GENETIC BASIS OF VITAMIN AND MINERAL LEVELS IN FRESH SWEET CORN KERNELS

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

Matheus Baseggio

August 2019

GENETIC BASIS OF VITAMIN AND MINERAL LEVELS IN FRESH SWEET CORN KERNELS

Matheus Baseggio, Ph. D.

Cornell University 2019

Nutritional deficiencies affect more than two billion people worldwide, with iron, zinc, vitamin E, and non-provitamin A carotenoids at risk of deficiency in the US. Sweet corn is widely consumed and does not usually provide significant amounts of these compounds. We assessed the natural variability of tocochromanols (vitamin E and antioxidants), carotenoids (provitamin A, lutein, zeaxanthin), and 15 minerals in fresh kernels (~21 DAP) of a sweet corn association panel that samples the genetic diversity of the US germplasm pool. For each phenotype, we performed a genome-wide association study to identify the genes involved in the genetic control of their quantitative variation and genomic prediction to provide insights into how best to enhance genetic gains in a sweet corn biofortification program.

In the first study, we identified significant associations of α-tocopherol (highest vitamin E activity) with *vte4*, as well as content and composition of tocotrienols (antioxidants) with *hggt1* and *vte1*, respectively. In the second study, we reported that β-carotene (provitamin A) was associated with *crtRB1*, and the relative flux between α- and β-branches of the carotenoid pathway was controlled by *lcyE*. For tocotrienols, we identified associations with two starch biosynthetic genes (*su1* and *sh2*) specific to sweet corn, and reported evidence for the involvement of *sh2*. In the third study, of the 15 studied minerals, iron and zinc were associated with *nas5* and cadmium with *hma3*. Weaker-effect associations specific for a location were

observed for calcium (WI) and nickel (NY), and these markers were within ± 250 kb of the genes *ras2* and *ptr2*, respectively.

Whole-genome prediction models had moderate prediction abilities for most of the phenotypes measured in the three studies, indicating that these models may be used for developing sweet corn lines with nutrient-dense kernels. Smaller marker datasets that target genes or quantitative trait loci associated with carotenoids or tocochromanols resulted in lower prediction abilities compared to the whole-genome set, but the inclusion of endosperm mutation type in the models increased the abilities for tocotrienols and certain carotenoids. Together, these studies represent the most extensive assessment of natural variation for vitamins and minerals in fresh sweet corn kernels and constitute a key step for improving the nutritional quality of sweet corn for human health.

BIOGRAPHICAL SKETCH

Matheus Baseggio was born in 1988 in Passo Fundo, Brazil. He grew up helping his father on the farm, and with all his family involved, becoming interested in agriculture was a natural progression. He went to school at Universidade de Passo Fundo, Brazil where he received his B.A. degree in Agronomic Engineering in 2010. The year before, 2009, as part of the requirements for graduation, he joined the Forage Extension Program at the Agronomy Department of the University of Florida as an intern during 4 months. While an intern, he enjoyed the exposure to dairy, beef, and hay production systems, and decided in summer of 2011 to enroll at the University of Florida to pursue graduate studies on forage management. In August 2013 he received a Master's degree from the Agronomy Department at UF, with focus on management practices to improve establishment of Tifton-85 bermudagrass.

Matt's experience with plant breeding and genetics started soon after he finished his M.S. degree, when he worked closely with Dr. Patricio Munoz at UF as a breeder assistant to help implement a new forage breeding program. After a year, Matt decided to pursue a Ph.D. degree in Plant Breeding and Genetics at Cornell University under the guidance of Dr. Michael Gore, with minor in Food Chemistry. During his program, he performed extensive field work and greatly developed his skills as a data analyst. As part of his extracurricular activities, Matt helped organize the Corteva Plant Sciences Symposium in 2017 and 2018. Matt also represented the Plant Breeding section at several national and regional meetings and received the Munger/Murphy Award for his outstanding performance in the areas of scholarship, research, and service as a graduate student.

To my family and friends

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my family, who always supported and encouraged me to fight for what I wanted; for being always present, for sparing no effort to provide the best possible environment for me to grow up, and of course, for the financial support, without which I would not be able to be where I am today.

I especially want to thank my advisor Dr. Michael Gore for his guidance throughout my Ph.D. program. His significant contribution to planning and accomplishment of my studies together with his involvement in guiding the writing of the dissertation are truly appreciated. I am thankful to him for providing me with the opportunity to come to Cornell. I would also like to thank other members of my committee, Dr. Gavin Sacks, and Dr. Margaret Smith for their willingness to provide advice and for reviewing my dissertation.

I would like to thank all the past and present members of the Gore lab. They all provided tremendous help with field experiments, data collection and analyses, and scientific writing. Thanks for all the ideas and critics during lab meetings, which made me improve my presentation skills. Special thanks to Nicholas Kaczmar, who greatly assisted in my field and lab experiments and provided a fun environment to work at.

I wish to thank the Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES-Brazil), the United States Department of Agriculture, and Plant Breeding and Genetics section for support of my graduate studies and research projects. Also, the faculty, staff, and students in the section were vital to making my Ph.D. at Cornell University an amazing experience.

I am also grateful for the field assistance and friendship of my fellow graduate students, in particular Al Kovaleski for his companionship. I would like to extend my sincerest thanks to

all my Brazilian and international friends, the ones I left in Brazil and all the friends I made here in the US who helped me keep strong during this period away from my family. Special thanks to my friends I made in Ithaca who always helped make my days even better.

Last but not least, thanks to God for my life and for all the opportunities He has granted me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

GENOME-WIDE ASSOCIATION AND GENOMIC PREDICTION MODELS OF TOCOCHROMANOLS IN FRESH SWEET CORN KERNELS[1]

INTRODUCTION

Tocochromanols, which include four tocopherols and four tocotrienols, are lipid-soluble compounds synthesized by photosynthetic organisms that function as powerful scavengers of lipid peroxy radicals and singlet oxygen quenchers (Kruk et al., 2005). In plants, tocochromanols are important for limiting the oxidation of storage lipids in the seed (Sattler et al., 2004) and providing protection against environmental stress (Liu et al., 2008). The saturated tail of tocopherols is derived from phytyl diphosphate, whereas the unsaturated tail of tocotrienols derives from geranylgeranyl diphosphate (Fig. 1). Within each of the two classes of tocochromanols, the four different chemical species (α, β, δ, and γ) are distinguished by the number and position of methyl groups on the aromatic ring [reviewed in DellaPenna and Mène-Saffrané (2011)]. In general, tocopherol species have greater vitamin E activity than their corresponding tocotrienol species, although tocotrienols tend to have greater antioxidant capacity (Sen et al., 2006). For both classes of compounds, vitamin E activity follows the order α > β > > γ > δ, with α-tocopherol having the highest vitamin E activity on a molar basis (Leth and Sondergaard, 1977).

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

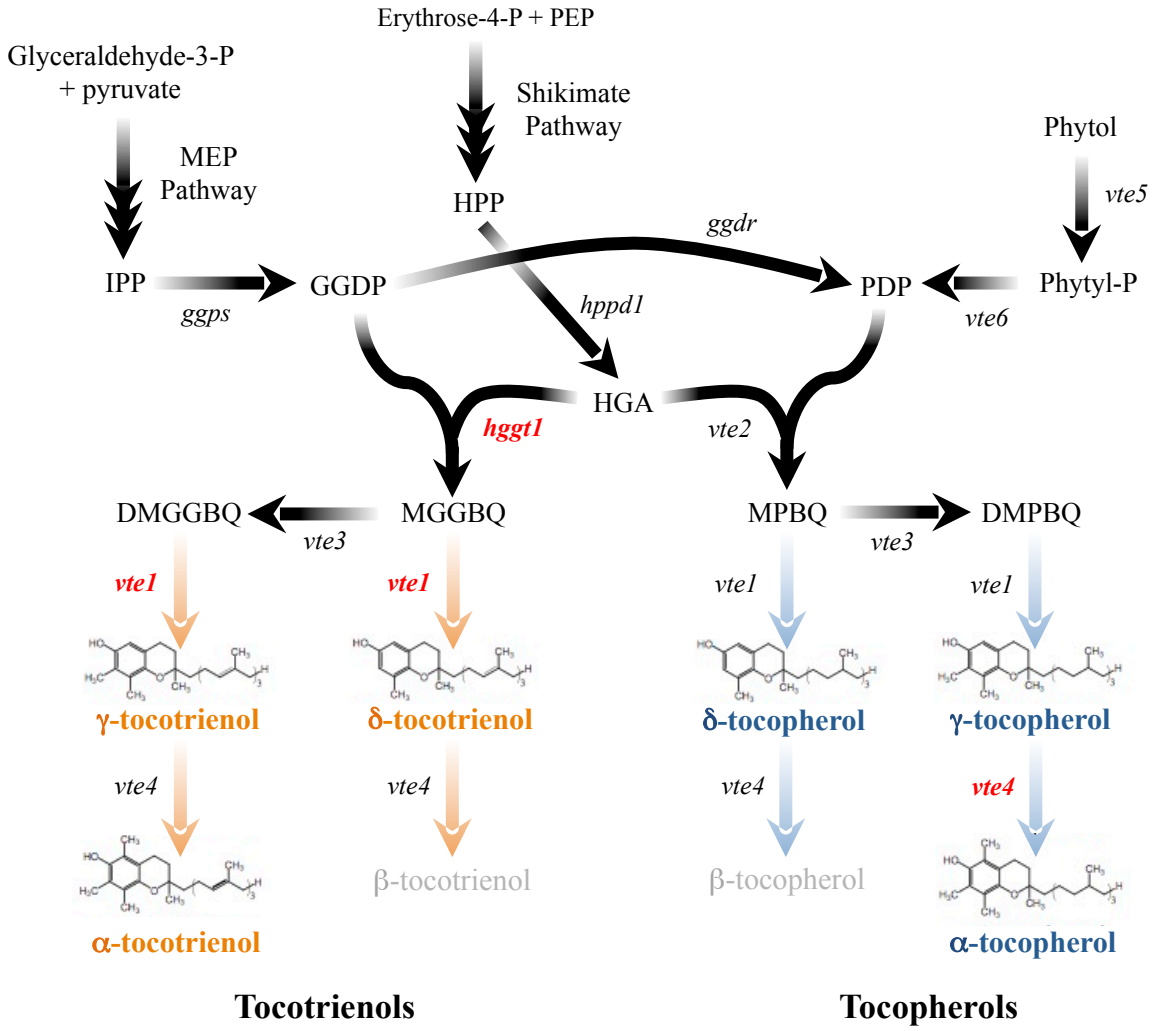

Figure 1.1. Tocochromanol biosynthetic pathway in maize. The six quantified compounds are shown in bolded orange (tocotrienols) or blue (tocopherols) text. The name of genes in bolded red text correspond to genes that are within ± 250 kb of the associated single nucleotide polymorphisms (SNPs) identified in our study for an adjacent compound or derivative trait. Compound abbreviations: DMGGBQ, 2,3-dimethyl-5-geranylgeranyl-1,4-benzoquinol; DMPBQ, 2,3-dimethyl-5-phytylbenzoquinol; GGDP, geranylgeranyl diphosphate; HGA, homogentisic acid; HPP, *p*-hydroxyphenylpyruvate; IPP, isopentenyl pyrophosphate; MGGBQ, 2-methyl-6-geranylgeranyl-1,4-benzoquinol; MPBQ, 2-methyl-6-phytyl-1,4-benzoquinol; PDP, phytyl diphosphate; Phytyl-P, phytyl monophosphate. Gene abbreviations: *ggdr*, *geranylgeranyl diphosphate reductase*; *gps*, *geranyl diphosphate synthase*; *hggt1*, *homogentisate geranylgeranyltransferase*; *hppd1*, *4-hydroxyphenylpyruvate dioxygenase*; *vte1*, *tocopherol cyclase*; *vte2*, *homogentisate phytyltransferase*; *vte3*, *MPBQ/MGGBQ methyltransferase*; *vte4*, *γ-tocopherol methyltransferase*; *vte5*, *phytol kinase*; *vte6*, *phytol phosphate kinase*. Pathway abbreviation: MEP, methylerythritol phosphate.

Like all vitamins in the human diet, vitamin E is required at recommended daily amounts to maintain optimal health (reviewed in Mene-Saffrane, 2017). Although clinical vitamin E deficiency is rare, affecting less than 1% of the US population (Centers for Disease Control, 2006), the prevalence of suboptimal dietary intake among individuals in the United States as measured by plasma α-tocopherol levels is surprisingly high, ranging from 43 to 87%, depending on age and ethnicity (Ford et al., 2006; McBurney et al., 2015). The substandard consumption of vitamin E in the diet has been associated with increased risk of cardiovascular diseases (Knekt et al., 1994; Kushi et al., 1996). Furthermore, limited evidence exists to support the association of vitamin E intake levels with other chronic diseases (Linus Pauling Institute, 2015).

Sweet corn is the third most abundantly consumed vegetable in the United States after tomato (*Solanum lycopersicum* L.) and potato (*Solanum tuberosum* L.) (USDA, 2018b) but the vitamin E level [100% α-tocopherol + 30% α-tocotrienol + 10% γ-tocopherol; Institute of Medicine (2000)] provided from a 100-g intake (one medium to large ear of sweet corn) is only about 2.2% of the recommended daily allowance (RDA) (Xie et al., 2017). Although the vitamin E content of sweet corn is lower than that of tomato (3.6% of the RDA for 100 g) but higher than that of potato (0.3% of the RDA for 100 g) (USDA, 2018a), considerable phenotypic diversity for tocopherols has been reported in a small panel of sweet corn lines and that diversity was highly stable across growing environments (Kurilich and Juvik, 1999; Ibrahim and Juvik, 2009). This suggests that natural variation for vitamin E could be harnessed in sweet corn breeding programs to help address vitamin E insufficiencies where this vegetable is frequently consumed.

The tocochromanol biosynthetic pathway has been fully elucidated and involves 36 enzymatic reactions that are conserved across plant species [reviewed in DellaPenna and Mène-Saffrané (2011); Fig. 1.1). The aromatic head group for all tocochromanols is homogentisic acid,

produced via the shikimate pathway, whereas the hydrophobic tail groups for tocochromanols are generated from isopentenyl pyrophosphate synthesized by the plastid-localized methylerithritol phosphate (MEP) pathway. Condensation of homogentisic acid with phytyl-diphosphate by homogentisate phytyltransferase (VTE2) or with geranylgeranyl diphosphate by homogentisate geranylgeranyltransferase (HGGT1) produces the committed precursors for tocopherols and tocotrienols, respectively, which, in turn, are methylated (VTE3 and VTE4) and cyclized (VTE1) in various sequences and combinations to yield the α, β, γ, and δ isoforms. Geranylgeranyl diphosphate for tocotrienol synthesis in maize endosperm is produced directly from isopentenyl pyrophosphate, whereas the phytyl-diphosphate used for tocopherol synthesis in the maize embryo is generated by an indirect route involving a chlorophyll-based cycle (Diepenbrock et al., 2017).

Through several genome-wide association studies (GWAS) of mature maize kernels over recent years, a number of genes responsible for natural variation in tocochromanols and vitamin E levels have been identified. Li et al. (2012) reported a strong association between *vte4* and α-tocopherol content in maize kernels, with deeper insights into this association provided by Lipka et al. (2013). Lipka et al. (2013) also demonstrated the much weaker association of *vte1*, *hggt1*, and an arogenate/prephenate dehydratase with grain tocotrienol levels. In the US maize nested association mapping (NAM) panel, eight genes among the 81 a priori genes in the genome that encode one of the 36 enzymatic reactions were identified to be associated with natural variation of tocochromanols in maize grain (Diepenbrock et al., 2017). Another six loci encoding novel activities also were identified, including, most notably, two protochlorophyllide reductases (*por1* and *por2*), which surprisingly explained the majority of tocopherol content variation in maize grain. Most recently, Wang et al. (2018) associated the genes involved in fatty acid biosynthesis,

protein import into the chloroplast, chlorophyll *b* degradation, and the regulation of chlorophyll biosynthesis with tocopherol grain traits.

These findings from GWAS of tocochromanols in maize grain at physiological maturity (dry kernel stage) serve as a starting point for identifying the genes responsible for quantitative variation of tocochromanols in developing kernels of fresh sweet corn. However, sweet corn constitutes a distinct subpopulation that has limited representation in maize diversity panels (Flint-Garcia et al., 2005; Romay et al., 2013). Indeed, Doebley et al. (1988) and Gerdes and Tracy (1994) suggested that sweet corn and dent corn represent two distinct breeding pools, with most sweet corn lines descended from three open-pollinated cultivars: 'Golden Bantam', 'Stowell's Evergreen', and 'Country Gentleman'. Therefore, the favorable alleles for increased tocochromanol content observed in dent corn studies may not be present at high frequency or even at all in the sweet corn germplasm pool. Additionally, Kurilich and Juvik (1999) and Xie et al. (2017) reported that tocopherols tend to increase as sweet corn kernels mature. Therefore, selecting for alleles of causal genes that are favorably expressed early in kernel development is critically important, especially given that fresh sweet corn is harvested around 18 to 21 d after pollination (DAP) (Jennings and McCombs, 1969).

All sweet corn has one or more mutations in the genes involved in the starch biosynthesis pathway that cause kernels to accumulate sugars in the endosperm in place of the starch accumulated in wild-type dent corn (Tracy, 1997). The homozygous *sh2* mutation results in the loss of an adenosine diphosphate-glucose pyrophosphorylase subunit and the accumulation of high levels of sucrose (Michaels and Andrew, 1986). The homozygous *su1* mutation disrupts a starch debranching enzyme that leads to higher levels of water-soluble phytoglycogen, along with increased sucrose, but at levels lower than *sh2* (Doehlert et al., 1993). Other mutations in

the starch pathway used singly or in combination for breeding sweet corn include but are not limited to *sugary enhancer1* (*se1*), *brittle2* (*bt2*), and *amylose-extender:dull:waxy* (*aeduwx*) (Hannah et al., 1993). Depending on the single mutation, pleiotropic effects related to compromised biosynthetic capacity can be exerted on other enzymes from the starch biosynthesis pathway [reviewed in Tetlow et al. (2004)]. Illustrative of an interaction with a key phytohormone, ethylene has been shown to impart pleiotropic effects on maize endosperm development of *sh2* lines (Young et al., 1997). Given the potential amplified connectedness between starch-deficient endosperm mutants and other biochemical pathways, the loci and alleles that control the levels of tocotrienols—the class of tocochromanols that are predominantly synthesized in the endosperm (Grams et al., 1970; Weber, 1987)—may differ in the high-sugar environment of sweet corn endosperm compared with those identified in studies of starch containing dent varieties.

In this study, we constructed a sweet corn association panel that captured the genetic diversity of temperate US breeding programs to dissect the genetic basis of natural variation for tocochromanol content in fresh kernels and develop genomic prediction models that could be used to enhance fresh sweet corn kernels for tocochromanol and vitamin E levels. We conducted (i) a GWAS to identify the genes involved in the genetic control of quantitative variation for tocochromanol levels in fresh (~21 DAP) kernels of sweet corn and (ii) genomic prediction studies to determine the optimal marker density needed to maximize predictive abilities for genomic selection in a sweet corn biofortification breeding program.

MATERIALS AND METHODS

Plant Materials and Experimental Design

We constructed an association panel of 411 diverse sweet corn inbred lines that were selected to sample the levels and patterns of genetic diversity found in the US sweet corn germplasm pool. The panel consisted of inbred lines homozygous for the starch-deficient endosperm mutations *su1*, *su1:se1*, *sh2*, *su1sh2*, *bt2*, and *aeduwx*. An additional 19 inbred lines included in the experiment were known at the time of inclusion or later confirmed (data not shown) not to be sweet corn. The inbred association panel was field-evaluated in the summers of 2014 and 2015 at Cornell University's Musgrave Research Farm in Aurora, NY. For each year, the panel was separated into three sets of varying numbers of lines based on plant height, and the sets were randomly partitioned into incomplete blocks. Within each set, each incomplete block of 20 experimental entries was augmented by the random assignment of two check plots depending on plant height. The incomplete blocks of sets 1 (short), 2 (medium), and 3 (tall) each included the two check lines, either 'We05407' and 'W5579', 'W5579' and 'Ia5125', or 'Ia5125' and 'IL125b', respectively. In addition, the positions of the sets within the field were randomized. Edge effects were reduced by planting a commercial sweet corn line around the perimeter of each replicate. Experimental units were one-row plots with a length of 3.05 m and inter-row spacing of 0.76 m. There was a 0.91-m alley at the end of each plot. In each plot, 24 kernels were planted and each plot was thinned to approximately 12 plants. Standard sweet corn cultivation practices for the Northeast US were followed. Weather data were obtained from an automated weather station (Spectrum Technologies, Inc., Aurora, IL) located within the field.

In both years, a single complete replication of the augmented incomplete block design experiment was used for measuring tocochromanol levels. In each plot, six plants were self-

pollinated by hand and the pollination dates were recorded. Two self-pollinated ears were hand-harvested from each plot at 400 growing degree-days (~21 DAP) as calculated via the NOAA 86/50 method (Barger, 1969), representing the immature milk stage of kernel development, when sweet corn is picked and eaten as a fresh vegetable. Immediately after harvest, whole ears were frozen in liquid N and shelled. For each sample, frozen kernels were randomly sampled and bulked across the two ears to produce a representative composite kernel sample, then placed in a cryogenic vial and maintained at −80°C. For each sample, 20 to 30 frozen kernels were ground to a fine powder in liquid N. Individual ground samples were transferred to a 1.5-mL tube cooled in liquid N, then transferred for storage at −80°C. Ground kernel samples were packed in dry ice and shipped to Michigan State University (East Lansing, MI) for extraction and measurement of tocochromanols.

Phenotypic Data Analysis

Tocochromanols were extracted from each ground sample, then quantified by high-performance liquid chromatography (HPLC) and fluorometry as previously described (Lipka et al., 2013), with 1 mg mL^{-1} of DL-α-tocopherol acetate added to the extraction buffer as an internal recovery control. The six quantified tocochromanol compounds were δ-tocotrienol (δT3), γ-tocotrienol (γT3), α-tocotrienol (αT3), δ-tocopherol (δT), γ-tocopherol (γT), and α-tocopherol (αT) in μg g^{-1} fresh kernel. Additionally, the following 14 sums, ratios, and proportions were calculated: total tocotrienols (total T3), total tocopherols (total T), total tocochromanols (total T3 + T), αT3/γT3, αT/γT, δT3/αT3, δT/αT, δT3/γT3, δT/γT, γT3/(γT3 + αT3), γT/(γT + αT), δT3/(γT3 + αT3), δT/(γT + αT), and total T/total T3.

To screen the raw HPLC data for significant outliers, we initially used the Box–Cox power transformation (Box and Cox, 1964) with a simple linear model with genotype, year, set

within year, block within set within year, and HPLC plate within year as fixed effects to identify the most appropriate transformation that corrected for unequal variances and the non-normality of error terms. The process was conducted using the MASS package in R version 3.2.3 (R Core Team, 2015) and tested lambda values ranging from –2 to +2 in increments of 0.5 before applying the optimal convenient lambda for each phenotype (Supplemental Table S1.1). Next, the full mixed linear model that allowed for the estimation of genetic effects separately from field design effects, following Wolfinger et al. (1997), was fitted for each phenotype in ASReml-R version 3.0 (Gilmour et al., 2009). The full model fitted was as follows:

$$Y_{ijklmnop} = \mu + check + year_j + set(year)_{jk} + block(set \times year)_{jkl} + genotype_m + genotype \times year_{jm} + plate(year)_{jn} + row(year)_{jo} + col(year)_{jp} + \varepsilon_{ijklmnop} \quad [1]$$

in which $Y_{ijklmnop}$ is an individual phenotypic observation, μ is the grand mean, $check_i$ is the fixed effect for the *i*th check, $year_j$ is the effect of the *j*th year, $set(year)_{jk}$ is the effect of the *k*th set within the *j*th year, $block(set \times year)_{jkl}$ is the effect of the *l*th incomplete block within the *k*th set within the *j*th year, $genotype_m$ is the effect of the *m*th experimental genotype (noncheck line), $genotype \times year_{jm}$ is the effect of the interaction between the *m*th genotype and *j*th year, $plate(year)_{jn}$ is the laboratory effect of the *n*th HPLC autosampler plate within the *j*th year, $row(year)_{jo}$ is the effect of the *o*th plot grid row within the *j*th year, $col(year)_{jp}$ is the effect of the *p*th plot grid column within the *j*th year, and $\varepsilon_{ijklmnop}$ is the residual error effect assumed to be independently and identically distributed according to a normal distribution with a mean of zero and the variance σ_ε^2. Except for the grand mean and check term, all other terms were modeled as random effects. Degrees of freedom were calculated via the Kenward–Rogers approximation

(Kenward and Roger, 1997). To detect significant outliers, Studentized deleted residuals (Neter et al., 1996) obtained from these mixed linear models were examined.

Once all outliers were removed for each tocochromanol phenotype, an iterative mixed linear model fitting procedure was conducted in ASReml-R version 3.0 (Gilmour et al., 2009) with the full model. Likelihood ratio tests were conducted to remove all terms from the model fitted as random effects that were not significant at $\alpha = 0.05$ (Littell et al., 2006) to generate a final, best fitted model for each phenotype (Supplemental Fig. S1.1). The final model for each tocochromanol phenotype was used to generate a best linear unbiased predictor (BLUP) for each genotype. The generated BLUPs were used in a GWAS and tocochromanol prediction models (Supplemental Table S1.2).

Variance component estimates from the reduced model were used to estimate heritability ($\hat{h}_l^2$) on a line-mean basis (Holland et al., 2003; Hung et al., 2012) for each tocochromanol phenotype, with the SE of the estimates calculated via the delta method (Lynch and Walsh, 1998; Holland et al., 2003). Pearson's *r* was used to estimate the degree of association between back-transformed BLUP values for each pair of tocochromanol traits at $\alpha = 0.05$ via the method 'pearson' from the function 'cor.test' in R. The back-transformed BLUP values were calculated with the inverse of the given convenient lambda and used to represent the true directionality of the relationship between traits (Supplemental Table S1.3).

DNA Extraction, Sequencing, and Genotyping

For each inbred line, a leaf tissue sample consisting of young leaves was collected from a single representative plant. The tissue samples were lyophilized and ground with a GenoGrinder (Spex SamplePrep, Metuchen, NJ). Total genomic DNA was isolated from powdered lyophilized leaf tissue with the DNeasy 96 Plant Kit (Qiagen Incorporated, Valencia, CA). The DNA

samples were sent for genotyping-by-sequencing (GBS) at the Cornell Biotechnology Resource Center (Cornell University, Ithaca, NY, USA) following the procedure of Elshire et al. (2011) with *Ape*KI as the restriction enzyme. Genotyping-by-sequencing libraries were constructed in 192- or 384-plex and sequenced on an NextSeq 500 or Illumina HiSeq 2500, respectively (Illumina Incorporated, San Diego, CA). Sequence data that supported the findings of this study have been deposited in the National Center of Biotechnology Information Sequence Read Archive under accession number SRP154923 and in BioProject under accession PRJNA482446.

With the raw GBS sequencing data, the genotypes at 955,690 high confidence single-nucleotide polymorphism (SNP) loci were called with the default parameters in the TASSEL 5 GBSv1 production pipeline with the ZeaGBSv2.7 Production TagsOnPhysicalMap file in B73 RefGen_v2 coordinates (AllZeaGBSv2.7_ProdTOPM_20130605.topm.h5, available at panzea.org, accessed 25 Sept. 2018) (Glaubitz et al., 2014). There were inbred lines from the sweet corn association panel that had also been included in the comprehensive genotyping study of the USDA-ARS North Central Regional Plant Introduction Station collection, conducted by Romay et al. (2013). Therefore, existing raw unimputed SNP genotypic data for 45 sweet corn lines (ZeaGBSv27_publicSamples_rawGenos_AGPv2–150114.h5, available at www.panzea.org, accessed 25 Sept. 2018) that had been phenotyped for tocochromanols in this study were used instead of generating a redundant GBS sequencing dataset. The SNP genotype calls from this study and those of the 45 lines from Romay et al. (2013) were combined and filtered to retain only biallelic SNPs with a call rate greater than 10% (i.e., the percentage of lines successfully genotyped per SNP), as specified by Romay et al. (2013). Missing SNP genotypes were imputed with FILLIN (Swarts et al., 2014) with an available set of maize haplotype donors that had a window size of 4 kb (AllZeaGBSv2.7impV5_AnonDonors4k.tar.gz, available at panzea.org,

accessed 25 Sept. 2018). This haplotype-based imputation method is not able to impute all missing data (Swarts et al., 2014) and thus some missing genotype data still remained and had to be filtered.

Upon completion of the imputation procedure, the inbred lines known to have *bt2* (n = 4) or *aeduwx* (n = 2) were removed from the dataset, allowing the panel to consist of the endosperm mutations most commonly found in the US sweet corn germplasm pool. In TASSEL version 5.2.39, additional quality filters imposed following haplotype-based imputation included removing SNPs with a call rate less than 70%, a minor allele frequency lower than 5%, heterozygosity greater than 10%, an inbreeding coefficient lower than 80%, or a mean read depth greater than 15. Additionally, lines with lower than a 40% call rate (i.e., the percentage of SNPs successfully genotyped for each line) were excluded. The imposition of these quality filters resulted in a final dataset of 174,996 high-quality SNPs scored on 384 lines that had a BLUP value for at least one tocochromanol trait. The raw unimputed SNP genotype calls for the 384 lines are available from the Dryad Digital Repository (https://doi.org/10.5061/dryad.jd5716f).

Genome-Wide Association Study

To conduct a GWAS for each phenotype that controlled for population structure and familial relatedness, a mixed linear model that included the population parameters previously determined approximation for enhanced computing efficiency (Zhang et al., 2010) was used to test for an association between the genotypes of each of the 174,996 SNPs and BLUP values in the R package GAPIT version 2017.08.18 (Lipka et al., 2012). The fitted mixed linear models included four principal components (PCs) (Price et al., 2006) and a kinship matrix based on VanRaden's Method 1 (VanRaden, 2008) that were calculated from a subset of 11,448 genome-wide SNP markers from the complete dataset that had not been imputed and had a call rate

higher than 90%, a minor allele frequency greater than 5%, heterozygosity less than 10%, an inbreeding coefficient greater than 80%, and a mean read depth lower than 15. Missing genotypes remaining for all SNP markers (subset and complete marker datasets) were conservatively imputed as a 'middle' (heterozygous) value in GAPIT. The optimal number of PCs to include as covariates in the mixed linear model was determined with the Bayesian information criterion (Schwarz, 1978). A likelihood-ratio-based R^2 statistic (Sun et al., 2010) denoted as R^2_{LR} was used to estimate the amount of phenotypic variation accounted for by the model. The method of Benjamini and Hochberg (1995) was used to account for multiple testing by controlling the false discovery rate (FDR) at 5%.

A chromosome-wide approach for implementing a multi-locus mixed-model (MLMM) (Segura et al., 2012) to resolve association signals involving large-effect genes has been previously described (Lipka et al., 2013). Briefly, the MLMM method used a stepwise mixed-model regression procedure with forward selection and backward elimination. In the first step of this chromosome-wide implementation, only SNP markers on the same chromosome with a major-effect gene were tested as explanatory variables for selection in the optimal model via the extended Bayesian information criterion (Chen and Chen, 2008). The impact of controlling for the influence of a large-effect gene on association signals was then assessed by reconducting the GWAS with MLMM-selected SNP markers included as covariates in mixed linear models.

Linkage Disequilibrium

The squared allele-frequency correlation (r^2) method of Hill and Weir (1988) was used to estimate linkage disequilibrium (LD) between pairs of SNP loci in TASSEL version 5.2.39 (Bradbury et al., 2007). The dataset of 174,996 high-quality SNP markers was used to estimate

LD, with the exception that the remaining missing SNP genotypes were not imputed with the 'middle' value prior to LD analysis.

Visual Classification of Lines for Endosperm Mutations

The sweet corn lines were evaluated for which endosperm mutations they possessed to help us better understand the differences in the content and composition of tocochromanols in fresh sweet corn kernels among the endosperm mutation group types. In 2014, two self-pollinated ears per plot were harvested at physiological maturity and dried to ~15% moisture content. For each plot, an image of two mature ears on the 1KK green background (https://wheatgenetics.org/download/category/21–1kk, accessed 25 Sept. 2018) was taken by hand with a digital camera (Sony DSC-W730, Sony Corporation, Tokyo, Japan). Of the 384 sweet corn inbred lines, 333 of them had images that allowed for visual classification of endosperm mutation type. To classify the inbred lines as having either the recessive *su1* or *sh2* endosperm mutation, the ears in each image were visually scored by one person (Matheus Baseggio) as having kernels with either one of two phenotypes: (i) wrinkled and glassy (*su1* mutation) or (ii) shrunken and opaque to translucent (*sh2* mutation) (Boyer and Shannon, 1983). Given the more visually complex kernel phenotypes that can result from double mutant combinations of endosperm genes (Boyer and Shannon, 1983), it was only possible to confidently score *su1sh2* inbreds as *sh2* and *su1:se1* inbreds as *su1*.

