for the same three light treatments, for a total of 24 experimental units. Resulting from a weather-delayed field planting, all six subplots of Ki11 and one subplot of OH7B with high-light treatment were late maturing and thus not harvested. A total of 75 mature grain samples harvested from individual plants across the 17 subplots were prepared and scored for tocochromanols with the same HPLC method as described above (Supplemental Table S5.3).

RNA isolation and expression data processing

For each subplot in the 2018 field experiment, 30 embryos were ground in liquid nitrogen, followed by subsampling 100–200 mg of ground tissue for RNA extraction using a modified hot borate method (Wan & Wilkins, 1994). RNA was DNase treated and checked for purity and quality using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Wilmington, DE). High quality RNA was used to construct Illumina TruSeq Stranded mRNA libraries (Illumina, San Diego, CA), which were sequenced on an Illumina HiSeq4000 (Illumina, San Diego, CA) at the Resource Technology Support Facility at Michigan State University.

RNAseq reads were trimmed using Cutadapt version 1.18 (Martin, 2011) with a quality cutoff score of 20 and a minimum read length of 30 nt. Trimmed RNAseq reads were aligned to the B73 RefGen_v4 reference genome (Jiao *et al.*, 2017) using HISAT2 version

2.1.0 (Kim *et al.,* 2019) with the following parameters: --min-intronlen 20' and --max-intronlen 60000 in stranded mode. Read alignments were used to generate Fragments per Kilobase of transcript per Million mapped reads (FPKMs) using Cufflinks2 v2.2.1 (Trapnell *et al.,* 2010) with the following parameters: --multi-read-correct, --max-bundle-frags 999999999, --min-intron-length 10, and --max-intron-length 60000 in stranded mode. Read counts per gene were calculated for a total of 39,498 genes using the count function of HTSeq version 0.6.1p1 (Anders *et al.,* 2014) with the following parameters: --format=bam, --order=pos, --minaqual=10, --idattr=ID, --type=gene, and --mode=union in stranded mode (Supplemental Table S5.4).

Statistical analyses of metabolites in maize grain

Prior to conducting statistical tests on metabolite profiles of mature grain samples harvested in 2018 and 2019, a robust outlier removal was performed at the subplot level, removing observations with median absolute deviation (MAD) greater than 3 (Leys *et al.,* 2013; Miller, 1991). After the outliers were removed, the mean of each subplot was calculated for subsequent statistical analyses (Supplemental Table S5.5).

We conducted statistical tests on metabolite profiles of mature grain (2018 and 2019) and embryo (2018) samples. The two mature grain metabolite datasets were analyzed as separate experiments, given the different number of genotypes evaluated in 2018 and 2019. Levene's test was initially conducted to test for equality of variances with the SAS PROC GLM procedure in SAS studio release 3.8 (https://www.sas.com/en_us/software/on-demand-for-academics.html). The genotype and treatment terms were tested separately for each

phenotype, with the most significant term ($\alpha = 0.05$) selected to correct for unequal variances (Supplemental Table S5.6). With the metabolite levels for each subplot from the 2018 and 2019 mature grain (averaged by subplot) and 2018 embryo samples, we screened the three datasets separately for outliers. Studentized deleted residuals (Neter *et al.,* 1996) were estimated in SAS PROC GLIMMIX procedure with correction for unequal variances of each phenotype in a mixed linear model (Equation 5.1) as follows (Federer & King, 2007; Steel *et al.,* 1997):

$$Y_{ijk} = \mu + \text{genotype}_i + \text{rep}_j + \text{genotype} \times \text{rep}_{ij} + \text{treatment}_k + \text{genotype} \times \text{treatment}_{ik} + \varepsilon_{ijk} \quad \text{(Equation 5.1)}$$

in which Y_{ijk} is an individual phenotypic observation; μ is the grand mean; genotype_i is the fixed effect of the ith genotype; rep_j is the random effect of the jth replicate; genotype × rep_{ij} is the random effect of the interaction between the ith genotype and jth replicate; treatment_k is the fixed effect of the kth treatment, genotype × treatment_{ik} is the fixed effect of the interaction between the ith genotype and kth treatment; and ε_{ijk} is the residual error effect assumed to be independently and identically distributed according to a normal distribution with mean zero and variance σ_ε^2, that is ~iid $N(0, \sigma_\varepsilon^2)$. No significant outlier was identified for any phenotype after a Bonferroni correction of $\alpha = 0.05$. An ANOVA test for the significance of model main effects and Tukey's Honestly Significant Difference (HSD) test for pairwise differences of metabolite levels among treatments within each genotype were subsequently performed using Equation 5.1 using the SAS PROC GLIMMIX procedure with a correction for unequal variances. For each metabolite phenotype, the Equation 5.1 model was used to separately generate BLUE values for the 2018 (mature grain and 24 DAP

embryo) and 2019 (mature grain) datasets (Supplemental Tables S5.7 and S5.8). A Pearson's correlation coefficient (*r*) was calculated between the BLUE values for each pair of metabolite (chlorophyll a and tocochromanols) phenotypes measured on the 24 DAP embryo samples with the function 'cor' in R version 3.5.1(R Core Team, 2021).

Expression data analysis

Within each of the seven genotypes, we tested for the differential expression of genes between treatments (high-light vs. light-deprivation; high-light vs. control; and light-deprivation vs. control) in the R package DESeq2 (Love *et al.*, 2014). Genes were declared to be differentially expressed with a false-discovery rate of 5% (Benjamini and Hochberg 1995). Enrichment analysis of GO terms was performed for all differentially expressed genes (DEGs) by the PANTHER classification system version 16.0 (Mi *et al.*, 2021) with the Fisher's Exact test at 5% FDR (Benjamini and Hochberg 1995).

To better satisfy the normality assumption of ANOVA and Tukey's HSD tests, a $\log_2$ transformation was performed on the normalized FPKM values of 126 *a priori* candidate genes (Supplemental Table S5.9) across the 42 embryo samples. Given the presence of normalized FPKM values equal to zero (7.5%), a small constant (1E−9) was first added to all FPKMs before applying the $\log_2$ transformation. ANOVA and Tukey's HSD tests were performed using packages 'car' (version 3.0-10) and 'emmeans' (version 1.5.2-1), respectively, in R version 3.5.1 (R Core Team, 2021).

