Combining Ability, Association Mapping, and Genomic Predictions for Provitamin A Carotenoid Concentrations in Tropical Maize (*Zea mays* L.)

By

Willy B. Suwarno

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Plant Breeding and Plant Genetics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2012

Date of final oral examination: 12/10/12

The dissertation is approved by the following members of the Final Oral Committee:

Kevin V. Pixley, Adjunct Associate Professor, Agronomy
Shawn M. Kaeppler, Professor, Agronomy
William F. Tracy, Professor, Agronomy
Cecile M. Ane, Associate Professor, Statistics
Raman Babu, Scientist, CIMMYT

ABSTRACT

Developing biofortified maize cultivars is a promising approach to overcome the widespread problem of vitamin A deficiency in the developing world. The objectives of a first study were to: (1) evaluate whether molecular marker-based genetic distance separation of maize lines into heterotic groups results in heterosis among groups that could further be developed into a useful heterotic pattern, and (2) assess gene action (general and specific combining ability, GCA and SCA) for grain yield and provitamin A concentrations among inbred lines representing putative heterotic groups. A second, association mapping study was conducted to (3) identify genes and genic regions controlling variation for carotenoid concentrations, (4) use additive linear models of selected SNP markers to predict carotenoid concentrations of lines in breeding programs, and (5) assess the suitability of association mapping analysis models using four association mapping panels. To address objectives (1) and (2), 21 lines were crossed following a modified North Carolina Design II with six sets, where sets 1-3 contained crosses between putative heterotic groups, and sets 4-6 were crosses within groups. The resultant 152 hybrids were evaluated in two-replicate trials at four environments in Mexico. Significant but small yield advantage of among versus within putative heterotic group crosses (0.4 Mg ha^{-1}, $P<0.05$) confirmed that genetic distance can be useful, but that further breeding work would be needed to develop these groups into a useful heterotic pattern. GCA effects were significant for both provitamin A concentration and grain yield, whereas SCA effects were significant only for grain yield, indicating that provitamin A concentration is controlled primarily by additive gene action. For objectives (3) and (4), association mapping identified the zeaxanthin epoxidase gene ($R^2=0.14$), and a significant marker ($R^2=0.10$) located close to the β-carotene hydroxylase gene,

CrtRB1, as important regions determining carotenoid phenotypes. Additive linear models using selected SNPs accurately predicted carotenoid concentrations of maize lines ($r \geq 0.8$, $P < 0.01$). For objective (5), the association mapping panels identified the phytoene synthase and the opaque-2 genes. Results of these field and molecular studies provided useful insights to enhance the effectiveness of provitamin A breeding efforts in maize.

ACKNOWLEDGEMENTS

Many people greatly contributed to my graduate career at the University of Wisconsin-Madison. On this occasion, I would like to deeply thank:

1. Dr. Kevin Pixley, for his advisory guidance and mentorship throughout my graduate study at the University of Wisconsin-Madison. His broad knowledge on applied maize breeding, leadership, and willingness to help will continue to inspire me throughout my scientific career.
2. Dr. Shawn Kaeppler, Dr. William Tracy, Dr. Cecile Ane, and Dr. Raman Babu for their useful suggestions and service on my committee.
3. Dr. Raman Babu for his guidance on molecular issues, particularly association mapping analyses.
4. Dr. Natalia Palacios for her expert advice on carotenoids, including access to the carotenoid association mapping panel, and to CIMMYT's Maize Nutritional Quality Laboratory staff, under the guidance of Dr. Natalia Palacios, for their assistance in carotenoid analyses.
5. Mr. German Minngram and field staff at CIMMYT's Tlaltizapan and Agua Fria research stations for their assistance in management of the nurseries and field trials.
6. Dr. Jose Crossa for sharing his knowledge and ideas on genomic prediction and association mapping, and Dr. Sherry Tanumihardjo for providing suggestion about biomarkers for vitamin A status.
7. HarvestPlus, CIMMYT, and the Directorate General of Higher Education of Indonesia for their generous funding support for my research projects and graduate study.

8. CIMMYT for access to research facilities, equipment, etc., all of which facilitated my research projects. I also thank the molecular breeding program of the CIMMYT's Global Maize Program for providing access to various association mapping data.
9. Current and former fellow graduate students and friends for their help and togetherness during our great time in Madison.
10. Dr. Hajrial Aswidinnoor, the late Dr. Sriani Sujiprihati, Dr. Sobir, Dr. Bambang Purwoko, the late Dr. Amris Makmur, the late Dr. Jajah Koswara, Dr. Syafrida Manuwoto, Dr. Muhamad Syukur, Dr. Agus Purwito, Dr. Ernan Rustandi, and colleagues at the Bogor Agricultural University, Indonesia, for inspiration that motivated and prepared me for my graduate study at UW and to embark on a career in plant breeding.
11. My family, especially my parents, Dr. Faiza Suwarno and Dr. Suwarno, for their continuous support and encouragement. I am also grateful to my wife, Dwi Wulansari, and children, Dafa, Ihsan, and Nisa for their support and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTAL TABLES

LIST OF SUPPLEMENTAL FIGURES

CHAPTER I

GENERAL INTRODUCTION

BACKGROUND AND RESEARCH RATIONALE

Maize is one of the most widely consumed staple foods and accounts for 30-50% of total caloric intake for many people in sub-Saharan Africa (Nuss and Tanumihardjo, 2010; Smale et al., 2011). While agriculture accounts for 20 to 40% of the gross domestic product in this region, there is a large yield gap between potential and achieved maize yields (Godfray et al., 2010), partly because of frequent droughts and pervasive low-input farming practices. Additionally, many people in this region suffer from malnutrition, including widespread vitamin A deficiency (VAD) (WHO, 2009; Nuss and Tanumihardjo, 2010). VAD can cause night blindness and possibly lead to corneal blindness, as well as cause stunted growth among affected children (West, 1991; West and Darnton-Hill, 2008). Development of high yielding maize varieties that are biofortified with biologically-usefully-high concentrations of provitamin A carotenoids in the grain is regarded as a key approach toward alleviating VAD in these regions (Ortiz-Monasterio et al., 2007; Pfeiffer and McClafferty, 2007).

Hybrid maize cultivars have the advantages of higher yield potential and better phenotypic appeal than open pollinated cultivars for many purposes (Dhillon, 1998). Hybrid maize breeding includes development of stable, vigorous, high-yielding inbred lines with the extensive evaluation of combining ability during the process of developing the lines, followed by use of selected inbred lines in development of improved hybrids (Singh, 1987). The value of any inbred line in hybrid breeding ultimately depends on its ability to combine with other lines to produce superior hybrids (Dhillon, 1998). Sprague and Tatum (1942) partitioned the combining

ability in single crosses into general and specific combing abilities. General combining ability (GCA) is "the average performance of an inbred line in a series of hybrid combinations", whereas specific combining ability (SCA) is "the deviation of the hybrid from the performance expected on the basis of general combining ability". Estimation of GCA and SCA effects has been widely applied in maize breeding programs to predict inbred line performance.

Characterization of maize lines for their combining ability is routinely conducted for numerous traits, including adaptation to drought and low N stress conditions (Betran et al., 2003; Medici et al., 2004), soil acidity (Welcker et al., 2005), aflatoxin accumulation (Williams et al., 2008), resistance to pathogens causing lodging (Moreno-Gonzalez et al., 2004), mite (Bynum et al., 2004), Striga hermonthica and Striga asiatica (Gethi and Smith, 2004), maize weevil (Dhliwayo et al., 2005), and many more. Combining ability of maize lines is also important for value-added traits, including nutritional characteristics, such as grain Fe and Zn density (Long et al., 2004), provitamins A, protein, oil and starch content, and grain yield.

The increase in size, vigor, or productivity of a hybrid plant over its parents is known as hybrid vigor or heterosis (Allard, 1960; Poehlman, 1983). The expression of heterosis (H) over mid-parent depends on the difference in allele frequency (y) of the parents and dominance effects (d) at various loci, that is $H = dy^2$ (Falconer and Mackay, 1996). Therefore, genetic diversity is very important in maize breeding, and inbred lines are typically developed from two or more genetically different populations to obtain high levels of heterosis in their hybrids (Singh, 1987). Successful development of maize inbred lines for use in hybrid formation is based on the identification and utilization of heterotic groups and patterns (Melani and Carena, 2005). Development of heterotic groups and assignment of experimental inbred lines into the

established heterotic groups before making hybrid crosses is time and cost efficient, because the number of crosses to be made and evaluated will decrease substantially. However, rather than naturally existing in the germplasm, strong evidence and widespread experience indicate that heterotic patterns are developed by plant breeders (Tracy and Chandler, 2006).

Molecular markers have been used to estimate genetic distance (GD) between maize inbred lines, study the extent of population structure, classify germplasm into heterotic groups, and predict future hybrid performance (Betran et al., 2003b; Oritz-Monasterio et al., 2007). Amplified fragment length polymorphism (AFLP) markers have been found efficient in assigning maize lines to heterotic groups and AFLP-based GD has been useful for predicting maize single cross performance for intra-population crosses of broad-based populations (Barbosa et al., 2003). Another marker system, simple sequence repeats (SSR), has also been used for GD-based grouping in maize (Barata and Carena, 2006; Barbosa et al., 2003; Phumichai et al., 2008) and wheat (Dreisigacker et al., 2005). Single Nucleotide Polymorphism (SNP) markers have been used more recently to perform such investigations, including the use of 449 unbiased SNPs to classify temperate and tropical/subtropical lines, yellow and white kernel lines, and dent and flint lines (Lu et al., 2009). Use of molecular-marker-determined GD to predict heterosis, however, has been a challenge because even though there is generally a significant correlation between GD and heterosis among inbred lines, the predictive ability of GD is largest for closely related lines and decreases for lines that have greater GD amongst them (Melchinger et al, 1999).

Introducing exotic germplasm into breeding programs can increase the genetic base from which elite commercial inbreds are derived (Glover et al., 2005). Previous studies have indicated the considerable potential of crossing subtropical and tropical germplasm (Vasal et al.,

1992) to achieve higher grain yields in the subtropics, or using maize landraces as sources to improve forage yield and quality in warm temperate areas (Bertoia et al., 2006). Breeding of maize in the Atlantic coast of Europe, for example, exploits heterosis between flint and dent germplasm (Soengas et al., 2006). Effective use of exotic germplasm in breeding programs, however, requires strategic introgression of the exotic material, respecting or building upon the already successful heterotic patterns. The use of molecular tools to assess the genetic relationships among elite and exotic lines can help guide the introgression of exotic germplasm and is one of the underlying objectives for the study of GDs and their association with hybrid performance.

Breeding for increased concentrations of provitamin A is promising because there is considerable genetic variation available for this trait in maize germplasm. Initial studies at the International Maize and Wheat Improvement Center (CIMMYT) revealed that among 1000 tropical maize genotypes, total provitamin A concentration varied from 0.24 to 8.80 μg g^{-1}, while the proportion of provitamin A to total carotenoids ranged between 5-30% (Ortiz-Monasterio et al., 2007). Further, the HarvestPlus, through CIMMYT and other partners, has been conducting extensive work on improving provitamin A level in elite maize lines, hybrids and synthetic populations. Classical and molecular breeding methods have been implemented, including use of various temperate and tropical sources with high concentrations of provitamin A, and marker assisted selection for reduced-function alleles of the lycopene epsilon-cyclase (*LcyE*) (Harjes et al., 2008) and β-carotene hydroxylase 1 (*CrtRB1*) (Yan et al., 2010; Babu et al., 2012) genes. Recently, in a number of improved inbred lines and populations, the concentration of provitamin A in the grain has reached 15-20 μg g^{-1} (N. Palacios, CIMMYT, pers. comm.). In terms of

breeding products, three-way crosses are preferred over single-cross hybrids in several maize consuming countries where VAD is prevalent, because seed production is less expensive while considerable yield potential and uniformity of the hybrids can be achieved.

Marker assisted selection (MAS) is regarded as a key approach for facilitating efficient breeding for high levels of provitamin A carotenoids in maize (Prasanna et al., 2010). Two genes, *CrtRB1* on chromosome 10 (Yan et al., 2010) and *LcyE* on chromosome 8 (Harjes et al. 2008), have been reported to affect provitamin A carotenoid concentrations in maize grain, where the former has larger effect than the latter (Babu et al., 2012). While MAS for favorable allele(s) of *CrtRB1* has been very helpful during development of outstanding high provitamin A maize lines, the carotenoid pathway is diverse and many genes play critical roles. Therefore, searching for favorable alleles for other important (rate-limiting) genes and use of genomic prediction tools are promising approaches to enhance selection efficiency.

Association mapping has been extensively used to dissect complex traits in maize (Yan et al., 2011), complementing the linkage mapping approach. Population structure and genetic relationships among lines are two major challenges that can cause spurious associations between markers and the trait of interest in association studies using collections of inbred lines. Some models have been developed to correct for population structure in mapping populations, including models with SNP marker and population structure information (Price et al., 2006), and mixed linear models which consider population and familial structure in addition to genetic marker information (Yu et al., 2006).

Use of a high marker density in association mapping is of great importance to capture rare variations in the genome. Most recently, the genotype-by-sequencing (GBS) platform

(Elshire et al., 2011) enables the possibility of genotyping using more than a million SNPs. Furthermore, while using a large, well-defined population such as the nested association mapping (NAM) panel in maize (McMullen et al., 2009) is ideal for general association studies, association panels comprised of elite germplasm from the respective breeding programs, such as CIMMYT's carotenoid association mapping (CAM), Drought Tolerant Maize for Africa (DTMA), and Improved Maize for African Soils (IMAS) panels (www.cimmyt.org) are particularly useful in identifying marker-trait associations that are relevant and of practical utility to the target breeding programs. The use of broad genotypic variation for specific traits of interest is expected to increase the power to identify rare variants.

The use of molecular plant breeding tools can enhance efficiency by replacing phenotypic with genotypic selection during some stages of the breeding process, thereby reducing overall phenotyping costs, biases (e.g. caused by environment factors and genotype by environment interactions), and measurement errors (Moose and Mum, 2008). Selection based on genomic predictions offers the opportunity to accurately predict carotenoid concentrations. With the increasing accessibility of high volume, low cost genotyping platforms, genomic selection can be practically of great use in breeding programs, especially for early-generation screening, when the number of families and progenies are typically large.

OBJECTIVES OF THE DISSERTATION RESEARCH

The objectives of this research were developed to inform and enhance the efficiency or effectiveness of maize provitamin A biofortification breeding efforts at CIMMYT and globally. The specific objectives of this research were to:

1. Evaluate whether molecular marker-based genetic distance separation of maize lines into heterotic groups results in heterosis among groups that could further be developed into a useful heterotic pattern.
2. Assess gene action (general and specific combining ability, GCA and SCA) for grain yield and total provitamin A concentrations among inbred lines representing putative heterotic groups from (1).
3. Identify genes and genic regions controlling variation for carotenoid concentrations.
4. Use additive linear models of selected SNP markers to predict carotenoid concentrations of lines in breeding programs.
5. Assess the suitability of association mapping analysis models using four association mapping panels.

FORMAT OF THE DISSERTATION AND OVERVIEW OF METHODS

The dissertation consist of a general abstract plus five chapters, of which chapters II, III and IV will each be modified for publication in refereed international journals.

- Chapter I contains a general introduction, background, rationale and objectives of the research.
- Chapter II addresses objectives 1 and 2. The chapter describes results of molecular marker-based genetic distance separation of maize lines into putative heterotic groups, and of replicated field trials of hybrids to study combining ability among inbred lines for grain yield and total provitamin A concentrations.
- Chapter III investigates objectives 3 and 4. The chapter discusses association mapping results identifying genes and genic regions controlling variation for carotenoid

concentrations, and development of additive linear models of selected SNP markers to predict carotenoid concentrations of lines.

- Chapter IV addresses objective 5. The chapter describes results of association mapping analyses for grain color and QPM binary phenotypes using four association mapping panels and their combined meta panel, and grouping of the lines based on the K-means clustering followed by the discriminant analysis of principal components (DAPC) methods.
- Chapter V contains conclusions of the research and future perspectives.

REFERENCES

Allard, R.W. 1960. Principles of Plant Breeding. John Willey and Sons. New York. 485p.

Babu, R., N.P. Rojas, S. Gao, J. Yan, and K. Pixley. 2012. Validation of the effects of molecular marker polymorphisms in LcyE and CrtRB1 on provitamin A concentrations for 26 tropical maize populations. Theoretical and Applied Genetics . Available at http://www.springerlink.com/index/10.1007/s00122-012-1987-3 (verified 12 October 2012).

Barata C., Carena M. 2006. Classification of North Dakota maize inbred lines into heterotic groups based on molecular and testcross data. Euphytica 151:339-349.

Barbosa A.M.M., Geraldi I.O., Benchimol L.L., Garcia A.A.F., Souza C.L., Souza A.P. 2003. Relationship of intra- and interpopulation tropical maize single cross hybrid performance and genetic distances computed from AFLP and SSR markers. Euphytica 130:87-99.