Marker-Based Classification of Lines for Endosperm Mutations

The visual classifications resulted in two binomial phenotypes (presence or absence of *su1*; presence or absence of *sh2*) that were used to train genomic prediction models separately for classifying the remaining 51 inbred lines. The 333 inbred lines with visual kernel scores for the presence or absence of *sh2* were randomly divided into two groups: 284 lines (85%) used as a set

for training and cross-validating the genomic prediction models and 49 lines (15%) comprising a test set to assess the error of the final selected model. An 85:15 split was also used for modeling the presence or absence of *su1*, with the exception that 15 previously known *su1sh2* inbreds were excluded. This resulted in 271 lines (85%) for training and cross-validating the genomic prediction models and 47 lines comprising the test set for the *su1* locus. Given that the implemented GBS approach did not target specific loci, statistical models were evaluated for their accuracy in predicting the presence or absence of the *su1* or *sh2* endosperm mutations with the following variable sized marker datasets: SNP markers ± 100, 250, 500, 750, or 1000 kb of *su1* (chromosome 4; 41,369,510–41,378,299 bp) or *sh2* (chromosome 3; 216,414,684–216,424,048 bp). The marker datasets consisted of markers selected from the same 174,996 high-quality SNP markers that were also used for GWAS. Prediction of the binomial phenotypes with SNP markers was evaluated with genomic best linear unbiased prediction (GBLUP) (Zhang et al., 2007; VanRaden, 2008). To conduct the GBLUP method, a realized relationship matrix based on VanRaden's Method 1 (VanRaden, 2008) calculated from SNP markers was fitted with the binomial family and logit link function in ASReml-R version 3.0 (Gilmour et al., 2009). For each locus, the realized relationship matrices were derived from the five sets of the variable-sized marker datasets.

The probability of a homozygous recessive genotype for each locus was obtained from the cumulative distribution function of the logistic distribution using the fivefold cross-validation approach reported in Owens et al. (2014). The average probability for each kernel phenotype was assessed by repeating this process 50 times, with inbred lines classified as having a homozygous recessive genotype for a given locus if the average probability was greater than 0.5. Sensitivity (the proportion of true positives) and specificity (the proportion of true negatives) were

calculated for each model, and the SNP marker dataset that maximized the sum of sensitivity and specificity was used to train the GBLUP model with all lines from the training set. The error rate of both final optimal models, which used the 1000-kb marker dataset for each locus, was then assessed on the test set (Supplemental Table S1.4). Next, the same two final models were used to predict the presence or absence of the *su1* and *sh2* endosperm mutations for the 51 uncharacterized lines and the previously known 15 *su1sh2* inbreds that had been excluded from the *su1* training and test sets.

Tocochromanol Prediction

The ability of SNP markers to predict each of the 20 tocochromanol phenotypes from the 384 inbred lines was evaluated with GBLUP (Zhang et al., 2007; VanRaden, 2008). The GBLUP method was conducted by calculating a realized relationship matrix based on VanRaden's Method 1 (VanRaden, 2008) from SNP markers, followed by model fitting in ASReml-R version 3.0 (Gilmour et al., 2009). The realized relationship matrices were derived from three different sets of SNPs that varied in marker number: genome-wide, pathway-level, and tocochromanol quantitative trait locus (QTL)-targeted. The genome-wide dataset included the 174,996 high-quality SNP markers, whereas the pathway-level dataset consisted of 4819 SNP markers within ± 250 kb of the 81 a priori candidate genes based on prior knowledge of the tocochromanol pathway and its precursors and on homology with *Arabidopsis thaliana* (L.) Heynh. (Diepenbrock et al., 2017; Supplemental Table S1.5). The tocochromanol QTL-targeted dataset included 946 SNP markers within ± 250 kb of the 14 a priori identified genes (Supplemental Table S1.6) underlying joint-linkage (JL) QTL associated with grain tocochromanol levels in the US maize NAM panel (Diepenbrock et al., 2017). The fivefold cross-validation approach described in Owens et al. (2014) was used to assess the predictive ability obtained for each

phenotype by assessing the Pearson's correlation between observed and genomic estimated breeding values. This process was repeated 50 times for each phenotype, with the mean of these correlations reported as the predictive ability. The same cross-validation folds allowed for a direct comparison among models. Additionally, the stratified sampling approach enabled each fold to be representative of the genotype frequencies for endosperm mutants (*su1*, *sh2*, and *su1sh2*) observed in the entire population. All prediction analyses were performed with and without a covariate accounting for the type of endosperm mutation (*su1*, *sh2*, or *su1sh2*).

RESULTS

Phenotypic Variation

Phenotypic variation for tocochromanol traits was evaluated in an association panel that was constructed to comprehensively represent the genetic diversity that exists in temperate US sweet corn breeding programs. The measurement of six tocochromanol compounds by HPLC in kernels sampled at the immature milk stage from 384 inbred lines revealed γ-species to be the most abundant, followed by α- and δ-species for both tocopherols and tocotrienols (Table 1.1). On average, the amount of γT3 moderately exceeded the sum total of all three tocopherol species in the sweet corn population. When lines were separated by the presence or absence of endosperm mutations through the combination of visual and marker-based classifications (Supplemental Table S1.2 and Supplemental Table S1.4), the average quantities of γT3 and δT3 were at significantly ($P < 0.0001$) greater levels in the *sh2* ($n = 76$) and *su1sh2* ($n = 19$) groups than in the *su1* ($n = 289$) group (Table 1.2). Interestingly, none of the other four tocochromanol compounds showed a similar pattern, with the exception that *sh2* had a significantly greater level of αT3 relative to the *su1* and *su1sh2* groups.

Table 1.1. Means and ranges for back-transformed best linear unbiased predictors (BLUPs) of 20 fresh kernel tocochromanol traits evaluated in the sweet corn association panel and estimated heritability ($\hat{h}_l^2$) on a line-mean basis across 2 yr.

		BLUPs			Heritabilities	
Trait[†]	**Lines**	**Mean**	**SD**[‡]	**Range**	**Estimate**	**SE**[§]
		——————μg g^{-1} fresh weight——————				
αT	383	1.53	0.70	0.33–4.72	0.87	0.01
αT3	384	1.91	0.43	1.04–3.99	0.68	0.03
δT	383	0.34	0.20	0.04–1.51	0.82	0.02
δT3	384	0.50	0.49	0.07–3.35	0.89	0.01
γT	383	8.67	2.65	1.82–17.4	0.78	0.02
γT3	384	10.41	4.92	2.11–36.96	0.84	0.02
Total T	383	9.61	2.80	2.21–23.48	0.79	0.02
Total T3	384	12.60	4.95	3.48–40.16	0.81	0.02
Total T + T3	383	22.67	6.00	8.34–49.67	0.81	0.02
αT/γT	383	0.20	0.13	0.034–0.98	0.87	0.01
δT/αT	384	0.28	0.24	0.015–1.63	0.85	0.02
δT/γT	384	0.04	0.02	0.015–0.18	0.88	0.01
δT/(γT + αT)	384	0.04	0.02	0.007–0.16	0.86	0.01
γT/(γT + αT)	383	0.84	0.09	0.513–0.97	0.86	0.01
αT3/γT3	384	0.23	0.13	0.061–0.95	0.83	0.02
δT3/αT3	383	0.27	0.25	0.030–1.59	0.87	0.01
δT3/γT3	383	0.04	0.02	0.012–0.18	0.87	0.01
δT3/(γT3 + αT3)	383	0.04	0.02	0.010–0.15	0.88	0.01
γT3/(γT3 + αT3)	384	0.82	0.07	0.528–0.94	0.82	0.02
Total T/Total T3	384	0.86	0.40	0.211–3.09	0.81	0.02

† αT, α-tocopherol; αT3, α-tocotrienol; δT, δ-tocopherol; δT3, δ-tocotrienol; γT, γ-tocopherol; γT3, γ-tocotrienol; Total T3, total tocotrienols; Total T, total tocopherols; Total T + T3, total tocochromanols.

‡ Standard deviation of the BLUPs.

§ Standard error of the heritabilities.

Table 1.2. Back-transformed estimated effects of endosperm mutation type for 20 fresh sweet corn kernel tocochromanol traits.

Trait†	*su1*	*sh2*	*su1sh2*	*P*-value‡
	——————μg g^{-1} fresh weight——————			
αT	1.36	1.39	1.48	0.763
αT3	1.84 b	1.98 a	1.70 b	**0.006**
δT	0.30	0.27	0.32	0.407
δT3	0.29 b§	0.72 a	0.69 a	**<0.0001**
γT	8.59	8.08	8.13	0.276
γT3	8.53 b	14.77 a	13.89 a	**<0.0001**
Total T	9.32	8.83	9.08	0.374
Total T3	10.52 b	16.58 a	15.68 a	**<0.0001**
Total T + T3	20.70 b	26.22 a	25.32 a	**<0.0001**
αT/γT	0.16	0.17	0.18	0.464
δT/αT	0.21	0.20	0.22	0.728
δT/γT	0.04	0.04	0.05	0.116
δT/(γT + αT)	0.03	0.03	0.04	0.287
γT/(γT + αT)	0.84	0.84	0.82	0.495
αT3/γT3	0.23 b	0.15 a	0.13 a	**<0.0001**
δT3/αT3	0.16 b	0.35 a	0.41 a	**<0.0001**
δT3/γT3	0.04 b	0.05 a	0.05 a	**<0.0001**
δT3/(γT3 + αT3)	0.03 b	0.04 a	0.04 a	**<0.0001**
γT3/(γT3 + αT3)	0.80 b	0.86 a	0.87 a	**<0.0001**
Total T/Total T3	0.89 b	0.53 a	0.58 a	**<0.0001**

† αT, α-tocopherol; αT3, α-tocotrienol; δT, δ-tocopherol; δT3, δ-tocotrienol; γT, γ-tocopherol; γT3, γ-tocotrienol; Total T3, total tocotrienols; Total T, total tocopherols; Total T + T3, total tocochromanols.

‡ *P*-value from a one-way ANOVA *F*-test for the endosperm mutation type effect. A bolded *P*-value indicates a statistically significant difference between two or more endosperm mutation type groups ($P < 0.05$).

§ Sweet corn lines grouped by endosperm mutation type having labels with the same letter are not significantly different according to the Tukey–Kramer honest significant difference test ($P < 0.05$). The test was only performed for traits that had a significant *F*-test.

Even though the six compounds are the product of a shared biosynthetic pathway, correlations between compounds only exceeded ~0.20 for two pairs of compounds (Supplemental Fig. S1.2). The correlation between δT3 and γT3 was 0.77, whereas the

correlation of δT with γT was 0.76. The nonexistent to weak correlations between other compound pairs imply the lack of a deeply shared genetic architecture, reflected by tocotrienols being synthesized predominantly in endosperm and tocopherols predominantly in embryo. As shown in Table 1.1, the six tocochromanol compounds and 14 sum, ratio, and proportion traits had an average heritability of 0.83, with estimates ranging from 0.68 (αT3) to 0.89 (δT3). These high heritabilities suggest that the extent of phenotypic variation for the tocochromanol traits is mostly under genetic control and would be responsive to selection in breeding programs.

Genome-wide Association Study

The genetic architecture of natural variation for tocochromanols in kernels harvested at the immature milk stage was dissected in an association panel of 384 sweet corn inbred lines scored with 174,996 high-quality SNP markers at genome-wide coverage. Even though sweet corn is a distinct subpopulation of maize (Romay et al., 2013), moderately weak patterns of population structure appeared to be defined by starch-deficient endosperm mutations within the association panel as inferred by a principal component analysis of the SNP genotypic data (Fig. 1.2). In the sweet corn association panel, the median LD (50th percentile) estimated with the 174,996 genome-wide SNP markers decayed to low (background) levels ($r^2 < 0.1$) by ~12 kb, but with a large variance in LD structure (Supplemental Fig. S1.3). Given the persistence of LD at higher percentile cutoffs, the potential importance of distant regulatory elements in the genetic control of kernel tocochromanols (Li et al., 2012), and to intersect with the GWAS results of Lipka et al. (2013), the candidate gene search space was limited to ± 250 kb (median $r^2 \leq 0.05$) of GWAS-detected SNP markers. Through implementation of a mixed linear model that controlled for population stratification and unequal relatedness, 336 unique SNPs were found to significantly associate with one or more phenotypes at a genome-wide FDR of 5%

(Supplemental Table S1.7 and Supplemental Fig. S1.4). At least one significant SNP was found on every maize chromosome, with the exception of chromosome 10, but 87.2% of the 336 unique SNPs were localized to chromosomes 2, 3, 4, and 5.

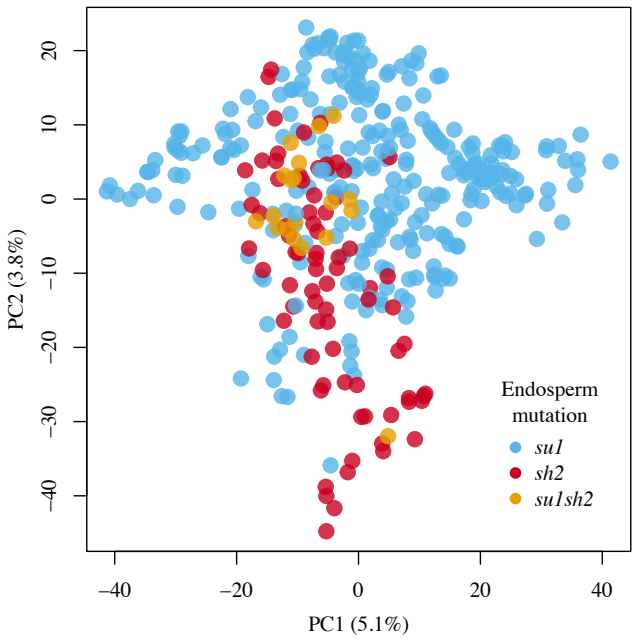

Figure 1.2. Principal component analysis of the sweet corn diversity panel. Genetic differentiation of 384 sweet corn inbred lines as revealed by the first two principal components from a principal component analysis of the single nucleotide polymorphism marker data.

The only significant association signals detected for αT were on chromosome 5 (Fig. 1.3A), with the signal peak defined by two SNPs (S5_200369243 and S5_200369213; *P*-values 2.40×10^{-8} and 3.10×10^{-8}, respectively). The two SNPs were separated by a distance of 30 bp, in perfect LD with each other, and located within an intron of the gene encoding γ-tocopherol methyltransferase (*vte4*, GRMZM2G035213), which catalyzes the conversion of γT to αT (Fig. 1.1). An additional significant SNP positioned more than 250 kb away from the start of the open reading frame for the *vte4* gene and in linkage equilibrium ($r^2 < 0.01$) with the other two significant intronic *vte4* SNPs was associated with αT and the αT/γT ratio (S5_200111824, *P*-values 5.38×10^{-7} and 1.65×10^{-7}, respectively) (Supplemental Fig. S1.5).

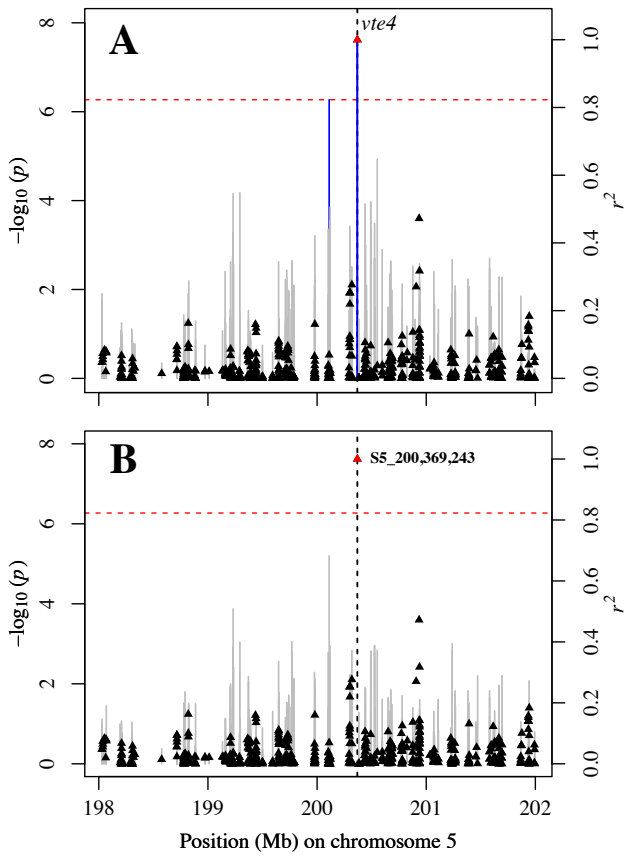

Figure 1.3. Genome-wide association study for α- tocopherol (αT) content in fresh kernels of sweet corn. (A) Scatter plot of association results from a mixed model analysis and linkage disequilibrium (LD) estimates (r^2). The vertical lines are $-\log_{10}$ *P*-values of single nucleotide polymorphisms (SNPs) and blue color represents SNPs that are statistically significant at a 5% false discovery rate (FDR). Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 200,369,243 bp (B73 RefGen_v2) on chromosome 5. The red horizontal dashed line indicates the $-\log_{10}$ *P*-value of the least statistically significant SNP at a 5% FDR. The black vertical dashed line indicates the genomic position of the *γ-tocopherol methyltransferase* gene (*vte4*). (B) Scatter plot of association results from a conditional mixed linear model analysis and LD estimates (r^2). The peak SNP (S5_200369243) from the optimal multilocus mixed-model was included as a covariate in the mixed linear model to control for the *vte4* effect.

When a chromosome-wide MLMM was used to clarify the association signals in this genomic interval better, the optimal model obtained by the MLMM for αT contained the peak SNP S5_200369243, whereas the peak SNP S5_200111824 was selected in the optimal model for the αT/γT ratio (Supplemental Table S1.8). For each of these two tocochromanol traits, there were no longer any significant associations at a 5% FDR when GWAS was conducted with their

MLMM-selected SNP as a covariate in the mixed linear model (Fig. 1.3B). Notably, the allele of the peak SNP S5_200369243 associated with higher levels of αT was fixed in *su1sh2* lines and all but two *sh2* lines (Supplemental Table S1.9). The results from this conditional analysis suggest that *vte4* and, potentially, a distant regulatory element are responsible for variation in the αT-related traits.

Within the pericentromeric region of chromosome 5, significant associations were identified between 40 SNPs that spanned a 3.3-Mb interval and at least one of the two tocotrienol-related traits δT3/(γT3 + αT3) and δT3/γT3 (Fig. 1.4A). The most significant SNP for δT3/γT3 was S5_131738084 (*P*-value 7.64×10^{-9}); S5_133512770 (*P*-value 3.55×10^{-8}) was the peak SNP for δT3/(γT3 + αT3). The latter of the two SNPs was located within an intron of the gene encoding tocopherol cyclase (*vte1*, GRMZM2G009785), an enzyme that catalyzes the synthesis of both γT3 and δT3. Two additional significant SNPs associated with both δT3/γT3 and δT3/(γT3 + αT3) (S5_133505829, *P*-values 5.58×10^{-8} and 1.67×10^{-7}; S5_133501992, *P*-values 6.83×10^{-8} and 1.94×10^{-7}, respectively) were also located within the *vte1* gene. To further resolve signals in the recombination-suppressed *vte1* region, all SNPs on chromosome 5 were considered when conducting the MLMM approach for the two tocotrienol traits. Only the peak SNP was selected in the optimal model for each trait. When each peak SNP was fitted separately as a covariate in the mixed linear model, all other significant associations in the vicinity of *vte1* were eliminated for δT3/(γT3 + αT3) and δT3/γT3 (Fig. 1.4B and Supplemental Fig. S1.6).

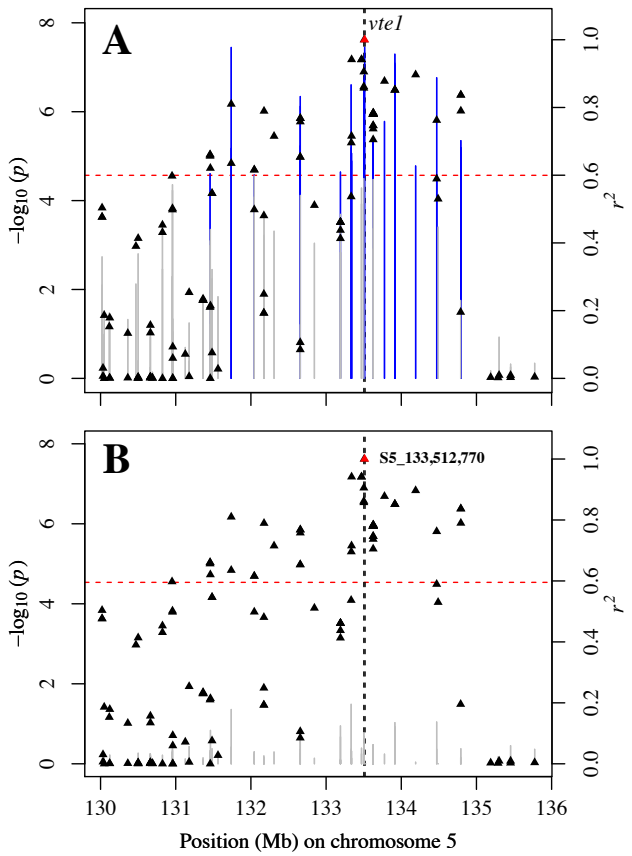

Figure 1.4. Genome-wide association study for the ratio of δ-tocotrienol (δT3) to the sum of γ-tocotrienol (γT3) and αT3 [δT3/(γT3 + αT3)] in fresh kernels of sweet corn. (A) Scatter plot of association results from a mixed linear model analysis and linkage disequilibrium (LD) estimates (r^2). The vertical lines are $-\log_{10}$ *P*-values of single nucleotide polymorphisms (SNPs) and blue color represents SNPs that are statistically significant at a 5% false discovery rate (FDR). Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 133,512,770 bp (B73 RefGen_v2) on chromosome 5. The red horizontal dashed line indicates the $-\log_{10}$ *P*-value of the least statistically significant SNP at a 5% FDR. The black vertical dashed line indicates the genomic position of the *tocopherol cyclase* gene (*vte1*). (B) Scatter plot of association results from a conditional mixed linear model analysis and LD estimates (r^2). The peak SNP (S5_133512770) from the optimal multilocus mixed-model was included as a covariate in the mixed linear model to control for the *vte1* effect.

Signals of association were also identified for tocotrienol traits in the pericentromeric region of chromosome 9. The start of the open reading frame for a gene encoding a homogentisate geranylgeranyltransferase (*hggt1*, GRMZM2G173358), the committed step in tocotrienol biosynthesis (Cahoon et al., 2003), was 138 kb downstream from a SNP

(S9_92345469, *P*-value 4.47×10^{-5}; Supplemental Fig. S1.7) on chromosome 9 that was significantly associated with γT3, the most abundant tocotrienol. Additionally, one SNP each was found to be significantly associated with total T3 (S9_91476108, *P*-value 4.28×10^{-5}) and the total T/total T3 ratio (S9_90663281, *P*-value 2.14×10^{-5}). Indicative of the long-range LD in this region, these two SNPs were in strong (S9_91476108, $r^2 = 0.74$) to moderate (S9_90663281, $r^2 = 0.34$) LD with SNP S9_92345469, although both SNPs were more than 1 Mb away from *hggt1*. Neither these nor any other SNPs were selected by the MLMM for the three tocotrienol traits. This was not unexpected, given that these were among the relatively weaker significant associations for tocochromanol traits in the sweet corn association panel.

Extensive association signals were identified for multiple tocotrienol traits that colocalized with two genes encoding enzymes involved in kernel starch biosynthesis, *Sh2* (adenosine diphosphate glucose pyrophosphorylase, large subunit; GRMZM2G429899) and *Su1* (isoamylase-type starch-debranching enzyme 1, GRMZM2G138060). Recessive mutations for either of these genes inhibit starch formation and increase sugar levels in the endosperm (Creech, 1965), the tissue in which tocotrienols are synthesized (Grams et al., 1970; Weber, 1987). On chromosome 3, 38 SNPs spanning a ~4 Mb interval that included *sh2* were significantly associated with at least one of eight tocotrienol-related traits (Fig. 1.5A). Similarly, 179 SNPs that covered a 10.2-Mb region encompassing *su1* on chromosome 4 were found to be significantly associated with one or more of seven tocotrienol-related traits (Fig. 1.5B). In concordance with the findings of Wilson et al. (2004), long-range patterns of LD defined both the *sh2* and *su1* genomic regions (Fig. 1.5A and B), thus limiting the mapping resolution. Of these 217 total SNPs from chromosomes 3 and 4, all but one of them were also significantly

associated with the type of endosperm mutation when used as a phenotype (*su1*, *sh2*, or *su1sh2*) for the 384 lines in GWAS (Supplemental Fig. S1.8).

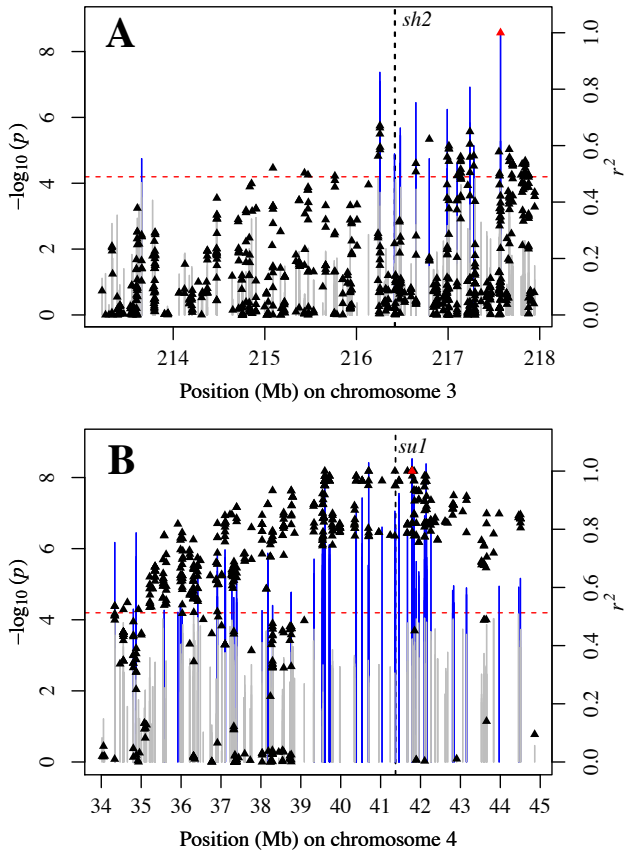

Figure 1.5. Genome-wide association study for δ-tocotrienol (δT3) content in fresh kernels of sweet corn. (A) Scatter plot of association results from a mixed linear model analysis and linkage disequilibrium (LD) estimates (r^2). The vertical lines are -$\log_{10}$ *P*-values of single nucleotide polymorphisms (SNPs) and blue color represents SNPs that are statistically significant at a 5% false discovery rate (FDR). Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 217,572,130 bp (B73 RefGen_v2) on chromosome 3. The red horizontal dashed line indicates the -$\log_{10}$ *P*-value of the least statistically significant SNP at a 5% FDR. The black vertical dashed line indicates the genomic position of the gene *shrunken2* (*sh2*). (B) Scatter plot of association results from a mixed linear model analysis and LD estimates (r^2). The vertical lines are -$\log_{10}$ *P*-values of SNPs and blue color represents SNPs that are statistically significant at a 5% FDR. Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 41,789,076 bp (B73 RefGen_v2) on chromosome 4. The black vertical dashed line indicates the position of the gene *sugary1* (*su1*).

Given the strong diffuse signals of association at *su1* and *sh2* and their potential contribution to distant complex associations (interchromosomal LD of $r^2 = 0.61$ between the

peak SNPs S3_216256039 and S4_41789076 at *sh2* and *su1*, respectively), GWAS was reconducted for all 20 tocochromanol traits with the type of endosperm mutation (*su1*, *sh2*, or *su1sh2*) as a covariate in the mixed linear model. When we controlled for association signals at *su1* and *sh2*, the significant association of SNPs (*P*-values 3.37×10^{-10} to 1.43×10^{-5}) encompassing *vte1* with δT3/(γT3 + αT3) and δT3/γT3 still remained (Supplemental Fig. S1.9 and Supplemental Table S1.10). Additionally, the two SNPs within *vte4* (S5_200369243 and S5_200369213) were still found to be significantly associated with αT (*P*-value 2.21×10^{-8} and 2.81×10^{-8}, respectively). In contrast, SNP S9_92345469 (*P*-value 6.34×10^{-5}) within 200 kb of *hggt1*, which was associated with γT3 when not controlling for endosperm mutation type was no longer significant at 5% FDR (Supplemental Fig. S1.9). Interestingly, the alleles of this SNP were not equally distributed between *su1* and *sh2* lines, such that all but one of the 76 *sh2* lines were fixed for the SNP allele associated with higher levels of γT3 (Supplemental Table S1.9).

To further clarify the association signal at *vte1* and of additional loci potentially masked by the endosperm effect, an MLMM analysis with endosperm mutation type as a covariate was conducted on a chromosome-wide level for tocotrienol-related traits. The optimal models obtained for δT3/γT3 and δT3/(γT3 + αT3) both included the SNP S5_131738084 (Supplemental Table S1.8), which was also the MLMM-selected peak SNP for δT3/γT3 when not controlling for endosperm mutation type. Although 1.76 Mb from *vte1*, this SNP was in high LD with four SNPs contained in *vte1* (r^2 = 0.60–0.81) that were significantly associated with δT3/γT3 and δT3/(γT3 + αT3) when controlling for endosperm mutation type. Additionally, the MLMM analysis resulted in the selection of the same single SNP (S5_214707875) for αT3/γT3 and γT3/(γT3 + αT3), representing a novel association for these two tocotrienol traits on

chromosome 5 (Supplemental Table S1.8). This SNP is within a gene encoding a zinc finger family protein (GRMZM2G178038).

With endosperm mutation type and the two MLMM-identified SNPs (S5_131738084 and S5_214707875) as covariates in the mixed linear model, GWAS was reconducted for all tocochromanol traits (Supplemental Table S1.11). It was found that one or more tocotrienol traits were significantly associated with a total of seven SNPs at 5% FDR (Supplemental Table S1.11). On chromosome 1, two SNPs (S1_279565998 and S1_279566000) in perfect LD and located within a gene encoding a peroxidase superfamily protein (GRMZM2G047456) were significantly associated (*P*-values 3.91×10^{-7} and 5.42×10^{-7}) with δT3/γT3. These two SNPs also had a slightly weaker association (FDR-adjusted *P*-value of 0.06) with δT3/(γT3 + αT3). Significant associations were also detected between total T3 and five SNPs (*P*-values 1.24×10^{-7}–1.19×10^{-6}) contained within (two SNPs) or ~225 to 590 kb away (three SNPs) from a gene that encoded an abscisic acid or stress-induced protein (HVA22, GRMZM2G311011) on chromosome 2. Additionally, one of these five SNPs was significantly associated with δT3 at the 5% FDR level. Although it is tempting to speculate on the biological involvement of these three associated genes, higher mapping resolution in combination with gene expression profiling and mutagenesis approaches are needed to assess the potential contribution of these identified novel loci to the genetic basis of tocotrienol traits more completely.

Prediction of Tocochromanols

The promise of genomic selection as an approach for the genetic improvement of fresh kernels for levels of tocochromanols and vitamin E in sweet corn breeding populations was evaluated. The predictive ability of whole-genome prediction (WGP) was assessed for all 20 tocochromanol phenotypes from the 384 inbred lines with the genome-wide dataset of 174,996

SNP markers. This analysis revealed a predictive ability of 0.49 averaged across the 20 phenotypes, with abilities ranging from 0.30 for γT/(γT + αT) to 0.68 for δT3/αT3 (Table 1.3). When all traits were considered, the correlation between heritabilities and predictive abilities was not statistically significant at a level of α = 5% (r = 0.18; P-value = 0.44). On average, tocotrienol-related traits had a higher predictive ability (average = 0.59) than tocopherol-related traits (average = 0.38). There was a strong, positive correlation (r = 0.65, P-value < 0.01) between the number of significant markers observed in GWAS at 5% FDR and predictive abilities (Table 1.3), which could partly account for the difference in predictive abilities between tocopherol and tocotrienol phenotypes.

Table 1.3. Predictive abilities of genomic prediction models using three marker sets as predictors and significant marker associations for 20 fresh sweet corn kernel tocochromanol traits.

Trait	GBLUP						GBLUP with endosperm mutation type covariate						Significant marker–trait associations[¶]
	Genome-wide[†]		Pathway-level[‡]		QTL targeted§		Genome-wide		Pathway-level		QTL targeted		
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
αT[#]	0.38	0.02	0.37	0.02	0.36	0.02	0.38	0.02	0.37	0.02	0.36	0.02	3
δT	0.40	0.03	0.34	0.02	0.27	0.02	0.39	0.03	0.33	0.02	0.26	0.03	0
γT	0.39	0.02	0.36	0.02	0.30	0.02	0.40	0.02	0.36	0.02	0.30	0.02	0
Total T	0.42	0.01	0.39	0.02	0.33	0.02	0.43	0.01	0.39	0.02	0.33	0.02	0
αT/γT	0.32	0.03	0.32	0.02	0.31	0.02	0.32	0.03	0.31	0.02	0.30	0.02	1
δT/αT	0.35	0.03	0.30	0.02	0.29	0.02	0.34	0.03	0.28	0.02	0.27	0.02	0
δT/γT	0.45	0.03	0.35	0.02	0.28	0.02	0.44	0.03	0.34	0.02	0.27	0.03	0
δT/(γT + αT)	0.41	0.03	0.30	0.02	0.24	0.02	0.40	0.04	0.29	0.02	0.23	0.03	0
γT/(γT + αT)	0.30	0.03	0.29	0.02	0.29	0.02	0.29	0.03	0.28	0.02	0.28	0.02	0
Average T	0.38		0.34		0.30		0.38		0.33		0.29		0.44
αT3	0.44	0.02	0.40	0.02	0.31	0.02	0.44	0.02	0.40	0.02	0.31	0.02	0
δT3	0.65	0.01	0.53	0.02	0.41	0.02	0.70	0.01	0.66	0.01	0.59	0.01	230
γT3	0.62	0.01	0.53	0.02	0.46	0.02	0.67	0.01	0.64	0.01	0.62	0.01	174
Total T3	0.61	0.01	0.52	0.02	0.44	0.02	0.65	0.01	0.63	0.01	0.59	0.01	165
αT3/γT3	0.59	0.01	0.49	0.02	0.46	0.02	0.61	0.01	0.55	0.01	0.56	0.01	5
δT3/αT3	0.68	0.01	0.55	0.02	0.46	0.02	0.71	0.01	0.64	0.01	0.61	0.01	83
δT3/γT3	0.57	0.01	0.47	0.02	0.27	0.03	0.61	0.01	0.55	0.02	0.41	0.02	0
δT3/(γT3 + αT3)	0.62	0.01	0.51	0.02	0.33	0.02	0.66	0.01	0.60	0.01	0.48	0.02	95
γT3/(γT3 + αT3)	0.57	0.01	0.49	0.02	0.45	0.02	0.59	0.01	0.53	0.02	0.54	0.02	2
Average T3	0.59		0.50		0.40		0.63		0.58		0.52		83.78
Total T3 + T	0.55	0.01	0.49	0.02	0.43	0.02	0.57	0.01	0.55	0.01	0.52	0.01	0
Total T/Total T3	0.49	0.02	0.38	0.02	0.28	0.02	0.56	0.01	0.52	0.01	0.49	0.01	120
Overall average	0.49		0.42		0.35		0.51		0.49		0.49		60.00

† 174,996 genome-wide markers.