RESULTS

Tocopherols in maize grain are responsive to differential light treatments

To assess the potential involvement of chlorophyll in the biosynthesis of tocopherols, experiments were conducted through the application of three light treatments (high-light, light-deprived, and control) on developing kernels of self-pollinated ears of several NAM parents in 2018 and 2019. For all measured tocopherol phenotypes (αT, δT, γT, and ΣT), significant genotype, treatment, and genotype × treatment effects ($\alpha = 0.05$) were observed for mature grain (2018-19) and embryo (2018) samples (Supplemental Table S5.10). Mature grain samples from light-deprived ears had near-zero tocopherol levels that were significantly lower than those from control and high light-treated ears of all analyzed NAM parents across the three datasets (Supplemental Table S5.11). However, there was a genotype-dependent increase or decrease of tocopherol levels in high-light samples compared to control samples (Figure 5.1). We also assessed the accumulation of chlorophyll a in the 42 pooled embryo samples from 2018, finding all three main effects to be significant. Near-zero and significantly lower chlorophyll a levels were observed in all light-deprived samples relative to control samples, whereas high light-treated samples had significantly higher chlorophyll a accumulation in B97, Ki11, and M37W (Figure 5.2).

A significant treatment effect was consistently observed for δT3 across the three datasets, but not for any of the other three tocotrienol phenotypes (αT3, γT3, and ΣT3) (Supplemental Table S5.10). Similar to tocopherols, a significant genotype effect was observed for the four tocotrienol phenotypes across all years and tissue sample types (embryo

and whole kernel), with the genotype × treatment effect also significant for most tocotrienol phenotypes. In contrast to the high-light and control treatments, the light-deprived 24 DAP embryo samples tended to have a significant increase in all tocotrienols for most of the seven NAM parents, but a similar response was not consistently observed for mature grain samples from both years. The high-light and control treatments did not produce clear patterns of response for the four tocotrienol phenotypes across both years and sample types (Supplemental Table S5.11).

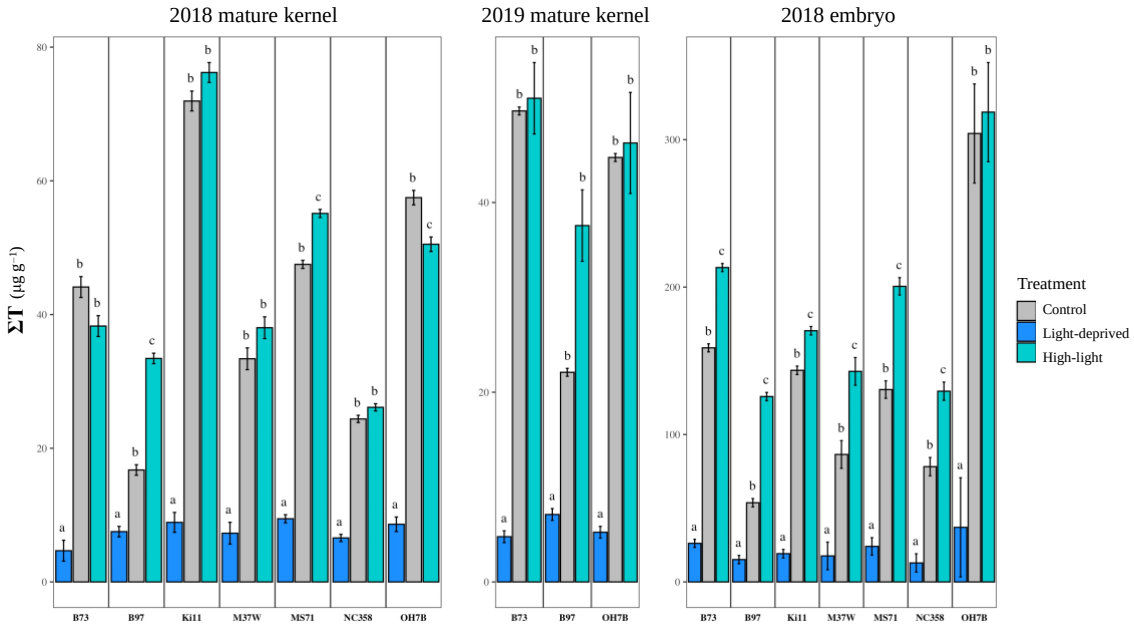

Figure 5.1. Bar plots showing BLUEs of ΣT (μg g^{-1}) in mature kernel samples from 2018 and 2019, and 24 DAP embryo samples from 2018. Samples with the same letter are not significantly different according to the Tukey's Honestly Significant Difference test ($P < 0.05$) that was conducted within each NAM parent.