Bertoia L., Lopez C.s., Burak R. 2006. Biplot Analysis of Forage Combining Ability in Maize Landraces. Crop Science 46:1346-1353.

Betran F.J., Beck D., Banziger M., Edmeades G.O. 2003. Genetic Analysis of Inbred and Hybrid Grain Yield under Stress and Nonstress Environments in Tropical Maize. Crop Science 43:807.

Betran F.J., Ribaut J.M., Beck D., Leon D.G.d. 2003b. Genetic diversity, specific combining ability, and heterosis in tropical maize under stress and nonstress environments. Crop Science 43:797-806.

Bynum E.D., Jr., Xu W., Archer T.L. 2004. Diallel Analysis of Spider Mite Resistant Maize Inbred Lines and F_1 Crosses. Crop Science 44:1535-1541.

Dhillon, BS. 1998. Maize. In: Hybrid Cultivar Development. S. S. Banga and S. K. Banga (Eds). Narosa Publishing House, New Delhi, India. p:282-315.

Dhliwayo T., Pixley K.V., Kazembe V. 2005. Combining Ability for Resistance to Maize Weevil among 14 Southern African Maize Inbred Lines. Crop Science 45:662-667.

Dreisigacker S., Melchinger A.E., Zhang P., Ammar K., Flachenecker C., Hoisington D., Warburton M.L. 2005. Hybrid performance and heterosis in spring bread wheat, and their relations to SSR-based genetic distances and coefficients of parentage. Euphytica 144:51-59.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J. a Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PloS one 6(5): e19379. Available at

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087801&tool=pmcentrez&rendertype=abstract (verified 18 July 2011).

Falconer, D.S. and T.F.C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longman, New York.

Gethi J.G., Smith M.E. 2004. Genetic Responses of Single Crosses of Maize to Striga hermonthica (Del.) Benth. and Striga asiatica (L.) Kuntze. Crop Science 44:2068-2077.

Glover M.A., Willmot D.B., Darrah L.L., Hibbard B.E., Zhu X. 2005. Diallel Analyses of Agronomic Traits Using Chinese and U.S. Maize Germplasm. Crop Science 45:1096-1102.

Godfray, H.C.J., J.R. Beddington, I.R. Crute, L. Haddad, D. Lawrence, J.F. Muir, J. Pretty, S. Robinson, S.M. Thomas, and C. Toulmin. 2010. Food security: the challenge of feeding 9 billion people. Science (New York, N.Y.) 327(5967): 812–8. Available at http://www.ncbi.nlm.nih.gov/pubmed/20110467 (verified 5 October 2012).

Harjes, C.E., T.R. Rocheford, L. Bai, T.P. Brutnell, C.B. Kandianis, S.G. Sowinski, A.E. Stapleton, R. Vallabhaneni, M. Williams, E.T. Wurtzel, J. Yan, and E.S. Buckler. 2008. Natural genetic variation in lycopene epsilon-cyclase tapped for maize biofortification. Science 319(5861): 330-3.

Long J.K., Banziger M., Smith M.E. 2004. Diallel Analysis of Grain Iron and Zinc Density in Southern African-Adapted Maize Inbreds. Crop Science 44:2019-2026.

Lu Y, J. Yan, C T. Guimaraes, S. Taba, Z. Hao, S. Gao, S. Chen, J. Li, S. Zhang, B. S. Vivek, C. Magorokosho, S. Mugo, D. Makumbi, S. N. Parentoni, T. Shah, T. Rong, J. H. Crouch, Y. Xu. 2009. Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theor Appl Genet DOI 10.1007/s00122-009-1162-7.

McMullen, M.D., S. Kresovich, H.S. Villeda, P. Bradbury, H. Li, Q. Sun, S. Flint-Garcia, J. Thornsberry, C. Acharya, C. Bottoms, P. Brown, C. Browne, M. Eller, K. Guill, C. Harjes, D. Kroon, N. Lepak, S.E. Mitchell, B. Peterson, G. Pressoir, S. Romero, M. Oropeza Rosas, S. Salvo, H. Yates, M. Hanson, E. Jones, S. Smith, J.C. Glaubitz, M. Goodman, D. Ware, J.B. Holland, and E.S. Buckler. 2009. Genetic properties of the maize nested association mapping population. Science (New York, N.Y.) 325(5941): 737–40. Available at http://www.ncbi.nlm.nih.gov/pubmed/19661427 (verified 12 July 2012).

Medici L.O., Pereira M.B., Lea P.J., Azevedo R.A. 2004. Diallel analysis of maize lines with contrasting responses to applied nitrogen. The Journal of Agricultural Science 142:535-541.

Melani M.D., Carena M.J. 2005. Alternative Maize Heterotic Patterns for the Northern Corn Belt. Crop Science 45:2186-2194.

Melchinger, A.E. 1999. Genetic diversity and heterosis. *In*: The Genetics and Exploitation of Heterosis in Crops. J.G. Coors and S. Pandey (eds). Madison, Wisconsin.

Moose, S. P. and R. H. Mumm. 2008. Molecular Plant Breeding as the Foundation for 21st Century Crop Improvement. Plant Physiology, July 2008, Vol. 147, pp. 969–977.

Moreno-Gonzalez J., Ares J.L.A., Ferro R.A., Ramarez L.C. 2004. Genetic and statistical models for estimating genetic parameters of maize seedling resistance to Fusarium graminearum Schwabe root rot. Euphytica 137:55-61.

Nuss E.T., S.A. Tanumihardjo. 2010. Maize: A paramount staple crop in the context of global nutrition. Compr Rev Food Sci Food Saf 9: 417–436.

Ortiz-Monasterio, J.I., N. Palacios-Rojas, E. Meng, K. Pixley, R. Trethowan, R.J. Pena. 2007. Enhancing the mineral and vitamin content of wheat and maize through plant breeding. Journal of Cereal Science 46: 293–307.

Pfeiffer, W.H., and B. McClafferty. 2007. HarvestPlus: Breeding Crops for Better Nutrition. Crop Sci. 47(Supplement_3): S-88

Phumichai C., Doungchan W., Puddhanon P., Jampatong S., Grudloyma P., Kirdsri C., Chunwongse J., Pulam T. 2008. SSR-based and grain yield-based diversity of hybrid maize in Thailand. Field Crops Research 108:157-162. DOI: 10.1016/j.fcr.2008.04.009.

Poehlman, J. M. 1983. Breeding Field Crops. Second ed. The Avi Publishing Company, Inc. Westport. 486p.

Prasanna, B.M., K. Pixley, M.L. Warburton, and C.-X. Xie. 2010. Molecular marker-assisted breeding options for maize improvement in Asia. Molecular Breeding 26(2): 339–356. Available at http://www.springerlink.com/index/10.1007/s11032-009-9387-3 (verified 19 September 2012).

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. Nature genetics 38(8): 904–9. Available at http://www.ncbi.nlm.nih.gov/pubmed/16862161 (verified 26 October 2012).

Singh, J. 1987. Field Manual of Maize Breeding Procedures. Food and Agriculture Organization of The United Nations. Rome.

Smale, M., D. Byerlee, and T. Jayne. 2011. Maize Revolutions in Sub-Saharan Africa. Policy Research Working Paper 5659. The World Bank.

Soengas P., Ordas B., Malvar R.A., Revilla P., Ordas A. 2006. Combining Abilities and Heterosis for Adaptation in Flint Maize Populations. Crop Science 46:2666-2669.

Sprague, G. F. and L. A. Tatum. 1942. General vs Specific Combining Ability in Single Cross of Corn. J. Am. Soc. Agron. 32:923-32.

Tracy, W.F. and M.A. Chandler. 2006. The historical and biological basis of the concept of heterotic patterns in corn bent dent maize. In: Lamkey, K.R. and Lee, M. Plant breeding: the Arnel R. Hallauer International Symposium. Blackwell Publishing.

Vasal S.K., Srinivasan G., Han G.C., Pandey S., et al. 1992. Heterosis and combining ability of CIMMYT's tropical x subtropical maize germplasm. Crop Science 32:1483.

Welcker C., The C., Andreau B., De Leon C., Parentoni S.N., Bernal J., Felicite J., Zonkeng C., Salazar F., Narro L., Charcosset A., Horst W.J. 2005. Heterosis and Combining Ability for Maize Adaptation to Tropical Acid Soils: Implications for Future Breeding Strategies. Crop Science 45:2405-2413.

West, K.P. 1991. Dietary vitamin A deficiency: effects on growth, infection, and mortality. Food and Nutr. Bull. 13(2).

West, K.P. and I. Darnton-Hill. 2008. Vitamin A deficiency. In: Semba, R.D., M.W. Bloem (eds). Nutrition and Health in Developing Countries. Second edition. Humana Press. New Jersey, USA.

WHO. 2009. Global prevalence of vitamin A deficiency in populations at risk 1995–2005. WHO Global Database on Vitamin A Deficiency. Geneva, World Health Organization.

Williams W.P., Windham G.L., Buckley P.M. 2008. Diallel Analysis of Aflatoxin Accumulation in Maize. Crop Science 48:134-138.

Yan, J., C.B. Kandianis, C.E. Harjes, L. Bai, E.-H. Kim, X. Yang, D.J. Skinner, Z. Fu, S. Mitchell, Q. Li, M.G.S. Fernandez, M. Zaharieva, R. Babu, Y. Fu, N. Palacios, J. Li, D. Dellapenna, T. Brutnell, E.S. Buckler, M.L. Warburton, and T. Rocheford. 2010. Rare genetic variation at Zea mays CrtRB1 increases beta-carotene in maize grain. Nature genetics 42(4): 322-7.

Yan, J., M. Warburton, and J. Crouch. 2011. Association Mapping for Enhancing Maize (L.) Genetic Improvement. Crop Science 51(2): 433. Available at https://www.crops.org/publications/cs/abstracts/51/2/433 (verified 18 July 2012).

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38(2): 203–8. Available at http://www.ncbi.nlm.nih.gov/pubmed/16380716 (verified 12 July 2012).

CHAPTER II

COMBINING ABILITY FOR GRAIN YIELD AND PROVITAMIN-A CAROTENOID CONCENTRATIONS IN TROPICAL MAIZE

ABSTRACT

Developing biofortified maize cultivars is one significant approach to overcome the widespread problem of vitamin A deficiency in the developing world, especially in sub-Saharan Africa. The objectives of this study were to: 1) evaluate whether marker-based separation of materials into heterotic patterns effectively maximized among group cross heterosis, 2) assess gene action (general and specific combining ability, GCA and SCA) for grain yield and provitamin A concentrations of hybrids among 21 tropical maize inbred lines representing the three proposed groups, and 3) to assess the degree of association between grain yield and provitamin A concentrations. The lines were crossed following a modified North Carolina Design II with six sets, where sets 1 - 3 contained crosses between putative heterotic groups (1x2, 1x3, and 2x3), and sets 4 - 6 were crosses within groups (1x1, 2x2, and 3x3). This resulted in 152 hybrids, after bulking reciprocals. The hybrids were evaluated at four environments in Mexico using an alpha-lattice design with two replications of one-row plots. The first plant in each plot was self-pollinated for provitamin A analysis. Significant but small yield advantage of among versus within putative heterotic group crosses (0.4 Mg ha^{-1}, $P<0.05$) suggests that further breeding work may be effective in developing useful heterotic groups from those putatively identified by maximizing genetic distances. A two-step approach, using genetic distance to predict, and SCA to further refine the assignment of lines into groups, appeared to be better than prediction of heterotic groups based on genetic distances alone. GCA and SCA were significant

for all traits, except SCA for provitamin A concentration, indicating that provitamin A concentration is controlled primarily by additive gene action. Grain yield was not significantly correlated with total provitamin A concentration, indicating that both traits could be improved simultaneously.

Keywords: combining ability, provitamin A, biofortification, heterotic groups

INTRODUCTION

Vitamin A deficiency (VAD) is a significant human health problem in developing countries, especially in sub-Saharan Africa and Southeast Asia (WHO, 2009). This problem, recently defined as a liver retinol reserve of <0.1 μmol g^{-1} liver (Tanumihardjo, 2011), is most severe among preschool-aged children and pregnant woman (Rice et al., 2004). Vitamin A is an essential micronutrient controlling several biological processes including vision, growth, and immunity. VAD can cause night blindness possibly leading to corneal blindness, as well as stunted growth among affected children (West, 1991; West and Darnton-Hill, 2008). Undersupply of vitamin A can compromise the immune system thereby increasing the risk of mortality from several infectious diseases, including measles, diarrhea, and malaria (Rice et al., 2004).

Because maize is a staple food in many parts of Africa, development of maize varieties that are biofortified with biologically-usefully-high concentrations of provitamin A carotenoids in the grain is regarded as a key approach toward alleviating VAD in these regions (Ortiz-Monasterio et al., 2007; Pfeiffer and McClafferty, 2007). Although there is no consensus about the exact ratio, recent reports indicate that bioconversion of β-carotene from maize to vitamin A occurs at a ratio of about 2.8 mg:1 mg (Howe and Tanumihardjo, 2006) or 3.2 mg:1 mg

(Muzhingi et al., 2011). Other carotenoids, lutein and zeaxanthin, are available in greater quantity in maize than β-carotene; however, although they have other health benefits such as antioxidants, they do not have provitamin A activity (Nuss and Tanumihardjo, 2010). Furthermore, a recent study indicated that provitamin A could enhance bioavailability of Fe, which is also an essential micronutrient whose deficiency causes widespread health problems (Pixley et al., 2011).

Breeding for increased concentrations of provitamin A is promising because there is considerable genetic variation available in maize germplasm. Initial CIMMYT studies revealed that among 1000 tropical maize genotypes, total provitamin A varied from 0.24 to 8.80 μg g^{-1}, while the proportion of provitamin A to total carotenoids ranged between 5-30% (Ortiz-Monasterio, 2007). Further, the HarvestPlus project has been conducting extensive work on improving provitamin A level in elite maize lines, hybrids and synthetic populations. Classical and molecular breeding methods have been implemented, including use of various temperate and tropical sources with high concentrations of provitamin A, and marker assisted selection for reduced-function alleles of the lycopene epsilon-cyclase (*LcyE*) (Harjes et al., 2008) and beta-carotene hydroxylase 1 (*CrtRB1*) (Babu et al., 2012; Yan et al., 2010) genes. Recently, in a number of improved inbred lines and populations, the concentration of provitamin A in the grain has reached 15-20 μg g^{-1} (Babu et al., 2012).

In terms of breeding products, three-way crosses are preferred over single-cross hybrids in several maize consuming countries where VAD is prevalent because seed production is less expensive while considerable uniformity of the hybrid plants can still be obtained. In hybrid development programs, understanding general and specific combining ability (GCA and SCA,

respectively) of and between lines, and forming and exploiting meaningful heterotic groups are key aspects for success. Egesel et al. (2003) evaluated a ten-parent diallel (45 temperate maize hybrids) and found that variation for carotenoids was more attributable to general than to specific combining ability effects, indicating a major role for additive gene action.

Molecular markers approaches have been used to estimate genetic distance between maize inbred lines, study the extent of population structure, and classify germplasm into heterotic groups (Oritz-Monasterio et al., 2007). Single Nucleotide Polymorphism (SNP) markers have been used more recently, including the use of 449 unbiased SNPs to distinguish temperate and tropical/subtropical lines, yellow and white kernel lines, and dent and flint lines (Lu et al., 2009). Use of markers to predict heterosis has been a challenge because even though there are correlations between genetic distances of inbred lines and heterosis, the predictive ability is low especially for lines that have greater genetic distances (Melchinger et al, 1999).

The objectives of this study were to: 1) evaluate whether marker-based separation of materials into heterotic patterns effectively maximized among group cross heterosis, 2) assess gene action (GCA and SCA) for grain yield and provitamin A concentrations of hybrids among 21 tropical maize inbred lines representing the three proposed groups, and 3) to assess the degree of association between grain yield and provitamin A concentrations.

MATERIALS AND METHODS

Molecular marker analysis and assignment of parent lines to heterotic groups

The SNPs were assayed at Kbiosciences, UK using the KasPar chemistry (www.kbioscience.co.uk). A total of 127 advanced (promising) inbred lines from the

CIMMYT/HarvestPlus maize provitamin A biofortification project (Supplemental Table II-1) were assigned to three putative heterotic groups developed using shared-allele genetic distances calculated from 402 SNP markers followed by Neighbor-joining tree construction (Figure II-1). Shared-allele genetic distance was calculated as one minus the proportion of shared to non-shared alleles across 402 SNPs for each pair of inbred lines. Based on pedigree distinctness within their respective putative heterotic group and on high provitamin A concentration in the grain, eight of the lines assigned to group 1, seven from group 2, and six from group 3 were further selected to form hybrids (Table II-1).