‡ 4819 markers within ± 250 kb of 81 a priori candidate genes.

§ 946 markers within ± 250 kb of 14 a priori genes underlying joint-linkage quantitative trait loci (QTL) associated with grain tocochromanol levels in the US maize nested association mapping panel.

¶ The number of significant marker associations for each trait in a genome-wide association study without covariates at a genome-wide false discovery rate of 5%.

αT, α-tocopherol; αT3, α-tocotrienol; δT, δ-tocopherol; δT3, δ-tocotrienol; γT, γ-tocopherol; γT3, γ-tocotrienol; Total T3, total tocotrienols; Total T, total tocopherols; Total T + T3, total tocochromanols; GBLUP, genomic best linear unbiased prediction.

Given the oligogenic nature of tocochromanol grain traits in maize, where most of the phenotypic variation is explained by a few moderate- to large-effect loci associated with biosynthetic pathways (Diepenbrock et al., 2017), pathway-level and tocochromanol QTL-targeted marker datasets were also evaluated for their predictive abilities. These two marker datasets included SNPs in proximity of either 81 a priori candidate genes from the precursor and core tocochromanol pathways in maize (Supplemental Table S1.5) or 14 genes underlying the QTL responsible for 56 to 93% of tocochromanol variation in maize grain (Supplemental Table S1.6; Diepenbrock et al., 2017). On average, the predictive abilities of both the pathway-level (0.42) and tocochromanol QTL-targeted (0.35) marker datasets for the 20 tocochromanol phenotypes were lower than that obtained with the genome-wide marker dataset (0.49; Table 1.3). Specifically, tocotrienol-related traits showed the highest accuracy reduction, with an average decrease of 9 percentage points (81 candidate gene set) and 19 percentage points (14 QTL gene set) relative to the genome-wide marker dataset. In contrast, there was an average decrease of only 4 to 8 percentage points for tocopherol-related traits across the two loci-focused marker sets, suggesting that the inclusion of additional genes is more critically needed to predict tocotrienol levels in sweet corn accurately.

In an effort to improve predictive ability, the endosperm mutation type (*su1*, *sh2*, or *su1sh2*) was assessed as a covariate in prediction models evaluating the marker datasets with three different levels of genome coverage. With the genome-wide marker dataset, the inclusion of the endosperm mutation type covariate improved predictive ability by 5 percentage points for both γT3 and δT3, the only two compounds that had a significant association with SNPs within and nearby *su1* and *sh2* (Supplemental Table S1.7). When the same covariate was included in prediction models with the pathway-level and tocochromanol QTL-targeted marker datasets,

there were similar improvements in accuracy for γT3 and δT3, but the increase in predictive abilities was higher and ranged from 11 to 18 percentage points across both marker sets. This allowed the predictive abilities of γT3 and δT3 with the pathway-level marker dataset to nearly equal that obtained with genome-wide markers when this covariate was not included. Conversely, the predictive abilities for the other three tocopherol compounds, which are synthesized in the embryo, and αT3 were essentially unchanged across all three marker datasets when the covariate for endosperm mutation type was included. Taken together, these results suggest that capturing the genetic information associated with *su1* and *sh2* is important to improving the predictive ability of γT3, δT3, and their related derivative traits when selecting in breeding populations that are segregating for both the *su1* and *sh2* endosperm mutations.

DISCUSSION

Improving the nutritional quality of fresh sweet corn through genetic improvement offers an avenue to help address vitamin E insufficiencies where this vegetable is frequently consumed. Such biofortification efforts would be enhanced by association studies that identify the loci underlying phenotypic variation for tocochromanol levels in sweet corn kernels. As a complement to GWAS for the genetic dissection of these nutritional kernel traits, the optimization of predictive abilities for genomic selection models with marker sets that are genome-wide or that more directly target genes controlling tocochromanol phenotypes would also provide insight into the genetic gains that could be expected under selection in a breeding program. In that light, we conducted a GWAS to identify the genetic controllers of natural variation for 20 tocochromanol kernel traits and, under different marker set scenarios, assessed the accuracy of genomic prediction models that could be used for the biofortification of sweet corn. This work represents the first GWAS conducted in a sweet corn association panel, the most

extensive assessment of natural variation for tocochromanol levels in fresh kernels, and the first genomic prediction analysis of tocochromanol traits in maize.

The extent of variation for α-, δ-, and γ- tocopherols and tocotrienols in fresh kernels from an association panel of diverse sweet corn lines was evaluated. The 3.8- to 47.8-fold range in variation, calculated as the maximum divided by the minimum BLUP value for each trait, revealed for these six tocochromanol compounds represents extensive phenotypic variability at the fresh kernel stage, with αT (the highest vitamin E activity compound) having a 14.3-fold range in variation. In contrast, there was a 30-fold higher range of variation for αT in physiologically mature dry kernels of temperate and tropical non-sweet corn lines comprising the Goodman–Buckler association panel (Lipka et al., 2013) relative to our sweet corn panel. The likely drivers explaining these differences are the higher levels of allelic diversity captured by the Goodman–Buckler association panel (Flint-Garcia et al., 2005) and the continued accumulation of tocochromanols to higher levels in the kernel beyond the ~21 DAP analyzed in this study (Kurilich and Juvik, 1999; Xie et al., 2017), the time point when sweet corn is typically harvested for consumption. Irrespective of these limitations, the observed wide range of phenotypic variation in the sweet corn association panel was found to be highly heritable ($\hat{h}_l^2$= 0.68–0.89) and capture a more biologically relevant topmost RDA of 4.4% for vitamin E (Supplemental Fig. S1.10).

Through a GWAS of the sweet corn association panel, significant associations for three core tocochromanol pathway genes (*vte4*, *vte1*, and *hggt1*) were identified at the genome-wide level, which are in agreement with GWAS results from prior studies of mature kernels in maize (Li et al., 2012; Lipka et al., 2013; Diepenbrock et al., 2017; Wang et al., 2018). The significant association between two nonindependent SNPs within *vte4* and αT content confirms the critical

role of this biosynthetic gene in compositional profiles for fresh sweet corn kernels. This association signal defined by two SNPs in complete LD was resolved to a single SNP selected by the MLMM, suggesting a lack of allelic heterogeneity, as was implicated for *vte4* in the Goodman–Buckler association panel (Lipka et al., 2013). However, this could be attributed to the relatively fewer number of scored SNPs and the expected lower haplotype diversity at *vte4* in the sweet corn association panel. Evidence was found to support the hypothesis of a distant upstream regulatory element at *vte4*, with an association signal ~170 kb away from a putative regulatory element previously shown to be associated with αT levels in maize grain (Li et al., 2012).

In concordance with Lipka et al. (2013) and Diepenbrock et al. (2017), moderately strong associations were identified between SNPs spanning a recombinationally suppressed pericentromeric region that encompassed *vte1* and tocotrienol traits. In these two previous studies, the clear attribution of the association signal to *vte1* was confounded by complex patterns of LD in the genomic interval. In our study, however, the optimal MLMM model was able to resolve the detected association signal to within the *vte1* gene for δT3/(γT3 + αT3). The enzyme encoded by *vte1*, tocopherol cyclase, converts 2,3-dimethyl-5-geranylgeranylbenzoquinol and 2-methyl-6-geranylgeranylbenzoquinol to γT3 and δT3, respectively, which is in agreement with the association of *vte1* with δT3/(γT3 + αT3) and δT3/γT3 in the sweet corn association panel. Furthermore, *vte1* is expressed at low levels in the tocotrienol-rich endosperm (Stelpflug et al., 2016) and therefore is more likely to be a limiting factor for tocotrienol than tocopherol biosynthesis (Lipka et al., 2013). Taken together, to date this is the strongest support that implicates *vte1* as contributing to the natural variation of tocotrienols in kernels.

A second gene related to tocotrienol biosynthesis, *hggt1*, was shown to be associated with tocotrienol traits. The HGGT enzyme catalyzes the first committed step of tocotrienol synthesis and is expressed strongly in the endosperm (Stelpflug et al., 2016), the site of tocotrienol accumulation (Grams et al., 1970; Weber, 1987). Within this pericentromeric region on chromosome 9, the most significant SNP associated with γT3 was 138 kb away from *hggt1*, which is within the range of physical distances for SNPs upstream of *hggt1* that had a significant association with γT3 and total tocotrienols in maize grain through a pathway-level analysis (Lipka et al., 2013). The two significant SNP associations detected at 1 and 1.8 Mb upstream of *hggt1* for total T and the total T/total T3 ratio, respectively, are most likely to have resulted from long-range LD patterns rather than very distant regulatory elements. In the US maize NAM panel, *hggt1* was shown to explain the highest phenotypic variation (24.0–40.2%) for δT3, γT3, and total T3. Although it was not the most significant locus for tocotrienol-related traits in this study and that of Lipka et al. (2013), the difference is probably because these two mapping panels have weaker statistical power because of their smaller sample size, rarer allele frequencies, and fewer scored SNP markers than the NAM panel.

In contrast to the GWAS of tocotrienols conducted by Lipka et al. (2013), which excluded the few sweet corn lines included in the Goodman–Buckler association panel, two genes involved in kernel starch biosynthesis (*su1* and *sh2*) were found to be associated with tocotrienol traits in our study. Given that *sh2* kernels (~80%) have a higher percentage of moisture relative to *su1* kernels (~75%) arising from differences in sugar and water-soluble polysaccharide content at the fresh-eating stage (Creech, 1965; Soberalske and Andrew, 1978), it was posited that the differences in the percentage of moisture could explain the association of *su1* and *sh2* with tocotrienols. However, the significant associations between these endosperm-

expressed genes and the tocotrienol-related traits still remained even after conducting a more stringent GWAS with a mixed linear model that had a kinship matrix and the first four PCs derived from all 174,996 SNP markers plus fresh kernel weight as a covariate (results not shown). When we considered the JL-QTL mapping results of three tocotrienol-related traits (γT3, total T3, and total T + T3) for grain from the US maize NAM panel, only one of the two sweet corn families ('B73 × P39') that segregate for the *su1* mutation had a significant allelic effect estimate for a JL-QTL with a support interval that included *su1* (Diepenbrock et al., 2017). However, this JL-QTL could not be resolved down to the gene level by GWAS in the NAM panel. Therefore, strong independent evidence is lacking for the implication of *su1* in the genetic control of tocotrienol levels in maize grain. The role of *sh2* in the regulation of tocotrienols, however, could not be assessed by Diepenbrock et al. (2017), because neither of the two sweet corn families segregated for the *sh2* mutation.

Of the six measured tocochromanol compounds in kernels, only γT3 and δT3 were at significantly greater levels ($P < 0.01$) for *sh2* and *su1sh2* lines relative to *su1* lines (Table 1.2). This could result from the unintentional fixation of causal alleles associated with the increased levels of these two tocotrienols, which might have arisen early in the breeding process if only a limited number of highly related *sh2* lines were used as donor parents (Tracy, 1997). Indeed, *hggt1* (S9_92345469) and *vte1* (S5_131738084) alleles that increased the level of γT3 were fixed in all *sh2* and *su1sh2* lines, as inferred by the peak SNPs, with the exception of one *sh2* line that had the weaker *hggt1* allele (Supplemental Table S1.9). The involvement of *hggt1* and *vte1* would directly influence γT3 levels, as HGGT1 condenses homogentisic acid and GGDP to produce 2-methyl-6-geranylgeranylbenzoquinol, which, after being methylated to 2,3-dimethyl-5-geranylgeranyl-1,4-benzoquinol by VTE3, is then converted to γT3 by VTE1 (Fig. 1.1). Given

that *hggt1* and *vte1* do not explain all of the variation between *su1* versus *sh2* and *su1sh2* for these two tocotrienols (Supplemental Table S1.8 and Supplemental Table S1.9), it is likely that one or more of the other nine genes identified to be underlying JL-QTL for γT3 and δT3 levels in maize grain (Diepenbrock et al., 2017) account for the missing heritability. Further exploration to evaluate the genetic contribution of these nine undetected genes and their variant allele frequencies within groups of *su1*, *sh2*, and *su1sh2* lines would be enhanced through the higher statistical power that would be attainable with a larger, more densely genotyped sweet corn association panel.

The genes most strongly associated with both γT3 and δT3 are the two endosperm-expressed genes, *su1* and *sh2*. However, neither of these genes was associated with tocopherol (embryo) traits (Grams et al., 1970; Weber, 1987). If either of these genes is responsible for the greater levels of γT3 and δT3 in kernels of *sh2* and *su1sh2* lines, it is most plausibly driven by the increased sugar content in the endosperm, especially that of the *su1sh2* and *sh2* genotypes that have two- to threefold and seven- to eightfold more sucrose at 20 DAP than *su1* and dent corn genotypes, respectively (Creech, 1965). In leaf tissue from *A. thaliana* plants grown on media supplemented with 3% sucrose, Hsieh and Goodman (2005) observed moderate increases in the expression of MEP pathway genes, including *1-deoxy-d-xylulose 5-phosphate synthase*, which is a gene (*dxs2*) that underlies a major JL-QTL for levels of γT3, δT3 and total T3 (but not tocopherol traits) in grain from the US maize NAM panel (Diepenbrock et al., 2017). Therefore, we hypothesize that the increased sucrose concentration in kernels of *su1sh2* and *sh2* lines stimulates the synthesis of tocotrienols in the endosperm through upregulation of the MEP pathway that provides isopentenyl pyrophosphate for biosynthesis of the tocotrienol tail groups. This could synergistically enhance tocotrienol production in the presence of the strongly

expressed *hggt1* allele that is essentially fixed in lines with the *sh2* mutation. Taken together, with the lack of evidence for an association between *su1* and tocotrienols in the US maize NAM panel and the relatively modest accumulation of sucrose, γT3, and δT3 in kernels of *su1* lines, *sh2* becomes the most probable genetic contributor to tocotrienol levels through its production of high sucrose in the endosperm. This hypothesis would be further supported if the association between *su1* and tocotrienols is eventually found to be spurious because of the high interchromosome LD ($r^2 = 0.61$) between *su1* and *sh2*.

Through the implementation of WGP via the GBLUP method, moderate (tocopherols) to moderately high (tocotrienols) predictive abilities were shown for tocochromanol phenotypes in the panel, suggesting that genomic selection could be used to improve genetic gain for tocochromanols and vitamin E in sweet corn breeding programs. The pathway-level and tocochromanol QTL-targeted marker datasets were found to have lower average predictive abilities than those from WGP, although tocochromanol traits are mostly explained by several moderate- to large-effect loci in the NAM panel (Diepenbrock et al., 2017). The most probable explanation for these lower predictive abilities is that the genotyped SNP markers (common variants) at the targeted loci were not in strong LD with causative variants. In support of this theory, the number of significantly associated markers at a genome-wide FDR of 5% from GWAS had a strong, positive correlation with the predictive abilities of tocochromanols. Additionally, 11 of the 14 causal loci controlling grain tocochromanols have significant allelic effect estimates in at least one of the two sweet corn families of the US maize NAM panel (Diepenbrock et al., 2017) but these would have escaped detection in models used for GWAS and WGP if the causal variants at these loci were rarer, weaker effects in the less densely genotyped, smaller sweet corn association panel. These findings are in contrast to the work of

Owens et al. (2014), who showed that an eight gene QTL-targeted set is as effective as genome-wide markers for the prediction of the highly oligogenic carotenoid grain traits in the Goodman–Buckler association panel. However, four of these eight genes were detected via GWAS in the Goodman–Buckler association panel, thus ensuring the capture of large-effect genes that are critical for modifying grain carotenoid composition in the prediction model. Conversely, only 2 of the 14 causal genes underlying QTL associated with grain tocochromanols (Diepenbrock et al., 2017) were identified in the sweet corn association panel.

CONCLUSIONS

We found natural variation for αT (the tocochromanol with the highest vitamin E activity) in sweet corn kernels at the fresh-eating stage to be predominantly under the genetic control of *vte4*, while *vte1* and *hggt1* are involved in controlling the content and composition of tocotrienols. Of the two starch biosynthesis genes found to associate with tocotrienols, the strongest evidence exists for the involvement of *sh2* rather than *su1* in modifying tocotrienol levels. However, additional experiments are needed to develop and evaluate a set of near isogenic lines that capture the different alleles of *su1* and *sh2* in several genetic backgrounds to determine the contribution of these two genes, if any, to heritable differences in tocotrienol levels. The majority of lines with the *sh2* mutation are fixed for alleles at *vte4*, *vte1*, and *hggt1* that collectively increase the levels of αT and γT3. In light of this finding, targeted resequencing and characterization of allelic variation at these three and the other undetected loci previously identified are needed to better assess if the extant sweet corn germplasm pool captures the most favorable variants that exist for maize as a species, especially given that sweet corn experienced a postdomestication genetic bottleneck and recent founder events (Tracy, 1997; Whitt et al., 2002). Whether it be through selection on existing or introgressed allelic variation in breeding

programs, our work constitutes an important step for the necessary genomics-assisted breeding efforts to enhance vitamin E to a level that meets or exceeds an RDA of 4.4% for 100 g of fresh sweet corn kernels.

SUPPLEMENTAL INFORMATION

Supplemental Table S1.1: Lambda values used in Box-Cox transformation of 20 fresh kernel tocochromanol traits.

Supplemental Table S1.2: Transformed best linear unbiased predictors of the 20 fresh kernel tocochromanol traits.

Supplemental Table S1.3: Back-transformed best linear unbiased predictors of the 20 fresh kernel tocochromanol traits.

Supplemental Table S1.4: Comparison of genomic prediction models for the presence/absence of two endosperm mutations (*su1* and *sh2*) using marker data sets with different levels of coverage.

Supplemental Table S1.5: Genomic information (RefGen_v2) for the 81 *a priori* candidate genes.

Supplemental Table S1.6: Genomic information (RefGen_v2) for the 14 *a priori* genes underlying joint-linkage quantitative trait loci associated with grain tocochromanol levels in the US maize nested association mapping panel.

Supplemental Table S1.7: Statistically significant results from a genome-wide association study of 20 fresh kernel tocochromanol traits.

Supplemental Table S1.8: Multi-locus mixed-model results from an analysis of tocochromanol traits for chromosomes 3, 4, and 5 with and without endosperm mutation type as a covariate.

Supplemental Table S1.9: Back-transformed effect estimates for *vte4*, *vte1*, *hggt1*-related SNPs selected with an optimal multi-locus mixed-model.

Supplemental Table S1.10: Statistically significant results from a genome-wide association study of 20 fresh kernel tocochromanol traits when including endosperm mutation type as a covariate in the mixed linear model.

Supplemental Table S1.11: Statistically significant results from a genome wide association study of 20 fresh kernel tocochromanol traits in sweet corn when using endosperm mutation type and the two SNPs selected by multi-locus mixed-models (S5_131738084 and S5_214707875) as covariates in the mixed linear model.

Supplemental Fig. S1.1: Sources of variation for tocochromanol traits in fresh sweet corn kernels.

Supplemental Fig. S1.2: Correlation matrix for back-transformed BLUPs of the 20 tocochromanol fresh kernel traits.

Supplemental Fig. S1.3: Linkage disequilibrium estimates in the sweet corn diversity panel.

Supplemental Fig. S1.4: Genome-wide association study of 20 fresh kernel tocochromanol traits in sweet corn.

Supplemental Fig. S1.5: Genome-wide association study for the ratio of α-tocopherol to γ-tocopherol in fresh kernels of sweet corn.

Supplemental Fig. S1.6: Genome-wide association study for the ratio of δ-tocotrienol to γ-tocotrienol in fresh kernels of sweet corn.

Supplemental Fig. S1.7: Genome-wide association study for γ-tocotrienol content in fresh kernels of sweet corn.

Supplemental Fig. S1.8: Genome-wide association study for endosperm mutation type of physiologically mature sweet corn kernels.

Supplemental Fig. S1.9: Genome-wide association study of 20 fresh kernel tocochromanol traits in sweet corn with endosperm mutation type (*su1*, *sh2*, or *su1sh2*) included as a covariate.

Supplemental Fig. S1.10: Distribution of the percentage of the recommended daily allowance (RDA) for vitamin E provided by inbred lines from the sweet corn association panel.

REFERENCES

Barger, G.L. 1969. Total growing degree days. Weekly Weather & Crop Bull. 56(18):10.

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.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. J. R. Stat. Soc. Series B Stat. Methodol. 26:211–252.

Boyer, C.D., and J.C. Shannon. 1983. The use of endosperm genes for sweet corn improvement. In: J. Janick, editor, Plant breeding reviews. Springer, Boston, MA. p.139–161.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. doi:10.1093/bioinformatics/btm308

Cahoon, E.B., S.E. Hall, K.G. Ripp, T.S. Ganzke, W.D. Hitz, and S.J. Coughlan. 2003. Metabolic redesign of vitamin E biosynthesis in plants for tocotrienol production and increased antioxidant content. Nat. Biotechnol. 21:1082–1087. doi:10.1038/nbt853

Centers for Disease Control. 2006. 2nd national report on biochemical indicators of diet and nutrition in the US population. Centers for Disease Control. https://www.cdc.gov/nutritionreport/pdf/Fat.pdf (accessed 25 Sept. 2018).

Chen, J., and Z. Chen. 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95:759–771. doi:10.1093/biomet/asn034

Creech, R.G. 1965. Genetic control of carbohydrate synthesis in maize endosperm. Genetics 52:1175–1186.

DellaPenna, D., and L. Mène-Saffrané. 2011. Vitamin E. In: F. Rebeille and R. Douce, editors, Advances in Botanical Research. Elsevier Ltd., Amsterdam, The Netherlands. p. 179–227.

Diepenbrock, C.H., C.B. Kandianis, A.E. Lipka, M. Magallanes-Lundback, B. Vaillancourt, E. Gongora-Castillo, et al. 2017. Novel loci underlie natural variation in vitamin E levels in maize grain. Plant Cell 29:2374–2392. doi:10.1105/tpc.17.00475

Doebley, J., J.F. Wendel, J.S.C. Smith, C.W. Stuber, and M.M. Goodman. 1988. The origin of cornbelt maize: The isozyme evidence. Econ. Bot. 42:120–131. doi:10.1007/BF02859042

Doehlert, D.C., T.M. Kuo, J.A. Juvik, E.P. Beers, and S.H. Duke. 1993. Characteristics of carbohydrate metabolism in sweet corn (sugary-1) endosperms. J. Am. Soc. Hortic. Sci. 118:661–666.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:E19379. doi:10.1371/journal.pone.0019379

Flint-Garcia, S.A., A.C. Thuillet, J. Yu, G. Pressoir, S.M. Romero, S.E. Mitchell, et al. 2005. Maize association population: A high-resolution platform for quantitative trait locus dissection. Plant J. 44:1054–1064. doi:10.1111/j.1365-313X.2005.02591.x

Ford, E.S., R.L. Schleicher, A.H. Mokdad, U.A. Ajani, and S. Liu. 2006. Distribution of serum concentrations of alpha-tocopherol and gamma-tocopherol in the US population. Am. J. Clin. Nutr. 84:375–383. doi:10.1093/ajcn/84.2.375

Gerdes, J.T., and W.F. Tracy. 1994. Diversity of historically important sweet corn inbreds as estimated by RFLPs, morphology, isozymes, and pedigree. Crop Sci. 34:26–33. doi:10.2135/cropsci1994.0011183X003400010004x

Gilmour, A.R.G., B.B. Cullis, R. Thompson, and D. Butler. 2009. Asreml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, et al. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One 9:E90346. doi:10.1371/journal.pone.0090346

Grams, G.W., C.W. Blessin, and G.E. Inglett. 1970. Distribution of tocopherols within the corn kernel. J. Am. Oil Chem. Soc. 47:337–339. doi:10.1007/BF02638997

Hannah, L.C., M. Giroux, and C. Boyer. 1993. Biotechnological modification of carbohydrates for sweet corn and maize improvement. Sci. Hortic. (Amsterdam) 55:177–197. doi:10.1016/0304-4238(93)90031-K

Hill, W.G., and B.S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. 33:54–78. doi:10.1016/0040-5809(88)90004-4

Holland, J.B., W.E. Nyquist, and C.T. Cervantes-Martínez. 2003. Estimating and interpreting heritability for plant breeding: An update. In J. Janick, editor, Plant Breeding Reviews 2. John Wiley and Sons, Hoboken, NJ. p. 9–112.

Hsieh, M.H., and H.M. Goodman. 2005. The Arabidopsis IspH homolog is involved in the plastid nonmevalonate pathway of isoprenoid biosynthesis. Plant Physiol. 138:641–653. doi:10.1104/pp.104.058735

Hung, H.Y., C. Browne, K. Guill, N. Coles, M. Eller, A. Garcia, 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. doi:10.1038/hdy.2011.103

Ibrahim, K.E., and J.A. Juvik. 2009. Feasibility for improving phytonutrient content in vegetable crops using conventional breeding strategies: Case study with carotenoids and tocopherols in sweet corn and broccoli. J. Agric. Food Chem. 57:4636–4644. doi:10.1021/jf900260d

Institute of Medicine. 2000. Dietary reference intakes for vitamin C, vitamin E, selenium, and carotenoids. The National Academies Press, Washington, DC.

Jennings, P.H., and C.L. McCombs. 1969. Effects of sugary-1 and shrunken-2 loci on kernel carbohydrate contents, phosphorylase and branching enzyme activities during maize kernel ontogeny. Phytochemistry 8:1357–1363. doi:10.1016/S0031-9422(00)85898-7

Kenward, M.G., and J.H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53:983. doi:10.2307/2533558

Knekt, P., A. Reunanen, R. Jarvinen, R. Seppanen, M. Heliovaara, and A. Aromaa. 1994. Antioxidant vitamin intake and coronary mortality in a longitudinal population study. Am. J. Epidemiol. 139:1180–1189. doi:10.1093/oxfordjournals.aje.a116964

Kruk, J., H. Hollander-Czytko, W. Oettmeier, and A. Trebst. 2005. Tocopherol as singlet oxygen scavenger in photosystem II. J. Plant Physiol. 162:749–757. doi:10.1016/j.jplph.2005.04.020

Kurilich, A.C., and J.A. Juvik. 1999. Quantification of carotenoid and tocopherol antioxidants in Zea mays. J. Agric. Food Chem. 47:1948–1955. doi:10.1021/jf981029d

Kushi, L.H., A.R. Folsom, R.J. Prineas, P.J. Mink, Y. Wu, and R.M. Bostick. 1996. Dietary antioxidant vitamins and death from coronary heart disease in postmenopausal women. N. Engl. J. Med. 334:1156–1162. doi:10.1056/NEJM199605023341803

Leth, T., and H. Sondergaard. 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. 107:2236–2243. doi:10.1093/jn/107.12.2236

Li, Q., X. Yang, S. Xu, Y. Cai, D. Zhang, Y. Han, et al. 2012. Genome-wide association studies identified three independent polymorphisms associated with alpha-tocopherol content in maize kernels. PLoS One 7:E36807. doi:10.1371/journal.pone.0036807

Linus Pauling Institute. 2015. Vitamin E. Oregon State Univ. http://lpi.oregonstate.edu/mic/vitamins/vitamin-E (accessed 25 Sept. 2018).

Lipka, A.E., M.A. Gore, M. Magallanes-Lundback, A. Mesberg, H. Lin, T. Tiede, et al. 2013. Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. G3 (Bethesda) 3:1287–1299. doi:10.1534/g3.113.006148

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, et al. 2012. GAPIT: Genome association and prediction integrated tool. Bioinformatics 28:2397–2399. doi:10.1093/bioinformatics/bts444

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. p. 733–756.

Liu, X., X. Hua, J. Guo, D. Qi, L. Wang, Z. Liu, et al. 2008. Enhanced tolerance to drought stress in transgenic tobacco plants overexpressing VTE1 for increased tocopherol production from Arabidopsis thaliana. Biotechnol. Lett. 30:1275–1280. doi:10.1007/s10529-008-9672-y

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

McBurney, M.I., E.A. Yu, E.D. Ciappio, J.K. Bird, M. Eggersdorfer, and S. Mehta. 2015. Suboptimal serum alpha-tocopherol concentrations observed among younger adults and those depending exclusively upon food sources, NHANES 2003–2006. PLoS One 10:E0135510. doi:10.1371/journal.pone.0135510

Mene-Saffrane, L. 2017. Vitamin E biosynthesis and its regulation in plants. Antioxidants (Basel) 7(1):2. doi:10.3390/antiox7010002

Michaels, T.E., and R.H. Andrew. 1986. Sugar accumulation in shrunken-2 sweet corn kernels. Crop Sci. 26:104–107. doi:10.2135/cropsci1986.0011183X002600010025x

Neter, J., M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models, McGraw-Hill, Boston, MA.

Owens, B.F., A.E. Lipka, M. Magallanes-Lundback, T. Tiede, C.H. Diepenbrock, C.B. Kandianis, 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. doi:10.1534/genetics.114.169979

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909. doi:10.1038/ng1847

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

Romay, M.C., M.J. Millard, J.C. Glaubitz, J.A. Peiffer, K.L. Swarts, T.M. Casstevens, et al. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14:R55. doi:10.1186/gb-2013-14-6-r55

Sattler, S.E., L.U. Gilliland, M. Magallanes-Lundback, M. Pollard, and D. DellaPenna. 2004. Vitamin E is essential for seed longevity and for preventing lipid peroxidation during germination. Plant Cell 16:1419–1432. doi:10.1105/tpc.021360

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464. doi:10.1214/aos/1176344136

Segura, V., B.J. Vilhjalmsson, A. Platt, A. Korte, U. Seren, Q. Long, et al. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44:825–830. doi:10.1038/ng.2314

Sen, C.K., S. Khanna, and S. Roy. 2006. Tocotrienols: Vitamin E beyond tocopherols. Life Sci. 78:2088–2098. doi:10.1016/j.lfs.2005.12.001

Soberalske, R.M., and R.H. Andrew. 1978. Gene effects on kernel moisture and sugars of near-isogenic lines of sweet corn. Crop Sci. 18:743–746. doi:10.2135/cropsci1978.0011183X001800050012x

Stelpflug, S.C., R.S. Sekhon, B. Vaillancourt, C.N. Hirsch, C.R. Buell, N. de Leon, et al. 2016. An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. Plant Genome 9:1–16. doi:10.3835/plantgenome2015.04.0025

Sun, G., C. Zhu, M.H. Kramer, S.S. Yang, W. Song, H.P. Piepho, et al. 2010. Variation explained in mixed-model association mapping. Heredity (Edinb) 105:333–340. doi:10.1038/hdy.2010.11

Swarts, K., H. Li, J.A.R. Navarro, D. An, M.C. Romay, S. Hearne, et al. 2014. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7:1–12. doi:10.3835/plantgenome2014.05.0023

Tetlow, I.J., M.K. Morell, and M.J. Emes. 2004. Recent developments in understanding the regulation of starch metabolism in higher plants. J. Exp. Bot. 55:2131–2145. doi:10.1093/jxb/erh248

Tracy, W.F. 1997. History, genetics, and breeding of supersweet (shrunken2) sweet corn. In: J. Janick, editor, Plant Breeding Reviews 17. John Wiley and Sons, Hoboken, NJ. p.189–236.

USDA. 2018a. National nutrient database for standard reference. Nutrient Data Laboratory, Beltsville Human Nutrition Research Center. https://ndb.nal.usda.gov/ndb/search/ (accessed 25 Sept. 2018).

USDA. 2018b. Vegetables 2017 summary. USDA NASS. http://usda.mannlib.cornell.edu/usda/nass/VegeSumm//2010s/2018/VegeSumm-02-13-2018.pdf (accessed 25 Sept. 2018).