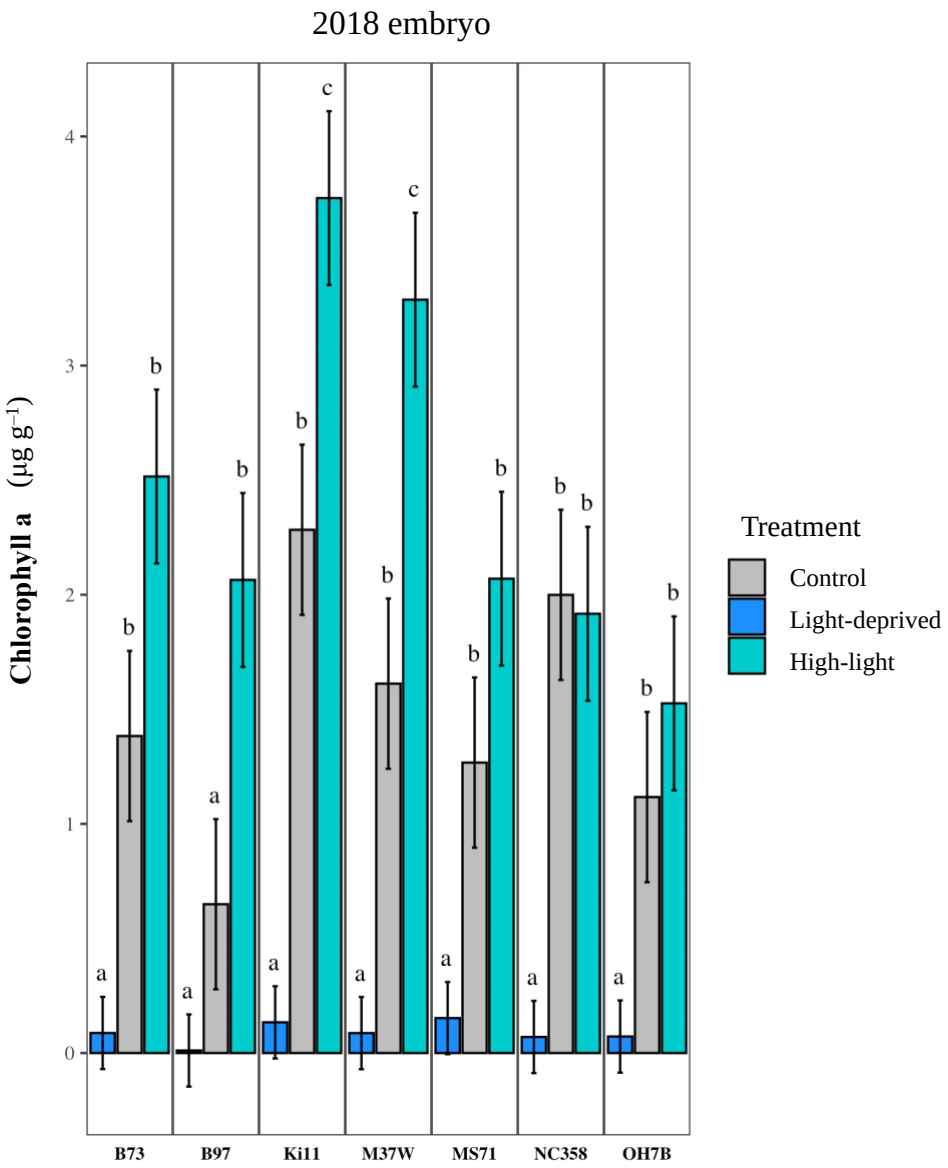

Figure 5.2. Bar plots showing BLUEs of chlorophyll (μg g^{-1}) in 24 DAP embryo samples from 2018. Samples with the same letter are not significantly different according to the Tukey's Honestly Significant Difference test ($P < 0.05$) that was conducted within each NAM parent.

Gene expression profile analysis of developing embryos identifies genotype-dependent responses

The transcriptomes of the 24 DAP embryo samples were analyzed to understand how the expression profile of genes in the chlorophyll and tocopherol biosynthesis pathways respond to differential light conditions. A total of 595 DEGs were detected at 5% FDR in a total of 21 pairwise treatment comparisons at the genotype level, with most of the genes differentially expressed in at least two comparisons (Supplemental Figure S5.1, Supplemental Table S5.12). Notably, a gene (Zm00001d016826) encoding a protein with high amino acid identity (73%) to MAINTENANCE OF PHOTOSYSTEM II (PSII) UNDER HIGH LIGHT 1 (MPH1; AT5G07020) in Arabidopsis was found to be differentially expressed in nine pairwise comparisons involving four different NAM parents (B97, Ki11, M37W, MS71). In Arabidopsis, this gene contributes to the maintenance of PSII homeostasis upon exposure to photoinhibitory light conditions by participating in the protection and stabilization of PSII (J. Liu & Last, 2015a, 2015b). An enrichment analysis of the 595 DEGs showed 28 enriched GO terms (FDR-corrected *P*-value < 0.05); however, neither enrichment for vitamin E biosynthetic (GO: 0010189) nor chlorophyll biosynthetic (GO: 0015995) processes was observed (Supplemental Table S5.13). The most significantly enriched GO term was chromatin assembly or disassembly (GO:0006333), with a 10.92-fold change and *P*-value of 2.58E−10.

We performed ANOVA to test the significance of model main effects for the transcript abundance of the 126 *a priori* pathway genes in 24 DAP embryo samples. All but 17 of the 126 *a priori* genes did not show a significant genotype effect ($P > 0.05$); however, only 17

genes had an expression profile that responded to the light treatment (Supplemental Table S5.14). Included among these 17 genes were three genes from the chlorophyll degradation pathway (Zm00001d034523; Zm00001d051567; and Zm00001d031934, *chao1*), five genes from the chorismate/tyrosine synthesis pathway (Zm00001d021611, Zm00001d029391, Zm00001d041700, Zm00001d047896, and Zm00001d052797), three genes from the tocochromanol core pathway (Zm00001d006657, *w3*; Zm00001d031071, *vte3*; and Zm00001d046558, *hggt1*), four genes from the chlorophyll synthesis pathway (Zm00001d026603, *chlh1*; Zm00001d018034, *ggh1*; Zm00001d014715, and Zm00001d053807), one gene from the tyrosine degradation pathway (Zm00001d045610) and one gene from the IPP pathway (Zm00001d038170, *dxs1*). Interestingly, the gene with the most significant treatment effect (Zm00001d051567) encodes a protein involved in chlorophyll degradation that has 61% amino acid identity to an Arabidopsis alpha/beta-hydrolases superfamily protein (AT4G36530) that belongs to a phylogenetic clade that does not include chlorophyll dephytylase1 or pheophytinase (Lin *et al.*, 2016). However, Zm00001d051567 is only 15% identical at the amino acid sequence level to the alpha/beta hydrolase coded by maize *vte7* (Zm00001d006778). The expression level of this gene was consistently the highest in high-light and lowest in light-deprived samples across the seven NAM parents.