Parent lines and formation of hybrids

The selected lines (Table II-1) were crossed to each other following a modified North Carolina Design II (Hallauer et al., 2010) using six sets of crosses, where sets 1 to 3 contained crosses between heterotic groups (1x2, 1x3, and 2x3) and sets 4 to 6 were crosses within groups (1x1, 2x2, 3x3). The attempted number of crosses were 8 x 7 lines = 56 hybrids for set 1; 8 x 6 lines = 48 hybrids for set 2; 7 x 6 lines = 42 hybrids for set 3; and 3 x 3 lines = 9 hybrids each for sets 4, 5, and 6. The within-group crosses were designed by sub-dividing each group to maximize within-group distances, in an attempt to avoid crosses among very closely related lines. The crosses were performed during June–October 2010 at CIMMYT's Tlaltizapan experimental station, Mexico, and January – May 2011 at CIMMYT's Agua Fria experimental station, Mexico. A total of 156 hybrids were obtained after bulking reciprocals, where set 1 to 6 consisted of 47, 45, 38, 9, 8, and 9 hybrids, respectively.

Field experiments

The hybrids were evaluated at four locations in Mexico: 121 hybrids obtained from Tlaltizapan 2010 crosses and three check cultivars were evaluated at Agua Fria, Puebla (AF) (20°32´N, 97°28´ W; 110 m above sea level (masl); average annual temperature 22°C; average annual precipitation 1200 mm) during winter 2010-2011, whereas 137 hybrids resulted from crosses in AF 2011 along with three check cultivars were evaluated in Tlaltizapan, Morelos (18°41´ N, 99°07´ W; 945 masl; average annual temperature 23.5°C; average annual precipitation 840 mm) conventional tillage (TL) (six tillage operations), Tlaltizapan conservation agriculture (TLCA) (one tillage operation, to reform planting beds), and Mexico's National Institute of Forestry, Agriculture and Livestock Research (INIFAP) Celaya, Guanajuato (CE) research station (20°26'N, 103°19'W; 1750 masl; average annual temperature 19°C; average annual precipitation 700 mm) during summer 2011. The experimental design was an alpha-lattice with two replications and 14 incomplete blocks in each replication (there were only 13 incomplete blocks at AF due to lesser number of hybrids evaluated). The plot size was 5 m x 1 row and plant densities were approximately 66,670 plants ha^{-1} (0.75 m between rows and 0.20 m between plants within a row), except at CE, where density was 90,000 plants ha^{-1}. The first plant in each plot at AF, TL, and TLCA was self-pollinated for carotenoids analysis.

Twenty-one parents of the hybrids along with three CIMMYT maize lines (Table II-1) were also evaluated in a separate trial at TLCA. The experimental design was an alpha-lattice with two replications and 4 incomplete blocks in each replication. The plot size and plant density were the same as those of the hybrids. The first plant in each plot was self-pollinated for

carotenoids analysis. Data from AF during winter 2010 were used as a second environment for carotenoid data for the inbred parent lines.

Traits measured for the hybrid and inbred trials included: grain yield (Mg ha^{-1}, adjusted to 12.5% moisture content), anthesis date (d), plant height (cm), and carotenoids concentration in grain (μg g^{-1}), including lutein, zeaxanthin, β-cryptoxanthin, β-carotene (all-trans), and total provitamin A. Total provitamin A concentration was calculated as β-carotene (all-trans + 9-cis + 13-cis isomers) + 0.5(β-cryptoxanthin).

Analysis of carotenoids in maize kernels

The carotenoids analyses were conducted at the CIMMYT maize quality laboratory. Random samples of 20-30 seeds were frozen at -80°C until grinding to a fine powder (0.5 μm), followed by the CIMMYT laboratory protocols for carotenoids analysis, including extraction, separation, and quantification by HPLC (Galicia et al., 2008). Carotenoids measured were lutein, zeaxanthin, β-cryptoxanthin, and β-carotene (all-trans, 9-cis, and 13-cis isomers).

Statistical analyses

Mixed model analyses were performed for each trait using the linear model:

$$Y = \mu + Env + Rep(Env) + Block(Rep \times Env) + Set + GCA_1(Set) + GCA_2(Set) + SCA(Set) + Env \times Set + Env \times [GCA_1(Set)] + Env \times [GCA_2(Set)] + Env \times [SCA(Set)] + \varepsilon$$

Where μ = grand mean, Env = environment, Rep = replication, GCA_1 and GCA_2 = general combining ability of parent-1 and parent-2, respectively, SCA = specific combining ability, and ε = experimental error. Set, GCA, and SCA were considered as fixed effects, whereas

environment, replication, block, and all interactions involving these factors were random. To understand the effects of hybrid and hybrid x environment interaction, the same model aggregating Set, GCA_1(Set), GCA_2(Set), SCA(Set) as "hybrid" and their interaction with the environment as "hybrid x environment" was fitted to the data.

In the presence of significant hybrid x environment interaction, Spearman rank correlation analysis among environments was performed and if the correlations were significant, combined analyses across environments were conducted. The mixed model and correlation analyses were performed in SAS using the MIXED and CORR procedures, respectively, while mean squares and F-tests for random effects were obtained using the GLM procedure (SAS Institute Inc., 2004) and the appropriate error term based on their respective Type III estimated mean squares. The GCA values of the lines within sets and the SCA values of hybrids within sets were estimated using fitted values obtained from the linear model above.

Heterosis effects were estimated on entry mean basis for total provitamin A concentration. Parents' provitamin A data were averaged by lines across two sites and then integrated in the hybrids mean dataset. High-parent heterosis (HPH) of each hybrid was calculated as the percentage difference between the hybrid mean and the best parent (Falconer and Mackay, 1996). Comparison of HPH among between and within putative heterotic group mating was performed using independent samples T test in SAS TTEST procedure (SAS Institute Inc., 2004). Equality of variances among these two samples was assessed and Satterthwaite approximation was used to estimate pooled degrees of freedom if the variances were not equal.

Three models for constituting heterotic groups (HG) among the 21 lines were compared: (1) three-HG model developed using 402 SNP genetic distance (which served as a basis of all analyses in this dissertation), (2) three-HG model based on SCA, and (3) two-HG model based on SCA. Models based on SCA were developed using the SCA effect estimate obtained from a model fit without set effects, which resulted in a 21 lines x 21 lines SCA matrix with above and below diagonal having the same values, approaching a half-diallel scheme. The values were transformed by adding the absolute value of the smallest SCA estimate to make all values in the matrix positive, which was necessary for subsequent analyses. From a total of 231 expected values in the half matrix including self-pollinations (1x1, 2x2, … 21x21), 73 missing values (crosses not available in the Design II which would have been formed if using a half-diallel design) were estimated using the average of row and column means for each respective cell. Hierarchical clustering analysis using complete linkage method was then performed using R (R Development Core Team, 2008), and the grouping information obtained was used as prior information for the Discriminant Analysis of Principal Components (DAPC) in R (Jombart et al., 2010). DAPC was mainly utilized to calculate the membership probability of each line in its heterotic group using k = 2 and 3 for two and three heterotic group models, respectively. This analysis resulted in two heterotic patterns (consisting of two and three putative heterotic groups, respectively) which were then validated using the existing grain yield data of hybrids evaluated in four environments.

Repeatability of the traits was calculated on the entry-mean basis using the following formula:

$$R = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GE}^2}{e} + \frac{\sigma^2}{re}}$$

where σ_G^2 is genotypic variance, σ_{GE}^2 is genotypic x environment variance, σ^2 is error variance, e is number of environments, and r is number of replications. Variances were estimated using REML method in SAS MIXED procedure.

A selection of promising hybrids, with outstanding grain yield and total provitamin A concentrations, was identified among hybrids with data for three environments (CE, TL, TLCA) for grain yield and two environments (TL, TLCA) for total provitamin A (a total of 131 hybrids). The selected hybrids had both grain yield above the mean + LSD 0.05 and total provitamin A above the mean.

RESULTS

Analyses of Variance

The hybrid and hybrid x environment interaction effects were significant for all traits except the hybrid x environment interaction for β-cryptoxanthin (Table II-2 and II-3). Spearman rank correlations among locations were significant for grain yield (minimum r=0.3**) and total provitamin A concentration (minimum r=0.6**), indicating that hybrid x environment interactions were not of an extreme crossover-type, and therefore combined analyses across locations were performed.

Further partitioning of the hybrid effect showed that the set effect was significant for agronomic traits, as well as for β-cryptoxanthin and total provitamin A concentrations (Table II-2 and II-3). Moreover, general and specific combining ability (GCA and SCA) effects were

significant for grain yield, anthesis date, and plant height, signaling the importance of both additive and non-additive gene action in determining the inheritance of these traits (Table II-2). The GCA main effects were also significant for all carotenoid traits evaluated, but SCA effects were not significant for zeaxanthin, β-carotene, and total provitamin A concentrations (Table II-3). These results indicate that carotenoid concentrations including the total provitamin A concentration are controlled by genes with mostly additive gene action. It is noteworthy that theoretical expectation is that heterozygosity and hence heterosis may be reduced in F_2 relative to the F_1 generation; however, the trait of interest for this research is total provitamin A in F_2 grain because consumers eat F_2 grain and not F_1 seed.

Environmental effect was significant for all traits except β-cryptoxanthin (Table II-2 and II-3). Average hybrid grain yield was largest when more days elapsed before flowering (among the summer season plantings) at the cooler, higher elevation site, Celaya (Figure II-2). Average plant height and total provitamin A concentrations at Agua Fria were lower than at Tlaltizapan. The two crop management treatments, normal and conservation agriculture at Tlaltizapan, did not result in significant differences for average grain yield, anthesis date, or total provitamin A concentrations, but plant height was shorter at TLCA (Figure II-2). Grain yield, total provitamin-A, and genetic distance among inbred parents for 156 hybrids evaluated are presented in Supplemental Table II-2.

Comparisons among mating sets

The between group matings (average across set 1, 2, and 3) differed significantly from within group matings (set 4, 5, and 6) for agronomic traits but not for carotenoid concentrations (Table II-4). The difference was positive for grain yield (0.4 Mg ha^{-1}) and plant height (5.1 cm),

and negative for anthesis date (-0.7 d). These results indicate greater hybrid vigor of the between group matings, in which heterosis led to higher grain yield, higher plant height, and earlier flowering.

Mating of putative heterotic group 1 with 2 and 3 resulted in significantly higher average grain yield (1.0 Mg ha^{-1}) and plant height (13.1 cm) compared to crosses within group 1 itself, while putative heterotic groups 2 and 3 did not demonstrate such evidence. Mating of group 2 with the other two groups resulted in earlier flowering (3 days on average), which is a common expression of heterosis, and was not the case for groups 1 and 3.

As intended from the formation of the putative heterotic groups, the average genetic distances among parental lines in sets of between HG crosses (sets 1-3, 0.31) were significantly larger than those in sets of within group matings (sets 4-6, 0.22). While there were no significant differences for average genetic distance among hybrids of sets 1, 2, and 3, set 4 had significantly larger average genetic distance than sets 5 and 6. On average, genetic distances were significantly correlated with grain yield ($r=0.4^{**}$); however, this correlation was significant for sets 3, 4, and 6 ($r=0.6^{**}$, 0.7^{*}, and 0.8^{**}, respectively) and not significant for sets 1, 2, and 5 (r=-0.1, 0.3, 0.6, respectively). Unlike for grain yield, genetic distances among parental lines were not correlated with provitamin A concentrations in hybrids (average across sets, r=0.0), except in set 6 ($r=0.8^{**}$) (Figure II-3). This could be partly due to evaluation of F_2 bulked seeds rather than F_1.

Combining ability, heterosis, and heterotic pattern

On average, there was no significant correlation between estimates of GCA effects for total provitamin A concentrations and grain yield (r=-0.1, n=21); however, a few lines had

positive average GCA effects for both traits (data not shown). Among these lines, only line 2 showed high average correlations across sets for both grain yield (r=0.8) and total provitamin A concentrations (r=0.6), and therefore appeared to be exceptionally promising for use in breeding. Additionally, line 6 had consistent, largest positive GCA for total provitamin A concentrations, and therefore could be useful for a provitamin A source in developing outstanding hybrids.

The average high-parent heterosis (HPH) for provitamin A concentration within mating sets ranged from 3% to 21%. Although the putative heterotic group 1 x 2 matings had significantly larger HPH than crosses among lines in groups 1x1 and 2x2 (P = 0.02), matings of among putative heterotic group lines on average did not exhibit greater HPH than within group matings (P = 0.08). This agrees with the evidence of non-significant SCA effects for total provitamin A concentration (Table II-3) and with non-significant correlation between grain yield and provitamin A concentration (Table II-6). However, as mentioned above, heterosis and SCA effects for provitamin A are underestimated because of evaluation of F_2 seeds rather than F_1.

Figure II-4 illustrates that the three-heterotic group (3-HG) model for grain yield was more appropriate for the inbred lines studied than the 2-HG model. Not surprisingly, because SCA was based on actual rather than predictive data (GD's), the grain yield difference of between versus within HG crosses was larger in SCA 3-HG (1.2 Mg ha^{-1}) than GD 3-HG (0.4 Mg ha^{-1}). Line membership in heterotic group 2 was consistently assigned using these two methods, except for line 14. Assignment of lines 1, 5, 7, 8 (GD 3-HG group 1) and 16, 19, 21 (GD 3-HG group 3), however, were classified differently in the SCA 3-HG model (Figure II-5). Incidentally, the SCA 2-HG model mainly combined the SCA 3-HG group-1 and SCA 3-HG

group-3 into one group, while the second SCA 2-HG group consisted mainly of lines assigned to group-2 by the SCA 3-HG model (Figure II-5).

Correlations among phenotypic traits

Grain yield had a small positive but significant correlation with β-cryptoxanthin ($r=0.2$, $P<0.05$), but was not correlated with total provitamin A concentration (Table II-6). Among carotenoid traits, lutein was negatively correlated with zeaxanthin ($r=-0.3$, $P<0.01$), and there was no significant correlation between β-cryptoxanthin and β-carotene ($r=-0.1$). β-carotene ($r=0.8$, $P<0.01$) but not β-cryptoxanthin ($r=0.1$, not significant) was strongly correlated with total provitamin A concentration, reflecting the greater influence of β-carotene than β-cryptoxanthin on total provitamin A concentration.

The phenotypic correlation of total provitamin A concentration with lutein was $r=-0.3$ ($P<0.01$), which was opposite to its correlation with zeaxanthin ($r=0.3$, $P<0.01$) (Table II-6). This reflects the α- vs. β-branch association with provitamin A, and thus the rationale behind *LcyE*'s value to provitamin A breeding.

Selection of promising hybrids for both grain yield and provitamin A

Hybrid 6x10 was perhaps the most outstanding experimental hybrid overall, with grain yield of 7.7 Mg ha^{-1} and total provitamin A concentration 20.7 μg g^{-1} (Table II-7). Its grain yield was not significantly different from the most high-yielding candidate (18x19, 8.5 Mg ha^{-1}). The best hybrid check, CIMMYT's single-cross 'CML451/CML486' (yellow kernel), had significantly more grain yield (10.1 Mg ha^{-1}) than all experimental hybrids, whereas the three-way cross commercial hybrid check, 'Jabali' (white kernel, from Monsanto), had similar grain

yield (7.9 Mg ha^{-1}) to the best experimental hybrids (Table II-7). As expected, the provitamin A concentration of the hybrid checks (3.8 and 0.5 μg g^{-1}) was much smaller than that of the experimental hybrids (average of 11.7 μg g^{-1}). Interestingly, line 2 (total pro-A 8.7 μg g^{-1}) appeared to be outstanding because it occurred in 5 of the 11 selected crosses (Table II-7).

DISCUSSION

Carotenoids composition in most yellow maize grain, ordered from highest to lowest concentrations, comprises lutein and zeaxanthin, β-carotene, α-carotene, and β-cryptoxanthin (USDA Natl. Nutrient Database, ndb.nal.usda.gov). Results from this study agreed with this general trend, with the most available carotenoids being lutein and zeaxanthin (41%), followed by β-carotene (37%), and β-cryptoxanthin (22%) (α-carotene was not measured). β-cryptoxanthin, which has half of vitamin A activity compared to β-carotene, was available in larger amount (on average of 4.9 μg g^{-1}) compared to 40 diverse maize genotypes (0.55 μg g^{-1}) reported by Kurilich and Juvick (1999) and 25 hybrids derived from ten parents (2.5 μg g^{-1}) reported by Egesel et al. (2003). These differences may have resulted from selection for total provitamin A which has been done during development of the inbred lines in the CIMMYT/HarvestPlus program, or may be due to general differences between temperate (mainly used in other reported studies) and tropical germplasm (used herein).

The study of combining ability was important for understanding gene action affecting provitamin A carotenoid concentrations. While Egesel et al. (2003) also reported significant GCA effects for β-carotene and β-cryptoxanthin, their finding of significant SCA effects for these carotenoids differed from results in this study. With significant GCA and non-significant

SCA effects, these results suggest that additive gene action was mostly responsible for determining provitamin A concentration. This agrees with previous studies at molecular level: allelic variation for *LcyE* and *CrtRB1* genes was associated with β-carotene concentration, in which both genes accounted for 52% of phenotypic variation and their combination effects were mainly additive (Harjes et al., 2008; Yan et al*.,* 2010).