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980

Wang, H., S. Xu, Y. Fan, N. Liu, W. Zhan, H. Liu, et al. 2018. Beyond pathways: Genetic dissection of tocopherol content in maize kernels by combining linkage and association analyses. Plant Biotechnol. J. doi:10.1111/pbi.12889

Weber, E.J. 1987. Carotenoids and tocols of corn grain determined by HPLC. J. Am. Oil Chem. Soc. 64:1129–1134. doi:10.1007/BF02612988

Whitt, S.R., L.M. Wilson, M.I. Tenaillon, B.S. Gaut, and E.S. Buckler. 2002. Genetic diversity and selection in the maize starch pathway. Proc. Natl. Acad. Sci. USA 99:12959–12962. doi:10.1073/pnas.202476999

Wilson, L.M., S.R. Whitt, A.M. Ibanez, T.R. Rocheford, M.M. Goodman, and E.S. Buckler. 2004. Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719–2733. doi:10.1105/tpc.104.025700

Wolfinger, R., W.T. Federer, and O. Cordero-Brana. 1997. Recovering information in augmented designs, using SAS PROC GLM and PROC Mixed. Agron. J. 89:856–859. doi:10.2134/agronj1997.00021962008900060002x

Xie, L., Y. Yu, J. Mao, H. Liu, J.G. Hu, T. Li, et al. 2017. Evaluation of biosynthesis, accumulation and antioxidant activityof vitamin E in sweet corn (Zea mays L.) during kernel development. Int. J. Mol. Sci. 18. doi:10.3390/ijms18122780

Young, T.E., D.R. Gallie, and D.A. DeMason. 1997. Ethylene-mediated programmed cell death during maize endosperm development of wild-type and shrunken2 genotypes. Plant Physiol. 115:737–751. doi:10.1104/pp.115.2.737

Zhang, Z., E. Ersoz, C.-Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, et al. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42:355–360. doi:10.1038/ng.546

Zhang, Z., R.J. Todhunter, E.S. Buckler, and L.D. Van Vleck. 2007. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. J. Anim. Sci. 85:881–885. doi:10.2527/jas.2006-656

CHAPTER 2

GENETIC BASIS AND GENOMIC PREDICTION MODELS FOR CAROTENOID LEVELS IN FRESH SWEET CORN[2]

INTRODUCTION

Carotenoids are fat-soluble compounds synthesized by plants that play a critical role in photosynthesis, acting as photoprotectants, antioxidants, and accessory pigments for light harvesting (reviewed in Cuttriss et al., 2011). In the human body, provitamin A carotenoids, such as α-carotene, β-carotene, and β-cryptoxanthin, can be converted to retinol (vitamin A). Lutein and zeaxanthin are the main compounds in the macula of the eye (Bone et al., 1993), which is the area responsible for sharp central vision. These two non-provitamin A carotenoids absorb excess blue and ultraviolet light entering the eye, protecting against subsequent cellular damage (reviewed in Krinsky et al., 2003). Their routine consumption has been associated with reduced risk of progression to late stage age-related macular degeneration (AMD) (Chew et al., 2014; Wu et al., 2015), the leading cause of blindness among older adults in the developed world (Congdon et al., 2004; Friedman et al., 2004). In the US, the 2005-2008 National Health and Nutrition Examination Survey (NHANES; National Center for Health Statistics, 2018) estimated that AMD affects ~6.5% of the population aged 40 years and older (Klein et al., 2011). Prevalence of AMD in North America is estimated to increase substantially in the next decades, rising to ~18-25 million cases by 2050 (Rein et al., 2009; Wong et al., 2014).

[2] Baseggio, M., M. Murray, M. Magallanes-Lundback, N. Kaczmar, J. Chamness, E.S. Buckler, M.E. Smith, D. DellaPenna, W.F. Tracy, and M.A. Gore. To be submitted to *The Plant Genome*

Given that humans cannot synthesize carotenoids, these must be acquired through dietary intake, particularly from fruits and vegetables. According to the 2015-2016 NHANES survey, intake of lutein and zeaxanthin combined for adults in the US average 1.7 mg day^{-1} (National Health and Nutrition Examination Survey, 2016). Sweet corn is the third most commonly consumed vegetable in the US (USDA, 2018) and one of the few food sources of zeaxanthin in the human diet (Rodriguez-Amaya, 2001). Few studies have been conducted to evaluate the variability of carotenoids in fresh sweet corn kernels and they only analyzed a limited number of genotypes present in the US (Kurilich and Juvik, 1999; Ibrahim and Juvik, 2009).

The carotenoid biosynthetic pathway has been characterized in *Arabidopsis thaliana* and is highly conserved in plants (reviewed in DellaPenna and Pogson, 2006; Cuttriss et al., 2011). Carotenoid production relies on the methylerythritol phosphate (MEP) pathway to generate the two isoprene isomers isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), which undergo several reactions resulting in the major carotenoid precursor, geranylgeranyl diphosphate (GGDP; Fig. 2.1). The committed step of the carotenoid pathway is the subsequent formation of phytoene by condensing two GGDP molecules via phytoene synthase (PSY). Conversion of phytoene into lycopene involves two desaturase and two isomerase enzymes, and the further cyclization of lycopene gives rise to the carotenoid diversity based on the end group. Two reactions with lycopene β-cyclase (LCY-β) produces β-carotene, whereas the addition of one β-ring by LCY-β and one ε-ring by lycopene ε-cyclase (LCY-ε) produces α-carotene. β-Carotene and α-carotene can be further converted into xanthophylls by the hydroxylation of the β-ring or/and ε-ring, leading to β-cryptoxanthin and further zeaxanthin or zeinoxanthin and lutein, respectively. Zeaxanthin, together with antheraxanthin and violaxanthin, is part of the xanthophyll cycle, which plays a key role protecting against photoinhibition by dissipating

excess excitation energy via non-photochemical quenching (reviewed in Jahns and Holzwarth, 2012). Additionally, violaxanthin can serve as precursor for biosynthesis of abscisic acid, an essential plant hormone associated with seed dormancy and response to abiotic stresses (reviewed in Kermode, 2005; Kundu and Gantait, 2017). Carotenoids are also precursors of strigolactones, phytohormones with diverse signaling activities that are produced via a pathway including carotenoid cleavage dioxygenases (CCD) (Alder et al., 2012).

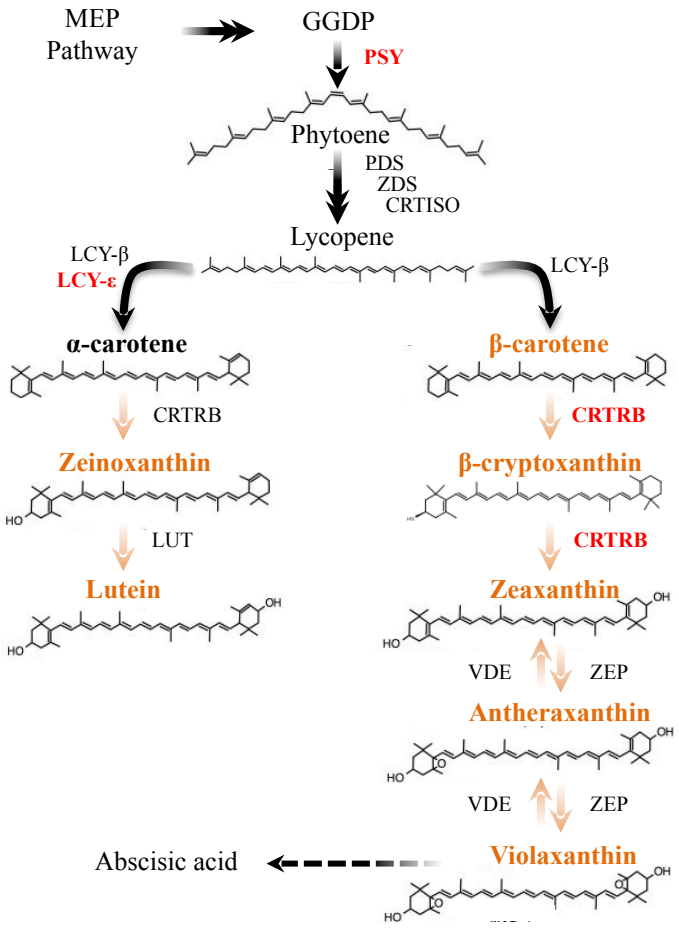

Figure 2.1. Carotenoid biosynthetic pathway in maize. The seven quantified compounds are shown in bolded orange text. The name of enzymes in bolded red text are produced by genes that are within ± 250 kb of the associated single nucleotide polymorphisms (SNPs) identified in our study. Compound abbreviations: GGDP, geranylgeranyl diphosphate; MEP, methylerythritol phosphate. Enzyme abbreviations: CRTRB, β-carotene hydroxylase; CRTISO, carotenoid isomerase; LCY-β, lycopene β-cyclase; LCY-ε, lycopene ε-cyclase; LUT, cytochrome P450 ε-ring hydroxylase; PDS, phytoene desaturase; PSY, phytoene synthase; VDE, violaxanthin de-epoxidase; ZEP, zeaxanthin epoxidase; ZDS, ζ-carotene desaturase.

Genome-wide association studies (GWAS) of mature maize kernels identified several genes responsible for the natural variation of carotenoids. Wong et al. (2004) reported that *phytoene synthase* (*psy1*) and *ζ-carotene desaturase* (*zds*), both genes encoding enzymes catalyzing early steps in the carotenoid pathway, were associated with accumulation of carotenoids, and positive selection resulted in minimal sequence variation for *psy1* within yellow maize germplasm (Palaisa et al., 2004; Fu et al., 2013). Assessment of the phenotypic variation for carotenoid profiles showed that lycopene ε-cyclase alters flux between α-carotene and β-carotene branches, explaining up to 58% of the variation and threefold difference in provitamin A carotenoids (Harjes et al., 2008). Yan et al. (2010) demonstrated that *β-carotene hydroxylase 1* (*crtRB1*) was associated with β-carotene concentration in mature maize kernels and that *crtRB1* alleles associated with increased β-carotene were rare in frequency in that panel. In addition to previously associated genes *lcyE* and *crtRB1*, Owens et al. (2014) first reported *zeaxanthin epoxidase* (*zep1*) and *cytochrome P450 ε-ring hydroxylase* (*lut1*) to impact carotenoid composition.

These findings from maize kernels at mature stage, however, may not directly translate to sweet corn kernels at fresh stage. Carotenoid levels and gene expression change as the kernel matures (Kurilich and Juvik, 1999; Vallabhaneni and Wurtzel, 2009; Song et al., 2015; Calvo-Brenes et al., 2019) and therefore selection of alleles responsible for increased carotenoid accumulation at fresh stage (~21 DAP) is needed. In addition to a lower genetic variability compared to dent corn (Romay et al., 2013), sweet corn has one or more mutations in the starch biosynthesis pathway, resulting in a greater sugar accumulation in the endosperm (Tracy, 1997), where most of the carotenoids are located (Weber, 1987). High levels of sugar have been previously associated with chromoplast biogenesis and increased carotenoid accumulation in

citrus fruits, tobacco floral nectaries, and *Arabidopsis* seedlings, roots, and tubers (Iglesias et al., 2001; Horner et al., 2007; Flores-Pérez et al., 2010; Li et al., 2012).

Although GWAS is a powerful tool to identify additional key genes that affect carotenoids in maize kernels, genomic selection could also be employed in order to detect favorable genomic signatures for selection, thus enhancing the genetic gain per unit of time (Meuwissen et al., 2001; Owens et al., 2014). Owens et al. (2014) reported that a small set of markers targeting candidate genes underlying quantitative trait loci (QTL) associated with carotenoid biosynthesis and retention was as effective for predicting carotenoid traits as genome-wide markers. Marker-assisted selection for favorable *lcyE* and *crtRB1* alleles has been successfully conducted to increase β-carotene levels in mature kernels of maize (Pixley et al., 2013; Yang et al., 2018). Allele mining of exotic germplasm produced 34 high-carotenoid lines, with particular increase in lutein and zeaxanthin levels (Burt et al., 2011). O'Hare et al. (2015) reported a fivefold increase in zeaxanthin of tropical sweet corn lines by traditional breeding when selecting for higher total carotenoid and greater proportion of zeaxanthin compared to other carotenoids in the kernel. They also observed a concomitant increase in β-carotene accumulation and a color change from yellow to a more yellow-orange.

In this study, we used a sweet corn association panel to determine the controllers of natural carotenoid variation in fresh kernels and develop a prediction model with potentially higher predictive ability to facilitate the genomics-assisted breeding of fresh sweet corn with higher levels of provitamin A, lutein, and zeaxanthin. Therefore, we conducted a GWAS to identify key genes and favorable alleles associated with increased carotenoid levels in fresh [~21 days after pollination (DAP)] sweet corn kernels, and genomic prediction studies to determine

the extent of marker density required to maximize predictive abilities for the establishment of a genomic selection breeding program for nutritional quality in sweet corn.

MATERIALS AND METHODS

Plant Materials and Experimental Design

We field-evaluated an association panel of 416 diverse sweet corn inbred lines, which samples the allelic diversity of temperate US sweet corn breeding programs (Baseggio et al., 2019), at Cornell University's Musgrave Research Farm in Aurora, NY during the 2014 and 2015 growing seasons. The sweet corn inbred lines included in this panel were homozygous for the following starch-deficient endosperm mutations: *sugary1* (*su1*), *sugary1:sugary enhancer1* (*su1se1*), *shrunken2* (*sh2*), *sugary1:shrunken2* (*su1sh2*), *brittle2* (*bt2*), and *amylose-extender:dull:waxy* (*aeduwx*). Also included in the experiment were 20 non-sweet corn inbred lines and four repeated check sweet corn inbred lines. The association panel was grown in an augmented incomplete block design and fresh kernel samples (harvested at 400 growing degree days or ~21 DAP) were produced and processed as described previously (Baseggio et al., 2019). Not all plots had harvestable ears because some inbred lines had poor agronomic performance or matured too late. Ground kernel samples were shipped on dry ice to Michigan State University (East Lansing, MI) for extraction and quantification of carotenoids.

Phenotypic Data Analysis

Carotenoids were extracted from each ground sample and quantified by high-performance liquid chromatography (HPLC) and fluorometry, with 1 mg mL^{-1} of β-apo-8'-carotenal as an internal recovery control as previously described (Owens et al., 2014). The seven carotenoid compounds measured in 859 kernel samples from 401 sweet corn, 19 dent, 4 check

inbred lines were antheraxanthin, β-carotene, β-cryptoxanthin, lutein, violaxanthin, zeaxanthin, and zeinoxanthin in μg g^{-1} fresh kernel. Given the difficulties associated with identifying and measuring low-abundant carotenoids, lycopene, α-carotene, δ-carotene, and other unidentified carotenes were summed to comprise the 'other carotenes' phenotype. Additionally, a series of 11 sums and ratios from Owens et al. (2014) were calculated with minor modifications as follows: zeinoxanthin/lutein, β-cryptoxanthin/zeaxanthin, β-carotene/β-cryptoxanthin, β-carotene/(β-cryptoxanthin+zeaxanthin), α-xanthophylls (sum of lutein and zeinoxanthin), β-xanthophylls (sum of antheraxanthin, β-cryptoxanthin, violaxanthin, and zeaxanthin), β-xanthophylls/α-xanthophylls, total carotenes (sum of β-carotene and other carotenes), total xanthophylls (sum of α- and β-xanthophylls), total carotenes/total xanthophylls, and total carotenoids (sum of the seven carotenoid compounds and other carotenes).

Sweet corn inbred lines homozygous for the recessive null allele of the *y1* gene that encodes phytoene synthase 1 have carotenoids in the embryo but essentially none in an endosperm with a genetic background-dependent white to pale yellow color (Buckner, 1990). Given that, we identified and removed white/pale yellow endosperm lines that had a sample in at least one year with very low levels of total carotenoids quantified by HPLC (total carotenoids: < 5.57 μg g^{-1}, 2014; < 5.98 μg g^{-1}, 2015) and simultaneously confirmed with visual scoring of kernel color by one person (Matheus Baseggio). This was done to completely control for the very strong genetic signal at *y1* associated with the Mendelian inherited presence (yellow/orange kernel color) vs. absence (white kernel color) of endosperm carotenoids (Owens et al., 2014) and allow for the exclusive study of quantitative variation for carotenoid levels. The visual scoring was based on images of two immature (~21 DAP) ears per plot on the 1KK green background (https://wheatgenetics.org/download/category/21–1kk) collected with a hand-held digital camera

(Sony DSC-W730, Sony Corporation, Tokyo, Japan) in 2015. As a result, 345 sweet corn (*n*=322), dent (*n*=19), and check (*n*=4) inbred lines with a range from light to dark yellow endosperm color remained.

The levels of zeinoxanthin and compounds comprising 'other carotenes' were below the lower limit of detection for HPLC in two and 27 samples, respectively. Consequently, the values for these samples were approximated within each year by uniform random variables ranging from zero to the minimum detected value for a given carotenoid phenotype similar to as described by Owens et al. (2014). The imputation of missing data with this approach allowed for a maximal sample size to be used in the quantitative analysis of these two phenotypes.

The raw HPLC data of the 19 carotenoid phenotypes from the 345 inbred lines were assessed for normality and screened for significant outliers following the method of Baseggio et al. (2019). Briefly, the Box-Cox power transformation (Box and Cox, 1964) implemented with the "boxcox" function from the MASS package in R version 3.2.3 (R Core Team, 2015) was used with a simple linear model including genotype, year, set within year, block within set within year, and HPLC autosampler plate within year as fixed effects to select an optimal convenient lambda for each phenotype (Supplemental Table S2.1). Next, the full mixed linear model 1 of Baseggio et al. (2019) that estimated genetic effects separately from field design effects was fitted for each transformed phenotype in ASReml-R version 3.0 (Gilmour et al., 2009). The fitted full model included the following terms: grand mean, check, year, set within year, block within set within year, genotype (non-check line), interaction between genotype and year, HPLC autosampler plate within year, plot grid row within year, plot grid column within year, and residual error following a normal distribution with mean 0 and variance σ^2. All terms were modeled as random effects except for the grand mean and check term. For each phenotype, detected outliers were excluded

based on the Studentized deleted residuals (Neter et al., 1996) generated from the fitted mixed linear model.

An iterative mixed linear model fitting procedure was performed with the full model described above in ASReml-R version 3.0 (Gilmour et al., 2009) on each transformed, outlier screened phenotype as described previously (Baseggio et al., 2019). Briefly, terms fitted as random effects were tested with likelihood ratio tests (Littell et al., 2006), and those not significant at $\alpha = 0.05$ were removed from the model. This resulted in the selection of a final, best fitted model for each phenotype that was then used to generate a best linear unbiased predictor (BLUP) for each genotype. In total, 322 sweet corn inbred lines had BLUPs for at least one of the 19 carotenoid phenotypes, but six inbred lines known to possess *aeduwx* or *bt2* were removed, resulting in a data set of 316 inbred lines having endosperm mutations (*su1*, *su1se1*, *sh2*, and *su1sh2*) occuring at a higher frequency in the association panel.

Heritability ($\hat{h}_l^2$) on a line-mean basis (Holland et al., 2003; Hung et al., 2012) was estimated using the variance components from the best fitted model, and standard errors of the estimates were calculated using the delta method (Lynch and Walsh, 1998; Holland et al., 2003). For each pairwise comparison of phenotypes, Pearson's correlation coefficient (r) was used to assess the strength of association ($\alpha = 0.05$) between back-transformed BLUP (Supplemental Table S2.2) values using the 'cor.test' function in R.

DNA Extraction, Sequencing, and Genotyping

Of the retained 316 sweet corn inbred lines with BLUPs, 293 had available raw genotyping-by-sequencing (GBS) data that were generated as described previously (Baseggio et al., 2019). In brief, the GBS procedure of Elshire et al. (2011) with *Ape*KI was used to construct multiplexed libraries that were sequenced on an NextSeq 500 or Illumina HiSeq 2500 (Illumina

Incorporated, San Diego, CA, USA) at the Cornell Biotechnology Resource Center (Cornell University, Ithaca, NY, USA). All raw GBS sequencing data are available from the National Center of Biotechnology Information Sequence Read Archive under accession number SRP154923 and in BioProject under accession PRJNA482446.

The construction of the SNP marker data set for the quantitative genetic analysis of the carotenoid phenotypes followed that of Baseggio et al. (2019) with minor modifications. Briefly, the genotypes of single-nucleotide polymorphisms (SNPs) at 955,690 high confidence loci were called based on the raw GBS sequencing data using the production pipeline in TASSEL 5 GBSv1 with the ZeaGBSv2.7 Production TagsOnPhysicalMap file (available at panzea.org, accessed 19 Nov 2018) in B73 RefGen_v2 coordinates (Glaubitz et al., 2014). To increase the number of lines with both SNP marker and carotenoid data, we merged raw unimputed SNP genotype calls for an additional 16 sweet corn inbred lines from Romay et al., 2013 (ZeaGBSv27_publicSamples_rawGenos_AGPv2-150114.h5, available at panzea.org, accessed 19 Nov 2018) with those from this study prior to any SNP filtering steps. Initial filtering on the combined raw unimputed SNP data from the 309 inbred lines consisted of removing SNPs having a minor allele observed in only one line (singletons and doubletons) and retaining biallelic SNPs with a call rate greater than 10%. Additionally, heterozygous genotype calls with an allele balance score (lowest allele read depth/total read depth) less than 0.3 or greater than 0.7 were set to missing. When two or more samples per line were available, the SNP genotype calls from replicated samples were merged if the identical-by-state (IBS) values from all sample pairwise comparisons exceeded 0.99 as in Romay et al. (2013), and SNP genotypes were set to missing if discordant between replicated samples. If replicated samples had IBS values below

this conservative threshold, the sample with the highest SNP call rate was selected to represent the inbred line.

The near complete imputation of missing SNP genotypes was performed using FILLIN (Swarts et al., 2014) with an available set of maize haplotype donors having a window size of 4 kb (available at panzea.org, accessed 19 Nov 2018). Given that the imputation method is unable to impute all missing genotypes (Swarts et al., 2014), additional filtering was needed related to the remaining missing data. Even after imputation, an inbred line still had a SNP call rate less than 40%, thus it was excluded from further analysis. In Tassel 5 version 20180802, we used a set of filters to further enhance the quality of the imputed data set by removing SNPs with a call rate less than 70%, a minor allele frequency (MAF) lower than 5%, heterozygosity greater than 10%, an inbreeding coefficient lower than 80%, or a mean read depth greater than 15. The final, complete SNP marker data set consisted of 172,486 high-quality SNP markers scored on 308 sweet corn inbred lines having a BLUP value for one or more carotenoid phenotypes.

Genome-wide association study

To conduct a GWAS for each carotenoid phenotype, a univariate mixed linear model was used to test each of the 172,486 SNP markers for association with BLUP values from the 308 inbred lines (Supplemental Table S2.3) in the GEMMA software version 0.97 (Zhou and Stephens, 2014). The mixed linear model accounted for population stratification and familial relatedness by including principal components (PCs) (Price et al., 2006) and a genomic relationship (kinship) matrix based on VanRaden's method 1 (VanRaden, 2008) calculated in the R package GAPIT version 2017.08.18 (Lipka et al., 2012). The PCs and kinship were calculated based on 12,559 unimputed SNPs—a genome-wide subset of the complete marker data set—that had a call rate higher than 90%, MAF greater than 5%, heterozygosity less than 10%, inbreeding

coefficient greater than 80%, and mean read depth lower than 15. Missing genotypes remaining in both SNP marker data sets were conservatively imputed as heterozygous in GAPIT. The Bayesian information criterion (BIC) (Schwarz, 1978) based on the maximum likelihood estimates of model parameters from GEMMA was used to determine the optimal number of PCs to include as covariates in the mixed linear model. Similarly, the BIC was used to determine whether to also include endosperm mutation type (*su1*, *sh2*, or *su1sh2*), which had been previously scored on the 308 inbred lines by Baseggio et al. (2019), as a covariate in the mixed linear model. This is because *su1* and *sh2* could be strongly associated with endosperm carotenoids (Weber, 1987) as was shown for levels of tocotrienols—a class of tocochromanols mostly found in the endosperm (Grams et al., 1970; Weber, 1987)—in fresh sweet corn kernels from the same association panel (Baseggio et al., 2019).

The likelihood-ratio-based R^2 statistic (R^2_{LR}) of Sun et al. (2010) was used to approximate the amount of phenotypic variation explained by a mixed linear model with or without a significant SNP detected in GWAS. The R^2_{LR} value of each model was calculated with the maximum log-likelihood of the model of interest fitted in GEMMA compared to the maximum log-likelihood of an intercept-only model fitted with the 'lm' function in R. For each phenotype, *P*-values (Wald test) of SNPs tested in GEMMA were adjusted to control the false-discovery rate (FDR) at a level of 5% with the Benjamini–Hochberg multiple test correction (Benjamini and Hochberg, 1995) available in the 'p.adjust' function of R version 3.2.3 (R Core Team, 2015). To identify candidate genes, the search interval was limited to ± 250 kb of the physical position of SNP markers significantly associated with a carotenoid phenotype, which follows the distance at which genome-wide linkage disequilibrium (LD) decays to nominal levels in this association panel (Baseggio et al., 2019).

The multi-locus mixed-model (MLMM) approach of Segura et al. (2012) was used for exploring the potential existence of allelic heterogeneity and uncovering novel associations. We implemented the MLMM approach to control for the influence of major-effect loci on an individual chromosome basis as described previously (Lipka et al., 2013). The extended BIC (Chen and Chen, 2008) was used in the selection of the optimal model. The control of major-effect loci was also assessed by reconducting GWAS with MLMM-selected SNPs included as covariates in the mixed linear model of GEMMA.

Carotenoid prediction

The prospect of genomic selection for breeding sweet corn with increased carotenoids was assessed in the 308 inbred lines using a single kernel genomic best linear unbiased prediction (GBLUP) model (Zhang et al., 2007; VanRaden, 2008). In the R package GAPIT version 2017.08.18 (Lipka et al., 2012), method 1 from VanRaden (2008) was used to calculate genomic relationship matrices derived from three different SNP datasets varying in the number of markers: carotenoid QTL-targeted, pathway-level, and genome-wide. The carotenoid QTL-targeted set consisted of 628 SNPs within ± 250 kb of eight *a priori* identified candidate genes underlying QTL associated with carotenoid biosynthesis and retention, while the pathway-level set had 4,689 SNPs within ± 250 kb of 60 *a priori* candidate genes (including the 8 genes from the QTL-targeted set) involved in the biosynthesis and cleavage of carotenoids (Owens et al., 2014; Supplemental Table S2.4). The 172,486 high-quality SNP markers comprised the genome-wide set. These three genomic relationship matrices were used individually as a random effect for prediction of carotenoid phenotypes with the function ‘emmreml’ (single kernel) in the EMMREML R package (Akdemir and Okeke, 2015).

A five-fold cross-validation approach was used to estimate the predictive ability of a model for each carotenoid phenotype by calculating the Pearson's correlation between observed BLUP and genomic estimated breeding values as described in Baseggio et al. (2019). The predictive ability of each model was based on a mean of correlations from 50 replicates of the five-fold cross-validation scheme. Each fold consisted of genotype frequencies for endosperm mutants (*su1*, *sh2*, and *su1sh2*) that were representative of the association panel, and the identical cross-validation folds were used across different models. Similar to genomic prediction of tocochromanol traits (Baseggio et al., 2019), endosperm mutation type (*su1*, *sh2*, or *su1sh2*) was evaluated as a covariate in prediction models.

RESULTS

Phenotypic variation

We conducted a quantitative assessment of carotenoid levels in fresh (immature milk stage) kernels harvested from an association panel of 308 sweet corn inbred lines with endosperm color ranging from light to dark yellow. The measurement of carotenoids by HPLC revealed that lutein and zeaxanthin represented about 65% of total carotenoids in the kernel, while the other five carotenoid phenotypes individually accounted for less than 10% of the total (Table 2.1). The two specifically measured compounds with provitamin A activity, β-carotene and β-cryptoxanthin, had similar concentrations, and when summed only represented approximately 6% of total carotenoids. When separating inbred lines according to their endosperm mutation type, three (antheraxanthin, β-cryptoxanthin, and lutein) of the seven individual compounds had an average amount shown to be at a significantly ($P < 0.05$) greater level in the *sh2* ($n = 46$) group than in the *su1* ($n = 245$) group (Table 2.2).

With the exception of β-carotene and β-cryptoxanthin (r = 0.17), the BLUP values of each carotenoid compound had a Pearson's correlation stronger than 0.5 with those of its immediate precursor in the carotenoid pathway (Supplemental Fig. S2.1). Correlations between β-carotene and other compounds were very weak (r = -0.05 to 0.17). In contrast, β-cryptoxanthin had relatively much stronger correlations with all other xanthophyll compounds (r = 0.55 to 0.76) but violaxanthin (r = 0.18). The estimates of heritability on a line-mean basis for the 19 carotenoid compound, sum, and ratio traits ranged from 0.75 for antheraxanthin to 0.94 for the ratio of α- to β-xanthophylls, with an average of 0.84. Such high heritability estimates suggest that these phenotypes would be amenable to genetic dissection and prediction in this sweet corn association panel.

Table 2.1. Means and ranges for back-transformed best linear unbiased predictors (BLUPs) of 19 fresh kernel carotenoid traits evaluated in the sweet corn association panel and estimated heritability ($\hat{h}_l^2$) on a line-mean basis across two years.

		BLUPs			Heritabilities	
Trait	**Lines**	**Mean**	**SD†**	**Range**	**Estimate**	**SE‡**
		——— μg g^{-1} fresh weight ———				
Antheraxanthin	308	1.22	0.30	0.39-2.07	0.75	0.03
β-carotene	308	0.54	0.35	0.16-2.83	0.90	0.01
β-cryptoxanthin	308	0.47	0.29	0.11-2.29	0.86	0.02
Lutein	308	5.82	3.04	0.64-19.39	0.92	0.01
Violaxanthin	308	0.99	0.21	0.56-2.47	0.76	0.03
Zeaxanthin	308	4.84	1.75	1.62-10.71	0.81	0.02
Zeinoxanthin	308	1.20	1.04	0.06-8.88	0.90	0.01
Other carotenes	307	1.38	0.73	0.29-4.94	0.79	0.02
α-xanthophylls	308	7.10	3.86	0.73-28.23	0.92	0.01
β-xanthophylls	308	5.16	1.82	1.72-11.15	0.78	0.03
Total xanthophylls	308	14.43	4.73	5.54-38.23	0.78	0.03
Total carotenes	307	1.97	0.89	0.76-6.01	0.85	0.02
Total carotenoids	308	16.48	5.28	7.40-43.95	0.79	0.03
β-carotene/β-cryptoxanthin	308	1.41	1.11	0.34-8.17	0.79	0.02
β-carotene/	308	0.12	0.13	0.03-0.88	0.85	0.02

(β-cryptoxanthin+zeaxanthin)						
β-cryptoxanthin/zeaxanthin	308	0.09	0.03	0.03-0.22	0.85	0.02
Zeinoxanthin/Lutein	308	0.21	0.12	0.04-0.80	0.89	0.01
β-/α-xanthophylls	307	1.35	0.84	0.32-6.29	0.94	0.01
Total carotenes/xanthophylls	308	0.14	0.06	0.05-0.45	0.78	0.02

† Standard deviation of the BLUPs.
‡ Standard error of the heritabilities.

Table 2.2. Back-transformed estimated effects of endosperm mutation type for 19 fresh kernel carotenoid traits.

Trait	*su1*	*sh2*	*su1sh2*	*P*-value†
	—— μg g^{-1} fresh weight ——			
Antheraxanthin	1.15 b‡	1.32 a	1.25 ab	**0.003**
β-carotene	0.47	0.48	0.41	0.434
β-cryptoxanthin	0.38 b	0.51 a	0.48 ab	**0.002**
Lutein	4.52 b	8.57 a	5.90 b	**<0.0001**
Violaxanthin	0.97	0.98	1.03	0.419
Zeaxanthin	4.64	4.91	4.83	0.574
Zeinoxanthin	0.82	1.11	1.11	**0.036**
Other carotenes	1.19 b	1.91 a	1.57 a	**<0.0001**
α-xanthophylls	5.49 b	10.09 a	7.19 ab	**<0.0001**
β-xanthophylls	4.78	5.06	5.15	0.470
Total xanthophylls	13.00 b	17.33 a	15.40 ab	**<0.0001**
Total carotenes	1.76 b	2.58 a	2.07 ab	**<0.0001**
Total carotenoids	14.89 b	20.03 a	17.57 ab	**<0.0001**
β-carotene/β-cryptoxanthin	1.24 a	0.96 b	0.87 b	**0.002**
β-carotene/(β-cryptoxanthin+zeaxanthin)	0.10	0.09	0.08	0.299
β-cryptoxanthin/zeaxanthin	0.08 b	0.10 a	0.09 ab	**<0.0001**
Zeinoxanthin/Lutein	0.20 a	0.16 b	0.21 ab	**0.028**
β-/α-xanthophylls	1.30 a	0.76 b	0.92 b	**<0.0001**
Total carotenes/Total xanthophylls	0.14	0.14	0.13	0.621

† *P*-value from one-way ANOVA *F*-test for the endosperm mutation type effect. Bolded *P*-value indicates a statistically significant difference between two or more endosperm mutation type groups ($P < 0.05$).

‡ Sweet corn lines grouped by endosperm mutation type having labels with the same letter are not significantly different according to the Tukey-Kramer honest significant difference test ($P < 0.05$). The test was only performed for traits that had a significant *F*-test.

Genome-wide association study

The association panel of 308 sweet corn inbred lines having yellow endosperm kernels at the fresh-eating stage, which had been scored with 172,486 genome-wide SNP markers, was used to elucidate the genetic basis of natural variation for carotenoids in fresh kernels. Through the implementation of a univariate mixed linear model that accounted for population structure, relatedness, and type of endosperm mutation, we identified 108 unique SNPs that were significantly associated with from one to four phenotypes at a genome-wide FDR of 5%. The 108 SNPs were distributed across seven chromosomes, with the vast majority (92.59%) of them located on chromosomes 2, 8, and 10 (Supplemental Fig. S2.2).