We identified 16 genes showing a significant genotype × treatment effect (Supplemental Table S5.14), including two *protochlorophyllide reductases* (Zm00001d001820, *por3*; and Zm00001d013937, *por2*), *light-harvesting complex-like protein 3* (Zm00001d015094, *lil3*), and *geranylgeranyl reductase* (Zm00001d018034, *ggr*),

all from the chlorophyll biosynthesis pathway. Providing a potential connection to tocopherol synthesis, *lil3* encodes a protein that anchors both POR and GGR, forming a light-harvesting complex for chlorophyll biosynthesis (Hey *et al.,* 2017; Mork-Jansson *et al.,* 2015). In addition, four genes (Zm00001d003830, *adt6*; Zm00001d010190, *got5*; Zm00001d021611; and Zm00001d029391) from the chorismate/tyrosine synthesis pathway, two genes from the tyrosine degradation pathway (Zm00001d007462, *tat1*; and Zm00001d033555), two genes from the 3,8-divinyl-chlorophyllide biosynthesis I pathway (Zm00001d011387; and Zm00001d026603, *chlh1*), one gene from the IPP pathway (Zm00001d045383, *dxs3*), one gene from the chlorophyll degradation pathway (Zm00001d034523), one gene from the chlorophyll cycle (Zm00001d042026, *phao1*), and one gene from the core tocochromanol pathway (Zm00001d039491, *vte2*) were identified to have a significant genotype × treatment effect. Of the 16 genes, five genes also had a significant genotype × treatment effect, which belong to either the 3,8-divinyl-chlorophyllide biosynthesis I (Zm00001d018034 and Zm00001d026603), chlorophyll degradation (Zm00001d034523), or chorismate/tyrosine synthesis (Zm00001d029391 and Zm00001d021611) pathways.

In addition to ANOVA, Tukey's HSD tests were conducted on the 126 *a priori* pathway genes between light treatment pairs of each NAM parent. However, the changes in gene expression levels in response to light treatments were predominantly genotype-dependent (Supplemental Table S5.14). Overall, we detected 42, 36, and 35 significant genes in the light-deprived vs. control, high-light vs. control, and light-deprived vs. high-light comparisons across the seven NAM parents, respectively. This resulted in the detection of 60 unique genes from nine pathways across all tested 21 pairwise comparisons. Of the seven

genes in the prenyl group synthesis pathway, only *vte5*, which encodes a kinase that catalyzes the first step of phytol phosphorylation into phytyl-monophosphate (Valentin *et al.*, 2006), was detected. Compared to high-light conditions, the expression level of *vte5* was significantly decreased in the embryo of B73 and B97 under light-deprived conditions.

In contrast to the in the prenyl group synthesis pathway, we observed an above average percentage (> 48.76%) of genes from three chlorophyll-related pathways: 12 (57.1%) genes from the 3,8-divinyl-chlorophyllide biosynthesis I, five genes (83.3%) from the chlorophyll cycle, and five genes (62.5%) from tetrapyrrole biosynthesis I. Consistent with the observed significant genotype × treatment effect for *por2*, the transcript abundance of *por2* was significantly elevated in the high-light samples of B73 relative to control samples, but this response was not observed in the other six NAM parents. Although *por1* did not have significant genotype × treatment effect, the transcript abundance of *por1* was significantly increased in high-light and decreased in light-deprived samples of B73.

DISCUSSION

Tocopherols belong to a family of lipophilic compounds exhibiting vitamin E and antioxidant activities that are essential for human health and plant fitness. Although the core tocopherol biosynthesis pathway has been well-elucidated in Arabidopsis, the mechanism that regulates the supply of the phytol group, a key precursor for tocopherol synthesis, is still to be confirmed in non-photosynthetic tissues such as maize grain. To that end, we conducted experiments with several NAM parents having contrasting *por1* and *por2* allelic effects to test the impact of three light treatments (high-light, light-deprivation, and control) on

tocochromanol and chlorophyll a levels in the developing kernels of self-pollinated ears. Our results from metabolite and expression profile analyses further strengthen the connection between the chlorophyll and tocopherol biosynthesis pathways in the synthesis of tocopherols in maize grain.

We detected significant treatment and genotype × treatment effects for all four tocopherol phenotypes and a significantly lower accumulation of tocopherols in light-deprived samples across both years and tissue samples. In contrast, significant treatment and genotype × treatment effects were not consistently detected for tocotrienol phenotypes, as there was no consistent increase or decrease of tocotrienol levels in light-deprived samples compared to control and high-light samples. If the significant differences observed for tocopherol levels between light-deprived samples and control samples were caused by abiotic stresses such as temperature or humidity changes posed by the light treatments, we would expect a similar decreased accumulation in other grain metabolites such as tocotrienols (Jones *et al.*, 1981; L. Wang *et al.*, 2012). As such, the decrease in grain tocopherols is most likely due to a strong attenuation of tocopherol synthesis from the imposed light-deprived treatment. Interestingly, there were still very low levels of tocopherols measured for the light-deprived grain samples, which could be attributed to several possibilities, including a trace amount of light leakage to the ear, the accumulation of tocopherols or their precursors prior the the application of the treatment at 12 DAP, and/or from the availability of GGDP-derived PDP that does not rely on chlorophyll-derived phytol (DellaPenna & Mène-Saffrané, 2011).

In contrast to the light-deprived mature grain samples, tocopherol levels in high-light nature grain samples were either higher or lower depending on the NAM parent relative to

control ears. To better understand these patterns, we focused more on the embryo samples that had been measured for the levels of tocopherols and chlorophyll a, which had a significantly higher level of ΣT in high-light samples in all but one (OH7B) NAM parent compared to the control samples. Even though all but one (NC358) of the NAM parents had high-light embryo samples with an increased level of chlorophyll a compared to the control, only three parents (B97, Ki11, and M37W) had a significant difference in chlorophyll a content between high-light and control samples. Collectively, these results imply that the high-light treatment increased the level of chlorophyll a in maize grain, which could then be used as a substrate to produce a higher level of phytol for the elevated synthesis of tocopherols.

In addition to tocochromanol and chlorophyll a levels, we conducted an ANOVA of transcript abundances for 126 *a priori* genes in 24 DAP embryos from the evaluated NAM parents. Of the 126 *a priori* genes, 28 were found to have significant genotype × treatment and/or treatment effects, including an overrepresentation of genes from 3,8-divinyl-chlorophyllide biosynthesis I (32%), tocochromanol (33%), chlorophyll degradation (30%), tetrapyrrole biosynthesis I (25%), and tyrosine degradation (25%) pathways (Supplemental Table S5.14). In contrast, none of the genes from the prenyl group synthesis (0%) pathway and only two genes (13.3%) from the IPP synthesis pathway had significant genotype × treatment or treatment effects. The two significant genes from the IPP synthesis pathway were *dxs1* (treatment) and *dxs3* (genotype × treatment), which were identified to associate with γT3 in the Ames panel (Table 4.2) and plastochromanol-8 in the U.S. NAM panel (Diepenbrock *et al.,* 2017), respectively, but not with tocopherols. Taken together, we demonstrated that the expression of genes in IPP pathway and prenyl group synthesis are not majorly light-

mediated, whereas more genes from the chlorophyll synthesis and degradation pathways are light-responsive in maize developing embryos.