With regard to genotype x environment interaction for β-carotene, Menkir and Maziya-Dixon (2004) reported this was non-significant in a study of 17 genotypes evaluated in three locations and two years. Further, although Egesel et al. (2003) found that GCA x year interaction for β-carotene was statistically significant, it was of little practical importance (0.75% of the total variation). These reports, conclusions by (Pfeiffer and McClafferty, 2007) and the findings suggest that provitamin A expression is more influenced by genotype and environment than by genotype x environment effects.

The difference of average shared-allele genetic distance among within and between groups mating (0.09) was similar the one reported by Wen et al. (2012) using 498 maize accessions of Tuxpeno core and diverse temperate (US stiff stalk and non stiff stalk) and tropical germplasm (CIMMYT maize lines, heterotic group A and B), where the difference on average Modified Roger Distance of between and within group using 1,536 SNP markers was 0.06. Moreover, the average between groups genetic distance in this study (0.31) is similar to that reported by Hamblin et al. (2007) utilizing diverse 256 diverse maize inbred lines and 847 SNPs, where the largest shared-allele distance between lines was less than 0.4. These indicated that the differences in distances among and within groups were not extremely large even though the inbred lines analyzed are quite diverse. Moreover, correlation of genetic distance with grain

yield was generally larger for sets with least genetic distance, in agreement with Melchinger (1999).

A breeding program developing and exploiting newly defined heterotic groups would support the objective of CIMMYT-HarvestPlus's provitamin A biofortification breeding program to develop excellent three-way cross hybrids with high yield and provitamin A concentration. In such a program, use of lines from three different heterotic groups would reduce seed production costs while maximizing performance of the commercial three-way hybrid. On the other hand, in a study utilizing a diverse panel of tropical and subtropical lines from CIMMYT and IITA stress tolerance breeding programs, Wen et al. (2011) found considerable genetic variation within the existing CIMMYT heterotic groups A and B. Therefore, reclassification of working germplasm in CIMMYT breeding programs using methods such as molecular-marker-determined GD's, SCA analysis, and discriminant analysis of principal components (Jombart et al., 2010), as described herein, could be helpful for enhancing hybrid development programs. Results from this study also suggest that initial success in identifying three heterotic groups using genetic distances among lines can be further improved using SCA analysis. If applying the SCA approach, it is important to use an appropriate linear model to estimate values for the SCA matrix; for example, using a model without hybrid set effect resulted in more appropriate assignment of the lines into their respective heterotic groups than using the model with set effect.

The absence of significant relationship between grain yield and total provitamin A concentration reveals the opportunity to develop hybrids having both high grain yield potential and high provitamin A concentrations in the grain. Although the results did not confirm those of

Egesel et al. (2003) who reported a strong correlation between grain yield and zeaxanthin and total carotenoids, both arrived to the same conclusion that grain yield and total carotenoids could be improved simultaneously.

CONCLUSIONS

Provitamin A concentration was controlled primarily by additive gene action. Significant genotype x environment interaction for total provitamin A concentration represents a challenge to developing cultivars with widespread impact on vitamin A malnutrition; however, the magnitude of interaction was not large, which agrees with most published reports. Small but significant yield advantage of crosses among versus within putative heterotic groups formed by maximizing genetic distances, confirmed that molecular-marker-determined genetic distance, while not a panacea, can provide an effective starting point for further breeding work to develop useful heterotic groups.

ACKNOWLEDGEMENTS

Support for this work was provided by HarvestPlus (www.harvestplus.org), an international program that develops micronutrient-rich staple food crops to reduce hidden hunger among malnourished populations. Thanks go to German Mingramm for his technical support of all field activities, the field staff at CIMMYT and INIFAP research stations that hosted the trials, and the staff of CIMMYT's maize quality laboratory for conducting the carotenoid analyses. Funding support provided by the Directorate General of Higher Education of Indonesia for this PhD study is highly appreciated.

REFERENCES

Babu, R., N.P. Rojas, S. Gao, J. Yan, and K. Pixley. 2012. Validation of the effects of molecular marker polymorphisms in LcyE and CrtRB1 on provitamin A concentrations for 26 tropical maize populations. Theoretical and Applied Genetics. Available at http://www.springerlink.com/index/10.1007/s00122-012-1987-3 (verified 12 October 2012).

Egesel, C.O., J.C. Wong, R.J. Lambert, and T.R. Rocheford. 2003. Combining Ability of Maize Inbreds for Carotenoids and Tocopherols. Crop Sci. 43: 818-823.

Falconer, D.S. and T.F.C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longman, New York.

Galicia, L., E. Nurit, A. Rosales, N. Palacios-Rojas. 2009. Laboratory protocols 2008: Maize nutrition quality and plant tissue analysis laboratory. Mexico, D.F.: CIMMYT.

Hallauer, A.R., M.J. Carena, and J.B. Miranda Filho. 2010. Quantitative Genetics in Maize Breeding. Springer. Ames, Iowa.

Hamblin, M.T., M.L. Warburton, and E.S. Buckler. 2007. Empirical comparison of Simple Sequence Repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PloS one 2(12): e1367.

Harjes, C.E., T.R. Rocheford, L. Bai, T.P. Brutnell, C.B. Kandianis, S.G. Sowinski, A.E. Stapleton, R. Vallabhaneni, M. Williams, E.T. Wurtzel, J. Yan, and E.S. Buckler. 2008. Natural genetic variation in lycopene epsilon-cyclase tapped for maize biofortification. Science 319(5861): 330-3.

Howe, J.A., and S. A. Tanumihardjo. 2006. Evaluation of Analytical Methods for Carotenoid Extraction from Biofortified Maize (Zea mays sp.). J. Agric. Food Chem., 54(21): 7992-7997.

Jombart, T., Devillard, S. and Balloux, F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics [online serial] 11, 94 DOI: 10.1186/1471-2156-11-94

Kurilich, A.C., & Juvik, J.A. 1999. Quantification of carotenoid and tocopherol antioxidants in Zea mays. Journal of agricultural and food chemistry, 47(5): 1948-1955.

Lu, Y., J. Yan, C.T. Guimarães, S. Taba, Z. Hao, S. Gao, S. Chen, J. Li, S. Zhang, B.S. Vivek, C. Magorokosho, S. Mugo, D. Makumbi, S.N. Parentoni, T. Shah, T. Rong, J.H. Crouch, and Y. Xu. 2009. Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theoretical and applied genetics, 120(1): 93–115.

Melchinger, A.E. 1999. Genetic diversity and heterosis. *In*: The Genetics and Exploitation of Heterosis in Crops. J.G. Coors and S. Pandey (eds). Madison, Wisconsin.

Menkir, A. and B. Maziya-Dixon. 2004. Influence of genotype and environment on β-carotene content of tropical yellow-endosperm maize genotypes. Maydica 49: 313–318.

Muzhingi, T., Gadaga, T. H., Siwela, A. H., Grusak, M. A., Russell, R. M., & Tang, G. 2011. Yellow maize with high b-carotene is an effective source of vitamin A in healthy Zimbabwean men. Am J Clin Nutr 94: 510-519.

Nuss E.T., S.A. Tanumihardjo. 2010. Maize: A paramount staple crop in the context of global nutrition. Compr Rev Food Sci Food Saf 9: 417–436.

Ortiz-Monasterio, J.I., N. Palacios-Rojas, E. Meng, K. Pixley, R. Trethowan, R.J. Pena. 2007. Enhancing the mineral and vitamin content of wheat and maize through plant breeding. Journal of Cereal Science 46: 293–307.

Ortiz, R., S. Taba, V.H.C. Tovar, M. Mezzalama, Y. Xu, J. Yan, and J.H. Crouch. 2010. Conserving and Enhancing Maize Genetic Resources as Global Public Goods–A Perspective from CIMMYT. Crop Science 50(1): 13.

Pfeiffer, W.H., and B. McClafferty. 2007. HarvestPlus: Breeding Crops for Better Nutrition. Crop Sci. 47(Supplement_3): S-88

Pixley, K.P., N. Palacios-Rojas, and R.P. Glahn. 2011. The usefulness of iron bioavailability as a target trait for breeding maize (*Zea mays* L.) with enhanced nutritional value. Field Crops Research 123(2): 153-160.

Rice, A.L, K.P. West, R.E. Black. 2004. Vitamin A deficiency. In: Ezzati, M, A.D. Lopez, A. Rodgers, C.J.L. Murray (eds). Comparative Quantification of Health Risks: Global and Regional Burden of Desease Attributable to Selected Major Risk Factors. Volume 1. World Health Organization, Geneva.

SAS Institute Inc. 2009. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.

Tanumihardjo, S. 2011. Vitamin A: biomarkers of nutrition for development. Am J Clin Nutr 94(suppl): 658S–665S.

Wen, W., J.L. Araus, T. Shah, J. Cairns, G. Mahuku, M. Bänziger, J.L. Torres, C. Sánchez, J. Yan. 2011. Molecular Characterization of a Diverse Maize Inbred Line Collection and its Potential Utilization for Stress Tolerance Improvement. Crop Sci. 51: 2569–2581.

Wen, W., J. Franco, V.H. Chavez-Tovar, J. Yan, and S. Taba. 2012. Genetic characterization of a core set of a tropical maize race Tuxpeño for further use in maize improvement. PloS one 7(3): e32626.

West, K.P. 1991. Dietary vitamin A deficiency: effects on growth, infection, and mortality. Food and Nutr. Bull. 13(2).

West, K.P. and I. Darnton-Hill. 2008. Vitamin A deficiency. In: Semba, R.D., M.W. Bloem (eds). Nutrition and Health in Developing Countries. Second edition. Humana Press. New Jersey, USA.

WHO. 2009. Global prevalence of vitamin A deficiency in populations at risk 1995–2005. WHO Global Database on Vitamin A Deficiency. Geneva, World Health Organization.

Yan, J., C.B. Kandianis, C.E. Harjes, L. Bai, E.-H. Kim, X. Yang, D.J. Skinner, Z. Fu, S. Mitchell, Q. Li, M.G.S. Fernandez, M. Zaharieva, R. Babu, Y. Fu, N. Palacios, J. Li, D. Dellapenna, T. Brutnell, E.S. Buckler, M.L. Warburton, and T. Rocheford. 2010. Rare genetic variation at Zea mays CrtRB1 increases beta-carotene in maize grain. Nature genetics 42(4): 322-7.

Table II-1. Twenty-one lines selected for use in hybrids formation as representatives of three putative heterotic groups formed by maximizing genetic diversity and minimizing within-group pedigree similarity among 127 lines using 402 SNP markers.

Putative heterotic group	Line No.	Pedigree	Anthesis date (d)	Total Pro-A ($\mu g\ g^{-1}$)[+]
1	1	(P591c4 1y2 GEN F12-1-1-1-B-B-B-B//P591c4 1y2 GEN F12-1-1-1-B-B-B/KUI carotenoid syn-FS25-3-2-B)-B-24-3-B	58	9.5
1	2	[[GQL5/[GQL5/CML202]F2-1sx]-3-1-2-B/[BETASYN]BC1-2-3-1/KUI+SC55SYN#]-B-B-B-12-B-B	58	8.4
1	3	[CML488/[BETASYN]BC1-2-2-3/KUI+SC55SYN#]-B-B-B-4-1-B	60	7.2
1	4	[[89[G27/TEWTSRPool]#-278-2-X-B/[COMPE2/P43SR//COMPE2]F#-20-1-1]-B-31-1-B-2-#/[BETASYN]BC1-6-1-3/KUISYN#]-B-B-B-5-B-B	60	6.4
1	5	(Florida A plus Syn-FS2-2-1-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-1(MAS:L4H1)-2	61	11.4
1	6	(KUI carotenoid syn-FS17-3-2-B-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-4(MAS:L4H1)-2-B-B	60	18.2
1	7	SAM4(ProA)BC1/KUISyn#-1-39-1-3-B-B	62	6.6
1	8	(MAS[MSR/312]-117-2-2-1-B-B-B/[BETASYN]BC1-4-2-1/KUISYN#-B)-B-2-3-B-B-B	60	7.4
2	9	([[[K64R/G16SR]-39-1/[K64R/G16SR]-20-2]-5-1-2-B*4/CML390]-B-38-1-B-7-#/[BETASYN]BC1-1-1-1-#-B//[[[K64R/G16SR]-39-1/[K64R/G16SR]-20-2]-5-1-2-B*4/CML390]-B-38-1-B-7-#/[BETASYN]BC1-1-1-1-#/CML297)-B-24-1-B	62	4.5
2	10	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-10-2-1-#/CML-305-B)-B-9-1-B	61	10.7
2	11	[[CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1-B//[CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1/CML-300-B]-1-33-B	60	4.4
2	12	(P72c1xCML297 x CL-02410-3-1-1-B/CML297)-B-2-3-1-B	63	8.6
2	13	(CML297/(CML489/[BETASYN]BC1-2-#/CML297)-B-24)-19-2	64	9.8
2	14	(CML297/(P72c1xCML297 x CL-02410-3-1-1-B/CML297)-B-16)-21-1	64	9.1
2	15	([[[NAW5867/P30SR]-40-1/[NAW5867/P30SR]-114-2]-16-2-2-B-2-B/CML395-6]-B-20-1-B-3-#/[BETASYN]BC1-3-1-1-#/CML300)-9-2-2-B-B	62	6.3
3	16	(Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-5-#-B-B//Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-5-#-B/CML304)-5-1-B	60	6.2
3	17	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-10-2-1-#/CML-304-B)-B-4-1-B	60	8.3
3	18	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#/CML-304-B)-B-13-1	60	6.2
3	19	(CML489/[BETASYN]BC1-2-#/CML300)-B-26-1-1-B	66	6.0
3	20	(KUI carotenoid syn-FS17-3-1-B-B/(CML239/GWIC)-1-7TL-1-1-1//CML300)-4-2-B	58	5.5
3	21	(Carotenoid Syn3-FS5-1-5-B-B/CML355//CML300)-4-1	61	9.3

[+] Total provitamin A concentration = all-trans β-carotene + 0.5(β-cryptoxanthin).

Table II-2. Summary of mixed model analysis of variance for grain yield (Mg ha^{-1}, 12.5% H_20), anthesis date (d), and plant height (cm).

Source of variation	DF	Mean square		
		Grain yield	Anthesis date	Plant height
Environments, Env	3	2,204.1 **	15,251.8 **	33,865.7 **
Replications, Rep / Env	4	6.6 ns	54.6 **	3,578.9 **
Blocks / Rep x Env	102	1.9 **	4.5 **	231.3 **
Hybrid	155	6.4 **	24.0 **	680.1 **
Set	5	14.5 *	320.0 **	4,423.1 **
GCA P1 / Set	25	11.3 **	24.0 **	567.4 **
GCA P2 / Set	22	14.0 **	32.1 **	1,882.4 **
SCA / Set	103	2.4 **	3.0 *	216.7 **
Env x Hybrid	373	2.4 **	3.4 **	168.2 **
Env x Set	15	1.9 ns	9.3 ns	239.3 ns
Env x GCA P1 / Set	70	3.3 **	4.3 **	188.3 *
Env x GCA P2 / Set	64	3.2 **	4.8 **	230.2 **
Env x SCA / Set	224	1.6 **	2.4 ns	133.8 ns
Error	426++	0.9	2.0	113.0
Grand mean		6.2	66.7	234.5
Range		3.0 – 8.5	60.5 – 84.0	195.0 – 262.0
CV (%)		15.1	2.1	4.5
Repeatability (%)		69.1	89.7	80.7

+ Set 1 = putative heterotic groups 1x2, Set 2 = 1x3, Set 3 = 2x3, Set 4 = 1x1, Set 5 = 2x2, and Set 6 = 3x3.

++ Error degrees of freedom is 424 for grain yield due to two missing values.

*, ** Significant at $P < 0.05$ and $P < 0.01$, respectively; ns: not significant.

GCA=general combining ability, SCA=specific combining ability, P1=parent-1, P2=parent-2

Table II-3. Summary of mixed model analysis of variance for carotenoid concentrations (μg g^{-1}).