The most significant association was identified for the ratio of β-carotene to β-cryptoxanthin+zeaxanthin on chromosome 10 (Fig. 2.2A). The peak SNP locus (S10_135801334; *P*-value 1.11 x 10^{-13}) for this association signal was located within the open reading frame (ORF) of a gene that encodes GRAS-transcription factor 22 (*grass22*, GRMZM2G173429), but ~255 kb away from *crtRB1* (*β-carotene hydroxylase* 1, GRMZM2G152135)--a gene encoding a nonheme dioxygenase that hydroxylates *β*-rings of carotenoids. This SNP was also the peak association for β-carotene (*P*-value 2.04 x 10^{-11}) and the ratio of β-carotene to β-cryptoxanthin (*P*-value 3.03 x 10^{-11}), while S10_135683780 had the strongest association with violaxanthin (*P*-value 7.94 x 10^{-7}). In total, 61 SNPs spanning a 5.09-Mb interval (133.7-138.8 Mb) on chromosome 10 were significantly associated with one or more of these four phenotypes (Supplemental Table S2.5), with 38 of the 61 associated SNPs (*P*-values 3.00 x 10^{-12} to 1.13 x 10^{-5}) located ± 250 kb of the ORF for *crtRB1*.

We used a chromosome-wide multi-locus mixed-model (MLMM) procedure to better resolve the complex of association signals within the 5.09-Mb region on chromosome 10. Each

optimal model for β-carotene, β-carotene/(β-cryptoxanthin+zeaxanthin), and β-carotene/β-cryptoxanthin included the peak SNP S10_135801334 (Supplemental Table S2.6). Additionally, the optimal models obtained for β-carotene and β-carotene/(β-cryptoxanthin+zeaxanthin) selected a second SNP (S10_136086332) that was located ~26 kb away from *crtRB1* and in very weak LD (r^2 = 0.11) with S10_135801334. Indicative of relatively weaker significant associations, no SNPs were selected by the MLMM for violaxanthin. When GWAS was reconducted with either one or two MLMM-selected SNPs, depending on the phenotype, included as covariates in the mixed linear model for β-carotene and its two derived phenotypes, all other signals at this 5.09-Mb segment and elsewhere on chromosome 10 were no longer significant at a genome-wide FDR of 5% (Fig. 2.2B). Additionally, 24 SNPs on chromosomes 1, 2, 3, and 8 that were associated with β-carotene/β-cryptoxanthin and/or β-carotene/(β-cryptoxanthin+zeaxanthin) were no longer significant. Conversely, only a single SNP (S6_58455321) from within the pericentromeric region of chromosome 6 remained significantly associated (*P*-value 1.83 x 10^{-7}) with β-carotene (Supplemental Fig. S2.3).

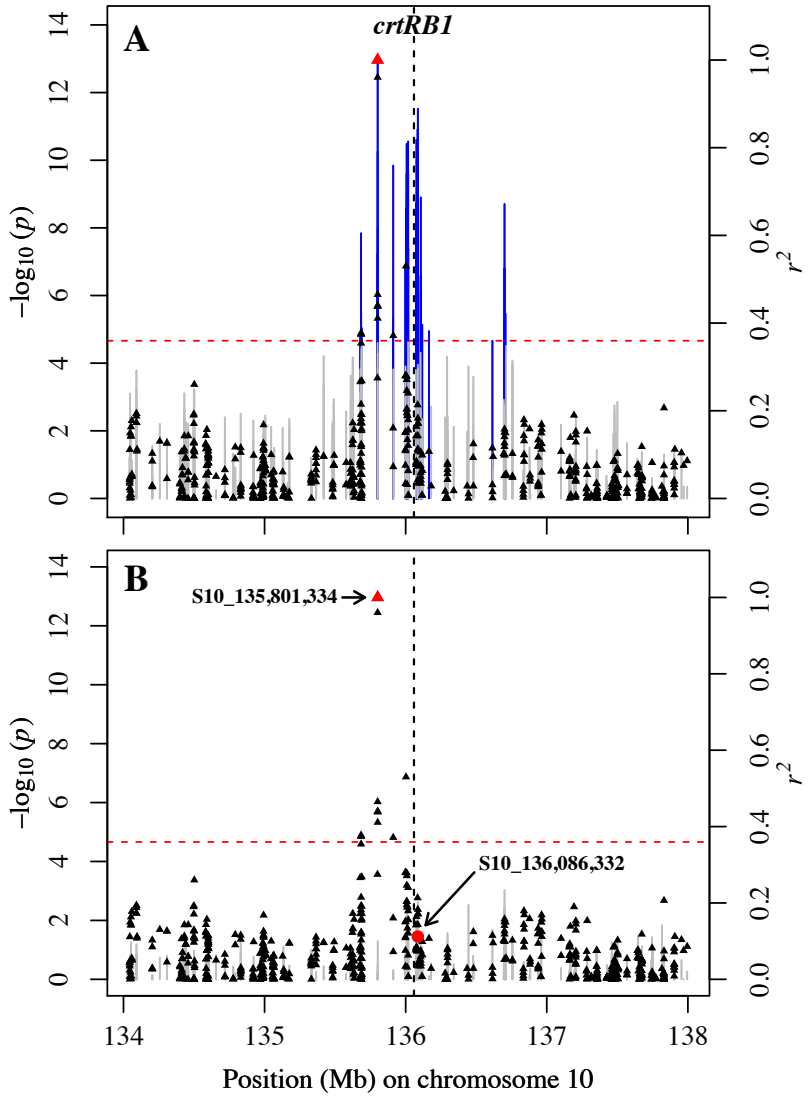

Figure 2.2. Genome-wide association study for the ratio of β-carotene to the sum of β-cryptoxanthin and zeaxanthin [β-carotene/(β-cryptoxanthin+zeaxanthin)] in fresh kernels of sweet corn. (A) Scatter plot of association results from a mixed model analysis and linkage disequilibrium (LD) estimates (r^2). The vertical lines are $-\log_{10}$ *P*-values of single nucleotide polymorphisms (SNPs) and blue color represents SNPs that are statistically significant at a 5% false discovery rate (FDR). Triangles are the r^2 values of each SNP relative to the peak SNP (indicated as a red triangle) at 135,801,334 bp (B73 RefGen_v2) on chromosome 10. The red horizontal dashed line indicates the $-\log_{10}$ *P*-value of the least statistically significant SNP at a 5% FDR. The black vertical dashed line indicates the genomic position of the *β-carotene hydroxylase 1* (*crtRB1*) gene. (B) Scatter plot of association results from a conditional mixed linear model analysis and LD estimates (r^2). The SNPs (S10_135801334, red triangle; and S10_136086332, red circle) from the optimal multi-locus mixed-model were included as covariates in the mixed linear model to control for the *crtRB1* effect.

The *lcyE* gene (GRMZM2G012966) on chromosome 8 had a SNP (S8_138888278) within its ORF that significantly associated with the ratio of β- to α-xanthophylls (Fig. 2.3A; *P*-value 1.01 x 10^{-12}). The *lcyE* gene encodes lycopene ε-cyclase, which has an enzymatic activity that influences flux down the α- versus β-branches of the carotenoid pathway (Cunningham et al., 1996). An additional nine SNPs spanning a 3.69-Mb interval on chromosome 8, including two SNPs located within the ORF of *lcyE* (S8_138888328 and S8_138888990; *P*-values 7.32 x 10^{-9} and 5.41 x 10^{-8}, respectively), as well as three SNPs from chromosome 9 were found to be associated with β-xanthophylls/α-xanthophylls. However, only the peak SNP S8_138888278 was selected in the optimal model obtained by the MLMM for the xanthophyll ratio phenotype (Supplemental Table S2.6). When the peak SNP was fitted as a covariate in the mixed linear model, all other SNPs from chromosomes 8 and 9 were no longer significantly associated with β-xanthophylls/α-xanthophylls at a 5% FDR (Fig. 2.3B). Interestingly, only one line each in the *sh2* and *su1sh2* endosperm mutation type groups were homozygous for the allele of the peak SNP associated with a larger average value of the β- to α-xanthophylls ratio (i.e., greater amount of β-xanthophylls), whereas the same SNP allele was found to be homozygous at a relatively higher frequency (12.7%) among *su1* lines (Supplemental Table S2.7).

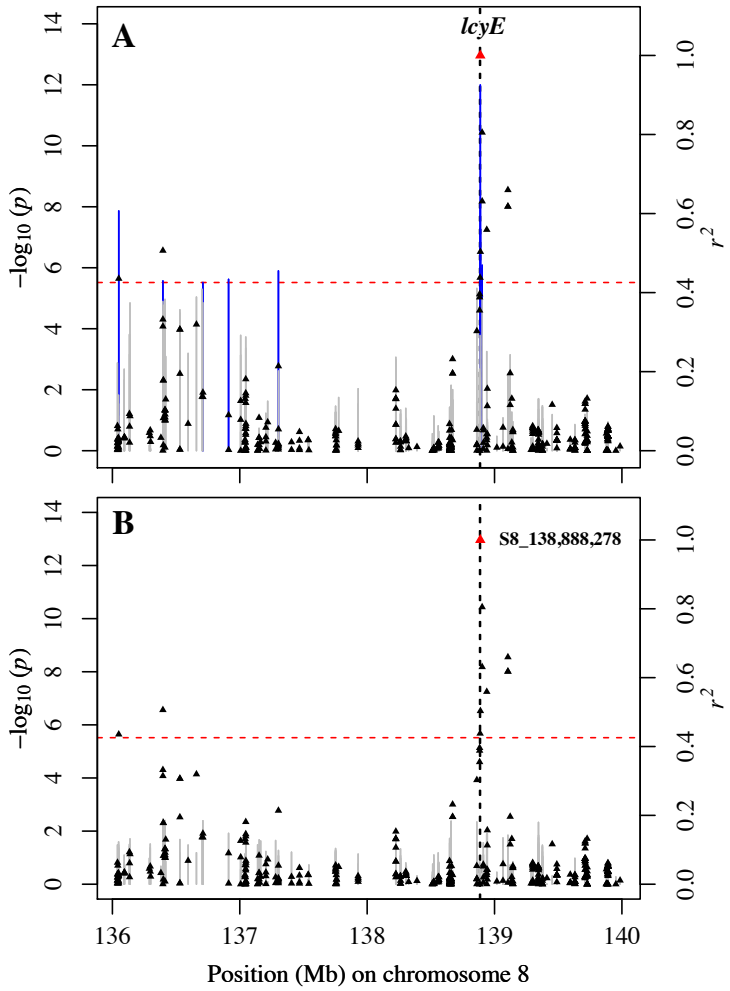

Figure 2.3. Genome-wide association study for the ratio of β- to α-xanthophylls in fresh kernels of sweet corn. (A) Scatter plot of association results from a mixed model analysis and linkage disequilibrium (LD) estimates (r^2). The vertical lines are $-\log_{10}$ *P*-values of single nucleotide polymorphisms (SNP) and blue color represents SNPs that are statistically significant at a 5% false discovery rate (FDR). Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 138,888,278 bp (B73 RefGen_v2) on chromosome 8. The red horizontal dashed line indicates the $-\log_{10}$ *P*-value of the least statistically significant SNP at a 5% FDR. The black vertical dashed line indicates the genomic position of the *lycopene ε-cyclase* gene (*lcyE*). (B) Scatter plot of association results from a conditional mixed linear model analysis and LD estimates (r^2). The SNP from the optimal multi-locus mixed-model (S8_138888278) was included as a covariate in the mixed linear model to control for the *lcyE* effect.

We detected seven SNPs covering a 748.29-Kb interval on chromosome 2 that were significantly associated with total xanthophylls. Of these SNPs, six of them were also associated with total carotenoids. The peak association signal for both traits was SNP S2_222880454 (*P*-values 2.74 x 10^{-8}, total xanthophylls; 1.15 x 10^{-7}, total carotenoids), which was located within the ORF of a putative seryl-tRNA synthetase (GRMZM2G172101). Through a MLMM analysis on a chromosome-wide level, this peak SNP (S2_222880454) was selected in the optimal model obtained for total xanthophylls and carotenoids. Notably, the peak SNP was found to be ~172 Kb from a gene encoding a member of the SWEET (Sugars Will Eventually be Exported Transporter) sucrose-efflux transporter family (*sweet14a*; GRMZM2G094955) that is strongly expressed in the endosperm of 16-24 DAP developing maize kernels (Stelpflug et al., 2016). In a conditional univariate mixed model analysis that included the peak SNP as a covariate, SNPs on chromosome 2 were no longer found to be associated with xanthophyll and carotenoid totals (Supplemental Fig. S2.4). This conditional analysis resulted in the detection of six additional SNPs on chromosome 1 (peak SNP S1_290953298; *P*-value 6.52 x 10^{-7}) significantly associated with total carotenoids at an FDR of 5% (Supplementary Table S2.8), with two of the six SNPs located within a gene that encodes an alkaline galactosidase (*aga4*, GRMZM2G077181). However, within ± 250 kb of these six SNPs there were no genes with an encoded protein that had an obvious role in the genetic control of kernel carotenoid levels.

Carotenoid prediction

To evaluate the potential of genomic selection for enhancing the level of carotenoids in fresh kernels, we evaluated whole-genome prediction (WGP) using the 172,486 SNP markers for all 19 carotenoid phenotypes that had been measured on the 308 sweet corn inbred lines. The average prediction ability across the 19 carotenoid phenotypes was 0.52, with a range in abilities

of 0.28 for violaxanthin to 0.73 for lutein (Table 2.3). The two measured compounds with provitamin A activity, β-carotene and β-cryptoxanthin, had moderately high prediction abilities of 0.49 and 0.61, respectively. A moderately strong positive correlation (r_{sp} = 0.47, *P*-value = 0.041) was found between heritability estimates and prediction abilities for the 19 carotenoid phenotypes. Conversely, predictive abilities had essentially no correlation (r_{sp} = -0.036, *P*-value = 0.883) with the number of significant markers detected in GWAS at 5% FDR.

Table 2.3. Predictive abilities of genomic prediction models for 19 fresh kernel carotenoid traits using three marker sets as predictors with or without endosperm mutation type.

	GBLUP†			GBLUP with endosperm mutation type covariate		
Trait	QTL targeted‡	Pathway- level§	Genome-wide¶	QTL targeted	Pathway- level	Genome-wide
Antheraxanthin	0.35 (0.03)	0.35 (0.03)	0.45 (0.02)	0.39 (0.02)	0.41 (0.02)	0.48 (0.02)
β-carotene	0.41 (0.03)	0.48 (0.03)	0.49 (0.04)	0.41 (0.03)	0.47 (0.04)	0.49 (0.04)
β-cryptoxanthin	0.36 (0.03)	0.49 (0.03)	0.61 (0.02)	0.39 (0.02)	0.53 (0.02)	0.63 (0.02)
Lutein	0.54 (0.02)	0.64 (0.02)	0.73 (0.01)	0.65 (0.01)	0.72 (0.01)	0.75 (0.01)
Violaxanthin	0.25 (0.03)	0.27 (0.03)	0.28 (0.03)	0.25 (0.02)	0.27 (0.03)	0.27 (0.03)
Zeaxanthin	0.24 (0.03)	0.35 (0.03)	0.42 (0.03)	0.24 (0.03)	0.36 (0.03)	0.44 (0.03)
Zeinoxanthin	0.42 (0.02)	0.49 (0.02)	0.54 (0.02)	0.43 (0.02)	0.50 (0.02)	0.54 (0.02)
Other carotenes	0.21 (0.03)	0.29 (0.03)	0.49 (0.02)	0.42 (0.02)	0.46 (0.02)	0.52 (0.02)
α-xanthophylls	0.53 (0.02)	0.64 (0.02)	0.71 (0.01)	0.63 (0.01)	0.70 (0.01)	0.73 (0.01)
β-xanthophylls	0.26 (0.03)	0.38 (0.03)	0.46 (0.03)	0.26 (0.03)	0.40 (0.03)	0.47 (0.03)
Total xanthophylls	0.41 (0.02)	0.56 (0.02)	0.67 (0.02)	0.52 (0.02)	0.64 (0.01)	0.70 (0.02)
Total carotenes	0.23 (0.03)	0.34 (0.03)	0.53 (0.02)	0.40 (0.02)	0.47 (0.03)	0.55 (0.02)
Total carotenoids	0.40 (0.02)	0.56 (0.02)	0.68 (0.02)	0.52 (0.02)	0.65 (0.01)	0.71 (0.02)
β-carotene/β-cryptoxanthin	0.48 (0.02)	0.51 (0.03)	0.54 (0.03)	0.48 (0.02)	0.53 (0.03)	0.55 (0.03)
β-carotene/(β-cryptoxanthin+Zeaxanthin)	0.44 (0.03)	0.46 (0.04)	0.42 (0.05)	0.43 (0.03)	0.46 (0.04)	0.42 (0.05)
β-cryptoxanthin/Zeaxanthin	0.35 (0.02)	0.44 (0.02)	0.54 (0.02)	0.40 (0.02)	0.50 (0.02)	0.55 (0.02)
Zeinoxanthin/Lutein	0.26 (0.03)	0.25 (0.03)	0.33 (0.03)	0.27 (0.03)	0.26 (0.03)	0.33 (0.03)
β-/α-xanthophylls	0.57 (0.02)	0.57 (0.02)	0.62 (0.02)	0.65 (0.01)	0.63 (0.02)	0.64 (0.01)
Total carotenes/Total xanthophylls	0.30 (0.03)	0.32 (0.03)	0.39 (0.04)	0.29 (0.04)	0.31 (0.03)	0.38 (0.04)
Average	0.37	0.44	0.52	0.42	0.49	0.53

† Genomic best linear unbiased prediction.
‡ 628 markers within ± 250 kb of eight *a priori* genes underlying joint-linkage quantitative trait loci associated with grain carotenoid biosynthesis and retention.
§ 4,689 markers within ± 250 kb of 60 *a priori* candidate genes.
¶ 172,486 genome-wide markers.

Carotenoid grain traits in maize show patterns of oligogenic inheritance (Wong et al., 2004; Chander et al., 2008; Kandianis et al., 2013), with variability for content and composition mostly under the genetic control of several moderate- to large-effect loci involved in the synthesis or cleavage of carotenoids (Owens et al., 2014). In that light, we evaluated the predictive ability of two marker data sets that included SNPs within ± 250 kb of eight candidate genes underpinning QTL associated with variation for carotenoid levels in maize grain (carotenoid QTL-targeted) or 60 candidate genes involved in carotenoid biosynthesis and retention in maize (pathway-level) (Supplemental Table S2.4). When compared to the genome-wide marker data set, on average, the prediction abilities of the 19 phenotypes were 15 and 8 percentage points lower for the carotenoid QTL-targeted and pathway-level marker sets, respectively (Table 2.3). The prediction ability of β-carotene with the pathway-level set was 7 percentage points higher than that of the QTL-targeted set, but was essentially equivalent to the predictive ability of the genome-wide marker set (0.49). In contrast, the decrease in prediction abilities for β-cryptoxanthin, lutein, zeaxanthin, and total carotenoids with the carotenoid QTL-targeted dataset ranged from 10-28 percentage points compared to abilities obtained with the pathway-level and genome-wide marker sets.

Given that there were significant differences in variation among endosperm mutation type groups for more than half of the carotenoid phenotypes (Table 2.2), we evaluated the extent to which prediction abilities would improve from the inclusion of a covariate for the type of endosperm mutation in prediction models that varied for marker coverage of the genome. On average, prediction ability across the 19 phenotypes only increased by a single percentage point when including the endosperm mutation type covariate (Table 2.3) in the WGP model. Illustrative of the impact of including this covariate for both less dense marker data sets, the

improvement in predictive abilities ranged from 5 to 21 percentage points for the eight phenotypes with a highly significant endosperm mutation type effect (P <0.0001; Table 2.2) when using the carotenoid QTL-targeted marker data set, whereas the improvement for the same phenotypes was a slightly narrower range of 6 to 17 percentage points with the pathway-level marker set. The improvement to prediction abilities across both reduced marker data sets were far more modest or negligible for the phenotypes with a weaker significant ($0.0001 < P < 0.05$; range: 0 to 6 percentage points) or non-significant ($P > 0.05$; range: -1 to 2 percentage points) endosperm mutation type effect.

DISCUSSION

The consumption of sweet corn enhanced for carotenoids, especially lutein and zeaxanthin, has the potential to help reduce the risk of AMD that is prevalent among the elderly of Western societies (Congdon et al., 2004; Friedman et al., 2004). Identified loci associated with the genetic control of carotenoid levels and genomic selection models optimized for predictive abilities could be used together to accelerate progress in breeding for higher levels of carotenoids in sweet corn at the fresh-eating stage. To establish a key step for biofortification efforts in sweet corn, we conducted a GWAS to elucidate the genetic basis of natural variation for 19 carotenoid phenotypes in fresh kernels with a range of light to dark yellow endosperm color from a panel of 308 inbred lines. Additionally, the prediction ability of genomic prediction models varying in marker densities and the genes they target were tested on the same set of carotenoid phenotypes to provide insights into the potential effectiveness of genomic selection. To our knowledge, this work is the most comprehensive quantitative genetic analysis of carotenoid variation in fresh sweet corn kernels.

The two most abundant carotenoids found in fresh kernels were lutein and zeaxanthin, which is consistent with the previously studied carotenoid profiles of yellow kernels from maize (dent/flint/sweet corn) inbred lines (Kurilich and Juvik, 1999; Owens et al., 2014). The sweet corn association panel showed a 30.3- and 6.61-fold range in variation for lutein and zeaxanthin, respectively, and average zeaxanthin was as high as the maximum content observed for 200 tropical sweet corn breeding lines (O'Hare et al., 2015). Through selection, the same authors initially increased zeaxanthin content to about 11 μg g^{-1} at eating stage, which is similar to the maximum observed in our panel (10.7 μg g^{-1}), and further increased up to 30.7 μg g^{-1} (Calvo-Brenes et al., 2019), demonstrating the feasibility of conventional breeding for specific carotenoid compounds. The maximum β-carotene content in our panel (2.83 μg g^{-1}) was lower than the reported by Fanning et al. (2010) for 385 tropical sweet corn breeding lines at 20 DAP (4.72 μg g^{-1}) selected for greater zeaxanthin levels. O'Hare et al. (2015) highlighted that conventional breeding may be used to increase both of these key carotenoids concurrently. Given that, we expect that genomic selection for favorable alleles of genes identified in this study may improve β-carotene and zeaxanthin levels even further in this panel.

The univariate GWAS of 19 carotenoid traits in fresh sweet corn kernels identified significant associations with SNP markers close to two core carotenoid pathway genes (*crtRB1* and *lcyE*), similar to other studies using mature dent corn kernels (Owens et al., 2014; Suwarno et al., 2015; Azmach et al., 2018). Although Owens et al. (2014) reported a relatively weak signal at GWAS for *crtRB1*, with significant associations (FDR < 5%) only with two indel markers, our panel detected 38 significant SNPs within 250 kb from the gene. Similar to that study, our data set did not include any SNP marker within the coding region of *crtRB1*, but markers up to 12.6 kb away from the gene in this panel were able to capture relevant variation

associated with β-carotene and its ratios, as opposed to Owens et al. (2014). When using a chromosome-wide MLMM approach to clarify the signals of these associations, two SNPs were retained in the optimum model (S10_135801334 and S10_136086332), suggesting variation at *crtRB1* is responsible for the accumulation of β-carotene in fresh sweet corn kernels and the presence of complex LD paterns in the region. An analysis of the haplotypes showed that only 15 lines have the most favorable alleles for both markers (TT), and they accumulate twice as much β-carotene than the most common haplotype (GC) present in 209 lines (Supplemental Table S2.7). Differently from Yan et al. (2010), which reported favorable *crtRB1* genotypes showed 8.9-48% reduction in total carotenoids compared to unfavorable alleles at that locus, the two SNP markers in the proximity of *crtRB1* identified in this study were not associated with total carotenoids (Supplemental Tabel S2.7). The detection of *crtRB1* for total carotenoids in that study, however, was only observed for segregating populations but not association panels, which have better phenotypic resolution (Yan et al., 2010).

Significant associations were also identified between SNP markers close to *lcyE* and the ratio of β- to α-xanthophylls. The signal was further resolved to a single SNP in the ORF of the gene and ~1 kb upstream of a 3' indel reported to have 3.3-fold change in flux between branches (Harjes et al., 2008), as opposed to two different SNPs being selected by MLMM for the Goodman–Buckler association panel (Owens et al., 2014). Allele distribution at this locus (S8_138888278; C/T) in the Goodman–Buckler panel was relatively balanced (0.53/0.47), whereas it was favorable to the β-branch in a panel of 130 diverse yellow maize inbreds developed for high β-carotene content (0.15/0.85; Azmach et al., 2018). In contrast, only ~11% of lines in our panel have the allele that is associated with a higher β- to α-xanthophylls ratio, suggesting that increasing the frequency of this allele by breeding could improve the

accumulation of compounds from the β-branch, such as β-carotene and zeaxanthin. Although combined *lcyE* and *crtRB1* effects were reported to be largely additive (Yan et al., 2010), *lcyE* was not associated with β-carotene in this panel, even after controlling for *crtRB1*. Weaker but significant associations on chromosome 8 were identified between the ratio β-carotene/β-cryptoxanthin and three SNPs at ~171.5 Mb, which are ~191 kb from a SNP marker (S8_171705574) previously associated with zeaxanthin and β-xanthophylls (Owens et al., 2014). Although observed in two independent panels, the signals in our panel were no longer significant after controlling for *crtRB1* by including the two MLMM-selected SNPs at chromosome 10, not supporting the presence of another gene associated with carotenoids in that region.

Taken together, our results show that most of the *sh2* lines (44 out of 46) have the allele that favors a flux through the α-branch. The presence of a strongly expressed *lut1* allele would lead to the greater accumulation of lutein observed in those lines, which is similar to the difference between *su1* and *sh2* lines reported for tocotrienols (Baseggio et al., 2019). That could not be confirmed due to low power to identify significant associations of *lut1* and other small-effect loci with lutein levels. Kurilich and Juvik (1999) also reported difference between endosperm mutation group types and IL451b isolines (*Su1*, *su1*, *sh2*, *su1se1*), and observed that *su1se1* lines were consistently lower for all carotenoids, which could not be assessed in our panel due to a lack of a reliable phenotype and higher coverage genotype needed for classifying lines as having the *se1* mutation. Therefore, these lines will need to be further screened with a specific primer for *se1* to confirm their endosperm mutation type before drawing any conclusions.

Significant SNPs at chromosome 2 associated with total xanthophylls and total carotenoids are in the same 960-kb region with long-range patterns of LD (mean $r^2 = 0.74$) where Baseggio et al. (2019) reported associations with total tocotrienols in a similar panel,

including a common significant SNP in both studies (S2_222219441). This region was associated with both total tocotrienols and total carotenoids, which are accumulated in the endosperm and are moderately correlated ($r^2 = 0.59$). This could indicate the presence of a gene regulating the production of precursors, particularly geranylgeranyl diphosphate (GGDP), which is a substrate for both classes of compounds, or just being a consequence of population structure not being controlled by the inclusion of either kinship or PC in the model. These hypotheses and the other signals on chromosomes 1 and 6 will need to be further investigated by a combination of approaches, such as gene expression profiling and mutagenesis, in order to assess the potential contribution of novel loci to the genetic variation of carotenoids and tocotrienols.

Predictive abilities of whole-genome prediction using GBLUP method for the carotenoid phenotypes assessed in this panel were moderate to moderately high, and were consistently higher than the values reported for 14 phenotypes also measured using dent corn kernels at mature stage of 201 highly diverse temperate and tropical lines from the Goodman–Buckler association (Owens et al., 2014). The higher relatedness between the lines in our panel as well as its slightly greater size ($n = 308$) compared to the Goodman–Buckler panel could explain the greater predictive abilities reported here (Clark et al., 2012). The moderately high predictive abilities, particularly for lutein (0.75) and β-carotene (0.53), suggest that genomic selection could be implemented to improve provitamin A and non-provitamin A carotenoids that are associated with the prevention of AMD in sweet corn breeding programs.

In agreement with Baseggio et al. (2019) using a similar panel for predicting tocochromanol levels, predictive abilities for the models using the pathway-level or carotenoid QTL-targeted marker data sets were lower than those from whole-genome prediction, which is in contrast to what Owens et al. (2014) reported for the same phenotypes but using a more diverse

panel. The lower predictive abilities are probably due to the low marker coverage and weak LD between the genotyped SNP markers and the causative variants. In fact, predictive abilities using the carotenoid QTL-targeted SNP set for the phenotypes directly related to the genes detected in the panel were only slightly lower than whole-genome prediction. As an example, the difference between the two SNP sets were 2 percentage points for the ratio of β- to α-xanthophylls (*lcyE* influences flux between α- and β-branches).

Although Owens et al. (2014) reported a strong correlation between prediction accuracies and number of significant markers in GWAS, this correlation was not significant in our panel, and even lutein, the carotenoid phenotype with the highest predictive ability, showed no significant marker association in GWAS. This could be due to the small power to detect lower-effect loci in our panel. Despite not passing our thresholds in GWAS, these signals were still captured by the genomic prediction models, resulting in higher predictive abilities. Transcriptional networks could also be integrated to identify groups of genes showing subthreshold associations with carotenoids, and these could be combined in prediction models to potentially improve the gains in predictive ability (Chan et al., 2011; Owens et al., 2014; Schaefer et al., 2018).

Phenotypic variability in our panel shows promise to genomic-breeding programs for improving the carotenoid profile in fresh kernels. When considering the consumption of fruits and vegetables as recommended by the 2015-2020 Dietary Guidelines for Americans, which would provide ~6 mg day^{-1} of lutein and zeaxanthin (US Department of Health and Human Services and US Department of Agriculture, 2015; reviewed in Mares, 2016), nine lines from our panel provide at least 30% and up to 48% with only 100 g of fresh sweet corn (one medium to large ear) (Supplemental Fig. S2.5). For vitamin A, the Institute of Medicine (2000) recommends

the intake of 700 and 900 μg day^{-1} of retinol activity equivalents (RAE) for adult women and men, respectively. Considering the ratios of conversion of β-carotene (12:1) and β-cryptoxanthin (24:1) to RAE (Institute of Medicine, 2000), our panel can provide up to 2.8% (men) and 3.6% (women) of the recommended daily allowance (RDA) for vitamin A with a 100-g intake of fresh kernels. Although this seems low, total carotenoids in our study were consistently higher than other panels (Kurilich and Juvik, 1999; Ibrahim and Juvik, 2009; Fanning et al., 2010). Given that all carotenoid phenotypes were highly heritable ($\hat{h}_l^2$ = 0.75-0.94), selection for favorable alleles at key steps in the pathway could be targeted to increase specific compounds, particularly β-carotene (highest provitamin A efficiency), as already done by breeding programs for developing countries using dent corn (Azmach et al., 2018). Breeding and selection for lines with a weak allele at *lcyE* that favors the flux to the β-branch could significantly increase both β-carotene and zeaxanthin, which has been demonstrated successful for tropical sweet corn lines (O'Hare et al., 2015). Additionally, the higher levels of zeaxanthin are associated with darker orange kernels (O'Hare et al., 2015) and would enable more nutritive sweet corn lines to be visually-distinguishable and might serve as a marketing tool (O'Hare et al., 2014).

CONCLUSIONS

Our study showed that natural variation of β-carotene, a provitamin A carotenoid, in fresh sweet corn kernels is associated with two SNPs on chromosome 10, one of which is within *crtRB1*. Although rare, we identified 13 *su1* and two *sh2* lines with the favorable alleles associated with increased β-carotene at both loci and these could be used as donors in a sweet corn breeding program for provitamin A biofortification. Variation between carotenoids from the α- and β-branches was predominantly under control of *lcyE* with most of the lines having the allele that favors carotenoids from the α-branch (α-carotene, zeinoxanthin, and lutein). Starch

biosynthesis genes were found to associate with carotenoid phenotypes, similar to tocotrienols (Baseggio et al., 2019), and additional approaches are needed to further characterize a potential connection. Also, targeted resequencing of the identified genes (*crtRB1* and *lcyE*), as well as other undetected loci, is needed to better assess the genetic variation present at the sweet corn germplasm. Finally, prediction models using genome-wide markers showed promise for selecting best performing sweet corn lines for lutein, zeaxanthin, and provitamin A carotenoids to be used in a genomic-assisted breeding program, which in turn may provide improved sweet corn germplasm for human health.

SUPPLEMENTAL INFORMATION

Supplemental Table S2.1. Lambda values used in Box-Cox transformation of 19 fresh kernel carotenoid traits in sweet corn.

Supplemental Table S2.2. Back-transformed best linear unbiased predictors of the 19 fresh kernel carotenoid traits used for the genome-wide association study and genomic prediction for 308 sweet corn inbred lines.

Supplemental Table S2.3. Transformed best linear unbiased predictors of the 19 fresh kernel carotenoid traits used for the genome-wide association study and genomic prediction for 308 sweet corn inbred lines.

Supplemental Table S2.4. Genomic information for the 60 *a priori* candidate genes.

Supplemental Table S2.5. Statistically significant results from a genome-wide association study of 19 fresh kernel carotenoid traits in sweet corn.

Supplemental Table S2.6. Multi-locus mixed-model results from an analysis of carotenoid traits for chromosomes 2, 8 and 10.

Supplemental Table S2.7. Back-transformed effect estimates for *crtRB1* and *lcyE*-related SNPs selected with an optimal multi-locus mixed-model.

Supplemental Table S2.8. Statistically significant results from a genome-wide association study of fresh kernel carotenoid traits in sweet corn when using the SNPs selected by multi-locus mixed-models as covariates in the mixed linear model.