Through genetic mapping in the maize U.S. NAM (Diepenbrock *et al.* 2017) and Ames panels (Table 4.2), the *por1* and *por2* genes from the chlorophyll biosynthesis pathway were associated with grain tocopherol levels, implicating the importance of chlorophyll-derived phytol for tocopherol synthesis in the maize grain. In Diepenbrock *et al.* (2017), multiple lines of evidence were leveraged to posit that tocopherol synthesis in maize grain primarily depends on phytol released from a chlorophyll-based cycle as opposed to chlorophyll degradation. The level of pheophytin a, which is a metabolite indicative of chlorophyll degradation, was only weakly correlated with ΣT in developing embryos from maize NAM parents. Relatedly, a 1:800 chlorophyll:tocopherol molar ratio was observed in 30 DAP embryos, which stoichiometrically argued against the major provision of phytol from chlorophyll degradation. In contrast, chlorophylls a and b and chlorophyllide a were strongly correlated with ΣT in the embryo. In agreement with Diepenbrock *et al.* (2017), our metabolite and gene expression profiling results suggest that PDP for tocopherol synthesis in maize grain is produced primarily from the phosphorylation of phytol generated by a light-dependent chlorophyll-based cycle.

Our hypothesis on the main tissue source of chlorophyll-derived phytol comes from the near-zero tocopherol levels consistently observed for both types of light-deprived samples. As light deprivation was only applied to self-pollinated ears, the diminished tocopherol levels in developing kernels of light-deprived ears suggests that the phytol derived from the proposed chlorophyll-based cycle was provided by the ear itself instead of transported from

the flag leaf as proposed by Zhan *et al.* (2019). The majority of this phytol provision is unlikely from husk leaves either, as the husk leaves of light-deprived ears at 24 DAP, although more pale in color, were still visibly green. If husk leaves were the main source of phytol for tocopherol synthesis in grain, it is unlikely that we would have observed the near-zero tocopherol levels in light-deprived grain samples. Thus, we can infer that most of the phytol for the production of tocopherols is synthesized in maize grain and specifically in the embryo where tocopherol predominantly accumulates (Diepenbrock *et al.*, 2017; Grams *et al.*, 1970; Weber, 1987).

In summary, we provide additional evidence that a chlorophyll-based cycle is the main provider of phytol for tocopherol synthesis in maize grain. The validation of the important involvement of chlorophyll in tocopherol synthesis would suggest that previously identified *por1*, *por2* and *vte7* loci are important breeding targets that can be selected upon to generate sufficient phytol for tocopherol synthesis. These genes combined with breeding targets from core tocopherol pathway genes such as *vte3* and *vte4* could result in an optimized grain profile that has increased tocopherol content with a higher composition of α-tocopherol (highest vitamin E activity) for the improved benefit of human health and nutrition.

SUPPLEMENTAL INFORMATION

Supplemental Figure S5.1. A bar plot showing the number of significant pairwise comparisons each unique DEG was identified in.

Supplemental Table S5.1. Concentration (raw) of tocochromanols (μg g^{-1}) from mature grain samples of seven NAM parents from 2018.

Supplemental Table S5.2. Concentration (raw) of tocochromanols (μg g^{-1}) from 24 DAP embryo samples of seven NAM parents from 2018.

Supplemental Table S5.3. Concentration (raw) of tocochromanols (μg g^{-1}) from mature grain samples of three NAM parents from 2019.

Supplemental Table S5.4. Read counts of 39,498 genes (B73 RefGen_v4) for the 42 pooled embryo samples in 2018.

Supplemental Table S5.5. Average concentration (raw) of tocochromanols (μg g^{-1}) from 2018 and 2019 of mature grain samples of NAM parents after outlier removal.

Supplemental Table S5.6. *P*-values from Levene's test results for equality of variances.

Supplemental Table S5.7. Best linear unbiased estimators (BLUEs) of nine tocochromanol phenotypes from mature grains of seven NAM parents in 2018 and 2019.

Supplemental Table S5.8. Best linear unbiased estimators (BLUEs) of tocochromanols and chlorophyll a from the 24 DAP embryos of seven NAM parents in 2018.

Supplemental Table S5.9. FPKM (raw) of 126 *a priori* candidate genes in 24 DAP embryos.

Supplemental Table S5.10. *P*-values from ANOVA for tocochromanol and chlorophyll a levels in mature grains and 24 DAP embryos in 2018 and 2019.

Supplemental Table S5.11. *P*-values from Tukey's HSD tests for tocochromanol and chlorophyll a levels in mature grains and 24 DAP embryos in 2018 and 2019.

Supplemental Table S5.12. A list of differentially expressed genes in 24 DAP embryos in 2018 between pairwise treatment comparisons within each NAM parent.

Supplemental Table S5.13. GO term enrichment results from 595 DEGs from 24 DAP embryos.

Supplemental Table S5.14. *P*-values from ANOVA and Tukey's HSD test results of the 126 *a priori* candidate genes for the expression FPKM of 24 DAP embryos from 2018.

REFERENCE

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169.

Buhr, F., Lahroussi, A., Springer, A., Rustgi, S., von Wettstein, D., Reinbothe, C., & Reinbothe, S. (2017). NADPH:protochlorophyllide oxidoreductase B (PORB) action in *Arabidopsis thaliana* revisited through transgenic expression of engineered barley PORB mutant proteins. *Plant Mol. Biol.*, *94*(1-2), 45–59.

Collakova, E., & DellaPenna, D. (2003). The role of homogentisate phytyltransferase and other tocopherol pathway enzymes in the regulation of tocopherol synthesis during abiotic stress. *Plant Physiol.*, *133*(2), 930–940.

DellaPenna, D., & Mène-Saffrané, L. (2011). Vitamin E. In F. Rébeillé & R. Douce (Eds.), *Advances in Botanical Research* (Vol. 59, pp. 179–227). Elsevier Ltd., Amsterdam, The Netherlands.