Source of variation	DF	Mean square				
		Lutein	Zea-xanthin	β-Crypto-xanthin	β-Carotene	Total Pro-A
Environments, Env	2	918.7 **	1,798.1 **	11.7 ns	597.3 **	303.2 **
Replications, Rep / Env	3	2.8 *	1.3 ns	5.1 ns	0.9 ns	6.8 ns
Blocks / Rep x Env	76	0.8 ns	2.8 ns	1.4 *	0.7 **	2.8 *
Hybrid	155	4.8 **	19.8 **	8.8 **	9.3 **	35.0 **
Set	5	7.7 ns	30.2 ns	112.9 **	75.7 ns	213.2 **
GCA-P1 / Set	25	8.5 **	42.3 **	9.2 **	19.8 *	78.1 **
GCA-P2 / Set	22	7.1 **	37.8 **	9.8 **	6.6 **	36.6 **
SCA / Set	103	2.0 *	5.2 ns	1.8 **	1.8 ns	6.0 ns
Env x Hybrid	226	2.4 **	8.2 **	1.5 ns	5.3 **	7.5 **
Env x Set	10	3.2 ns	21.9 ns	2.6	16.8 ns	11.4 ns
Env x GCA-P1 / Set	46	3.7 **	10.8 **	2.2	9.9 **	12.5 **
Env x GCA-P2 / Set	42	1.8 ns	8.8 **	1.8	2.5 ns	3.2 ns
Env x SCA / Set	128	1.6 **	4.4 **	1.1	2.8 **	6.0 **
Error	244	0.9	2.5	1.0	0.5	1.9
Grand mean		3.5	5.9	4.9	3.9	10.8
Range		0.9 – 10.1	0.8 – 15.1	2.0 – 11.0	0.9 – 10.0	3.0 – 23.2
CV (%)		26.7	26.6	20.8	17.5	12.8
Repeatability (%)		57.1	67.4	88.9	55.5	84.9

+ Set 1 = putative heterotic groups 1x2, Set 2 = 1x3, Set 3 = 2x3, Set 4 = 1x1, Set 5 = 2x2, and Set 6 = 3x3.

++ Total provitamin A concentration = all-trans β-carotene + 9-cis β-carotene + 13-cis β-carotene + 0.5(β-cryptoxanthin).

*, ** Significant at $P < 0.05$ and $P < 0.01$, respectively; ns: not significant.

GCA=general combining ability, SCA=specific combining ability, P1=parent-1, P2=parent-2

Table II-4. Set means and group comparisons for grain yield (Mg ha^{-1}, 12.5% H_2O), anthesis date (d), plant height (cm), and carotenoid concentrations (µg g^{-1}).

Effects	Grain yield (Mg ha^{-1})	Anthesis date (d)	Plant height (cm)	Lutein (µg g^{-1})	Zea-xanthin (µg g^{-1})	β-Crypto-xanthin (µg g^{-1})	β-Carotene (µg g^{-1})	Total Pro-A^{++} (µg g^{-1})	Average genetic distance$^{\#\#}$
Between-group matings									
Set 1^{+}	6.4 $^{a\#}$	66.4 ^{cd}	233.1 ^{b}	3.6 $^{\wedge}$	6.5 $^{\wedge}$	5.6 ^{b}	4.3 $^{\wedge}$	12.4 ^{ab}	0.32 ^{a}
Set 2	6.1 ^{ab}	65.6 ^{d}	233.0 ^{b}	3.0	5.4	3.7 ^{d}	4.0	10.5 ^{bc}	0.31 ^{a}
Set 3	6.1 ^{ab}	68.5 ^{ab}	240.5 ^{a}	3.2	5.5	5.5 ^{bc}	2.8	9.0 ^{cd}	0.29 ^{a}
Average	6.2	66.9	235.5	3.3	5.8	4.9	3.7	10.6	0.31
Within-group matings									
Set 4	5.3 ^{b}	65.0 ^{d}	220.0 ^{c}	4.0	7.8	3.6 ^{d}	5.9	13.8 ^{a}	0.25 ^{b}
Set 5	6.1 ^{ab}	70.0 ^{a}	238.4 ^{ab}	4.0	5.4	7.8 ^{a}	2.7	10.0 ^{bcd}	0.21 ^{b}
Set 6	6.0 ^{ab}	67.6 ^{bc}	232.8 ^{ab}	3.7	6.2	4.4 ^{cd}	2.6	8.0 ^{d}	0.21 ^{b}
Average	5.8	67.5	230.4	3.9	6.5	5.3	3.7	10.6	0.22
Between vs within groups									
Average difference	0.4 *	-0.7 *	5.1 **	-0.6	-0.7	-0.3 ^{ns}	0.0	0.0 ^{ns}	0.09 **
Set 1 and 2 vs 4	1.0 **	1.0 *	13.1 **	-0.7	-1.8	1.0 **	-1.8	-2.4 **	0.07 **
Set 1 and 3 vs 5	0.2 ^{ns}	-2.6 **	-1.7 ^{ns}	-0.6	0.6	-2.2 **	0.9	0.7 ^{ns}	0.09 **
Set 2 and 3 vs 6	0.1 ^{ns}	-0.5 ^{ns}	4.0 ^{ns}	-0.6	-0.8	0.2 ^{ns}	0.8	1.7 *	0.09 **

$^{+}$ Set 1 = putative heterotic groups 1x2, Set 2 = 1x3, Set 3 = 2x3, Set 4 = 1x1, Set 5 = 2x2, and Set 6 = 3x3.

$^{++}$ Total provitamin A concentration = all-trans β-carotene + 9-cis β-carotene + 13-cis β-carotene + 0.5(β-cryptoxanthin).

$^{\wedge}$ Set effect was not significant in ANOVA (Table II-3); therefore post hoc comparisons were not performed.

$^{\#}$ Means followed by the same letters in the same column are not significantly different based on Tukey-Kramer test at $\alpha = 0.05$.

$^{\#\#}$ Average genetic distance among inbred parents within each set of hybrids.

*, ** Significant at $P < 0.05$ and $P < 0.01$, respectively; ns: not significant.

Table II-5. General combining ability (GCA) for total provitamin A carotenoid concentrations (Pro-A, μg g^{-1})[+] and grain yield (GY, Mg ha^{-1}) in six mating sets[++].

Set 1			Set 2			Set 3		
Lines	GCA		Lines	GCA		Lines	GCA	
	GY	Pro-A		GY	Pro-A		GY	Pro-A
1	0.0	-1.0	1	-0.7 **	-0.2	9	1.1 **	-1.2 *
2	0.6 *	1.0	2	0.5	-1.0	10	-0.2	0.5
3	0.8 **	-1.9 **	3	0.5	-2.0 **	11	0.0	-2.1 **
4	-0.3	-2.3 **	4	-0.2	-0.3	12	0.4	2.1 **
5	-0.7 **	-0.7	5	-0.3	-1.1	13	-0.9 **	1.0 **
6	-0.3	5.1 **	6	-0.4	5.0 **	14	0.0	0.1
7	-0.2	-0.2	7	0.5	-0.5	15	-0.4	-0.3
9	0.5 *	-1.4 *	8	0.3	0.0	16	-0.6 *	-1.7 **
10	0.2	0.2	16	-0.7 **	-2.4 **	17	0.7 **	0.8
11	-0.8 **	-1.7 **	17	0.3	1.2 *	18	1.0 **	0.3
12	0.1	0.9	18	1.1 **	0.2	19	-0.9 **	0.5
13	0.3	3.1 **	19	0	2.2 **	20	0.0	-1.2 *
14	-0.3	0.3	20	-0.7 **	-1.9 **	21	-0.2	1.3 *
15	0.0	-1.3 *	21	-0.1	0.8			

Set 4			Set 5			Set 6		
Lines	GCA		Lines	GCA		Lines	GCA	
	GY	Pro-A		GY	Pro-A		GY	Pro-A
1	0.5	-1.7	9	0.0	-0.2	16	-0.1	-1.6
2	0.4	1.7	10	-0.4	1.6	17	0.3	0.3
3	0.2	-1.7	11	0.3	-1.3	18	1.0 *	0.3
4	-0.6	0.0	12	-0.2	0.7	19	1.3 **	2.4 *
5	0.5	-1.8	13	-0.5	0.7	20	-1.2 **	-0.7
6	-1.0 *	3.5 **	14	0.8	-1.4	21	-1.2 **	-0.6

[+] Total provitamin A concentration = all-trans β-carotene + 9-cis β-carotene + 13-cis β-carotene + 0.5(β-cryptoxanthin).

[++] Set 1 = putative heterotic groups 1x2, Set 2 = 1x3, Set 3 = 2x3, Set 4 = 1x1, Set 5 = 2x2, and Set 6 = 3x3.

*, ** Significantly different from zero at $P < 0.05$ and 0.01, respectively.

Table II-6. Pearson phenotypic correlation coefficients among agronomic traits and carotenoid concentrations (154 df).

Trait	GY	AD	PH	Lut	Zea	β-Cry	β-Car
AD	0.0						
PH	0.5**	-0.3**					
Lut	-0.1	0.6**	-0.3**				
Zea	0.0	-0.6**	0.4**	-0.3**			
β-Cry	0.2*	0.1	0.3**	0.1	0.3**		
β-Car	0.0	-0.2*	0.0	-0.2*	0.1	-0.1	
Pro-A	0.0	-0.4**	0.2*	-0.3**	0.3**	0.1	0.8**

GY = grain yield, AD = anthesis date, PH = plant height, Lut = Lutein, Zea = Zeaxanthin, β-Cry = β-Cryptoxanthin, β-Car = β-Carotene, Pro-A = total provitamin A

++ Total provitamin A concentration = all-trans β-carotene + 9-cis β-carotene + 13-cis β-carotene + 0.5(β-cryptoxanthin).

*, ** Significant at $P < 0.05$ and $P < 0.01$, respectively; ns: not significant.

Table II-7. Top eleven hybrids selected simultaneously based on grain yield (GY, Mg ha^{-1}, 12.5% H_20) and total provitamin A concentrations (Pro-A, µg g^{-1}).

Hybrid[#]	GY (Mg ha^{-1})	Pro-A (µg g^{-1})	Average rank GY[##]	Average rank Pro-A[##]
6 x 10	7.7	20.7	23	7
2 x 13	7.8	18.8	24	11
2 x 12	7.6	15.6	30	25
3 x 10	7.8	14.5	26	27
3 x 13	7.9	13.9	16	35
2 x 15	7.6	13.0	45	42
2 x 9	8.1	12.6	40	52
3 x 12	7.9	12.4	19	61
2 x 18	8.0	12.3	12	59
7 x 9	7.8	12.3	25	48
18 x 19	8.5	11.8	11	51
CML451/CML486	10.1[*]	3.8	5	128
Jabali[+]	7.9	0.5	39	131
Selection average	7.9	14.3	25	38
Grand average[++]	6.3	11.7	66	66
LSD 0.05	1.2	3.2		

[#] Only hybrids with data in at least three environments for GY and two environments for Pro-A were used (a total of 131 hybrids). Reciprocals were bulked.

[##] Average ranks over three environments (CE, TL, TLCA) for GY (best average rank was 5, average was 66, worst was 129) and two environments (TL, TLCA) for Pro-A (best was 1, average was 66, worst was 131).

[*] Significantly higher than all other hybrids.

[+] Three-way cross, commercial maize hybrid from Monsanto.

[++] Average from all 131 candidates.

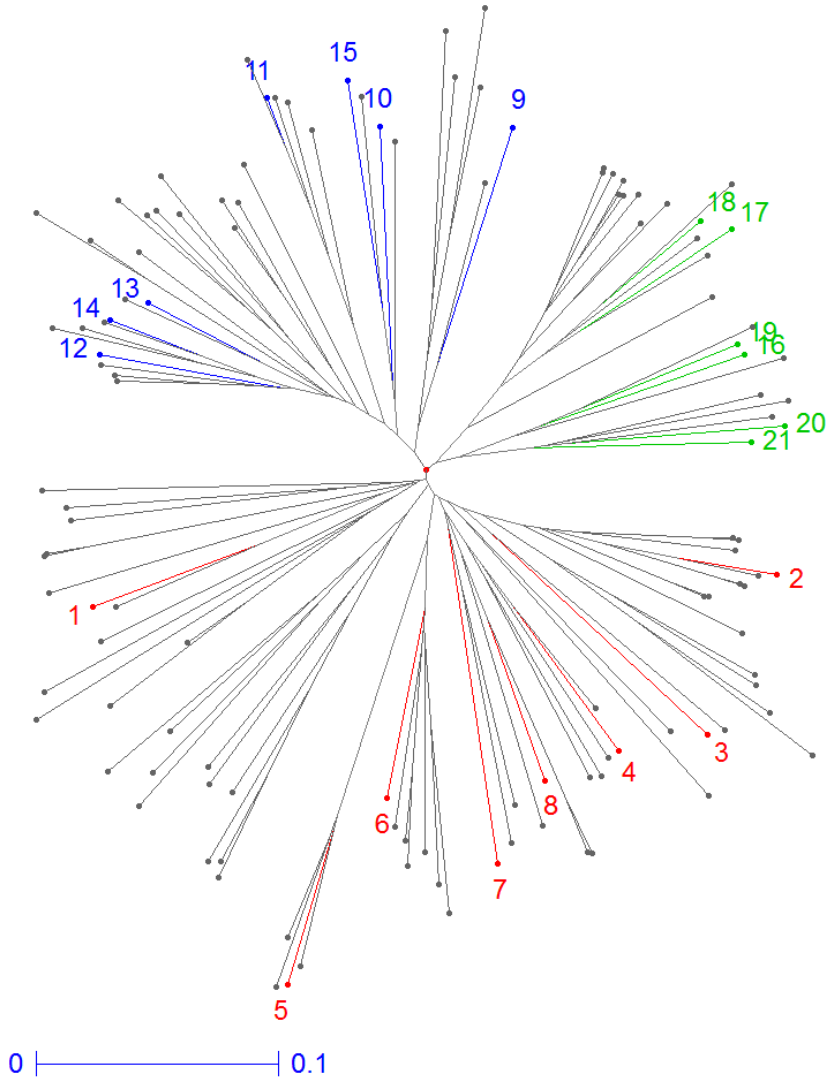


Figure II-1. A neighbor joining tree of 127 lines based on shared-allele distances from 402 SNPs. Red, blue, and green represents lines selected from putative heterotic group 1, 2, and 3, respectively.

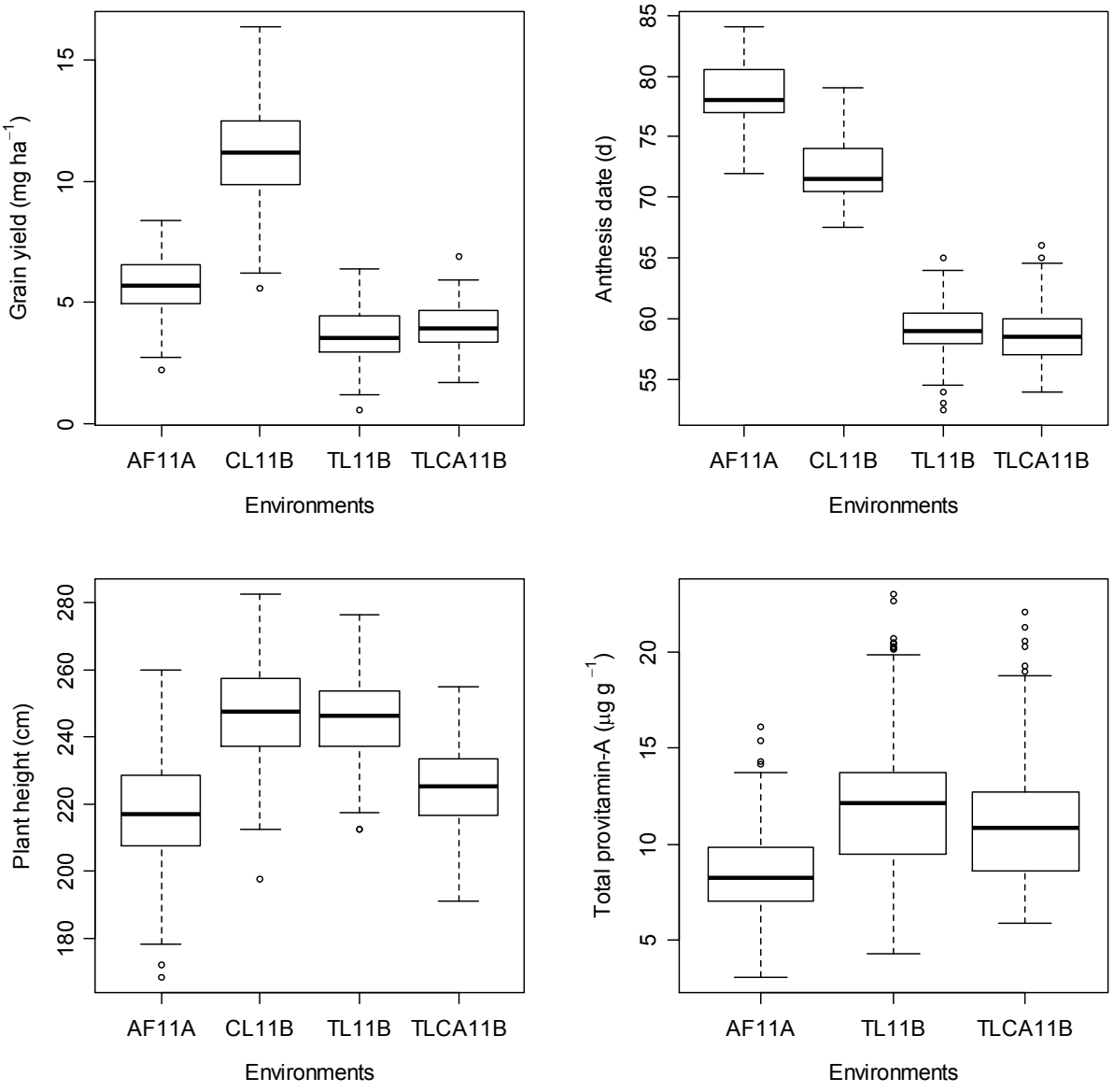


Figure II-2. Range of grain yield, anthesis date, plant height, and total provitamin A concentration values in the evaluation environments. Total provitamin A concentration = all-trans β-carotene + 9-cis β-carotene + 13-cis β-carotene + 0.5(β-cryptoxanthin). AF11A = Agua Fria, Winter 2010/2011, CL11B = Celaya, Summer 2011, TL11B = Tlaltizapan conventional tillage, Summer 2011, TLCA11B = Tlaltizapan conservation agriculture, Summer 2011.