Supplemental Fig. S2.1. Correlation matrix for BLUPs of the 19 carotenoid traits from fresh kernels in sweet corn.

Supplemental Fig. S2.2. Genome-wide association study of 19 fresh kernel carotenoid traits in sweet corn.

Supplemental Fig. S2.3. Genome-wide association study for β-carotene in fresh kernels of sweet corn.

Supplemental Fig. S2.4. Genome-wide association study for total xanthophylls in fresh kernels of sweet corn.

Supplemental Fig. S2.5. Distribution of the percentage of the recommended daily allowance (RDA) for vitamin A (retinol activity equivalent), lutein + zeaxanthin, and zeaxanthin provided by inbred lines from the sweet corn association panel.

REFERENCES

Akdemir, D. and U.G. Okeke. 2015. EMMREML: Fitting mixed models with known covariance structures. https://CRAN.R-project.org/package=EMMREML (accessed 6 April 2019).

Alder, A., M. Jamil, M. Marzorati, M. Bruno, M. Vermathen, P. Bigler, et al. 2012. The path from β-carotene to carlactone, a strigolactone-like plant hormone. Science 335: 1348-1351. doi:10.1126/science.1218094.

Azmach, G., A. Menkir, C. Spillane and M. Gedil. 2018. Genetic loci controlling carotenoid biosynthesis in diverse tropical maize lines. G3 (Bethesda) 8: 1049-1065. doi:10.1534/g3.117.300511.

Baseggio, M., M. Murray, M. Magallanes-Lundback, N. Kaczmar, J. Chamness, E.S. Buckler, et al. 2019. Genome-wide association and genomic prediction models of tocochromanols in fresh sweet corn kernels. Plant Genome 12. doi:10.3835/plantgenome2018.06.0038.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Stat. Soc. Ser. B. (Stat. Method.) 57: 289-300.

Bone, R.A., J.T. Landrum, G.W. Hime, A. Cains and J. Zamor. 1993. Stereochemistry of the human macular carotenoids. Invest Ophthalmol Vis Sci 34: 2033-2040.

Box, G.E.P. and D.R. Cox. 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B Stat. Soc. 26: 211-252.

Buckner, B. 1990. Cloning of the *y1* locus of maize, a gene involved in the biosynthesis of carotenoids. The Plant Cell 2: 867-876. doi:10.1105/tpc.2.9.867.

Burt, A.J., C.M. Grainger, M.P. Smid, Barry J. Shelp and E.A. Lee. 2011. Allele mining of exotic maize germplasm to enhance macular carotenoids. Crop Sci. 51: 991–1004.

Calvo-Brenes, P., K. Fanning and T. O'Hare. 2019. Does kernel position on the cob affect zeaxanthin, lutein and total carotenoid contents or quality parameters, in zeaxanthin-biofortified sweet-corn? Food Chem. 277: 490-495. doi:10.1016/j.foodchem.2018.10.141.

Chan, E.K.F., H.C. Rowe, J.A. Corwin, B. Joseph and D.J. Kliebenstein. 2011. Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. PLoS Biol. 9: e1001125. doi:10.1371/journal.pbio.1001125.

Chander, S., Y.Q. Guo, X.H. Yang, J. Zhang, X.Q. Lu, J.B. Yan, et al. 2008. Using molecular markers to identify two major loci controlling carotenoid contents in maize grain. Theoretical and Applied Genetics 116: 223-233. doi:10.1007/s00122-007-0661-7.

Chen, J. and Z. Chen. 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95: 759-771. doi:10.1093/biomet/asn034.

Chew, E.Y., T.E. Clemons, J.P. SanGiovanni, R.P. Danis, F.L. Ferris, M.J. Elman, et al. 2014. Secondary analyses of the effects of lutein/zeaxanthin on rge-Related macular degeneration progression. JAMA Ophthalmology 132: 142. doi:10.1001/jamaophthalmol.2013.7376.

Clark, S.A., J.M. Hickey, H.D. Daetwyler and J.H. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44: 4. doi:10.1186/1297-9686-44-4.

Congdon, N., B. O'Colmain, C.C.W. Klaver, Ronald Klein, B. Munoz, D.S. Friedman, et al. 2004. Causes and prevalence of visual impairment among adults in the United States. Archives of Ophthalmology 122: 477. doi:10.1001/archopht.122.4.477.

Cunningham, F.X., Jr., B. Pogson, Z. Sun, K.A. McDonald, D. DellaPenna and E. Gantt. 1996. Functional analysis of the beta and epsilon lycopene cyclase enzymes of *Arabidopsis* reveals a mechanism for control of cyclic carotenoid formation. Plant Cell 8: 1613-1626. doi:10.1105/tpc.8.9.1613.

Cuttriss, A.J., C.I. Cazzonelli, E.T. Wurtzel and B.J. Pogson. 2011. Carotenoids. 58: 1-36. doi:10.1016/b978-0-12-386479-6.00005-6.

DellaPenna, D. and B.J. Pogson. 2006. Vitamin synthesis in plants: Tocopherols and carotenoids. Annu. Rev. Plant Biol. 57: 711-738. doi:10.1146/annurev.arplant.56.032604.144301.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6: e19379. doi:10.1371/journal.pone.0019379.

Fanning, K.J., I. Martin, L. Wong, V. Keating, S. Pun and T. O'Hare. 2010. Screening sweetcorn for enhanced zeaxanthin concentration. J. Sci. Food Agric. 90: 91-96. doi:10.1002/jsfa.3787.

Flores-Pérez, Ú., J. Pérez-Gil, M. Closa, L.P. Wright, P. Botella-Pavía, M.A. Phillips, et al. 2010. PLEIOTROPIC REGULATORY LOCUS 1 (*PRL1*) integrates the regulation of sugar responses with isoprenoid metabolism in *Arabidopsis*. Molecular Plant 3: 101-112. doi:10.1093/mp/ssp100.

Friedman, D.S., B.J. O'Colmain, B. Muñoz, S.C. Tomany, C. McCarty, P.T. de Jong, et al. 2004. Prevalence of age-related macular degeneration in the United States. Arch Ophthalmol. 122: 564-572.

Fu, Z., Y. Chai, Y. Zhou, X. Yang, M.L. Warburton, S. Xu, et al. 2013. Natural variation in the sequence of PSY1 and frequency of favorable polymorphisms among tropical and temperate maize germplasm. Theor Appl Genet, 126: 923-935. doi:10.1007/s00122-012-2026-0.

Gilmour, A.R., B.J. Gogel, B.R. Cullis and T. R. 2009. ASReml user guide release 3.0.

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, et al. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9: e90346. doi:10.1371/journal.pone.0090346.

Grams, G.W., C.W. Blessin and G.E. Inglett. 1970. Distribution of tocopherols within the corn kernel. J. Am. Oil Chem.' Soc. 47: 337-339. doi:10.1007/bf02638997.

Harjes, C.E., T.R. Rocheford, L. Bai, T.P. Brutnell, C.B. Kandianis, S.G. Sowinski, et al. 2008. Natural genetic variation in *lycopene epsilon cyclase* tapped for maize biofortification. Science 319: 330-333. doi:10.1126/science.1150255.

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

Horner, H.T., R.A. Healy, G. Ren, D. Fritz, A. Klyne, C. Seames, et al. 2007. Amyloplast to chromoplast conversion in developing ornamental tobacco floral nectaries provides sugar for nectar and antioxidants for protection. Am J Bot 94: 12-24. doi:10.3732/ajb.94.1.12.

Hung, H.Y., C. Browne, K. Guill, N. Coles, M. Eller, A. Garcia, 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. doi:10.1038/hdy.2011.103.

Ibrahim, K.E. and J.A. Juvik. 2009. Feasibility for improving phytonutrient content in vegetable crops using conventional breeding strategies: case study with carotenoids and tocopherols in sweet corn and broccoli. J Agric Food Chem 57: 4636-4644. doi:10.1021/jf900260d.

Iglesias, D.J., F.R. Tadeo, F. Legaz, E. Primo-Millo and M. Talon. 2001. In vivo sucrose stimulation of colour change in citrus fruit epicarps: Interactions between nutritional and hormonal signals. Physiol. Plant. 112: 244-250.

Institute of Medicine. 2000. Dietary reference intakes for vitamin C, vitamin E, selenium, and carotenoids. The National Academies Press, Washington, DC.

Jahns, P. and A.R. Holzwarth. 2012. The role of the xanthophyll cycle and of lutein in photoprotection of photosystem II. Biochim. Biophys. Acta 1817: 182-193. doi:10.1016/j.bbabio.2011.04.012.

Kandianis, C.B., R. Stevens, W. Liu, N. Palacios, K. Montgomery, K. Pixley, et al. 2013. Genetic architecture controlling variation in grain carotenoid composition and concentrations in two maize populations. Theoretical and Applied Genetics 126: 2879-2895. doi:10.1007/s00122-013-2179-5.

Kermode, A.R. 2005. Role of abscisic acid in seed dormancy. J. Plant Growth Regul. 24: 319-344. doi:10.1007/s00344-005-0110-2.

Klein, R., C.-F. Chou, B.E.K. Klein, X. Zhang, S.M. Meuer and J.B. Saaddine. 2011. Prevalence of age-related macular degeneration in the US population. Archives of Ophthalmology 129: 75. doi:10.1001/archophthalmol.2010.318.

Krinsky, N.I., J.T. Landrum and R.A. Bone. 2003. Biologic mechanisms of the protective role of lutein and zeaxanthin in the eye. Annu. Rev. Nutr. 23: 171-201. doi:10.1146/annurev.nutr.23.011702.073307.

Kundu, S. and S. Gantait. 2017. Abscisic acid signal crosstalk during abiotic stress response. Plant Gene 11: 61-69. doi:10.1016/j.plgene.2017.04.007.

Kurilich, A.C. and J.A. Juvik. 1999. Quantification of carotenoid and tocopherol antioxidants in *Zea mays*. J Agric Food Chem 47: 1948-1955.

Li, L., Y. Yang, Q. Xu, K. Owsiany, R. Welsch, C. Chitchumroonchokchai, et al. 2012. The *Or* gene enhances carotenoid accumulation and stability during post-harvest storage of potato tubers. Molecular Plant 5: 339-352. doi:10.1093/mp/ssr099.

Lipka, A.E., M.A. Gore, M. Magallanes-Lundback, A. Mesberg, H. Lin, T. Tiede, et al. 2013. Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. G3 (Bethesda) 3: 1287-1299. doi:10.1534/g3.113.006148.

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, et al. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397-2399. doi:10.1093/bioinformatics/bts444.

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, N.C. p. 733-756.

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

Mares, J. 2016. Lutein and zeaxanthin isomers in eye health and disease. Annu. Rev. Nutr. 36: 571-602. doi:10.1146/annurev-nutr-071715-051110.

Meuwissen, T.H.E., B.J. Hayes and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

National Center for Health Statistics. 2018. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey: For the vision & eye health surveillance system. http://www.norc.org/PDFs/VEHSS/NHANESDataSummaryReportVEHSS.pdf (accessed 5 April 2019).

National Health and Nutrition Examination Survey. 2016. What we eat in America, NHANES 2013-2014. https://www.ars.usda.gov/ARSUserFiles/80400530/pdf/1516/Table_1_NIN_GEN_15.pdf (accessed 6 April 2019).

Neter, J., M.H. Kutner, C.J. Nachtsheim and W. Wasserman. 1996. Applied linear statistical models. McGraw-Hill, Boston.

O'Hare, T.J., K.J. Fanning and I.F. Martin. 2015. Zeaxanthin biofortification of sweet-corn and factors affecting zeaxanthin accumulation and colour change. Arch Biochem Biophys 572: 184-187. doi:10.1016/j.abb.2015.01.015.

O'Hare, T.J., I. Martin, K.J. Fanning, S. Kirchhoff, L.S. Wong, V. Keating, et al. 2014. Sweetcorn colour change and consumer perception associated with increasing zeaxanthin for the amelioration of age-related macular degeneration. Acta Horticulturae: 221-226. doi:10.17660/ActaHortic.2014.1040.30.

Owens, B.F., A.E. Lipka, M. Magallanes-Lundback, T. Tiede, C.H. Diepenbrock, C.B. Kandianis, 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. doi:10.1534/genetics.114.169979.

Palaisa, K., M. Morgante, S. Tingey and A. Rafalski. 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. Proc. Natl. Acad. Sci. USA 101: 9885-9890. doi:10.1073/pnas.0307839101.

Pixley, K., N.P. Rojas, R. Babu, R. Mutale, R. Surles and E. Simpungwe. 2013. Biofortification of maize with provitamin A carotenoids. 271-292. doi:10.1007/978-1-62703-203-2_17.

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904-909. doi:10.1038/ng1847.

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

Rein, D.B., J.S. Wittenborn, X. Zhang, A.A. Honeycutt, S.B. Lesesne, J. Saaddine, et al. 2009. Forecasting age-related macular degeneration through the year 2050: the potential impact of new treatments. Arch Ophthalmol 127: 533-540. doi:10.1001/archophthalmol.2009.58.

Rodriguez-Amaya, D.B. 2001. A guide to carotenoid analysis in foods. ILSI Human Nutrition Institute, Washington, D.C.

Romay, M.C., M.J. Millard, J.C. Glaubitz, J.A. Peiffer, K.L. Swarts, T.M. Casstevens, et al. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14: R55. doi:10.1186/gb-2013-14-6-r55.

Schaefer, R.J., J.-M. Michno, J. Jeffers, O. Hoekenga, B. Dilkes, I. Baxter, et al. 2018. Integrating coexpression networks with GWAS to prioritize causal genes in maize. The Plant Cell 30: 2922-2942. doi:10.1105/tpc.18.00299.

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

Segura, V., B.J. Vilhjalmsson, A. Platt, A. Korte, U. Seren, Q. Long, et al. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44: 825-830. doi:10.1038/ng.2314.

Song, J., D. Li, M. He, J. Chen and C. Liu. 2015. Comparison of carotenoid composition in immature and mature grains of corn (*Zea mays* L.) varieties. Int. J. Food Prop. 19: 351-358. doi:10.1080/10942912.2015.1031245.

Stelpflug, S.C., R.S. Sekhon, B. Vaillancourt, C.N. Hirsch, C.R. Buell, N. de Leon, et al. 2016. An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. Plant Gen. 9. doi:10.3835/plantgenome2015.04.0025.

Sun, G., C. Zhu, M.H. Kramer, S.S. Yang, W. Song, H.P. Piepho, et al. 2010. Variation explained in mixed-model association mapping. Heredity (Edinb) 105: 333-340. doi:10.1038/hdy.2010.11.

Suwarno, W.B., K.V. Pixley, N. Palacios-Rojas, S.M. Kaeppler and R. Babu. 2015. Genome-wide association analysis reveals new targets for carotenoid biofortification in maize. Theoretical and Applied Genetics 128: 851-864. doi:10.1007/s00122-015-2475-3.

Swarts, K., H. Li, J.A.R. Navarro, D. An, M.C. Romay, S. Hearne, et al. 2014. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Gen. 7: 1-12.

Tracy, W.F. 1997. History, genetics, and breeding of supersweet (*shrunken2*) sweet corn. 189-236. doi:10.1002/9780470650073.ch7.

US Department of Health and Human Services and US Department of Agriculture. 2015. 2015 – 2020 Dietary guidelines for Americans. https://health.gov/dietaryguidelines/2015/guidelines/ (accessed 8 April 2019).

USDA. 2018. Vegetables 2017 summary. National Agricultural Statistics Service, http://usda.mannlib.cornell.edu/usda/nass/VegeSumm//2010s/2018/VegeSumm-02-13-2018.pdf (accessed 12 June 2018).

Vallabhaneni, R. and E.T. Wurtzel. 2009. Timing and biosynthetic potential for carotenoid accumulation in genetically diverse germplasm of maize. Plant Physiol. 150: 562-572. doi:10.1104/pp.109.137042.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414-4423. doi:10.3168/jds.2007-0980.

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

Wong, J.C., R.J. Lambert, E.T. Wurtzel and T.R. Rocheford. 2004. QTL and candidate genes *phytoene synthase* and *zeta-carotene desaturase* associated with the accumulation of carotenoids in maize. Theor. Appl. Genet. 108: 349-359. doi:10.1007/s00122-003-1436-4.

Wong, W.L., X. Su, X. Li, C.M. Cheung, R. Klein, C.Y. Cheng, et al. 2014. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob Health 2: e106-116. doi:10.1016/S2214-109X(13)70145-1.

Wu, J., E. Cho, W.C. Willett, S.M. Sastry and D.A. Schaumberg. 2015. Intakes of lutein, zeaxanthin, and other carotenoids and age-related macular degeneration during 2 decades of prospective follow-up. JAMA Ophthalmology 133: 1415. doi:10.1001/jamaophthalmol.2015.3590.

Yan, J., C.B. Kandianis, C.E. Harjes, L. Bai, E.-H. Kim, X. Yang, et al. 2010. Rare genetic variation at *Zea mays crtRB1* increases β-carotene in maize grain. Nat. Genet. 42: 322-327. doi:10.1038/ng.551.

Yang, R., Z. Yan, Q. Wang, X. Li and F. Feng. 2018. Marker-assisted backcrossing of *lcyE* for enhancement of proA in sweet corn. Euphytica 214. doi:10.1007/s10681-018-2212-5.

Zhang, Z., R.J. Todhunter, E.S. Buckler and L.D. Van Vleck. 2007. Technical note: use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. J. Anim. Sci. 85: 881-885. doi:10.2527/jas.2006-656.

Zhou, X. and M. Stephens. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods 11: 407-409. doi:10.1038/nmeth.2848.

CHAPTER 3

GENETIC ANALYSES OF THE KERNEL IONOME FROM FRESH SWEET CORN [3]

INTRODUCTION

Insufficient consumption of essential micronutrients, including iron and zinc, may cause disorders ranging from mild anemia to neuropsychologic impairment (Sandstead, 2000; Miller, 2013). Although nutrient deficiency is far less prevalent in developed nations, approximately 10 million people are iron deficient in the US (Miller, 2013), and this affects mainly women, children, and adults older than 65 years (Guralnik, 2004; CDC, 2006; Clark, 2008). Inadequate iron intake is an important contributing factor in the development of iron deficiency and anemia (Clark, 2008). The US recommended dietary allowance for iron is 8 and 18 mg d^{-1} for men and women, respectively, yet the median intake by men is 16-18 mg d^{-1}, whereas for women is only 12 mg d^{-1} (Institute of Medicine, 2001). Zinc deficiency is more difficult to quantify, but it is estimated that 17.3% of the global population is at risk due to dietary inadequacy (Wessells and Brown, 2012).

Dietary surveys have estimated that 9 and 14% of US adults consume magnesium and potassium, respectively, at quantities below the recommended daily intake, and are therefore at risk of deficiency (Broadley and White, 2010). Conversely, significant decline of mean concentrations of iron, copper, potassium, and calcium in horticultural produce in the US has been reported since the mid twentieth century (Davis et al., 2004; White and Broadley, 2015). Sweet corn, the third most commonly consumed vegetable in the US (USDA, 2018b), usually

[3] Baseggio, M., M. Murray, G. Ziegler, N. Kaczmar, J. Chamness, E.S. Buckler, M.E. Smith, I. Baxter, W.F. Tracy, and M.A. Gore. To be submitted to *G3*

does not significantly contribute to the recommended daily amount of these nutrients (USDA, 2018a).

In plants, at least 17 elements are known to be essential to development (Taiz and Zeiger, 2010). In addition to carbon, hydrogen, and oxygen, which are derived from air and water, six other elements (nitrogen, phosphorus, potassium, calcium, magnesium, and sulfur) are required in large amounts, and these are called macronutrients. In smaller amounts, iron, manganese, boron, molybdenum, copper, zinc, chlorine, and cobalt are supplied by the soil and called micronutrients. Complex physiological and biochemical processes drive elemental uptake and accumulation in plants, and these processes are affected by physical, physiological, genetic, and environmental factors (Oktem et al., 2010; Asaro et al., 2016). Also, the full complement of mineral nutrients, called the ionome, varies as an integrated network (reviewed in Baxter, 2015). Understanding the genetic mechanisms of nutrient accumulation in the kernel, which represents the end point of a plant's development and is an important source of nutrients for humans, will support the breeding of more efficient plants with increased nutrient levels.

Metabolic pathways from absorption to accumulation have been characterized in plants, identifying several gene families, including nicotianamine synthase (*nas*; Zhou et al., 2013), zinc-regulated transporters (*zip*; Li et al., 2013), nicotianamine aminotransferase (*naat*; Kanazawa et al., 1994), deoxymugineic acid synthase (*dmas*; Bashir et al., 2006), natural resistance-associated macrophage protein (*nramp*; Jin et al., 2015), yellow stripe (*ys1;* Curie et al., 2001), and heavy-metal ATPase (*hma*; Sasaki et al., 2014). Jin et al. (2015) reported two genes encoding NRAMP that were considered to be candidates for natural accumulation of iron and zinc in maize kernels. Additionally, multiple loci that influence kernel nutrient bioavailability have previously been identified in maize (Lung'aho et al., 2011). Šimić et al.

(2012) detected QTL associated with the ratios of zinc, iron, and magnesium to phosphorus concentrations close to phytase genes, which breaks down phytate and releases bioavailable zinc, iron, magnesium, and phosphorus. Analysis of the genetic control of the fresh kernel ionome in sweet corn is still needed.

In this study, we assessed the natural quantitative variation for 15 elements in fresh (~21 DAP) kernels of a sweet corn association panel to determine the genetic controllers of elemental accumulation. We also developed whole-genome prediction models to enhance nutrient levels in fresh sweet corn and provide insights into the expected genetic gains in a sweet corn biofortification program.

MATERIALS AND METHODS

Plant Materials and Experimental Design

An association panel of 422 sweet corn inbred lines that samples the genetic diversity of sweet corn public breeding programs in the US was evaluated in 2014 and 2015 at Cornell University's Musgrave Research Farm in Aurora, NY and at the University of Wisconsin's West Madison Research Station in Verona, WI. The sweet corn lines included in this panel are homozygous for the following starch-deficient endosperm mutations: *sugary1* (*su1*), *sugary1:sugary enhancer1* (*su1se1*), *shrunken2* (*sh2*), *sugary1:shrunken2* (*su1sh2*), *brittle2* (*bt2*), and *amylose-extender:dull:waxy* (*aeduwx*) (Baseggio et al., 2019). The association panel was grown in an incomplete block design augmented with two (out of four) repeated check plots, and fresh kernel samples were harvested at 400 growing degree days (~21 DAP) as described previously (Baseggio et al., 2019). Kernel samples from 1559 plots were lyophilized for four days and shipped to Donald Danforth Plant Science Center (St. Louis, MO) for ionome profile assessment.

Phenotypic Data Analysis

The elemental analysis of single kernels was conducted using inductively coupled plasma mass spectrometry (ICP-MS) as described in Baxter et al. (2014). The 15 elements measured were boron, cadmium, calcium, copper, iron, magnesium, manganese, molybdenum, nickel, phosphorus, potassium, rubidium, sulfur, and zinc. Because the absorption of iron, zinc, and manganese is inhibited by phytate (reviewed in White and Broadley, 2009), and phosphorous concentration in the kernel is an indicator of phytate level (Raboy, 1997), ratios between phosphorus and the respective iron, zinc, and magnesium concentrations were calculated as a proxy for their bioavailability (Šimić et al., 2012).

Similar to Baxter et al. (2014), to screen for extreme outliers due to machine error and enable proper variance component estimation, values for a given element that were greater than 15 median absolute deviations (MAD) at each environment (location × year combination) were removed. Given that elements with low concentrations may produce negative values, these were set to missing if less than 1% of the samples for a given trait were negative. Next, a more robust outlier screening procedure was conducted using a mixed linear model that allowed for the estimation of genetic effects separately from field design effects, following Wolfinger et al. (1997). The full model fitted for each phenotype across locations in ASReml-R version 3.0 (Gilmour et al., 2009) was as follows:

$$Y_{ijklmnopqr} = \mu + check_i + env_j + set(env)_{jk} + block(set \times env)_{jkl} + geno_m + (geno \times env)_{jm} + ICP.run_n + sample_o + x_p\beta_{weight} + row(env)_{jq} + col(env)_{jr} + \varepsilon_{ijklmnopqr} \quad [2]$$

in which $Y_{ijklmnopq}$ is an individual concentration, μ is the grand mean, $check_i$ is the fixed effect

for the *i*th check, env_j is the effect of the *j*th environment, $set(env)_{jk}$ is the effect of the *k*th set within the *j*th environment, block$(set \times env)_{jkl}$ is the effect of the *l*th incomplete block within the *k*th set within the *j*th environment, $geno_m$ is the effect of the *m*th experimental genotype (noncheck line), $(geno \times env)_{jm}$ is the effect of the interaction between the *m*th genotype and *j*th environment, $ICP.run_n$ is the laboratory effect of the *n*th ICP run, $sample_o$ is *o*th kernel sample, x_p is the first order trend for kernel weight, $row(env)_{jq}$ is the effect of the *q*th plot grid row within the *j*th environment, $col(env)_{jr}$ is the effect of the *r*th plot grid column within the *j*th environment, and $\varepsilon_{ijklmnopqr}$ is the heterogeneous residual error effect within each environment with a first order autoregressive correlation structure among plots in both row and column directions. Except for the grand mean and check term, all other terms were modeled as approximately independent and identically distributed [$N(0, \sigma^2)$] random variables. Studentized deleted residuals (Neter et al., 1996) obtained from these mixed linear models and degrees of freedom calculated using the Kenward-Rogers approximation (Kenward and Roger, 1997) were used to detect significant outliers.

Once all outliers were removed, an iterative mixed linear model fitting procedure was conducted in ASReml-R version 3.0 (Gilmour et al., 2009) with the full model as described previously (Baseggio et al., 2019). Terms fitted as random effects and first order autoregressive error correlations were tested with likelihood ratio tests (Littell et al., 2006), and those not significant at $\alpha = 0.05$ were removed from the model. This resulted in the selection of a final, best fitted model for each element that was then used to generate a best linear unbiased predictor (BLUP) for each genotype. After removing six inbred lines known to possess mutations at *aeduwx* or *bt2* and those without genotype data, BLUPs from 401 sweet corn inbred lines having more common endosperm mutations [*su1*, *su1se1* (classified as *su1* here), *sh2*, and *su1sh2*] were

used for further analyses (Supplemental Table S3.1). Additionally, in order to assess element variation and single-nucleotide polymorphism (SNP) associations within each location (NY or WI), BLUPs for each ionomic compound were calculated independently within location using the full model.

Anthesis and silking (*i.e.*, days from planting to when 50% of plants had shed pollen or were at silking) were recorded for two reps at both locations in 2015. The same fitting procedure and full model used for ionomic phenotypes, except for the error being modeled as $N(0, \sigma^2)$, were used to calculate BLUPs for anthesis and silking for the same 401 genotypes. Both traits were tested as covariate in GWAS models and used to assess their association with kernel element concentration. Lines without flowering data ($n = 20$) were imputed with BLUP means from the remaining 381 lines.

Variance component estimates from the full model were used to estimate heritability on a line-mean basis (Holland et al., 2003; Hung et al., 2012) for each ionomic phenotype, where the mean residual variance across all environments was used as the residual error variance. Pearson's correlation coefficient (r) was used to estimate the degree of association between BLUP values of element concentrations and also with flowering time at $\alpha = 0.05$ via the method 'pearson' from the function 'cor.test' in R.

DNA Extraction, Sequencing, and Genotyping

Genotyping-by-sequencing (GBS) data from the 401 inbred lines with BLUPs were generated as described previously (Baseggio et al., 2019). In brief, the GBS procedure of Elshire et al. (2011) with *Ape*KI was used to construct multiplexed libraries that were sequenced on a NextSeq 500 or Illumina HiSeq 2500 (Illumina Incorporated, San Diego, CA, USA) at the Cornell Biotechnology Resource Center (Cornell University, Ithaca, NY, USA). All raw GBS

sequencing data are available from the National Center of Biotechnology Information Sequence Read Archive under accession number SRP154923 and in BioProject under accession PRJNA482446.

With the raw GBS sequencing data, the genotypes at 955,690 high confidence SNP loci were called with the default parameters in the TASSEL 5 GBSv1 production pipeline with the ZeaGBSv2.7 Production TagsOnPhysicalMap file (available at panzea.org, accessed 19 Nov 2018) in B73 RefGen_v2 coordinates (Glaubitz et al., 2014). Existing raw unimputed SNP genotypic data for 16 sweet corn lines (ZeaGBSv27_publicSamples_rawGenos_AGPv2-150114.h5, available at panzea.org, accessed 19 Nov 2018) with ionomics data from this study were merged with our data set prior to any SNP filtering steps. Initial filtering on the combined unimputed SNP data set consisted of removing SNPs having a minor allele observed in only one line (singletons and doubletons) and retaining biallelic SNPs with a call rate greater than 10%. Similar to Baseggio et al. (2019), heterozygous genotype calls with an allele balance score (lowest allele read depth/total read depth) less than 0.3 or greater than 0.7 were set to missing. For duplicated lines, the SNP genotype calls from samples that shared an accession name were merged if the identical-by-state (IBS) values from all sample pairwise comparisons exceeded 0.99 as in Romay et al. (2013), and SNP genotypes were set to missing if discordant between replicated samples. For accessions where IBS values were lower than this conservative threshold, the sample with the highest SNP call rate was selected to represent the inbred line.

Missing SNP genotypes were imputed with FILLIN (Swarts et al., 2014) using an available set of maize haplotype donors having a window size of 4 kb (available at panzea.org, accessed 19 Nov 2018). Given that this imputation method is unable to impute all missing genotypes (Swarts et al., 2014), additional filtering was conducted. In Tassel 5 version

20190321, quality filters following haplotype-based imputation included removing SNPs with a call rate less than 70%, a minor allele frequency (MAF) lower than 5%, heterozygosity greater than 10%, an inbreeding coefficient lower than 80%, or a mean read depth greater than 15. The final, complete SNP marker data set consisted of 163,573 high-quality SNP markers scored on 401 sweet corn inbred lines having a BLUP value for one or more ionomic phenotypes.

Genome-Wide Association Study

Prior to conducting a GWAS, the Box-Cox power transformation (Box and Cox, 1964) was performed on BLUPs of each ionomic compound with an intercept-only model to identify the most appropriate transformation that corrected for unequal variances and the non-normality of error terms. The process was conducted using the MASS package in R version 3.2.3 (R Core Team, 2015) and tested lambda values ranging from -2 to +2 in increments of 0.5 before applying the optimal convenient lambda for each phenotype (Supplemental Table S3.2).

To conduct a GWAS for each ionomic compound, a univariate mixed linear model was used to test each of the 163,573 SNP markers for association with Box-Cox transformed BLUP values from the 401 inbred lines (Supplemental Table S3.3) in the GEMMA software version 0.97 (Zhou and Stephens, 2014). The mixed linear model accounted for population stratification and familial relatedness by including principal components (PCs) (Price et al., 2006) and a genomic relationship (kinship) matrix based on VanRaden's method 1 (VanRaden, 2008), both of them calculated in the R package GAPIT version 2017.08.18 (Lipka et al., 2012) using a subset of 11,827 unimputed SNPs. This subset of the complete marker data set included SNPs with a call rate higher than 90%, MAF greater than 5%, heterozygosity less than 10%, inbreeding coefficient greater than 80%, and mean read depth lower than 15. Missing genotypes remaining in both SNP marker data sets were conservatively imputed as heterozygous in GAPIT. The

optimal number of PCs to include as covariates in the mixed linear model was determined with the Bayesian information criterion (BIC) (Schwarz, 1978). Similarly, the BIC was used to determine whether to also include BLUPs of flowering time and endosperm mutation type (*su1*, *sh2*, or *su1sh2*) as covariates in the mixed linear model. Distribution of inorganic nutrients vary in the sweet corn kernel (Cheah et al., 2019) and therefore mutations at *su1* and *sh2* could be associated with elements that accumulate in the endosperm as was shown for levels of tocotrienols—a class of tocochromanols mostly found in the endosperm (Grams et al., 1970; Weber, 1987)—in fresh sweet corn kernels from the same association panel (Baseggio et al., 2019). Kernel type of the remaining 17 inbred lines with ionomic phenotypes but without endosperm mutation type was estimated using the same model developed for the other 384 lines reported by Baseggio et al. (2019).

The likelihood-ratio-based R^2 statistic (R^2_{LR}) of Sun et al. (2010) was used to estimate the amount of phenotypic variation explained by the mixed linear model with or without a significant SNP detected in GWAS. The maximum log-likelihoods of the model of interest and the intercept-only model required for R^2_{LR} were fitted in GEMMA and with the 'lm' function in R, respectively. For each phenotype, *P*-values (Wald test) of SNPs tested in GEMMA were adjusted to control the false-discovery rate (FDR) at 5% with the Benjamini–Hochberg multiple test correction (Benjamini and Hochberg, 1995) available in the 'p.adjust' function of R version 3.2.3 (R Core Team, 2015). Given the distance to which genome-wide linkage disequilibrium decays to background levels in this association panel (Baseggio et al., 2019), candidate gene searches were limited to ± 250 kb of the physical position of SNP markers significantly associated with an ionomic phenotype.

The multi-locus mixed-model (MLMM) approach of Segura et al. (2012) was used to control for the influence of major-effect loci and exploring the potential existence of novel associations. The MLMM method was conducted on an individual chromosome basis as described previously (Lipka et al., 2013) and the extended BIC (Chen and Chen, 2008) was used in the selection of the optimal model. The control of major-effect loci was then assessed by reconducting GWAS with MLMM-selected SNPs included as covariates in the mixed linear model of GEMMA.