Diepenbrock, C. H., Kandianis, C. B., Lipka, A. E., Magallanes-Lundback, M., Vaillancourt, B., Góngora-Castillo, E., Wallace, J. G., Cepela, J., Mesberg, A., Bradbury, P. J., Ilut, D. C., Mateos-Hernandez, M., Hamilton, J., Owens, B. F., Tiede, T., Buckler, E. S., Rocheford, T., Buell, C. R., Gore, M. A., & DellaPenna, D. (2017). Novel loci underlie natural variation in vitamin E levels in maize grain. *Plant Cell*, *29*(10), 2374–2392.

Federer, W. T., & King, F. (2007). *Variations on Split Plot and Split Block Experiment Designs*. John Wiley & Sons.

Grams, G. W., Blessin, C. W., & Inglett, G. E. (1970). Distribution of tocopherols within the corn kernel. *J. Am. Oil Chem. Soc.*, *47*(9), 337–339.

Hey, D., Rothbart, M., Herbst, J., Wang, P., Müller, J., Wittmann, D., Gruhl, K., & Grimm, B. (2017). LIL3, a light-harvesting complex protein, links terpenoid and tetrapyrrole biosynthesis in *Arabidopsis thaliana*. *Plant Physiol.*, *174*(2), 1037–1050.

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.

Jones, R. J., Gengenbach, B. G., & Cardwell, V. B. (1981). Temperature effects on *in vitro* kernel development of maize. *Crop Sci.*, *21*(5), 761–766.

Keller, Y., Bouvier, F., d'Harlingue, A., & Camara, B. (1998). Metabolic compartmentation of plastid

prenyllipid biosynthesis--evidence for the involvement of a multifunctional geranylgeranyl reductase. *Eur. J. Biochem.*, *251*(1-2), 413–417.

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, *37*(8), 907–915.

Leth, T., & Søndergaard, H. (1977). Biological activity of vitamin E compounds and natural materials by the resorption-gestation test, and chemical determination of the vitamin E activity in foods and feeds. *J. Nutr.*, *107*(12), 2236–2243.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.*, *49*(4), 764–766.

Lin, Y.-P., & Charng, Y.-Y. (2021). Chlorophyll dephytylation in chlorophyll metabolism: a simple reaction catalyzed by various enzymes. *Plant Sci.*, *302*, 110682.

Lin, Y.-P., Lee, T.-Y., Tanaka, A., & Charng, Y.-Y. (2014). Analysis of an Arabidopsis heat-sensitive mutant reveals that chlorophyll synthase is involved in reutilization of chlorophyllide during chlorophyll turnover. *Plant J.*, *80*(1), 14–26.

Lin, Y.-P., Wu, M.-C., & Charng, Y.-Y. (2016). Identification of a chlorophyll dephytylase involved in chlorophyll turnover in Arabidopsis. *Plant Cell*, *28*(12), 2974–2990.

Lipka, A. E., Gore, M. A., Magallanes-Lundback, M., Mesberg, A., Lin, H., Tiede, T., Chen, C., Buell, C. R., Buckler, E. S., Rocheford, T., & DellaPenna, D. (2013). Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. *G3*, *3*(8), 1287–1299.

Liu, J., & Last, R. L. (2015a). MPH1 is a thylakoid membrane protein involved in protecting photosystem II from photodamage in land plants. *Plant Signal. Behav.*, *10*(10), e1076602.

Liu, J., & Last, R. L. (2015b). A land plant-specific thylakoid membrane protein contributes to photosystem II maintenance in Arabidopsis thaliana. *Plant J.*, *82*(5), 731–743.

Liu, X., Hua, X., Guo, J., Qi, D., Wang, L., Liu, Z., Jin, Z., Chen, S., & Liu, G. (2008). Enhanced tolerance to drought stress in transgenic tobacco plants overexpressing *VTE1* for increased tocopherol production from *Arabidopsis thaliana*. *Biotechnol. Lett.*, *30*(7), 1275–1280.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, *15*(12), 550.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12.

Mène-Saffrané, L., & Pellaud, S. (2017). Current strategies for vitamin E biofortification of crops. *Curr. Opin. Biotechnol.*, *44*, 189–197.

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.*, *49*(D1), D394–D403.

Miller, J. (1991). Reaction time analysis with outlier exclusion: bias varies with sample size. *Q. J. Exp. Psychol. A*, *43*(4), 907–912.

Mork-Jansson, A., Bue, A. K., Gargano, D., Furnes, C., Reisinger, V., Arnold, J., Kmiec, K., &

Eichacker, L. A. (2015). Lil3 Assembles with Proteins Regulating Chlorophyll Synthesis in Barley. *PloS One*, *10*(7), e0133145.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (Vol. 4). Irwin Chicago.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.*

Sattler, S. E., Gilliland, L. U., Magallanes-Lundback, M., Pollard, M., & DellaPenna, D. (2004). Vitamin E is essential for seed longevity and for preventing lipid peroxidation during germination. *Plant Cell*, *16*(6), 1419–1432.

Steel, R. G. D., Torrie, J. H., & Dickey, D. A. (1997). *Principles and Procedures of Statistics: A Biometrical Approach*. McGraw-Hill.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, *28*(5), 511–515.

Valentin, H. E., Lincoln, K., Moshiri, F., Jensen, P. K., Qi, Q., Venkatesh, T. V., Karunanandaa, B., Baszis, S. R., Norris, S. R., Savidge, B., Gruys, K. J., & Last, R. L. (2006). The *Arabidopsis* vitamin E pathway *gene5-1* mutant reveals a critical role for phytol kinase in seed tocopherol biosynthesis. *Plant Cell*, *18*(1), 212–224.

Vom Dorp, K., Hölzl, G., Plohmann, C., Eisenhut, M., Abraham, M., Weber, A. P. M., Hanson, A. D., & Dörmann, P. (2015). Remobilization of phytol from chlorophyll degradation is essential for tocopherol synthesis and growth of Arabidopsis. *Plant Cell*, *27*(10), 2846–2859.

Wan, C. Y., & Wilkins, T. A. (1994). A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (*Gossypium hirsutum* L.). *Anal. Biochem.*, *223*(1), 7–12.

Wang, L., Ma, H., Song, L., Shu, Y., & Gu, W. (2012). Comparative proteomics analysis reveals the mechanism of pre-harvest seed deterioration of soybean under high temperature and humidity stress. *J. Proteomics*, *75*(7), 2109–2127.