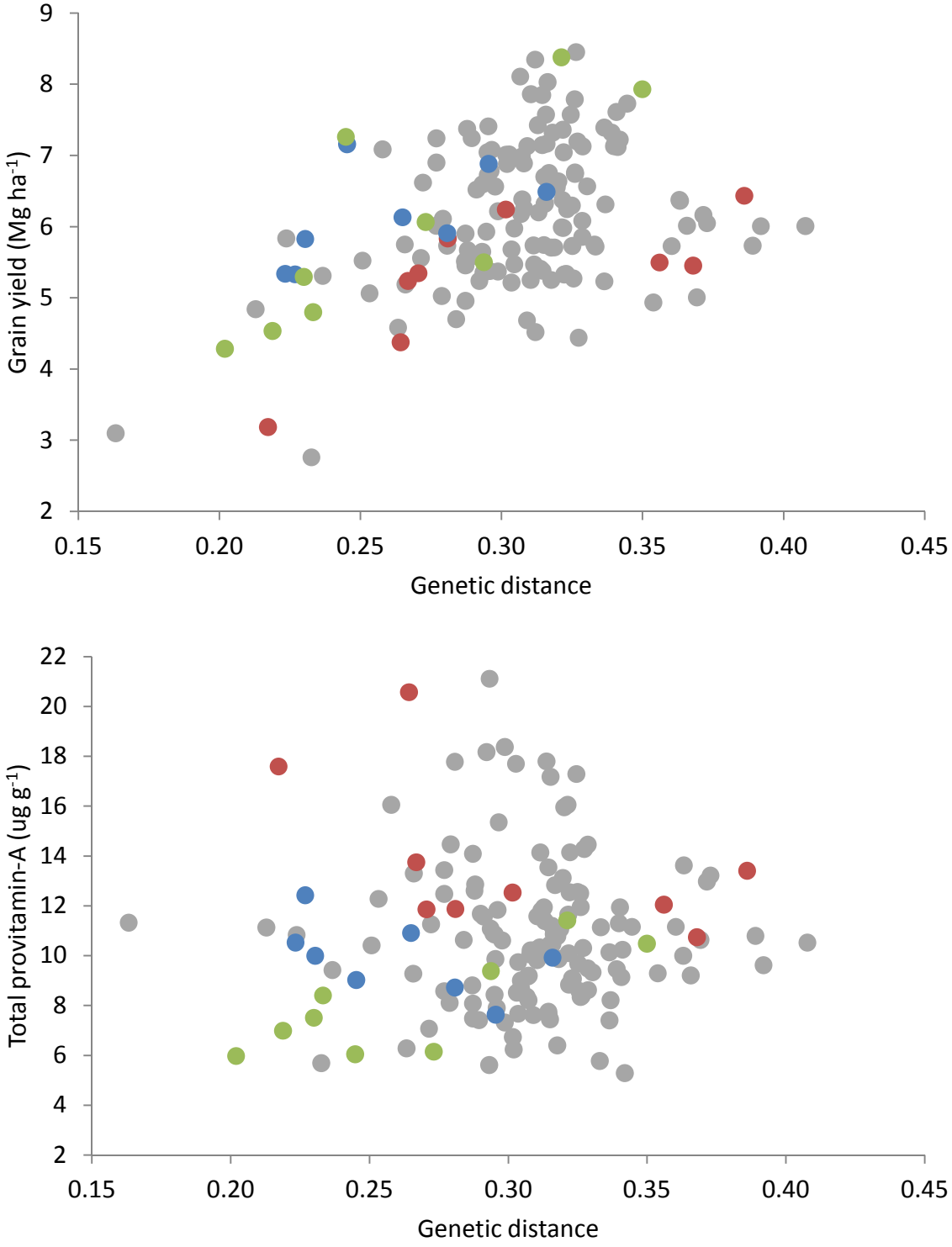


Figure II-3. Relationship between grain yield (top panel) and total provitamin A (bottom panel) with genetic distances from 402 SNP markers. Red, blue, and green are hybrids within heterotic group 1, 2, and 3, respectively, whereas between heterotic groups hybrids are in grey.

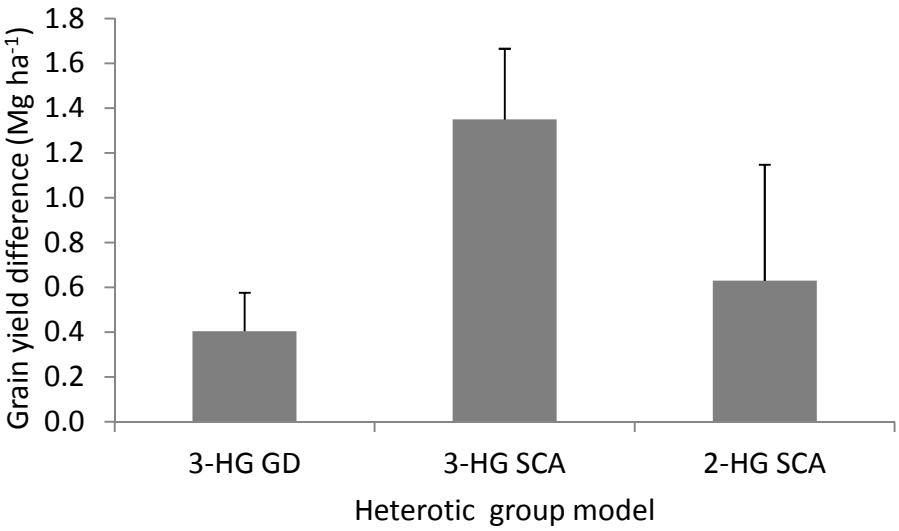


Figure II-4. Comparisons of three heterotic group models in terms of estimated grain yield difference among between and within groups mating. Three-HG GD = three heterotic groups based on 402 SNP genetic distance, 3-HG SCA and 2-HG SCA = three and two heterotic groups model based on specific combining ability, respectively. Error bars are standard error of the between versus within groups difference.

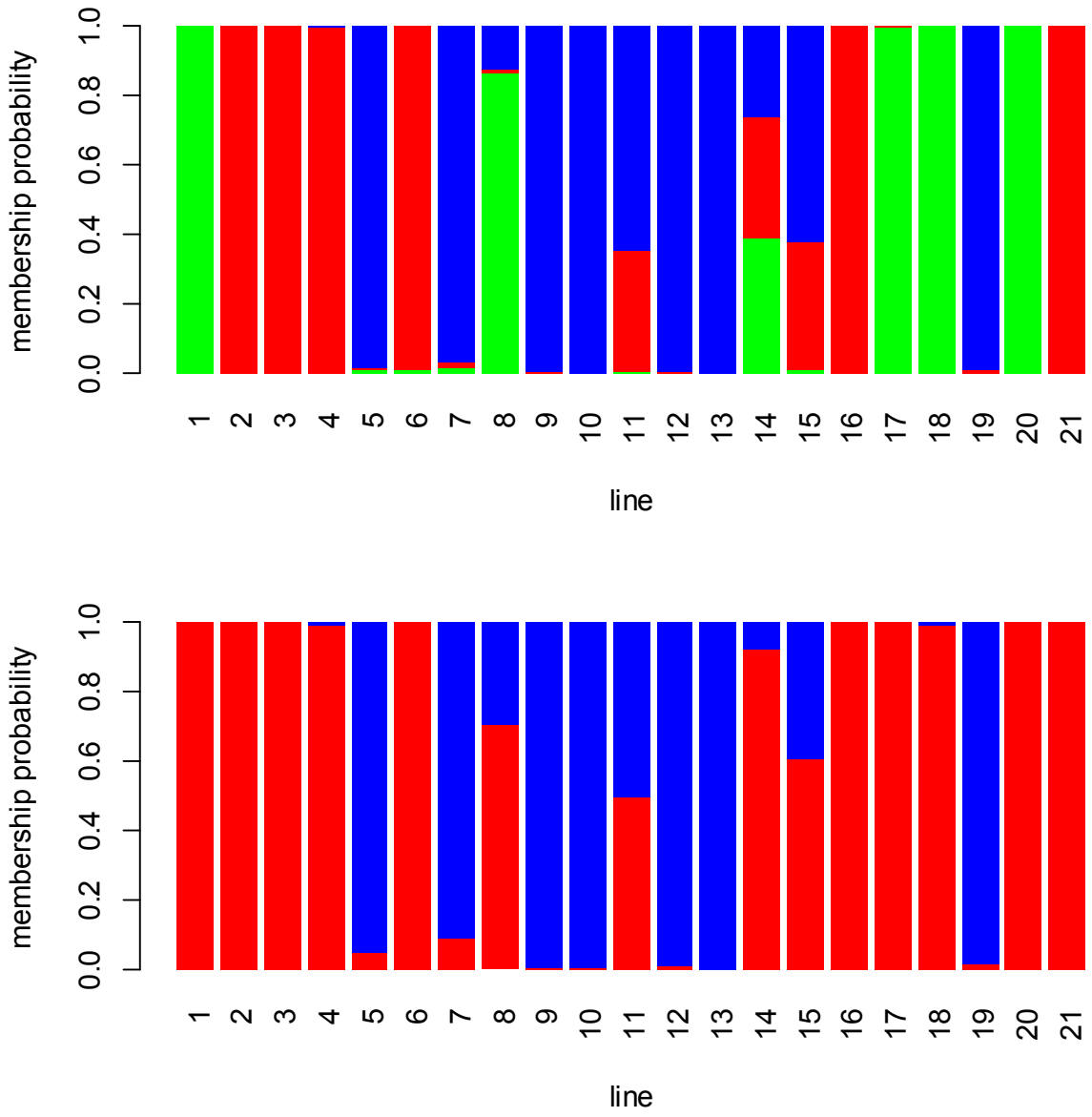


Figure II-5. Membership probability of each line as revealed by DAPC. Top and bottom panel are three and two heterotic groups model, respectively. 3-HG model: line 1-8 = group 1 (red); 9-15 = group 2 (blue), 16-21 = group 3 (green).

Supplemental Table II-1. List of 127 lines used in formation of putative heterotic groups.

Line No.	Stock No.	Pedigree
1	HP 459-506	(P591c4 1y2 GEN F12-1-1-1-B-B-B-B//P591c4 1y2 GEN F12-1-1-1-B-B-B/KUI carotenoid syn-FS25-3-2-B)-B-24-3
2	HP 421-37	[[GQL5/[GQL5/CML202]F2-1sx]-3-1-2-B/[BETASYN]BC1-2-3-1/KUI+SC55SYN#]-B-B-B-12-B
3	291-138	[CML488/[BETASYN]BC1-2-2-3/KUI+SC55SYN#]-B-B-B-4
4	HP 421-64	[[89[G27/TEWTSRPool]#-278-2-X-B/[COMPE2/P43SR//COMPE2]F#-20-1-1]-B-31-1-B-2-#/[BETASYN]BC1-6-1-3/KUISYN#]-B-B-B-5-B
5	HP 467-37	(Florida A plus Syn-FS2-2-1-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-1(MAS:L4H1)-2
6	HP 465-38	(KUI carotenoid syn-FS17-3-2-B-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-4(MAS:L4H1)-2
7	HP 472-3	SAM4(ProA)BC1/KUISyn#-1-39-1-3-B-B
8	HP 422-84	(MAS[MSR/312]-117-2-2-1-B-B-B/[BETASYN]BC1-4-2-1/KUISYN#-B)-B-2-2-B
9	HP 459-269	([[[K64R/G16SR]-39-1/[K64R/G16SR]-20-2]-5-1-2-B*4/CML390]-B-38-1-B-7-#/[BETASYN]BC1-1-1-1-#-B//[[[K64R/G16SR]-39-1/[K64R/G16SR]-20-2]-5-1-2-B*4/CML390]-B-38-1-B-7-#/[BETASYN]BC1-1-1-1-#/CML297)-B-24-1
10	HP 426-29	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-10-2-1-#/CML-305-B)-B-9-1
11	HP316-62	[[CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1-B//[CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1/CML-300-B]-1-33
12	HP 420-30	(P72c1xCML297 x CL-02410-3-1-1-B/CML297)-B-2-3-1
13	HP 434-50	(CML297/(CML489/[BETASYN]BC1-2-#/CML297)-B-24)-19-1
14	HP 434-250	(CML297/(P72c1xCML297 x CL-02410-3-1-1-B/CML297)-B-16)-21-1
15	HP 423-25	([[[NAW5867/P30SR]-40-1/[NAW5867/P30SR]-114-2]-16-2-2-B-2-B/CML395-6]-B-20-1-B-3-#/[BETASYN]BC1-3-1-1-#/CML300)-9-2-2-B
16	HP 427-39	(Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-5-#-B-B//Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-5-#-B/CML304)-5-1
17	HP 426-40	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-10-2-1-#/CML-304-B)-B-4-1
18	HP 426-63	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#/CML-304-B)-B-13-1
19	HP 420-8	(CML489/[BETASYN]BC1-2-#/CML300)-B-26-1-1
20	HP 466-69	(KUI carotenoid syn-FS17-3-1-B-B/(CML239/GWIC)-1-7TL-1-1-1//CML300)-4-2
21	HP 466-348	(Carotenoid Syn3-FS5-1-5-B-B/CML355//CML300)-4-1
22	HP 459-367	([GQL5/[GQL5/[MSRXPOOL9]C1F2-205-1(OSU23i)-5-3-X-X-1-B-B]F2-4sx]-8-6-B-B/[BETASYN]BC1-6-4-1-#-B//[GQL5/[GQL5/[MSRXPOOL9]C1F2-205-1(OSU23i)-5-3-X-X-1-B-B]F2-4sx]-8-6-B-B/[BETASYN]BC1-6-4-1-#/CML300)-B-18-4
23	HP 459-384	(P591c4 1y2 GEN F12-1-1-1-B-B-B-B//P591c4 1y2 GEN F12-1-1-1-B-B-B/CML297)-B-7-2
24	HP 421-29	[[GQL5/[GQL5/CML202]F2-1sx]-3-1-2-B/[BETASYN]BC1-2-3-1/KUI+SC55SYN#]-B-B-B-1-B
25	HP 427-8	(Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-5-#-B-B//Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-5-#-B/CML300)-8-1
26	HP 427-72	(Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-5-#-B-B//Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-5-#-B/CML304)-31-6
27	HP 483-96	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#-B//MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#/CML300)-5-5
28	HP 425-49	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#-B//MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#/CML300)-15-3
29	HP 425-18	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-10-2-1-#-B//MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-10-2-1-#/CML304)-5-1