Ionomics Prediction

We used a single kernel genomic best linear unbiased prediction (GBLUP) model (Zhang et al., 2007; VanRaden, 2008) to assess the ability of genome-wide SNP markers to predict each of the 18 ionomics phenotypes scored in the 401 inbred lines. The genomic relationship matrix was derived from the same SNP marker set as in GWAS (163,573 SNPs) using the R package GAPIT version 2017.08.18 (Lipka et al., 2012) with method 1 from VanRaden (2008). The kinship matrix was modeled as a random effect for prediction of the Box-Cox-transformed ionomic phenotypes as in GWAS with the function 'emmreml' in the EMMREML R package (Akdemir and Okeke, 2015).

A five-fold cross-validation approach was used to estimate the predictive ability of each phenotype by assessing the Pearson's correlation between observed and genomic estimated breeding values as described in Baseggio et al. (2019). The predictive ability of each model was based on a mean of correlations from 50 replicates of the five-fold cross-validation scheme, and each fold consisted of genotype frequencies for endosperm mutants (*su1*, *sh2*, and *su1sh2*) that were representative of the whole association population. Similar to genomic prediction of

tocochromanol traits (Baseggio et al., 2019), endosperm mutation type (*su1*, *sh2*, or *su1sh2*) was evaluated as a covariate in prediction models.

RESULTS

Phenotypic variation

We assessed variation in the ionomic profile of fresh kernels from an association panel of 401 sweet corn inbred lines. Kernel composition exhibited extensive phenotypic variation for the 15 elements (Table 3.1), with the ratio between the maximum and the minimum BLUP values in the panel varying from 1.3 to 6.4 for sulfur and cadmium, respectively. The macronutrients potassium, magnesium, phosphorus, and sulfur had average concentrations above 1,000 μg g^{-1}, with potassium being the element in highest concentration with values up to 11,313 μg g^{-1}. The average of calcium—the only other macronutrient profiled—was less than 100 μg g^{-1}. Cadmium was the element with the lowest average concentration, followed by nickel, which had an average of 0.21 μg g^{-1}. When separating inbred lines according to their endosperm mutation type (Table 3.2), for five (cadmium, copper, iron, manganese, and potassium) of the 15 elements, *sh2* lines (n = 78) had an average amount significantly greater ($P < 0.05$) than *su1* lines (n = 301).

Table 3.1. Means and ranges for best linear unbiased predictors (BLUPs) of 15 fresh kernel elements and three ratios evaluated in the sweet corn association panel and estimated heritability ($\widehat{h_l^2}$) on a line-mean basis across two years and two locations, as well as correlations between BLUPs from each location.

Trait	Lines	BLUPs			$\hat{h}_l^2$	Corr.‡
		Mean	SD†	Range		
		——— µg g^{-1} dry weight ———				
Boron	399	2.70	0.11	2.42-3.15	0.18	0.08
Cadmium	401	0.02	0.01	0.01-0.04	0.71	0.60
Calcium	400	99.79	10.70	70.89-145.13	0.42	0.27
Copper	401	4.42	0.75	2.69-7.05	0.83	0.68
Iron	401	21.40	2.05	16.46-30.31	0.66	0.50
Magnesium	401	1400.93	71.30	1210.59-1699.99	0.68	0.48
Manganese	401	9.12	1.19	5.89-13.64	0.77	0.54
Molybdenum	400	0.32	0.04	0.23-0.46	0.46	0.38
Nickel	401	0.21	0.03	0.14-0.37	0.59	0.46
Phosphorus	401	3410.47	164.17	2965.52-3856.27	0.65	0.47
Potassium	401	9768.17	467.70	8251.24-11313.15	0.69	0.52
Rubidium	401	3.27	0.25	2.74-4.18	0.55	-§
Strontium	401	0.50	0.04	0.41-0.70	0.49	0.27
Sulfur	401	1719.25	69.00	1512.78-1952.95	0.36	-
Zinc	401	24.64	2.28	18.63-32.36	0.63	0.45
Iron/Phosphorus	401	0.01	0.001	0.005-0.008	0.66	0.51
Magnesium /Phosphorus	401	0.41	0.021	0.338-0.476	0.77	0.60
Zinc/Phosphorus	401	0.01	0.001	0.006-0.009	0.69	0.48

† Standard deviation of the BLUPs.

‡ Correlation coefficient between BLUPs from the two locations.

§ Genotype term was not significant for one location.

Table 3.2. Estimated effects of endosperm mutation type for 15 fresh kernel elements and three ratios.

Trait	*su1*	*sh2*	*su1sh2*	*P*-value†
	μg g^{-1} dry weight			
Boron	2.696	2.691	2.720	0.56
Cadmium	0.016 b‡	0.018 a	0.017 ab	**0.04**
Calcium	99.836	99.319	100.901	0.82
Copper	4.308 b	4.717 a	4.985 a	**<0.01**
Iron	21.218 b	21.839 a	22.335 a	**<0.01**
Magnesium	1397.299	1405.866	1433.021	0.06
Manganese	8.997 b	9.383 a	9.875 b	**<0.01**
Molybdenum	0.323	0.322	0.322	0.95
Nickel	0.211	0.217	0.217	0.33
Phosphorus	3399.573 a§	3434.226 a	3475.413 a	**0.04**
Potassium	9715.742 b	9890.313 a	10052.469 a	**<0.01**
Rubidium	3.271	3.264	3.268	0.98
Strontium	0.498	0.495	0.511	0.21
Sulfur	1713.284 b	1727.783 b	1770.629 a	**<0.01**
Zinc	24.564	24.848	25.019	0.45
Iron/Phosphorus	0.0061	0.0062	0.0064	0.07
Magnesium /Phosphorus	0.4089	0.4068	0.4090	0.74
Zinc/Phosphorus	0.0072	0.0072	0.0072	0.93

† *P*-value from one-way analysis of variance (ANOVA) *F*-test for the endosperm mutation type effect. Bolded *P*-value indicates a statistically significant difference between two or more endosperm mutation type groups ($P < 0.05$).

‡ Sweet corn lines grouped by endosperm mutation type having labels with the same letter are not significantly different according to the Tukey-Kramer honest significant difference test ($P < 0.05$). The test was only performed for traits that had a significant *F*-test.

§ *F*-test showed significant effect of endosperm mutation type ($P < 0.05$), but Tukey-Kramer test did not show any difference between groups ($P > 0.05$).

Heritability values for the 15 elements were variable (Table 3.1), ranging from 0.18 (boron) to 0.83 (copper), with most of estimates greater than 0.50. Variance components revealed that genotype × environment interaction was moderate for most elements (Supplemental Fig. S3.1), which was also observed by the weak to moderate correlation between BLUPs of WI and NY, ranging from 0.08 (boron) to 0.68 (copper) (Table 3.1). Average BLUP values showed significant differences between locations for all compounds, except for zinc (Fig. 3.1). As for spatial variability within location, field design effects contributed less than 10% of total variation

(Supplemental Fig. S3.1).

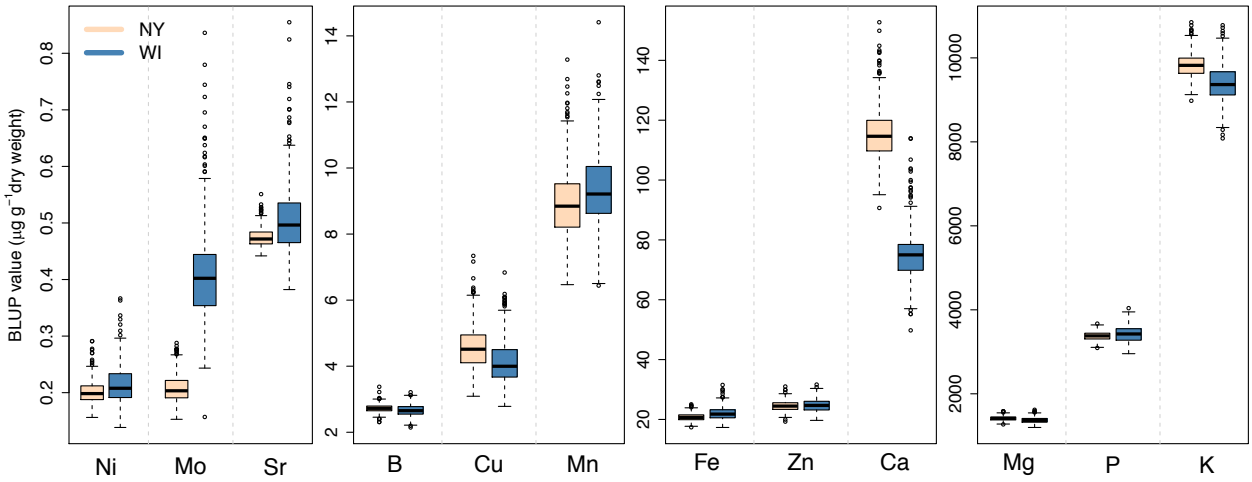

Figure 3.1. Distribution of BLUP values for 12 elemental concentrations that had significant genotypic variation in both locations in fresh kernels of 401 sweet corn inbred lines grown in NY and WI.

Element pairs with strong correlations ($r > 0.50$) included strontium/calcium, magnesium/phosphorus, phosphorus/zinc, and zinc/iron. Days to silking and anthesis were highly correlated ($r = 0.95$) and therefore the more heritable phenotype ($\hat{h}_l^2$ anthesis = 0.93) was used as flowering time for further analyses. Correlations between BLUPs were, on average, positive or non-significant, with the exception of the negative correlation between nine elements and anthesis ($-0.32 < r < -0.12$; Supplemental Fig. S3.2). Overall, molybdenum was the least correlated element, significantly correlated only with zinc. The correlations estimated using BLUPs within each location followed similar patterns.

Genome-Wide Association Study

The genetic basis of natural elemental concentrations and three ratios in the fresh kernels of 401 sweet corn inbred lines field-tested in two locations was dissected using 163,573 genome-wide SNP markers in a univariate mixed linear model that accounted for population structure, relatedness, type of endosperm mutation, and flowering time. Although the correlations were

significant (*P*-value < 0.05) between anthesis and 13 elemental concentrations, only models for calcium, iron, molybdenum, rubidium, and strontium included anthesis as covariate for GWAS. Across locations, four elements and one ratio were significantly associated with 268 unique SNPs, with two main regions: one on chromosome 2 for cadmium and one on chromosome 7 for iron, zinc, and the ratio of zinc to phosphorus (Supplemental Table S3.4).

The most extensive association was identified between cadmium and 185 SNP markers on chromosome 2 (Fig. 3.2A). The peak SNP (S2_157751802; *P*-value 9.10×10^{-24}) was within the open reading frame (ORF) of a gene that encodes a UDP-glycosyltransferase superfamily protein (GRMZM2G463996) and accounted for 18% of the phenotypic variance for this element (Supplemental Table S3.4). Significant SNPs (*P*-values 9.10×10^{-24} to 5.70×10^{-5}) encompassed a 35-Mb region and included several candidate genes, particularly heavy metal ATPase 3 (*hma3*; GRMZM2G175576) and NRAMP metal ion transporter (*nramp*; GRMZM2G366919). To better clarify the signals of association in this extensive genomic interval, a chromosome-wide multi-locus mixed-model procedure (MLMM) was conducted for cadmium. The resulting optimal model only included the peak SNP (S2_157751802; Supplemental Table S3.5), and when GWAS was reconducted with this MLMM-selected SNP included as a covariate in the mixed linear model, all other signals were no longer significant at a genome-wide FDR of 5% (Fig. 3.2B). Additionally, two SNPs on chromosomes 3 and 8 that were weakly associated with cadmium in the model without the MLMM-selected SNP were no longer significant when controlling for the large-effect locus on chromosome 2.

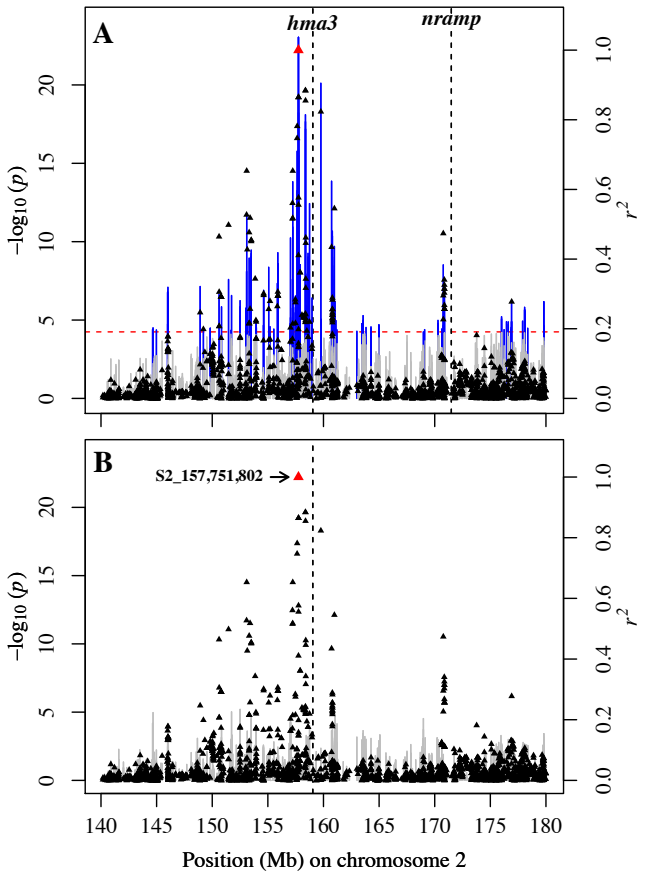

Figure 3.2. Genome-wide association study for cadmium level in fresh kernels of sweet corn. (A) Scatter plot of association results from a mixed model analysis and linkage disequilibrium (LD) estimates (r^2). The vertical lines are $-\log_{10}$ *P*-values of single nucleotide polymorphisms (SNP) and blue color represents SNPs that are statistically significant at a 5% false discovery rate (FDR). Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 157,751,802 bp (B73 RefGen_v2) on chromosome 2. The red horizontal dashed line indicates the $-\log_{10}$ *P*-value of the least statistically significant SNP at a 5% FDR. The black vertical dashed lines indicate the genomic positions of the heavy metal ATPase 3 gene (*hma3*) and the NRAMP metal ion transporter. (B) Scatter plot of association results from a conditional mixed linear model analysis and LD estimates (r^2). The SNP from the optimal multi-locus mixed-model (S2_157751802) was included as a covariate in the mixed linear model to control for the large effect.

Significant associations between zinc and 39 SNP markers at the end of chromosome 7 were identified (Fig. 3.3A), with the peak SNP (S7_174248412; *P*-value 1.67×10^{-10}) located in an intergenic region and ~18 kb away from a transcription factor (*bzip54*; GRMZM2G361847) that is highly expressed in the maize pericarp at 18 DAP (Stelpflug et al., 2016). This SNP was also ~200 kb away from a gene encoding *nicotianamine synthase 5* (*nas5*; GRMZM2G050108),

which is weakly expressed in the maize embryo from 16 to 24 DAP (Stelpflug et al., 2016). One SNP within the same region and ~31 kb away from the peak SNP for zinc was also associated with iron (S7_174279369; *P*-value 2.24×10^{-7}). When using a chromosome-wide MLMM procedure to better resolve the complex of association signals within the 1.19-Mb region on chromosome 7, only the peak SNP was selected in the optimal model for zinc and iron. In a conditional univariate mixed model analysis that included the peak SNP as a covariate for each trait, significant SNPs were no longer found to be associated with either zinc (Fig. 3.3B) or iron, including one SNP on chromosome 1 and six SNPs on chromosome 8 that were significant for zinc in the model without the MLMM-selected SNP.

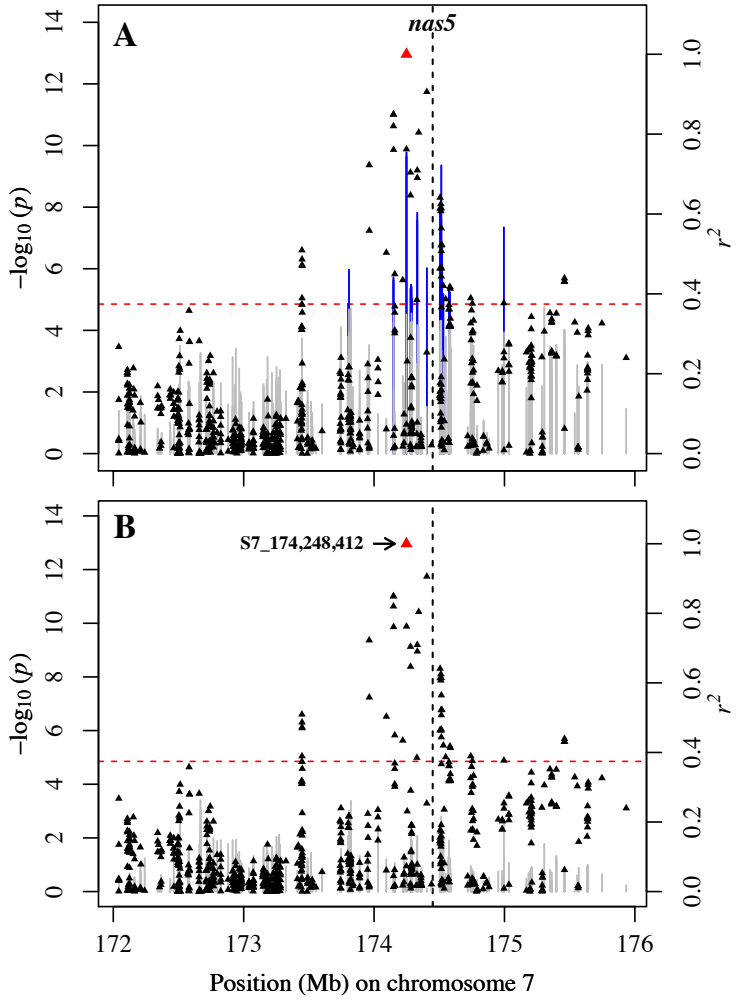

Figure 3.3. Genome-wide association study for zinc level in fresh kernels of sweet corn. (A) Scatter plot of association results from a mixed model analysis and linkage disequilibrium (LD) estimates (r^2). The vertical lines are $-\log_{10}$ *P*-values of single nucleotide polymorphisms (SNP) and blue color represents SNPs that are statistically significant at a 5% false discovery rate (FDR). Triangles are the r^2 values of each SNP relative to the peak SNP (indicated in red) at 174,248,412 bp (B73 RefGen_v2) on chromosome 7. The red horizontal dashed line indicates the $-\log_{10}$ *P*-value of the least statistically significant SNP at a 5% FDR. The black vertical dashed line indicates the genomic position of the nicotianamine synthase 5 gene (*nas5*). (B) Scatter plot of association results from a conditional mixed linear model analysis and LD estimates (r^2). The SNP from the optimal multi-locus mixed-model (S7_174248412) was included as a covariate in the mixed linear model to control for the large effect.

When the ratio of zinc to phosphorous concentrations was analyzed as a proxy for zinc bioavailability, significant associations were also observed for the same region of chromosome 7 and the peak SNP (S7_174515604) was ~65 kb from the *nas5* gene. In contrast to the analysis

for zinc, once this peak SNP, which was the only one selected in the MLMM procedure, was added as covariate in the mixed linear model, four SNPs on chromosome 1 were still significantly associated with zinc/phosphorous ratio. The peak SNP (S1_217134411; *P*-value 2.70×10^{-9}) was ~ 255 kb from a gene that encodes an ABC transporter (GRMZM2G142249) and ~180 kb from a gene encoding a heavy metal transport/detoxification protein (AC217910.3_FG001).

Additional associations were observed between two SNPs in complete LD on chromosome 3 and rubidium accumulation (*P*-value 5.12×10^{-7}), and both are located within the ORF of a gene that encodes an F-box/RNI-like superfamily protein (GRMZM2G138176). The optimal model from the MLMM procedure did not include any SNP, suggesting this was a relatively weaker association.

Element accumulation in the maize kernel is a complex quantitative trait that is highly dependent on the environment (Ziegler et al., 2017). In order to increase the reliability and consistency of signal detection and identify weaker associations that may be unique to a specific environment, BLUPs were also calculated within each location (NY and WI) and used in GWAS. In both locations, zinc was associated with SNPs at the end of chromosome 7 (Supplemental Table S3.4). The same peak SNP from the analysis across locations (S7_174248412) was the most significant association with zinc in NY (*P*-value 5.73×10^{-9}), whereas a SNP ~267 kb away (S7_174515604, r^2 between peak SNPs = 0.57) and within the ORF of a gene encoding a DnaJ heat shock protein (GRMZM2G113036) was the peak SNP in WI (*P*-value 9.34×10^{-8}), and these are ~200 and 65 kb apart, respectively, from *nas5*. In both locations, only the peak SNP was included in the best model using a chromosome-wide MLMM

approach and once GWAS was reconducted with the given SNP as a covariate, no other significant signal was observed.

Similar to the analysis across locations, two SNPs within the same region on chromosome 7 were significantly associated with iron in NY (Supplemental Table S3.4). Besides significant signals in this region, two SNPs on chromosome 8 were associated with the ratio of zinc to phosphorous in WI (peak SNP S8_163896298; *P*-value 4.84×10^{-9}), and these are 172 and 166 kb from genes encoding an ABC transporter (GRMZM5G884972) and a peptide transporter (GRMZM5G817436), respectively. Both SNPs were still significantly associated with the ratio even when conducting the GWAS with the MLMM-selected SNP from chromosome 7 (S7_174331432) as a covariate in the model (Supplemental Table S3.6).

Consistent with the across-location analysis, 89 SNPs at the same region on chromosome 2 were associated with cadmium accumulation in WI. The peak SNP (S2_159765450; *P*-value 4.51×10^{-17}), however, was more than 2 Mb away from the peak found when using across-locations BLUPs, but closer to *hma3*. Similar to previous results, this was the only SNP included in the best model using the MLMM procedure, and when conducting the conditional GWAS with this SNP in the model, all other signals were no longer significant.

Additional significant markers identified for a specific location included one SNP on chromosome 1 associated with molybdenum (S1_184627650; *P*-value 2.54×10^{-7}) and three SNPs on chromosome 9 associated with nickel (peak SNP S9_2213924; *P*-value 5.26×10^{-8}), both in NY, and three SNPs on chromosome 10 associated with calcium levels in WI (peak SNP S10_124069084; *P*-value 1.09×10^{-7}). The SNP on chromosome 1 is located less than 3 kb downstream from the ORF of a gene encoding a calcium-transporting ATPase (GRMZM2G118919), whereas the three SNPs on chromosome 9 are within a gene encoding a

peptide transporter (*ptr2*; GRMZM2G057611) and one SNP on chromosome 10 is within the ORF of a gene encoding a nucleic acid-binding protein (*pot1b*; GRMZM2G169648).

Ionomics prediction

The potential of genomic selection for modifying the levels of 15 nutrients and three ratios in sweet corn fresh kernels from 401 inbred lines was evaluated using whole-genome prediction models with a kinship matrix derived from 163,573 SNP markers. The average predictive ability, measured as the correlation between observed and genomic estimated breeding values, was 0.37, ranging from 0.17 (boron and rubidium) to 0.53 (copper) (Table 3.3). Iron and zinc, the two most deficient nutrients in human diets, had above average predictive ability, suggesting whole-genome prediction could be used to enhance these elements in fresh sweet corn kernels for human consumption. A strong positive correlation ($r = 0.81$, P-value < 0.001) was found between heritability estimates and predictive abilities for the 18 ionomic traits.

Elements are not distributed uniformly throughout the seeds (Cheah et al., 2019). Given that some are located mainly in the endosperm and that differences in concentration were observed among inbred line groups (*su1*, *sh2*, and *su1sh2*; Table 3.2), endosperm mutation type was included as a covariate in the prediction models to assess the effect on the predictive abilities. The inclusion of endosperm mutation type had a negligible effect for most of the elements.

Table 3.3. Predictive abilities of genomic prediction models for 15 fresh kernel ionomic elements and three ratios with or without endosperm mutation type.

Trait	GBLUP†	GBLUP with endosperm mutation type covariate
Boron	0.17 (0.03)	0.16 (0.03)
Cadmium	0.42 (0.02)	0.42 (0.02)
Calcium	0.28 (0.03)	0.27 (0.03)
Copper	0.53 (0.02)	0.54 (0.02)
Iron	0.46 (0.02)	0.46 (0.02)
Magnesium	0.41 (0.02)	0.41 (0.02)
Manganese	0.46 (0.02)	0.46 (0.02)
Molybdenum	0.24 (0.02)	0.23 (0.02)
Nickel	0.36 (0.02)	0.36 (0.02)
Phosphorus	0.46 (0.02)	0.46 (0.02)
Potassium	0.40 (0.02)	0.40 (0.02)
Rubidium	0.17 (0.04)	0.16 (0.04)
Strontium	0.24 (0.03)	0.23 (0.03)
Sulfur	0.27 (0.02)	0.27 (0.02)
Zinc	0.50 (0.02)	0.49 (0.02)
Iron/Phosphorus	0.45 (0.02)	0.44 (0.02)
Magnesium/Phosphorus	0.36 (0.02)	0.35 (0.02)
Zinc/Phosphorus	0.47 (0.02)	0.46 (0.02)
Average	0.37	0.36

† Genomic best linear unbiased prediction using 163,573 genome-wide SNP markers

DISCUSSION

Iron and zinc deficiencies are a worldwide problem and affect a large proportion of women, children, and older adults in the US (Clark, 2008). Sweet corn is the third most consumed vegetable in the US but usually does not significantly contribute to the daily-recommended amount of these nutrients (USDA, 2018a). Analysis of the kernel ionome is a

powerful tool to understand the complex regulatory mechanism of accumulation for essential elements and previous studies have identified several genes controlling natural variation of elements in mature maize kernels (Zhou et al., 2013; Benke et al., 2015; Hindu et al., 2018; Schaefer et al., 2018) and rice grain (Norton et al., 2014; Yang et al., 2018). Given that sweet corn is eaten as a fresh vegetable and the ionome profile changes throughout the plant development (Baxter et al., 2014; Cheah et al., 2019), we measured 15 elements and three ratios in fresh kernels of 401 sweet corn lines to identify the genetic controllers of elemental variation in immature kernels. The assessment of the full complement of elements can reveal the integrated networks that affect their accumulation (Asaro et al., 2017) and whole-genome prediction models can provide insights into the expected genetic gains in a biofortification breeding program.

Elemental concentrations in this panel showed variation comparable to other studies from maize kernels at maturity. Zinc and iron variability (18.6-32.4 μg g^{-1} and 16.5-30.3 μg g^{-1}, respectively) was similar to the nested association mapping (NAM) population using mature kernels from more than 5000 lines (Zn: 19.7-36.7 μg g^{-1}, Fe: 13.1-28.5 μg g^{-1}; Ziegler et al., 2017) and 49 tropical elite varieties (Zn: 16-25 μg g^{-1}, Fe: 17-24 μg g^{-1}; Oikeh et al., 2007), but lower than 1417 maize germplasm from CIMMYT core collection (Zn: 13-58 μg g^{-1}, Fe: 10-63 μg g^{-1}; Banziger and Long, 2000). Although the average and ranges for most of the elements were comparable to what was reported by Pietz et al. (1978) for 243 maize inbred lines, coefficients of variation (CV) in our panel were relatively lower, with only cadmium having a CV greater than 20%, which might be a result of the high relatedness of the sweet corn pool (Tracy, 1997). Toxic effects of metals to human health are not expected, considering that maximum concentration observed for each heavy metal is consistently lower than the guidelines

for safe limits (70, 56, and 35 mg day^{-1} and 1.75 mg $month^{-1}$ for zinc, iron, copper, and cadmium, respectively, for a 70-kg person; JECFA, 2017). As an example, a medium ear of sweet corn (100 g fresh weight) with the maximum concentration of cadmium found in this panel would have less than 5 % of the maximum tolerable intake for this element in a day.

As mutations in the starch biosynthetic pathway can potentially change the kernel ionome (Baxter et al., 2014), we also compared each element among endosperm mutation type groups (*su1*, *sh2*, and *su1sh2*). In agreement with what was observed for certain tocotrienols (Baseggio et al., 2019), which accumulate mainly in the endosperm (Weber, 1987), there were significant differences among endosperm mutation types for seven elements, and *sh2* lines usually had greater levels than *su1* lines. In another study, five of these seven elements also showed strong effect of *su1* mutation when compared to *Su1* genotypes (Baxter et al., 2014). Differences in our sweet corn panel, however, were not as pronounced given that comparisons did not include non-mutant lines. The differences also only represented a small portion of total element concentration (Table 3.3), as opposed to what was observed for total tocotrienols, for example, which accumulated 1.6 times more in *sh2* than *su1* lines (Baseggio et al., 2019).

Significant correlations between several pairs of elements suggest potential joint regulation of their uptake, transport, and/or accumulation in the fresh kernel. Similar to what was observed for mature maize kernels (Lung'aho et al., 2011; Asaro et al., 2017) and in agreement with their shared chemical and physiological properties, strong correlations ($r > 0.5$) were observed between calcium/strontium, magnesium/phosphorus, iron/zinc, and zinc/phosphorus. Moderate correlations were also observed between flowering time and 13 elements, but these are likely due to population structure. The same elements had similar correlations with the first

principal component (PC1) derived from genotypic information, and the latter was moderately correlated with anthesis (r = 0.41; Supplemental Fig. S3.2).

The univariate GWAS of 18 ionomic traits in fresh sweet corn kernels identified significant associations, particularly of iron and zinc with SNP markers on chromosome 7 and cadmium with SNPs on chromosome 2. Genetic correlation between iron and zinc has been previously reported, as 8 of 10 identified QTL were involved with both traits in maize grain (Jin et al., 2013). A GWAS using the NAM population also reported associations with a resample model inclusion probability (RMIP) of 0.05 for zinc and up to 0.91 with SNPs within ± 250 kb of the peak SNP for the given element in this study (Ziegler et al., 2017). The region on chromosome 7 that is associated with iron and zinc has several genes, with *nas5* as the most likely candidate.

Nicotianamine synthase (NAS) enzymes catalyze the reaction in which three S-adenyl-Met (SAM) molecules are conjugated into one nicotianamine (NA) molecule, a chelator of transition metals with roles in long- and short-transport of metal cations (von Wirén et al., 1999; Takahashi et al., 2003). In the *Poaceae*, NA is also a precursor in synthesis of mugineic-acid phytosiderophores (MA), chelators that dissolve iron in the rhizosphere followed by reabsorption of Fe-MA complexes by YS1 transporters (Mizuno et al., 2003). Mizuno et al. (2003) also showed that *ZmNAS1* and *ZmNAS2* were expressed only in iron-deficient maize roots, whereas *ZmNAS3*, which has 90% gene similarity to *ZmNAS5*, was negatively regulated by iron deficiency. This suggests that NA synthesized by *ZmNAS3* under sufficient iron conditions might form complexes with iron to be transported in and out of cells or organelles via Yellow Stripe-Like (YSL) transporters, similar to nongraminaceous plants (Inoue et al., 2003; Mizuno et al., 2003).

Studies overexpressing *nas* genes have shown effective iron and zinc biofortification in several plants. Johnson et al. (2011) reported up to four-fold and two-fold increases for iron and zinc concentration, respectively, by overexpressing one of three *nas* genes, and this enrichment was in phosphorous-free regions in the rice endosperm. Also, mutant lines with increased expression of *OsNAS3*, which shares 82% identity with *ZmNAS5*, resulted in polished rice with increased bioavailable iron, and this was able to reverse signs of iron-deficiency when fed to anemic mice (Lee et al., 2009).

The extensive region on chromosome 2 with SNPs associated with cadmium contains several genes that may be responsible for the genetic control of this metal level in fresh sweet corn kernels. The same region was associated with cadmium using the NAM population (Ziegler et al., 2017), with two highly significant SNPs located at positions 155,736,793 (*P*-value 1.59×10^{-20}) and 158,441,682 (*P*-value 5.08×10^{-18}). This region shows extensive linkage disequilibrium, with correlations between the peak SNP (S2_157751802) and SNP markers more than 2 Mb away greater than 0.80 (Fig. 3.1). Given that, and due to the presence of two other highly significant SNPs (*P*-values 3.95×10^{-7} and 9.50×10^{-6}) ~75 kb away from the gene encoding a heavy metal ATPase 3 protein (HMA3), this is a likely candidate. HMA3 is a root-specific metal transporter involved in cadmium detoxification by sequestering it into root vacuoles and limiting translocation of cadmium from roots to the shoots, resulting in lower accumulation of this metal in grains (Sasaki et al., 2014). A GWAS and transgenic complementation study revealed that *hma3* was the sole major locus responsible for cadmium levels in *Arabidopsis* leaves, and ten different haplotypes were identified (Chao et al., 2012). This protein has also been associated with transport of zinc, cobalt, and lead (Takahashi et al., 2012). Sasaki et al. (2014) reported that overexpression of *OsHMA3* reduced cadmium accumulation in the rice

grain, but shoot zinc level was maintained by up-regulating genes from the ZIP family, which are responsible for zinc uptake and translocation. This suggests that selection for lower cadmium accumulation to reduce possible toxicity should not influence zinc accumulation in the kernel, and this is in agreement with the lack of association of zinc with SNP markers on chromosome 2 in this panel.

Analysis of the kernel ionome within each environment identified three additional regions not observed across locations, and these were associated with molybdenum and nickel in NY and calcium in WI. The SNP associated with molybdenum is close to a gene encoding a calcium-transporting ATPase (*aca2*, GRMZM2G118919). Although Ca^{+2}-ATPases have been implicated in the transport of manganese and possibly zinc in addition to calcium (Huda et al., 2013), no study has shown them to associate with molybdenum.