Wang, P., & Grimm, B. (2021). Connecting chlorophyll metabolism with accumulation of the photosynthetic apparatus. *Trends Plant Sci.*, *26*(5), 484–495.

Weber, E. J. (1987). Carotenoids and tocols of corn grain determined by HPLC. *J. Am. Oil Chem. Soc.*, *64*(8), 1129–1134.

Wolf, G. (2005). The discovery of the antioxidant function of vitamin E: the contribution of Henry A. Mattill. *J. Nutr.*, *135*(3), 363–366.

Zhan, W., Liu, J., Pan, Q., Wang, H., Yan, S., Li, K., Deng, M., Li, W., Liu, N., Kong, Q., Fernie, A. R., & Yan, J. (2019). An allele of *ZmPORB2* encoding a protochlorophyllide oxidoreductase promotes tocopherol accumulation in both leaves and kernels of maize. *Plant J.*, *100*(1), 114–127.

Chapter 6 Conclusions

In this dissertation, I utilized the research tools from quantitative genetics, population genetics, and statistics, and conducted genetic studies on two of the Three Sisters: common bean and maize. Specifically, this dissertation has three topics: i) genomic characterization of the Native Seeds/SEARCH common bean collection and its seed coat patterns; and utilization of genomic tools to identify causal genes involved in the transport and/or biosynthesis of ii) elemental phenotypes and iii) tocochromanols (vitamin E and antioxidants) in maize grain.

In Chapter 2, I have characterized and established the importance of a unique set of nearly 300 common bean accessions from Native Seeds/SEARCH. This novel genetic resource was found to be underrepresented in current seed banks and it is likely to possess important alleles for biotic and abiotic stress tolerance that have been selected by local farmers over thousands of years. We have observed high pairwise F_{ST} among all three subpopulations within the panel and higher overall nucleotide diversity relative to within each subpopulation. Considering the overlapping collection locations of two of the three subpopulations, this suggested the presence of both natural selection, as well as local preference and artificial selection that shaped the population structure of domesticated common beans. We are also a step closer to understanding the complex genetic control of common bean seed coat patterns. We were able to identify four independent SNPs that were associated with seed coat patterns and five *myb* genes to be the putative causal genes

underlying the *C* locus. A potentially novel locus was also identified, controlling a specific partial coloring seed coat phenotype. The cloning of the *C* locus is currently underway by other research groups in the community and more research is warranted to understand the complex genetic control of the extremely diverse patterns and colors of common bean seed coats.

Through the analysis of 11 elemental phenotypes in maize grain, we observed moderate to high heritability for all elements and all positive pairwise correlations in the Ames panel. Through high-resolution genome-wide association study (GWAS) with 7.7 M SNP markers, we identified nine candidate genes for seven elements, including two novel associations between *rte2* with B, and *irt1* with Ni. A total of five available UniformMu mutants for three genes [UFMu-05386, UFMu-04494, and UFMu-11923 for Zm00001d011013 (*cap1*); UFMu-07153 for Zm00001d017429 (*ys1*); and UFMu-00305 for Zm00001d017427 (*ysl2*)] were obtained from the Maize Genetics Cooperation Stock Center and planted in 2020 for further functional validation of the identified genes. Not all GWAS signals were resolved to a gene level though. For example, we observed a significant association on chromosome 9 for nickel, and this genomic region overlaps with the association signal detected for nickel in the maize NAM panel (Ziegler *et al.,* 2017) and a sweet corn diversity panel (Baseggio *et al.,* 2021). Notably, as potentially important biofortification targets, we identified two genes that were significantly associated with zinc (*nas5* and *ysl2*) and iron (*nas5*) concentrations in the grain. The favorable alleles (alleles associated with higher zinc or iron mean concentrations) of *nas5* are occurring at low frequencies, especially in the Stiff Stalk and Tropical subpopulations within the Ames panel.

Whole-genome prediction models, using either Bayesian Ridge Regression (BRR) or BayesB, generated similar moderate predictive abilities for all 11 elements. Consistent with GWAS results, where copper, molybdenum, and nickel had the highest number of detected associations and the largest amount of phenotypic variation explained by the peak SNPs, the BayesB model slightly outperformed BRR for these three elements.

Aside from elements, vitamin E is also an important nutrient for both plant and human health. Our comprehensive study on tocochromanol levels in the maize grain identified a total of 13 causal genes. Two *a priori* pathway genes, *vte5* (*phytol kinase*) and *dxs1* (*1-deoxy-D-xylulose-5-phosphate synthase*) were found for the first time to be associated with tocochromanol concentrations in maize grain. Additionally, we demonstrated the first association of two genes *vte7* (*alpha/beta hydrolase*) and *samt1* (*S-adenosylmethionine transporter 1*) with tocopherols by integrating GWAS and transcriptome-wide association study (TWAS) results. In Diepenbrock *et al.* (2017), a chlorophyll-based cycle was proposed following the identification of two *protochlorophyllide reductases* (*por1* and *por2*) from the chlorophyll biosynthesis pathway, however, the key gene(s) catalyzing this cycle was not discovered. The identified *vte7* gene is the missing gene that can hydrolyze chlorophyll a to provide phytol group and regulate tocopherol levels in maize. The majority of the 13 identified causal genes were found to be under cis-regulatory control through expression quantitative trait loci (eQTL) mapping. Five trans-eQTL were identified for the 13 genes, and a gene underlying the trans-eQTL was proposed to play a regulatory role for *dxs2*, a known gene in the IPP pathway for tocotrienol tail group biosynthesis. This regulatory gene was *phytoene synthase 1* (*psy1*), a pathway gene catalyzing the first committed step in carotenoid

synthesis.

The experiment detailed in Chapter 5 provides valuable supporting evidence to the hypothesized chlorophyll-based cycle and its involvement in the synthesis of tocopherols. The observed near-zero tocopherol levels in the kernels of light-deprived ears strongly support the importance of chlorophyll in the provision of phytol for tocopherol synthesis in maize grain. This experiment, together with the ongoing experiments with *por1* and *por2* knockout mutants generated by CRISPR/Cas9-mediated editing, will potentially confirm the important roles of these two *por* genes in tocopherol synthesis and provide more insight into the interplay between chlorophyll and tocopherol pathways in the maize embryo.