Line No.	Stock No.	Pedigree
30	HP 425-62	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#-B//MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#/CML297)-6-3
31	HP 426-52	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#/CML-297-B)-B-5-2
32	HP 426-32	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-10-2-1-#/CML-305-B)-B-10-1
33	HP 426-306	[[CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1-B//[CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1/CML-305-B]-1-11-1
34	HP 426-173	([CML 312/MAS[MSR/312]-117-2]-B-91-3-B-B/[BETASYN]BC1-2-1-1-1/CML-297-B)-B-7-2
35	HP 426-280	[[[EV7992]C1F2-430-3-3-3-X-7-B-B/CML202]-6-2-2-3-B*3/[BETASYN]BC1-10-1-1-1-B//[[EV7992]C1F2-430-3-3-3-X-7-B-B/CML202]-6-2-2-3-B*3/[BETASYN]BC1-10-1-1-1/CML-305-B]-1-6-1
36	HP 465-1	KUI carotenoid syn-FS11-1-1-B-B-B-B-B-B
37	HP 465-2	KUI carotenoid syn-FS17-3-2-B-B-B-B-B-B
38	HP 465-3	KUI carotenoid syn-FS25-3-2-B-B-B-B-B-B
39	HP 465-4	Carotenoid Syn3-FS8-4-3-B-B-B-B-B-B
40	HP 465-5	Carotenoid Syn3-FS11-4-3-B-B-B-B-B-B
41	HP 465-6	Florida A plus Syn-FS2-2-1-B-B-B-B-B
42	HP 465-7	CML300
43	HP 465-8	CML305
44	HP376-65,76	CML297
45	HP 467-60	(CML297/(KU1409/DE3/KU1409)S2-18-2-B)-B-4(MAS:L4H1)-2
46	HP 467-61	(CML297/(KU1409/DE3/KU1409)S2-18-2-B)-B-4(MAS:L4H1)-3
47	HP 467-40	(Florida A plus Syn-FS2-2-1-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-3(MAS:L4H1)-2
48	HP 467-9	(KUI carotenoid syn-FS11-1-1-B-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-1(MAS:L4H1)-1
49	HP 467-5	(KUI carotenoid syn-FS11-1-1-B-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-2(MAS:L4H1)-3
50	HP 467-18	(KUI carotenoid syn-FS17-3-2-B-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-1(MAS:L4H1)-2
51	HP 465-39	(KUI carotenoid syn-FS25-3-2-B-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-1(MAS:L4H1)-1
52	HP 465-41	(KUI carotenoid syn-FS25-3-2-B-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-2(MAS:L4H1)-1
53	HP 420-24	(DRB-F2-60-1-1-1-BB/[BETASYN]BC1-9-#/CML297)-B-17-1-1
54	HP 420-63	(P72c1xCML297 x CL-02410-3-1-1-B/CML305)-B-8-1-1
55	HP 422-47	(CML300/CML486)-7-2-2-B
56	HP393-3	((DTPYC9-F65-2-3-1-1-B-BxDTPYC9-F65-2-2-1-1-B-B)xDTPYC9-F86-1-1-1-1-B-B-B)-B-B-7-1-B
57	HP 422-69	(CML305/CML486)-8-1-1-B
58	HP 462-87	[CML445/[BETASYN]BC1-2-3-3-1//CML445/[BETASYN]BC1-2-3-3/KUI+SC55SYN#]-9-1-1-1
59	HP391-12	([CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1/CML297)-5-1-2
60	HP 422-16	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-4-1-1-1/CML297)-6-2-1-B
61	HP 422-34	(CML297/CML486)-9-1-2-B
62	HP 462-24	[MAS[MSR/312]-117-2-2-1-B-B-B/[BETASYN]BC1-4-1-6-#//MAS[MSR/312]-117-2-2-1-B-B-B/[BETASYN]BC1-4-1-6/KUI+SC55SYN#]-32-1-2-4
63	HP 462-47	[CML445/[BETASYN]BC1-1-2-6-#/CML445//[BETASYN]BC1-1-2-6/FloridaASYN#]-25-2-1-1
64	HP 434-65	(CML297/(CML489/[BETASYN]BC1-2-#/CML297)-B-24)-31-2

Line No.	Stock No.	Pedigree
65	HP 434-184	(CML297/(P72c1xCML297 x CL-02410-3-1-1-B/CML297)-B-1)-1-4
66	HP 434-228	(CML297/(P72c1xCML297 x CL-02410-3-1-1-B/CML297)-B-16)-15-1
67	HP 467-8	(KUI carotenoid syn-FS11-1-1-B-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-2(MAS:L4H1)-6
68	HP 467-51	(Florida A plus Syn-FS2-2-1-B-B/(KU1409/DE3/KU1409)S2-18-2-B)-B-5(MAS:L4H1)-6
69	HP 466-92	(KUI carotenoid syn-FS17-3-1-B-B/CML356//CML305)-2-1
70	HP 466-169	(KUI carotenoid syn-FS17-3-2-B-B/CML353//CML300)-9-3
71	HP 466-188	(KUI carotenoid syn-FS17-3-2-B-B/(CML239/GWIC)-1-7TL-1-1-1//CML300)-9-1
72	HP 466-223	(KUI carotenoid syn-FS17-3-2-B-B/CML356//CML300)-2-5
73	HP 466-384	(Carotenoid Syn3-FS5-1-5-B-B/CML355//DTPYC9-F65-2-3-1-1-B-B-B-B)-4-3
74	HP 466-436	(Carotenoid Syn3-FS5-1-5-B-B/(CML356/GWIB)-1-23TL-1-2-1//DTPYC9-F65-2-3-1-1-B-B-B-B)-6-2
75	HP 466-330	(Carotenoid Syn3-FS5-1-5-B-B/CML353//CML486)-6-1
76	HP 482-21	[[[K64R/G16SR]-39-1/[K64R/G16SR]-20-2]-5-1-2-B*4/CML390]-B-38-1-B-7-#/[BETASYN]BC1-1-1-1-#-B-B-B-B-B
77	HP 433-43, HP 482-20	[[[NAW5867/P30SR]-43-2/[NAW5867/P30SR]-114-1]-9-3-3-B-1-B/CML395-1]-B-13-1-B-4-#/[BETASYN]BC1-8-1-1-1-B-B-B-B
78	HP 433-3	[[EV7992]C1F2-430-3-3-3-X-7-B-B/CML202]-6-2-2-3-B*3/[BETASYN]BC1-10-1-1-1-1-B-B-B-B-B
79	HP 433-30	[CML 312/MAS[MSR/312]-117-2]-B-91-3-B-B/[BETASYN]BC1-2-1-1-1-B-B-B-B-B
80	HP 482-1	[CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1-B-B-B-B-B-B
81	HP 433-4	[DTPYC9-F11-2-3-1-1-B-B x DTPYC9-F46-1-2-1-1-B]-B-2-3-B-B)-B-B-B
82	HP 433-5	[DTPYC9-F46-1-2-1-1-B x DTPYC9-F74-1-1-1-1-B-B]-B-4-3-B-B-B-B-B
83	HP 433-6	[DTPYC9-F65-2-2-1-1-B-B x DTPYC9-F46-1-2-1-1-B]-B-3-2-B-B-B-B-B
84	HP 433-7, HP 482-2	[DTPYC9-F65-2-3-1-1-B-B x DTPYC9-F65-2-2-1-1-B-B]-3-4-2-B-B-B-B-B
85	HP 482-3	[DTPYC9-F65-2-3-1-1-B-B x DTPYC9-F65-2-2-1-1-B-B]-6-3-3-B-B-B-B-B
86	HP 433-8	[DTPYC9-F74-1-1-1-1-B-B x DTPYC9-F65-2-2-1-1-B-B]-B-3-4-B-B-B-B-B
87	HP 433-39	[GQL5/[GQL5/CML202]F2-1sx]-3-1-2-B/[BETASYN]BC1-2-5-1-2-B-B-B
88	HP 433-9	[SAM4/BETASYN]BC2FS1-1-1-1-B-B-B-B-B-B
89	HP 433-10	[SAM4/BETASYN]BC2FS3-1-3-3-B-B-B-B-B-B
90	HP 390-17	[SAM4/BETASYN]BC2FS36-4-1-2-B-B-B-B-B
91	HP 433-36	[ZM305/BETASYN]BC2-133-1-2-B-B-B-B-B-B
92	HP 433-37	[ZM305/BETASYN]BC2-133-1-3-B-B-B-B-B-B
93	HP 433-38	[ZM305/BETASYN]BC2-182-1-2-B-B-B-B-B
94	HP 433-12	Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-16-2-3-1-2-B-B-B-B
95	HP390-20, 1060-1	Ac8730SR-##-124-1-5-B-1-#/[BETASYN]BC1-5-#-B-B-B-B
96	HP 433-20	Carotenoid Syn3-FS4-2-4-B-B-B-B-B-B
97	HP376-67, 78	CML304
98	HP243-125, HP310-19, HP390-24	CML451
99	HP 433-40	CML486
100	HP 433-13, HP 482-4	CML488/[BETASYN]BC1-15-5-B-B-B-B
101	HP 482-22, HP390-31, 1060-3	CML488/[BETASYN]BC1-15-7-1-1-1-B-B-B

Line No.	Stock No.	Pedigree
102	HP 482-5, HP390-32, HP310-20 HP376-47,57,64,75	CML489/[BETASYN]BC1-2-#-B-B-B-B-B
103	HP 433-42, HP 482-18	CML489/[BETASYN]BC1-5-2-1-B-B-B-B-B
104	HP 433-14, HP 482-6	CML489/[BETASYN]BC1-7-2-1-1-4-B-B-B-B
105	HP 433-16	KUI carotenoid syn-FS17-3-1-B-B-B-B-B-B
106	HP 433-22	MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-10-2-1-#-B-B-B-B-B
107	HP 482-9	MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-10-3-#-B-B-B-B-B
108	HP 433-23, HP 482-10	MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-11-3-1-#-B-B-B-B-B
109	HP 433-28	MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-3-7-1-1-B-B
110	HP 433-24	MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-4-1-1-1-B-B-B-B-B-B
111	HP 433-25	MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-6-1-1-1-B-B-B-B-B
112	HP 433-26	MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-6-1-2-1-7-B-B-B-B
113	HP 433-27	MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-9-3-1-1-2-B-B-B-B
114	HP 433-29	MAS[MSR/312]-117-2-2-1-B-B-B/[BETASYN]BC1-11-5-2-1-3-B-B-B-B
115	HP 433-31	OBATANPA-SRc1F3(balbulk1)-#bal/[BETASYN]BC1-54-1-2-B-B-B-B-B-B-B-B
116	HP 433-32, HP 482-11	OBATANPA-SRc1F3(balbulk1)-#bal/[BETASYN]BC1-67-1-2-B-B-B-B-B-B-B-B
117	HP390-59	(ZM305/[BETASYN]BC1)-29-1-1-B-B-B-B-B
118	HP 433-35, HP 482-14	(ZM305/[BETASYN]BC1)-29-3-4-B-B-B-B-B-B-B-B
119	HP391-49	([CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1/CML305)-5-2-2
120	HP 422-15	(MAS[206/312]-23-2-1-1-B-B-B/[BETASYN]BC1-4-1-1-1/CML297)-6-1-1-B
121	HP 422-41	(CML300/CML486)-2-5-2-B
122	HP391-8	([CML197/N3//CML206]-X-32-1-4-B*5/[BETASYN]BC1-4-4-4-1/CML297)-4-5-1
123	HP 422-28	(CML297/CML486)-5-1-2-B
124	HP 462-26	[MAS[MSR/312]-117-2-2-1-B-B-B/[BETASYN]BC1-4-1-6-#//MAS[MSR/312]-117-2-2-1-B-B-B/[BETASYN]BC1-4-1-6/KUI+SC55SYN#]-32-1-3-2
125	HP 463-8	[[GQL5/[GQL5/[MSRXPOOL9]C1F2-205-1(OSU23i)-5-3-X-X-1-B-B]F2-4sx]-8-4-B//[BETASYN]BC1-7-4-3-#/[GQL5/[GQL5/[MSRXPOOL9]C1F2-205-1(OSU23i)-5-3-X-X-1-B-B]F2-4sx]-8-4-B/[BETASYN]BC1-7-4-3/KUISYN#]-4-2-2-2
126	HP 463-37	[[GQL5/[GQL5/[MSRXPOOL9]C1F2-205-1(OSU23i)-5-3-X-X-1-B-B]F2-4sx]-8-6-B-B//[BETASYN]BC1-14-2-1-#/[GQL5/[GQL5/[MSRXPOOL9]C1F2-205-1(OSU23i)-5-3-X-X-1-B-B]F2-4sx]-8-6-B-B/[BETASYN]BC1-14-2-1/KUISYN#]-16-1-1-1
127	HP 423-13	([[[K64R/G16SR]-39-1/[K64R/G16SR]-20-2]-5-1-2-B*4/CML390]-B-38-1-B-7-#/[BETASYN]BC1-1-1-1-#/CML305)-9-2-2-B

Supplemental Table II-2. Least square means of grain yield (across four environments), total provitamin A concentration (three environments), and genetic distance among inbred parents for 156 hybrids evaluated.

Set	Parent-1	Parent-2	Hybrid	Grain yield (Mg ha^{-1})	Total Provitamin A (ug g^{-1})	Genetic distance[#]
1	1	9	1x9	7.41	9.86	0.30
1	1	10	1x10	7.38	12.61	0.29
1	1	11	1x11	5.25	9.83	0.31
1	1	12	1x12	7.24	12.48	0.28
1	1	14	1x14	5.36	11.09	0.29
1	2	9	2x9	8.03	10.87	0.32
1	2	10	2x10	6.75	12.83	0.32
1	2	11	2x11	5.86	14.46	0.33
1	2	12	2x12	7.15	13.55	0.31
1	2	13	2x13	7.58	17.29	0.32
1	2	14	2x14	7.16	11.94	0.34
1	2	15	2x15	7.13	11.30	0.34
1	3	9	3x9	7.85	7.75	0.31
1	3	10	3x10	7.43	11.94	0.31
1	3	11	3x11	6.31	8.22	0.34
1	3	12	3x12	7.57	11.19	0.32
1	3	13	3x13	7.79	12.50	0.33
1	3	14	3x14	7.73	11.16	0.34
1	3	15	3x15	6.57	9.33	0.33
1	4	9	4x9	5.74	10.33	0.31
1	4	10	4x10	6.70	10.46	0.32
1	4	11	4x11	6.22	7.32	0.30
1	4	12	4x12	6.01	13.44	0.28
1	4	13	4x13	5.59	11.68	0.29
1	4	14	4x14	5.72	11.14	0.33
1	4	15	4x15	7.22	5.29	0.34
1	5	9	5x9	5.71	11.08	0.32
1	5	10	5x10	5.73	10.79	0.39
1	5	11	5x11	5.01	10.64	0.37
1	5	12	5x12	6.16	12.98	0.37
1	5	13	5x13	6.36	13.62	0.36
1	5	14	5x14	5.73	11.16	0.36
1	5	15	5x15	6.01	10.53	0.41
1	6	9	6x9	7.09	16.06	0.26

Set	Parent-1	Parent-2	Hybrid	Grain yield (Mg ha^{-1})	Total Provitamin A (ug g^{-1})	Genetic distance[#]
1	6	10	6x10	7.02	17.70	0.30
1	6	11	6x11	5.33	14.15	0.32
1	6	12	6x12	5.73	17.78	0.28
1	6	13	6x13	6.59	21.12	0.29
1	6	14	6x14	5.37	18.38	0.30
1	6	15	6x15	5.99	16.06	0.32
1	7	9	7x9	7.20	10.31	0.33
1	7	10	7x10	5.71	10.78	0.32
1	7	11	7x11	6.08	9.49	0.33
1	7	12	7x12	5.93	10.89	0.29
1	7	13	7x13	6.64	15.96	0.32
1	7	14	7x14	6.56	13.13	0.32
1	7	15	7x15	6.05	13.22	0.37
2	1	16	1x16	5.38	7.90	0.30
2	1	17	1x17	4.70	10.63	0.28
2	1	18	1x18	6.26	10.21	0.31
2	1	19	1x19	4.44	14.28	0.33
2	1	20	1x20	4.68	7.61	0.31
2	1	21	1x21	6.52	11.57	0.29
2	2	16	2x16	5.38	7.61	0.31
2	2	17	2x17	7.04	12.54	0.32
2	2	18	2x18	7.86	11.58	0.31
2	2	19	2x19	7.61	9.14	0.34
2	2	20	2x20	5.74	7.44	0.32
2	2	21	2x21	5.47	9.00	0.30
2	3	16	3x16	5.25	6.40	0.32
2	3	17	3x17	7.16	10.48	0.32
2	3	18	3x18	8.45	8.39	0.33
2	3	19	3x19	7.12	10.24	0.34
2	3	20	3x20	5.68	7.67	0.30
2	3	21	3x21	5.90	8.08	0.29
2	4	16	4x16	5.03	8.11	0.28
2	4	17	4x17	6.37	11.65	0.32
2	4	18	4x18	7.36	10.10	0.32
2	4	19	4x19	6.20	11.39	0.31
2	4	20	4x20	4.58	6.28	0.26
2	4	21	4x21	5.46	14.09	0.29
2	5	16	5x16	4.94	9.29	0.35

Set	Parent-1	Parent-2	Hybrid	Grain yield (Mg ha^{-1})	Total Provitamin A (ug g^{-1})	Genetic distance[#]
2	5	17	5x17	6.01	9.62	0.39
2	5	18	5x18	6.38	10.00	0.36
2	5	20	5x20	6.01	9.21	0.37
2	5	21	5x21	5.34	9.09	0.32
2	6	16	6x16	4.52	11.61	0.31
2	6	17	6x17	6.32	17.18	0.32
2	6	18	6x18	7.08	15.35	0.30
2	6	19	6x19	5.41	17.80	0.31
2	6	20	6x20	5.19	13.30	0.27
2	6	21	6x21	5.24	18.17	0.29
2	7	17	7x17	6.33	10.41	0.32
2	7	18	7x18	7.39	10.14	0.34
2	7	20	7x20	5.22	9.74	0.30
2	7	21	7x21	7.32	9.87	0.32
2	8	16	8x16	7.01	6.74	0.30
2	8	17	8x17	6.78	11.84	0.30
2	8	18	8x18	6.56	10.61	0.30
2	8	19	8x19	5.47	14.14	0.31
2	8	20	8x20	5.51	8.81	0.29
2	8	21	8x21	6.62	11.26	0.27
3	9	16	9x16	6.88	6.24	0.30
3	9	17	9x17	7.13	8.62	0.33
3	9	18	9x18	8.11	8.36	0.31
3	9	19	9x19	6.90	8.58	0.28
3	10	16	10x16	6.18	9.17	0.31
3	10	17	10x17	5.84	10.84	0.22
3	10	18	10x18	5.75	9.28	0.27
3	10	19	10x19	5.52	10.41	0.25
3	10	20	10x20	6.39	9.20	0.31
3	10	21	10x21	6.25	9.07	0.32
3	11	16	11x16	5.27	8.59	0.33
3	11	17	11x17	7.24	7.41	0.29
3	11	18	11x18	6.99	8.21	0.31
3	11	19	11x19	5.65	5.62	0.29
3	11	20	11x20	5.75	5.78	0.33
3	12	16	12x16	7.05	8.43	0.30
3	12	17	12x17	6.74	11.95	0.33
3	12	18	12x18	8.35	11.82	0.31