The same association with nickel was observed in the NAM population, where the most significant SNP (position: 2,175,252; *P*-value 5.88×10^{-66}; RMIP = 0.46) was only 38.7 kb away from the peak in this study. The three SNPs on chromosome 9 associated with nickel are within the ORF of a gene encoding a peptide transporter (*ptr2*). Peptide transporters have a wide range of substrates, particularly nitrate, malate, and histidine (Ouyang et al., 2010), among which histidine binds to nickel with high affinity (Ashrafi et al., 2011) and has been involved in detoxification and translocation in some nickel hyperaccumulators (Krämer et al., 1996). Chiang et al., 2004) has demonstrated the capacity of cotransport of protons and peptides for *ptr2* and Richau et al. (2009) reported exogenous histidine enhanced nickel loading into the xylem in an hyperaccumulator plant, but the transporters of the histidine-bound nickel have not yet been identified. Although *ptr2* has not been directly associated with nickel transport, it may be involved in transporting histidine-nickel complexes.

The peak SNP associated with calcium in WI was 230 and 28 kb away from a gene encoding a Ras-related protein (*ras2*; GRMZM2G173878) and a heavy metal transport protein (*hma*; GRMZM2G169726), respectively. Members of the Ras superfamily function as signaling switches in vital cellular processes, such as cell proliferation and cell cycle (reviewed in Aspenström, 2004). The activation of Ras-like GTPases is regulated by calcium-dependent mechanisms and certain small GTPases were shown to influence Ca^{+2} channels, regulating the influx of extracellular calcium into the cells (Trimmer, 2002). Although no association with calcium was observed within this region in the NAM population (Ziegler et al., 2017), a SNP (*P*-value 8.32×10^{-10}; RMIP = 0.14) ~25.6 kb away from the peak in this study was significantly associated with strontium, another alkaline earth metal that was highly correlated with calcium in this panel ($r = 0.72$). Also, both elements have similar physical and chemical properties and their transport and distribution in maize are alike (Seregin and Kozhevnikova, 2004). Taken together, the signal observed for calcium, if not related to *ras2*, may be due to associations with other correlated elements in the ionome such as strontium.

Given the high correlation between pairs of elements observed in this panel (Supplemental Fig. S3.2), a multivariate GWAS could be conducted to increase the power to detect genes controlling elements with similar biochemical properties, similar approach as used by Pauli et al. (2018) for the cotton seed ionome and Carlson et al. (2019) for seed fatty acid composition in oat seed. Taiz and Zeiger (2010) suggested two groups of elements depending on their chemical reactions in the plant, the 'ionomic group' (calcium, potassium, magnesium, and manganese) and the 'redox group' (copper, iron, molybdenum, nickel, and zinc). In fact, the same association on chromosome 7 was identified for both iron and zinc in this study, and

therefore using a multivariate approach may help us identify additional genes responsible for absorption, transport or accumulation of more than one element.

Predictive abilities of whole-genome prediction using the GBLUP method for the ionomic phenotypes assessed in this panel were low to moderate, consistent with their heritability. In fact, heritability estimates were highly correlated with average predictive abilities for the 18 phenotypes (r = 0.81), in agreement with Manickavelu et al. (2017). Predictive abilities for iron (0.46) and zinc (0.50) are in agreement with a study using 330 wheat lines (0.51 for both iron and zinc) with the same GBLUP model and a cross-validation approach that borrowed information within environment, between lines across environments, and among correlated environments (Velu et al., 2016). When using cross-validation similar to this study, where genotypes being estimated have not been evaluated in any environment, their predictive abilities were slightly lower (0.45 for iron and 0.36 for zinc). Manickavelu et al. (2017) observed predictive abilities of 0.31 and 0.36 for iron and zinc, respectively, using GBLUP at one location. The same authors also reported relatively high and moderate predictive abilities for macro and micronutrients, respectively, but this trend was not observed in the present study. Crossa et al. (2010) reported different predictive abilities for the same population in different environments, suggesting that genotype × environment interaction is an important component of genetic variability. Given that genotype × environment was observed in this panel and different associations were identified for specific locations, we hypothesize that explicitly modeling genotype × environment may increase predictive abilities to help identify the best lines for the given ionomic phenotype.

CONCLUSIONS

Natural variation of iron and zinc in fresh sweet corn kernels is most likely controlled by

nas5, whereas cadmium accumulation is due to *hma3*. Within-location analysis showed associations for nickel and molybdenum in NY and for calcium in WI, and candidate genes include *ras2* for calcium and *ptr2* for nickel. Although we have shown evidence for these genes, additional experiments are needed to further support their involvement in elemental accumulation in fresh sweet corn kernels. Starch biosynthesis genes were found to mildly associate with certain elements, but differences between line groups with different mutations were small. Prediction models showed promise, as predictive abilities were moderate for elements of interest, indicating that they can be used in a genomic-assisted breeding program to select the best-performing sweet corn lines with improved iron and zinc for human nutrition.

SUPPLEMENTAL INFORMATION

Supplemental Table S3.1. Best linear unbiased predictors of the 18 fresh kernel ionomic traits used for the genome-wide association study and genomic prediction for 401 sweet corn inbred lines.

Supplemental Table S3.2. Lambda values used in Box-Cox transformation of 19 fresh kernel ionomic traits in sweet corn.

Supplemental Table S3.3. Transformed best linear unbiased predictors of the 18 fresh kernel ionomic traits used for the genome-wide association study and genomic prediction for 401 sweet corn inbred lines.

Supplemental Table S3.4. Statistically significant results from a genome-wide association study of 18 fresh kernel ionomic traits in sweet corn.

Supplemental Table S3.5. Multi-locus mixed-model results from an analysis of ionomic traits for chromosomes 1, 2, 7, 9, and 10.

Supplemental Table S3.6. Statistically significant results from a genome wide association study of fresh kernel ionomic traits in sweet corn when using SNPs selected by multi-locus mixed-models as covariates in the mixed linear model.

Supplemental Fig. S3.1. Sources of variation for ionomic traits in fresh sweet corn kernels.

Supplemental Fig. S3.2. Correlation matrix for BLUPs of the 15 ionomic traits and three ratios from fresh kernels in sweet corn. Also included BLUPs for days to anthesis and the first two principal components.

REFERENCES

Akdemir, D. and U.G. Okeke. 2015. EMMREML: Fitting mixed models with known covariance structures. https://CRAN.R-project.org/package=EMMREML (accessed 6 April 2019).

Asaro, A., B. Dilkes and I. Baxter. 2017. Multivariate analysis reveals environmental and genetic determinants of element covariation in the maize grain ionome. bioRxiv. doi:10.1101/241380.

Asaro, A., G. Ziegler, C. Ziyomo, O.A. Hoekenga, B.P. Dilkes and I. Baxter. 2016. The interaction of genotype and environment determines variation in the maize kernel ionome. G3 (Bethesda) 6: 4175-4183. doi:10.1534/g3.116.034827.

Ashrafi, K., J.T. Murphy, J.J. Bruinsma, D.L. Schneider, S. Collier, J. Guthrie, et al. 2011. Histidine protects against zinc and nickel toxicity in *Caenorhabditis elegans*. PLoS Genet. 7: e1002013. doi:10.1371/journal.pgen.1002013.

Aspenström, P. 2004. Integration of signalling pathways regulated by small GTPases and calcium. Biochim. Biophys. Acta 1742: 51-58. doi:10.1016/j.bbamcr.2004.09.029.

Banziger, M. and J. Long. 2000. The potential for increasing the iron and zinc density of maize through plant-breeding. Food Nutr. Bull. 21: 397-400.

Baseggio, M., M. Murray, M. Magallanes-Lundback, N. Kaczmar, J. Chamness, E.S. Buckler, et al. 2019. Genome-wide association and genomic prediction models of tocochromanols in fresh sweet corn kernels. Plant Genome 12. doi:10.3835/plantgenome2018.06.0038.

Bashir, K., H. Inoue, S. Nagasaka, M. Takahashi, H. Nakanishi, S. Mori, et al. 2006. Cloning and characterization of *deoxymugineic acid synthase* genes from graminaceous plants. J. Biol. Chem. 281: 32395-32402. doi:10.1074/jbc.M604133200.

Baxter, I.R. 2015. Should we treat the ionome as a combination of individual elements, or should we be deriving novel combined traits? Journal of Experimental Botany 66(8):2127–2131.

Baxter, I.R., G. Ziegler, B. Lahner, M.V. Mickelbart, R. Foley, J. Danku, 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. doi:10.1371/journal.pone.0087628.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Stat. Soc. Ser. B. (Stat. Method.) 57: 289-300.

Benke, A., C. Urbany and B. Stich. 2015. Genome-wide association mapping of iron homeostasis in the maize association population. BMC Genet. 16: 1. doi:10.1186/s12863-014-0153-0.

Box, G.E.P. and D.R. Cox. 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B Stat. Soc. 26: 211-252.

Broadley, M.R. and P.J. White. 2010. Eats roots and leaves. Can edible horticultural crops address dietary calcium, magnesium and potassium deficiencies? Proc Nutr Soc 69: 601-612. doi:10.1017/S0029665110001588.

Carlson, M.O., G. Montilla-Bascon, O.A. Hoekenga, N.A. Tinker, J. Poland, M. Baseggio, et al. 2019. Multivariate genome-wide association analyses reveal the genetic basis of seed fatty acid composition in oat (*Avena sativa* L.). bioRxiv. doi:10.1101/589952.

CDC. 2006. 2nd National report on biochemical indicators of diet and nutrition in the US population. https://www.cdc.gov/nutritionreport/pdf/Fat.pdf (accessed 8 June 2018).

Chao, D.Y., A. Silva, I. Baxter, Y.S. Huang, M. Nordborg, J. Danku, et al. 2012. Genome-wide association studies identify *heavy metal ATPase3* as the primary determinant of natural variation in leaf cadmium in *Arabidopsis thaliana*. PLoS genetics 8: e1002923. doi:10.1371/journal.pgen.1002923.

Cheah, Z.X., P.M. Kopittke, S.M. Harper, T.J. O'Hare, P. Wang, D.J. Paterson, M.D. de Jonge, and M.J. Bell. 2019. *In situ* analyses of inorganic nutrient distribution in sweetcorn and maize kernels using synchrotron-based X-ray fluorescence microscopy. Annals of Botany 123: 543–556.

Chen, J. and Z. Chen. 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95: 759-771. doi:10.1093/biomet/asn034.

Chiang, C.-S., G. Stacey, Y.-F. Tsay. 2004. Mechanisms and functional properties of two peptide transporters, AtPTR2 and fPTR2. The journal of biological chemistry. 279:30150-30157.

Clark, S.F. 2008. Iron deficiency anemia. Nutrition in Clinical Practice 23: 128-141. doi:10.1177/0884533608314536.

Crossa, J., G.d.l. Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, et al. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713-724. doi:10.1534/genetics.110.118521.

Curie, C., Z. Panaviene, C. Loulergue, S.L. Dellaporta, J.F. Briat and E.L. Walker. 2001. Maize *yellow stripe1* encodes a membrane protein directly involved in Fe(III) uptake. Nature 409: 346-349. doi:10.1038/35053080.

Davis, D.R., M.D. Epp and H.D. Riordan. 2004. Changes in USDA food composition data for 43 garden crops, 1950 to 1999. J Am Coll Nutr 23: 669-682.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6: e19379. doi:10.1371/journal.pone.0019379.

Gilmour, A.R., B.J. Gogel, B.R. Cullis and T. R. 2009. ASReml user guide release 3.0.

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, et al. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9: e90346. doi:10.1371/journal.pone.0090346.

Grams, G.W., C.W. Blessin and G.E. Inglett. 1970. Distribution of tocopherols within the corn kernel. J. Am. Oil Chem. Soc. 47: 337-339. doi:10.1007/bf02638997.

Guralnik, J.M. 2004. Prevalence of anemia in persons 65 years and older in the United States: Evidence for a high rate of unexplained anemia. Blood 104: 2263-2268. doi:10.1182/blood-2004-05-1812.

Hindu, V., N. Palacios-Rojas, R. Babu, W.B. Suwarno, Z. Rashid, R. Usha, et al. 2018. Identification and validation of genomic regions influencing kernel zinc and iron in maize. Theoretical and Applied Genetics 131: 1443-1457. doi:10.1007/s00122-018-3089-3.

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

Huda, K.M.K., M.S.A. Banu, R. Tuteja and N. Tuteja. 2013. Global calcium transducer P-type Ca^{2+}-ATPases open new avenues for agriculture by regulating stress signalling. J. Exp. Bot. 64: 3099-3109. doi:10.1093/jxb/ert182.

Hung, H.Y., C. Browne, K. Guill, N. Coles, M. Eller, A. Garcia, 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. doi:10.1038/hdy.2011.103.

Inoue, H., K. Higuchi, M. Takahashi, H. Nakanishi, S. Mori and N.K. Nishizawa. 2003. Three rice nicotianamine synthase genes, *OsNAS1*, *OsNAS2*, and *OsNAS3* are expressed in cells involved in long-distance transport of iron and differentially regulated by iron. The Plant Journal 36: 366-381. doi:10.1046/j.1365-313X.2003.01878.x.

Institute of Medicine. 2001. Dietary reference intakes for vitamin A, vitamin K, arsenic, boron, chromium, copper, iodine, iron, manganese, molybdenum, nickel, silicon, vanadium, and zinc. National Academy Press, Washington (DC).

JECFA. 2017. Evaluations of the joint FAO/WHO expert committee on food additives. http://apps.who.int/food-additives-contaminants-jecfa-database/search.aspx (accessed 13 April 2019).

Jin, T., J. Chen, L. Zhu, Y. Zhao, J. Guo and Y. Huang. 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. doi:10.1186/s12863-015-0176-1.

Jin, T., J. Zhou, J. Chen, L. Zhu, Y. Zhao and Y. Huang. 2013. The genetic architecture of zinc and iron content in maize grains as revealed by QTL mapping and meta-analysis. Breeding Science 63: 317-324. doi:10.1270/jsbbs.63.317.

Johnson, A.A.T., B. Kyriacou, D.L. Callahan, L. Carruthers, J. Stangoulis, E. Lombi, et al. 2011. Constitutive overexpression of the OsNAS gene family reveals single-gene strategies for effective iron- and zinc-biofortification of rice endosperm. PLoS ONE 6: e24476. doi:10.1371/journal.pone.0024476.

Kanazawa, K., K. Higuchi, N.-K. Nishizawa, S. Fushiya, M. Chino and S. Mori. 1994. Nicotianamine aminotransferase activities are correlated to the phytosiderophore secretions under Fe-deficient conditions in *Gramineae*. J. Exp. Bot. 45: 1903-1906. doi:10.1093/jxb/45.12.1903.

Kenward, M.G. and J.H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53: 983. doi:10.2307/2533558.

Krämer, U., J.D. Cotter-Howells, J.M. Charnock, A.J.M. Baker and J.A.C. Smith. 1996. Free histidine as a metal chelator in plants that accumulate nickel. Nature 379: 635-638. doi:10.1038/379635a0.

Lee, S., U.S. Jeon, S.J. Lee, Y.K. Kim, D.P. Persson, S. Husted, et al. 2009. Iron fortification of rice seeds through activation of the nicotianamine synthase gene. Proceedings of the National Academy of Sciences 106: 22014-22019. doi:10.1073/pnas.0910950106.

Li, S., X. Zhou, Y. Huang, L. Zhu, S. Zhang, Y. Zhao, et al. 2013. Identification and characterization of the zinc-regulated transporters, iron-regulated transporter-like protein (ZIP) gene family in maize. BMC Plant Biol. 13: 114. doi:10.1186/1471-2229-13-114.

Lipka, A.E., M.A. Gore, M. Magallanes-Lundback, A. Mesberg, H. Lin, T. Tiede, et al. 2013. Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. G3 (Bethesda) 3: 1287-1299. doi:10.1534/g3.113.006148.

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, et al. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397-2399. doi:10.1093/bioinformatics/bts444.

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, N.C. p. 733-756.

Lung'aho, M.G., A.M. Mwaniki, S.J. Szalma, J.J. Hart, M.A. Rutzke, L.V. Kochian, et al. 2011. Genetic and physiological analysis of iron biofortification in maize kernels. PLoS One 6: e20429. doi:10.1371/journal.pone.0020429.

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

Manickavelu, A., T. Hattori, S. Yamaoka, K. Yoshimura, Y. Kondou, A. Onogi, 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. doi:10.1371/journal.pone.0169416.

Miller, J.L. 2013. Iron deficiency anemia: A common and curable disease. Cold Spring Harb Perspect Med 3. doi:10.1101/cshperspect.a011866.

Mizuno, D., K. Higuchi, T. Sakamoto, H. Nakanishi, S. Mori and N.K. Nishizawa. 2003. Three nicotianamine synthase genes isolated from maize are differentially regulated by iron nutritional status. Plant Physiol. 132: 1989-1997. doi:10.1104/pp.102.019869.

Neter, J., M.H. Kutner, C.J. Nachtsheim and W. Wasserman. 1996. Applied linear statistical models. McGraw-Hill, Boston.

Norton, G.J., A. Douglas, B. Lahner, E. Yakubova, M.L. Guerinot, S.R.M. Pinson, et al. 2014. Genome wide association mapping of grain arsenic, copper, molybdenum and zinc in rice (*Oryza sativa* L.) grown at four international field sites. PLoS ONE 9: e89685. doi:10.1371/journal.pone.0089685.

Oikeh, S.O., A. Menkir, B. Maziya-Dixon, R. Welch and R.P. Glahn. 2007. Genotypic differences in concentration and bioavailability of kernel-iron in tropical maize varieties grown under field conditions. J. Plant Nutr. 26: 2307-2319. doi:10.1081/pln-120024283.

Oktem, A., A.G. Oktem and H.Y. Emeklier. 2010. Effect of nitrogen on yield and some quality parameters of sweet corn. Commun. Soil Sci. Plant Anal. 41: 832-847. doi:10.1080/00103621003592358.

Ouyang, J., Z. Cai, K. Xia, Y. Wang, J. Duan and M. Zhang. 2010. Identification and analysis of eight peptide transporter homologs in rice. Plant Sci. 179: 374-382. doi:10.1016/j.plantsci.2010.06.013.

Pauli, D., G. Ziegler, M. Ren, M.A. Jenks, D.J. Hunsaker, M. Zhang, et al. 2018. Multivariate analysis of the cotton seed ionome reveals a shared genetic architecture. G3 (Bethesda) 8: 1147-1160. doi:10.1534/g3.117.300479.

Pietz, R.I., J.R. Peterson, C. Lue-Hing and L.F. Welch. 1978. Variability in the concentration of twelve elements in corn grain. Journal of Environment Quality 7: 106. doi:10.2134/jeq1978.00472425000700010021x.

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904-909. doi:10.1038/ng1847.

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

Raboy, V. 1997. Accumulation and storage of phosphate and minerals. In: L. BA and V. IK, editors, Cellular and molecular biology of plant seed development. Kluwer Academic Publishing., Dordrecht (The Netherlands). p. 441–447.

Richau, K.H., A.D. Kozhevnikova, I.V. Seregin, R. Vooijs, P.L.M. Koevoets, J.A.C. Smith, V.B. Ivanov, and H. Schat. 2009. Chelation by histidine inhibits the vacuolar sequestration of nickel in roots of the hyperaccumulator *Thlaspi caerulescens*. New Phytologist 183: 106–116.

Romay, M.C., M.J. Millard, J.C. Glaubitz, J.A. Peiffer, K.L. Swarts, T.M. Casstevens, et al. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14: R55. doi:10.1186/gb-2013-14-6-r55.

Sandstead, H.H. 2000. Causes of iron and zinc deficiencies and their effects on brain. J Nutr 130: 347s-349s.

Sasaki, A., N. Yamaji and J.F. Ma. 2014. Overexpression of OsHMA3 enhances Cd tolerance and expression of Zn transporter genes in rice. J. Exp. Bot. 65: 6013-6021. doi:10.1093/jxb/eru340.

Schaefer, R.J., J.-M. Michno, J. Jeffers, O. Hoekenga, B. Dilkes, I. Baxter, et al. 2018. Integrating coexpression networks with GWAS to prioritize causal genes in maize. The Plant Cell 30: 2922-2942. doi:10.1105/tpc.18.00299.

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

Segura, V., B.J. Vilhjalmsson, A. Platt, A. Korte, U. Seren, Q. Long, et al. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44: 825-830. doi:10.1038/ng.2314.

Seregin, I.V. and A.D. Kozhevnikova. 2004. Strontium transport, distribution, and toxic effects on maize seedling growth. Russian Journal of Plant Physiology 51: 215-221. doi:10.1023/B:RUPP.0000019217.89936.e7.

Šimić, D., S. Mladenović Drinić, Z. Zdunić, A. Jambrović, T. Ledenčan, J. Brkić, et al. 2012. Quantitative trait loci for biofortification traits in maize grain. J. Hered. 103: 47-54. doi:10.1093/jhered/esr122.

Stelpflug, S.C., R.S. Sekhon, B. Vaillancourt, C.N. Hirsch, C.R. Buell, N. de Leon, et al. 2016. An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. Plant Gen. 9. doi:10.3835/plantgenome2015.04.0025.

Sun, G., C. Zhu, M.H. Kramer, S.S. Yang, W. Song, H.P. Piepho, et al. 2010. Variation explained in mixed-model association mapping. Heredity (Edinb) 105: 333-340. doi:10.1038/hdy.2010.11.

Swarts, K., H. Li, J.A.R. Navarro, D. An, M.C. Romay, S. Hearne, et al. 2014. Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Gen. 7: 1-12.

Taiz, L. and E. Zeiger. 2010. Plant physiology. 5th ed. Sinauer Associates, Sunderland, MA.

Takahashi, M., Y. Terada, I. Nakai, H. Nakanishi, E. Yoshimura, S. Mori, et al. 2003. Role of nicotianamine in the intracellular delivery of metals and plant reproductive development. Plant Cell 15: 1263-1280.

Takahashi, R., K. Bashir, Y. Ishimaru, N.K. Nishizawa and H. Nakanishi. 2012. The role of heavy-metal ATPases, HMAs, in zinc and cadmium transport in rice. Plant Signaling & Behavior 7: 1605-1607. doi:10.4161/psb.22454.

Tracy, W.F. 1997. History, genetics, and breeding of supersweet (*shrunken2*) sweet corn. 189-236. doi:10.1002/9780470650073.ch7.

Trimmer, J.S. 2002. Unexpected cross talk: Small GTPase regulation of calcium channel trafficking. Science Signaling. doi:10.1126/stke.2002.114.pe2.

USDA. 2018a. National nutrient database for standard reference. Nutrient Data Laboratory, Beltsville Human Nutrition Research Center, https://ndb.nal.usda.gov/ndb/search/ (accessed 20 July 2018).

USDA. 2018b. Vegetables 2017 summary. National Agricultural Statistics Service, http://usda.mannlib.cornell.edu/usda/nass/VegeSumm//2010s/2018/VegeSumm-02-13-2018.pdf (accessed 12 June 2018).

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414-4423. doi:10.3168/jds.2007-0980.

Velu, G., J. Crossa, R.P. Singh, Y. Hao, S. Dreisigacker, P. Perez-Rodriguez, et al. 2016. Genomic prediction for grain zinc and iron concentrations in spring wheat. Theoretical and Applied Genetics 129: 1595-1605. doi:10.1007/s00122-016-2726-y.

von Wirén, N., S. Klair, S. Bansal, J.-F. Briat, H. Khodr, T. Shioiri, et al. 1999. Nicotianamine chelates both FeIII and FeII. Implications for metal transport in plants. Plant Physiol. 119: 1107-1114. doi:10.1104/pp.119.3.1107.

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

Wessells, K.R. and K.H. Brown. 2012. Estimating the global prevalence of zinc deficiency: Results based on zinc availability in national food supplies and the prevalence of stunting. PLoS ONE 7: e50568. doi:10.1371/journal.pone.0050568.

White, P.J. and M.R. Broadley. 2009. Biofortification of crops with seven mineral elements often lacking in human diets - iron, zinc, copper, calcium, magnesium, selenium and iodine. New Phytol. 182: 49-84. doi:10.1111/j.1469-8137.2008.02738.x.

White, P.J. and M.R. Broadley. 2015. Historical variation in the mineral composition of edible horticultural products. The Journal of Horticultural Science and Biotechnology 80: 660-667. doi:10.1080/14620316.2005.11511995.

Wolfinger, R., W.T. Federer and O. Cordero-Brana. 1997. Recovering information in augmented designs, using SAS PROC GLM and PROC Mixed. Agron. J. 89: 856. doi:10.2134/agronj1997.00021962008900060002x.

Yang, M., K. Lu, F.-J. Zhao, W. Xie, P. Ramakrishna, G. Wang, et al. 2018. Genetic basis of rice ionomic variation revealed by genome-wide association studies. The Plant Cell: tpc.00375.02018. doi:10.1105/tpc.18.00375.

Zhang, Z., R.J. Todhunter, E.S. Buckler and L.D. Van Vleck. 2007. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. J. Anim. Sci. 85: 881-885. doi:10.2527/jas.2006-656.

Zhou, X., S. Li, Q. Zhao, X. Liu, S. Zhang, C. Sun, et al. 2013. Genome-wide identification, classification and expression profiling of *nicotianamine synthase* (*NAS*) gene family in maize. BMC Genomics 14: 238. doi:10.1186/1471-2164-14-238.

Zhou, X. and M. Stephens. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods 11: 407-409. doi:10.1038/nmeth.2848.

Ziegler, G., P. Kear, D. Wu, C. Ziyomo, A. Lipka, M. Gore, et al. 2017. Elemental accumulation in kernels of the maize nested association mapping panel reveals signals of gene by environment interactions. bioRxiv. doi:10.1101/164962.

CHAPTER 4

GENERAL CONCLUSIONS

Nutritional deficiencies are a worldwide problem and affect mainly children, women, and adults over 65 years. Although these deficiencies are far less prevalent in the developed nations, surprisingly large proportions of the US population still do not obtain the daily recommended amount of several nutrients, particularly iron, zinc, vitamin E, and certain carotenoids. Given that sweet corn is the third most commonly consumed vegetable in the US and previous work has shown it to possess high variability for these compounds, we assessed the natural accumulation of tocochromanols, carotenoids, and minerals in fresh kernels of a sweet corn association panel. Genome-wide association studies for these traits enabled us to understand the genetic basis of nutrient accumulation at eating stage and genomic prediction models provided insights of the genetic gains expected in a sweet corn breeding program for increased nutritional quality.

Our study showed that variation for α-tocopherol, the tocochromanol with the highest vitamin E activity, was mostly under the genetic control of *vte4* (γ-tocopherol methyltransferase). Content and composition of tocotrienols, which have greater antioxidant capacity, in fresh sweet corn kernels are controlled by *hggt1* (homogentisate geranylgeranyltransferase) and *vte1* (tocopherol cyclase), respectively. Additionally, we first demonstrated an association of two starch biosynthetic genes (*su1* and *sh2*) specific to sweet corn with tocotrienols, with strong evidence for the involvement of *sh2*. This hypothesis will be further investigated by using near isogenic lines with different alleles for *su1*, *sh2*, and *se1* in two backgrounds (Oh43 and IL451b) from different sources (GRIN, Dr. William Tracy, and Dr. John Juvik). Fresh and mature kernels from these lines will be sampled following the same procedure of this study to assess the effect of those alleles in tocotrienol accumulation. In addition, RNA samples will be collected to

evaluate differential gene expression and determine how those genes modify, if they do, biochemical processes in the kernel. A more thorough assessment of the alleles for the genes identified in this study and others previously reported to associate with tocochromanols will be conducted with a targeted sequencing platform such as rhAmpSeq (Beltz et al., 2018). This will enable us to better assess the allelic variation to determine if sweet corn germplasm captures the most favorable variants that exist for maize and to select more accurately the best lines for these traits.

Whole-genome prediction models had moderate to high predictive abilities for tocochromanol phenotypes, and the inclusion of endosperm mutation type in the models increased the abilities even further for tocotrienol compounds. Although tocochromanols are controlled by a few moderate- to large-effect loci, marker data sets that target 14 genes from previously reported QTL or 81 genes of the precursor and core tocochromanol pathways had on average smaller predictive abilities than whole-genome marker sets, and this reduction was more evident for tocotrienols. Whole-genome models were used to select about 25 lines with increased accumulation of α-tocopherol, total tocotrienols, and total tocochromanols, and these were crossed mostly within the same endosperm mutation group and between lines with complementary alleles at the loci identified in this study. We conducted over 150 crosses, which were self-pollinated for one generation, and four of those families will be planted and evaluated to validate our models.

The analysis of carotenoids showed that variation of β-carotene, the most efficient provitamin A carotenoid, was associated with *crtRB1* (β-carotene hydroxylase). Favorable alleles at this locus were rare but present in both endosperm mutation groups (*su1* and *sh2*). Variation between α- and β-branches of the carotenoid pathway was controlled mostly by *lcyE* (lycopene ε-

cyclase) and given that only a few lines have the allele that favors the β-branch, marker-assisted selection may be performed to increase its presence in the panel and thus improve accumulation of β-carotene (provitamin A) and zeaxanthin (reduced risk of advanced-AMD). Similar to tocotrienols, starch biosynthesis genes were associated with certain carotenoids and therefore, these will also be evaluated with the near-isogenic lines. Given the evidence of lower carotenoid content for double mutant lines *su1se1*, sequencing at the *se1* locus will need to be conducted to clarify any possible role of this gene. Prediction models achieved moderate (β-carotene) and moderately high (lutein) predictive abilities and thus were also used to select the lines for the crosses mentioned previously.

Our study on the kernel ionome at the fresh eating stage of sweet corn identified key regions associated mainly with iron, zinc, and cadmium accumulation. Iron and zinc were associated with markers in proximity with *nas5* (nicotianamine synthase) and cadmium with *hma3* (heavy metal ATPase). Additional loci were associated with nickel and molybdenum in NY and calcium in WI, indicating the presence of gene × environment interaction. From these associations, we identified *ras2* (Ras-like GTPase) and *ptr2* (peptide transporter) as candidates for the variability of calcium and nickel, respectively, but further studies are needed to confirm their involvement. Also, a multivariate analysis using highly correlated ionomic phenotypes together may reveal novel associations not identified in the univariate models. We developed whole-genome prediction models with moderate predictive abilities that may now be used for selecting sweet corn lines with improved kernel element composition, such as increased iron and zinc and lower cadmium. We also showed that starch biosynthesis genes were associated with certain element concentrations, but this information did not improve predictive abilities when included in the models. So far, these models were not used for selecting the lines for crossing but

may be used in the future to develop even more nutritious sweet corn lines, with increased vitamins, antioxidants, and minerals associated with human health.

As with any experiment, there is always room for improvement and things we would have done differently. Although we do have soil samples for one year at one location, I would consider taking multiple samples in each environment. This information could then be interpolated to estimate soil elemental concentration across the field and be used to assess any direct association with kernel element levels. Soil samples would help us determine if the changes observed in the kernel ionome were due to localized environment and thus improve our genetic mapping. Similarly, weather data, which are available, may be used to help explain the differences between locations, as element absorption in the roots is highly associated with soil moisture.

Another thing I would do differently is to better assess the endosperm mutations at starch biosynthesis genes present in each inbred line of the panel beforehand. This would have saved field space and labor, given that the panel had about 20 lines without any endosperm mutation (dent corn), which were grown, pollinated, sampled, and measured, but were not used in our final analysis. This information would also help when estimating concentrations by endosperm mutation type, enabling to distinguish *su1se1* from *su1* lines, which we were not able to do visually or using sparse genetic markers. I would also make sure to have genotypic data for all the lines phenotyped. Although we did our best to match phenotypes and genotypes, we still had about 5-10 lines without genotypic information that could not be used for GWAS. Also, collecting samples for RNA sequencing would greatly help us dissect the genetic signals we identified in GWAS and understand how the mutations at the starch biosynthesis genes interact with the tocochromanol and carotenoid pathways. Due to intensive labor and high costs,

however, RNA sequencing should be conducted using only a subset of lines, such as the two sets of four isolines we plan to analyze in the near future.

As for the field, I wish we had conducted a herbicide trial during the first years to evaluate what products could be used without any damage in the panel. This would allow us to control the weeds much more easily and save a lot of manual labor. Other than that, better monitoring the electric fence is essential, as it stopped working last year and we lost several crosses due to the raccoons feeding from the sweet corn ears.

Finally, this study represents the most extensive assessment of natural variation for tocochromanols (vitamin E and antioxidants), carotenoids (provitamin A, lutein, and zeanxanthin), and element levels required for human health and nutrition (e.g. iron and zinc) in fresh kernels of sweet corn. Through this quantitative genetic analysis we determined key genes responsible for the genetic control of these traits and provided insights into the genetic gains expected using genomic prediction. We have established a key step for improving the nutritional quality of fresh sweet corn kernels and started a new genomics-assisted breeding program to help address nutritional deficiencies reported in the US. First crosses have been performed and continuing work will be done by future students in the Gore Lab, and as a result, we expect to have improved germplasm available for breeders to convert high yielding and locally adapted germplasm to highly nutritious sweet corn lines.

REFERENCES

Beltz, K., D. Tsang, J. Wang, S. Rose, Y. Bao, Y. Wang, K. Larkin, S. Rupp, D. Schrepfer, K. Datta, K. Gunderson, C. Sailor, S. Hansen, J. Dobosy, L. Lewis, A. Menezes, J. Walder, M. Behlke, and C. Chen. 2018. A high-performing and cost-effective SNP genotyping method using rhPCR and universal reporters. Advances in Bioscience and Biotechnology 9:497-512.