As with any research, many additional experiments can provide interesting insights for further exploration. For example, I have collected seed coat scans of all the common bean accessions in our study. If the intricacies of irregular pattern recognition for the seed coats can be resolved through image processing algorithms, it would be an intriguing addition to the study and can provide more granularity to the genetic mapping of seed coat patterns and colors.

For the maize elemental study, apart from the gene validation work mentioned above, it would be valuable to have detailed soil sample data from the field, apart from the existing low-resolution soil type data. As soil nutrient availability is a major environmental factor affecting element uptake and accumulation in plants (Marschner & Rengel, 2012), this additional dataset would greatly help with the removal of environmental effects and improving the statistical power for genetic mapping. In addition, the Ames panel for the elemental study was planted in only one location over two consecutive years, contributing to

why I did not observe a strong G×E effect for phenotypes that was evident in multiple location trials (Asaro *et al.,* 2016; Fikas *et al.,* 2019; Ziegler *et al.,* 2017). The grain elemental data collected in multiple environments (location × year) would not only allow me to estimate the QTL×E interactions, but it would also be interesting to see the performance of whole-genome prediction models when incorporating these interactions. Both analyses can provide valuable insights for breeding efforts. Furthermore, a haplotype-based association analysis could be performed on the *nas5* and *ysl2* loci, which could help with the identification of the best combination of favorable haplotypes for breeding.

In Chapter 4, the expression profile of genes in developing grain was obtained at only one time point (23 days after pollination). However, changes in gene expression over kernel development, especially for *a priori* pathway genes, would provide us with a far deeper insight into the genetic control of tocochromanol accumulation on a time axis. Although a massive undertaking, sampling the kernels in a time series at earlier and later developmental stages would provide more depth for elucidating the regulatory network of tocochromanol accumulation. Similarly, spatial variation of genetic control could be explored through the dissection of developing kernels and individual evaluation of tocochromanols and gene expression profile in the endosperm and embryo. The independently analyzed embryo and endosperm data could shed light on the potential flux of precursors between kernel tissues for tocochromanol biosynthesis. We could also perform two variant annotation analyses for all SNPs within 1 Mb of the 13 causal genes identified in this study, genomic evolutionary rate profiling (GERP) and SNP effect analysis (SnpEff). GERP is a tool for prediction of the evolutionary constraint of a variant through multi-species alignments (Rodgers-Melnick *et al.,*

2015), and SnpEff is a program for variant annotation and coding effect prediction (Cingolani *et al.*, 2012). Together, both analyses could provide functional insights into the 13 causal genes in tocochromanol biosynthesis.

For the experiment detailed in Chapter 5, in retrospect, a trial study should have been performed to explore the best sampling and dissection methods for developing kernels that would have the least impact on gene expression profile. More importantly, the remnant endosperm and whole kernel samples after sampling should have been preserved in 2018, which later proved to be a major missing piece that prompted the repeat of the experiment in 2019. In addition, because of the lethality of double CRISPR/Cas9 knockout mutants of *por1* and *por2*, overexpression of the two *por* genes with embryo-specific promoters would also provide valuable information on the function of the two genes that are independent of whole-plant changes.

In summary, this dissertation provided important knowledge of the Native Seed/SEARCH common bean collection and comprehensive assessments of the natural variation of elements and tocochromanols in maize grain. I hope that the knowledge gained from these studies would be utilized in breeding programs and transform into elite cultivars with high yield, high stress tolerance, and balanced nutritional profiles, and help alleviate the hunger and hidden hunger problems, which is still prevalent in a lot of regions in the world.

REFERENCES

Asaro, A., Ziegler, G., Ziyomo, C., Hoekenga, O. A., Dilkes, B. P., & Baxter, I. (2016). The interaction of genotype and environment determines variation in the maize kernel ionome. *G3*, *6*(12), 4175–4183.

Baseggio, M., Murray, M., Wu, D., Ziegler, G., Kaczmar, N., Chamness, J., Hamilton, J. P., Buell, C. R., Vatamaniuk, O. K., Buckler, E. S., Smith, M. E., Baxter, I., Tracy, W. F., & Gore, M. A. (2021). Genome-wide association study suggests an independent genetic basis of zinc and cadmium concentrations in fresh sweet corn kernels. *G3*. https://doi.org/10.1093/g3journal/jkab186

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92.

Diepenbrock, C. H., Kandianis, C. B., Lipka, A. E., Magallanes-Lundback, M., Vaillancourt, B., Góngora-Castillo, E., Wallace, J. G., Cepela, J., Mesberg, A., Bradbury, P. J., Ilut, D. C., Mateos-Hernandez, M., Hamilton, J., Owens, B. F., Tiede, T., Buckler, E. S., Rocheford, T., Buell, C. R., Gore, M. A., & DellaPenna, D. (2017). Novel loci underlie natural variation in vitamin E levels in maize grain. *The Plant Cell*, *29*(10), 2374–2392.

Fikas, A. A., Dilkes, B. P., & Baxter, I. (2019). Multivariate analysis reveals environmental and genetic determinants of element covariation in the maize grain ionome. *Plant Direct*, *3*(5), e00139.

Grams, G. W., Blessin, C. W., & Inglett, G. E. (1970). Distribution of tocopherols within the corn kernel. *Journal of the American Oil Chemists' Society*, *47*(9), 337–339.

Marschner, P., & Rengel, Z. (2012). Chapter 12 - Nutrient Availability in Soils. In P. Marschner (Ed.), *Marschner's Mineral Nutrition of Higher Plants (Third Edition)* (pp. 315–330). Academic Press.

Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., Mitchell, S. E., Li, C., Li, Y., & Buckler, E. S. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(12), 3823–3828.

Weber, E. J. (1987). Carotenoids and tocols of corn grain determined by HPLC. *Journal of the American Oil Chemists' Society*, *64*(8), 1129–1134.

Ziegler, G., Kear, P. J., Wu, D., Ziyomo, C., Lipka, A. E., Gore, M., Hoekenga, O., & Baxter, I. (2017). Elemental accumulation in kernels of the maize nested association mapping panel reveals signals of gene by environment interactions. In *bioRxiv* (p. 164962). https://doi.org/10.1101/164962