Set	Parent-1	Parent-2	Hybrid	Grain yield (Mg ha^{-1})	Total Provitamin A (ug g^{-1})	Genetic distance[#]
3	12	19	12x19	5.06	12.28	0.25
3	12	20	12x20	5.98	8.84	0.32
3	12	21	12x21	6.11	14.47	0.28
3	13	16	13x16	2.76	5.69	0.23
3	13	17	13x17	6.30	12.56	0.32
3	13	18	13x18	6.89	10.05	0.31
3	13	19	13x19	3.10	11.33	0.16
3	13	20	13x20	6.97	8.52	0.30
3	13	21	13x21	5.67	12.86	0.29
3	14	17	14x17	7.13	10.21	0.31
3	14	18	14x18	6.73	10.85	0.30
3	14	19	14x19	5.31	9.42	0.24
3	14	20	14x20	5.98	8.70	0.30
3	14	21	14x21	5.56	7.07	0.27
3	15	16	15x16	4.96	7.48	0.29
3	15	17	15x17	7.32	9.46	0.34
3	15	18	15x18	6.76	8.34	0.33
3	15	19	15x19	4.84	11.13	0.21
3	15	20	15x20	5.23	7.41	0.34
3	15	21	15x21	5.73	9.69	0.33
4	1	2	1x2	6.24	12.53	0.30
4	1	3	1x3	5.84	11.87	0.28
4	1	4	1x4	5.35	11.86	0.27
4	5	2	5x2	6.43	13.41	0.39
4	5	3	5x3	5.45	10.74	0.37
4	5	4	5x4	5.50	12.05	0.36
4	6	2	6x2	4.38	20.58	0.26
4	6	3	6x3	5.24	13.75	0.27
4	6	4	6x4	3.19	17.59	0.22
5	9	12	9x12	5.34	10.53	0.22
5	9	13	9x13	5.83	10.00	0.23
5	9	14	9x14	7.16	9.03	0.25
5	10	12	10x12	6.13	10.91	0.26
5	10	13	10x13	5.33	12.43	0.23
5	11	12	11x12	6.49	9.92	0.32
5	11	13	11x13	5.91	8.72	0.28
5	11	14	11x14	6.88	7.64	0.30
6	17	16	17x16	6.07	6.15	0.27

Set	Parent-1	Parent-2	Hybrid	Grain yield (Mg ha^{-1})	Total Provitamin A (ug g^{-1})	Genetic distance[#]
6	17	19	17x19	7.93	10.49	0.35
6	17	20	17x20	4.80	8.41	0.23
6	18	16	18x16	7.26	6.05	0.24
6	18	19	18x19	8.38	11.42	0.32
6	18	20	18x20	5.30	7.51	0.23
6	21	16	21x16	4.54	6.99	0.22
6	21	19	21x19	5.50	9.38	0.29
6	21	20	21x20	4.29	5.97	0.20

[#] Shared allele distance across 402 SNP markers.

CHAPTER III

ASSOCIATION MAPPING AND GENOMIC PREDICTION FOR CAROTENOID CONCENTRATIONS IN MAIZE GRAIN

ABSTRACT

Genome-wide association studies have been used extensively to identify allelic variation for genes controlling important phenotypes in plants. The objectives of this study were to identify genes and genic regions controlling natural variation for carotenoid concentrations, and to build additive linear models from the high density SNP marker data set to predict carotenoid concentrations in breeding programs. Inbred lines differed significantly for concentrations of all carotenoids ($P<0.01$). The population structure of the association mapping panel was best classified into four clusters based on the K-means method. The proportion of phenotypic variation due to population structure differed among traits; 5-7% for zeaxanthin and lutein:zeaxanthin ratio, 17-18% for lutein and β-cryptoxanthin, and 26-28% for β-carotene and total provitamin A concentrations. Association mapping using a model with SNP markers and ten principal components identified the zeaxanthin epoxidase (*ZEP*) gene on chromosome 2 ($R^2=0.14$), and a significant marker on chromosome 10 ($R^2=0.10$) located close to the β-carotene hydroxylase gene, *CrtRB1*, as important regions controlling carotenoids, among other significant marker associations with carotenoid phenotypes. Additive linear models using selected SNP markers successfully predicted carotenoid concentrations as indicated by large correlations between observed and predicted values ($r\geq0.8$, $P<0.01$); good accuracies were obtained for β-carotene and total provitamin A concentrations where the observed-predicted correlations were

0.83 and 0.80, respectively, and the root mean square error (RMSE) was 1.01 and 2.21 μg g^{-1}. These results suggest that genomic prediction has potential value for increasing the efficiency of maize provitamin A biofortification breeding programs.

Keywords: association mapping, genomic prediction, population structure

INTRODUCTION

Maize is one of the most widely consumed staple foods, especially for many people of sub-Saharan Africa and Latin America (Nuss and Tanumihardjo, 2010). While vitamin A deficiency is prevalent in this region, maize biofortification with high levels of provitamin A carotenoids in the grain is a promising solution to overcome this problem (Graham et al., 2001). HarvestPlus, through the International Maize and Wheat Improvement Center (CIMMYT) and other partners, has been breeding maize hybrids and open pollinated varieties with increased total provitamin A carotenoid concentrations since 2004 (Pfeiffer and McClafferty, 2007). Currently, the target of 15 μg g^{-1} of provitamin A carotenoids has been achieved in some breeding lines and populations, and three outstanding hybrids with total provitamin A carotenoid concentration more than 7 μg g^{-1} have been officially released for commercialization in Zambia in 2012.

Marker assisted selection is regarded as a key approach for facilitating efficient breeding for high levels of provitamin A carotenoids in maize (Prasanna et al., 2010). Two genes, β-carotene hydroxylase-1 (*CrtRB1*) on chromosome 10 (Yan et al., 2010) and lycopene epsilon-cyclase (*LcyE*) on chromosome 8 (Harjes et al. 2008), have been reported to affect provitamin A carotenoid concentrations in maize grain. *CrtRB1* was found to explain a 15-fold change in the β-carotene to β-cryptoxanthin ratio (Yan et al., 2010) and more recently, 2-10 fold effect of

CrtRB1-3'TE polymorphism on increasing the β-carotene and total provitamin A concentrations has been demonstrated in 26 tropical maize populations (Babu et al., 2012). Therefore, the *CrtRB1*'s favorable allele is very meaningful to select in the breeding process. Although not as significant as *CrtRB1*, *LcyE* was also found to have considerable influence on provitamin A carotenoid concentrations, where 58% of variation in α- versus β-branches in the carotenoid pathway is explained by four regions within this gene (Harjes et al., 2008).

While marker assisted selection for favorable allele(s) of *CrtRB1* has been very helpful during development of outstanding high provitamin A maize cultivars, the carotenoid pathway is diverse and many genes play critical roles. Therefore, searching for favorable alleles for other important (rate-limiting) genes and use of genomic prediction tools are promising approaches to enhance selection efficiency. Association mapping has been used extensively in the past few years to identify genes controlling important phenotypes in plants, such as flowering time in maize (*Zea mays* L.) (Ducrocq et al., 2008; Buckler et al., 2009), Arabidopsis (*Arabidopsis thaliana* L.) (Brachi et al., 2010), barley (*Hordeum vulgare* L.) (Stracke et al., 2009), and ryegrass (*Lolium perenne* L.) (Skøt et al., 2005); yield and its components in rice (*Oryza sativa* L.) (Agrama et al., 2007); and quality traits in potato (*Solanum tuberosum* L.) (D'hoop et al., 2007) and cotton (*Gossypium hirsutum* L.) (Abdurakhmonov et al., 2008).

Association mapping has a few significant advantages relative to linkage mapping, such as the ability to include wider variance for the trait of interest and obtaining finer mapping results (reviewed in Yu & Buckler, 2006). Whereas linkage mapping uses a structured population (for example, F_2, backcross, or recombinant inbred lines (RIL)), association mapping relies on historical recombination that has occurred over many years. In cross-pollinated species such as

maize, linkage disequilibrium (LD) decays faster than in self-pollinated plant species (Abdurakhmonov and Abdukarimov, 2008); therefore, greater historical recombination and higher mapping resolutions is expected for cross-pollinated species than for self-pollinated species, but more markers are required to capture allelic variations.

Population structure and familial relationships among lines are two major challenges that can cause spurious associations between markers and the trait of interest in association studies using collections of inbred lines. Several statistical methods have been used to increase the precision of association mapping results; one of the most recent is using a mixed linear model involving genotype (G) and population structure (Q) as fixed effects and familial relatedness (K) as a random effect to optimize the control of type I and II error rates in declaring significance of the markers (Yu et al., 2006). Correcting for population structure can be accomplished by principal component analysis (PCA) or through Bayesian approaches such as STRUCTURE (Pritchard et al., 2000), while familial relatedness can be represented in the model by a pair-wise kinship matrix (Yu et al., 2006).

Besides population and family structures, other important considerations for association mapping are marker density, population size, and genetic control of the trait of interest. Use of a high marker density data set in association mapping is of great importance to capture rare variations in the genome. Most recently, the genotype-by-sequencing (GBS) platform (Elshire et al., 2011) enables the possibility of single nucleotide polymorphism (SNP) genotyping using more than a million SNPs. This platform offers the opportunity to obtain dense coverage of the genome, but the call rate (proportion of non-missing genotypes at a marker) is typically small. Moreover, population size affects the power to identify polymorphisms associated with the trait

of interest. For ten quantitative trait loci (QTL) linkage mapping studies using F_2 mapping population, the power to detect QTL increased from 0.1-0.4 using population size of 100 to 0.5-0.9 for population size of 500, depending on the heritability of the trait (reviewed in Bernardo, 2002). While using a large, well-defined population such as the nested association mapping (NAM) panel in maize (McMullen et al., 2009) is ideal for general association studies, association panels comprised of elite germplasm from breeding programs such as CIMMYT's carotenoid association mapping (CAM), Drought Tolerant Maize for Africa (DTMA), and Improved Maize for African Soils (IMAS) panels (www.cimmyt.org) are particularly useful in identifying marker trait associations that are of practical utility to the target breeding programs. The use of broad genotypic variation for specific traits of interest is expected to increase the power to identify rare variants. While association mapping for qualitative traits (controlled by few genes with large effects), such as endosperm color or tryptophan content in maize, is a straight-forward exercise, the same for quantitative traits (many genes, small effects) which are characterized by low heritability and large genotype by environment interaction effects (GxE), requires adoption of more sophisticated statistical models (Yu and Buckler, 2006).

Selection based on genomic predictions offers the opportunity to efficiently modify carotenoid concentrations. With the increasing accessibility of high volume-low cost genotyping platforms, genomic selection can be of great use in breeding programs, especially for early-generation screening. Because genomic selection is a recent tool for plant breeding, few applications have been reported to date; however, some results in maize and barley have demonstrated its potential. In maize, a recent study in European maize revealed that genomic prediction using 960 SNPs across six segregating populations and using random regression best

linear unbiased prediction produced correlations between observed and predicted values of 0.81 for grain moisture and 0.36 for grain yield (Zhao et al., 2012). Multi-collinearity, or multiple linear relationships among SNP markers, is an important challenge in building models for genomic predictions, and several statistical tools are available to account for these, including stepwise regression, ridge regression, and others (reviewed in Heffner et al., 2009; Heslot et al., 2012).

The objectives of this study were to identify genes and genic regions controlling natural variation for carotenoid concentrations, and to build additive linear models from the high density SNP marker data set for use in predicting carotenoid concentrations in breeding programs.

MATERIAL AND METHODS

Phenotype Data

CIMMYT's carotenoid association mapping (CAM) panel, which is an expansion of Yan et al.'s (2010) CAM panel of about 200 lines, consisting of 435 diverse tropical, subtropical, and temperate maize lines, was used in this study. The lines were grown in two years (summer 2010 and summer 2011) in Tlaltizapan, Morelos, Mexico (18°41′ N, 99°07′ W; 945 m above sea level; average annual temperature 23.5°C; average annual precipitation 840 mm) using a single replication of one row, five meters plots. Two to six plants were self-pollinated and the seeds were bulked for carotenoids analysis at the CIMMYT maize quality laboratory. Random samples of 20-30 seeds were kept frozen at -80°C until being ground to a fine powder (0.5 μm), followed by the CIMMYT laboratory protocols for carotenoids analysis, including extraction, separation, and quantification by HPLC (Galicia et al., 2008). Lutein, zeaxanthin, β-

cryptoxanthin, β-carotene, and total provitamin A concentrations (equal to 0.5(β-cryptoxanthin)+β-carotene) were measured and reported in μg g^{-1} of kernel dry weight.

Genotype Data

Two SNP marker platforms were used: 55K and GBS, consisting of around 55,000 and 680,000 markers, respectively. The 55K genotyping used the MaizeSNP50 Genotyping BeadChip from Illumina (catalog is available at www.illumina.com) and was done at Syngenta facility, Slater, IA, and the GBS genotyping was conducted at the Institute for Genomic Diversity, Cornell University, Ithaca, NY. SNP markers having call rate greater than 0.85 and minor allele frequency greater than 0.05 from the 55K and GBS data sets were selected and all missing genotype data were then imputed using TASSEL software (www.maizegenetics.net/tassel/). This resulted in final numbers of SNP markers for each data set of 38,421 and 103,466, respectively; the two data sets were also combined to constitute a 55K+GBS data set consisting of 141,887 SNPs.

Statistical Analysis

An analysis of variance with years and inbred lines as random effects was performed for each carotenoid trait. All response variables (y) were transformed to ln(y+1) prior to analyses to approach normality of residuals and equality of variances assumptions. Distributions of phenotypic values before and after transformation are presented in Supplemental Figure III-1. Pearson phenotypic correlation coefficients among carotenoid concentrations were calculated using inbred line least square means in ln(y+1) scale, and Spearman rank correlation coefficients between years were calculated to evaluate consistency of phenotypes across the two years.

ANOVA and correlation analyses were conducted using SAS/STAT software 9.2 (SAS Institute, 2009).

The population structure for the GBS data set was evaluated using the K-means clustering method, followed by the discriminant analysis of principal components (DAPC) (Jombart et al., 2010) using the *adegenet* package in R (R Core Team, 2012). The grouping of the lines based on DAPC was then used for labeling principal component analysis (PCA) results using the 55K, GBS, and 55K+GBS data sets. Ten principal components from the PCA using the 55K+GBS data set were used as covariates in linear models for association mapping analyses. The PCA was performed using the method implemented in the EIGENSTRAT software, in which number of principal components of 1, 2, 5, and 10 produce similar results, and 10 principal components are recommended as a default value for this software (Price et al., 2006).

Individual SNP-based association tests were conducted using the correlation/trend method (Weir, 2008) on the combined 55K+GBS data set using SNP & Variation Suite v7.6 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com). Two association mapping models were used:

$$Y = SNP*\beta + \varepsilon \text{ (Model 1)}$$

$$Y = SNP*\beta + PC*\alpha + \varepsilon \text{ (Model 2)}$$

where Y = response variable (least square means of carotenoids phenotype), SNP = SNP marker, and PC = principal component coordinate from the PCA. The phenotypic values (y) were transformed to ln(y+1) scale for all carotenoids. Association mapping model evaluations were based on visual observation of the probability-probability (P-P) plots, which are the plots of

observed P-values versus expected P-values under the null hypothesis that there is no association between marker and phenotype. The false discovery rate (FDR) method according to Storey (2002) was used to correct P-values for multiple testing using a significance threshold FDR < 0.005 ($-\log_{10}(FDR) > 2.3$) for all association mapping results. The positions of significant markers were then compared to the positions of candidate genes as found in the maize sequence database (www.maizesequence.org) (Table III-1).

Prediction of carotenoids phenotypes using SNP markers was conducted using the combined 55K+GBS data set. For each carotenoid, the 200 most-significant markers (unadjusted correlation/trend P-value < 0.005) from the association mapping Model 2 were selected as candidate markers. A full model using these 200 SNPs was fitted to the complete data set and then the stepwise model selection procedure, using the Akaike information criterion (AIC) as goodness-of-fit parameter (Yamashita, 2007), was used to build the best additive model for each carotenoid. Ten-fold cross validation was used to test this model using the 'DAAG' package in R (R Core Team, 2012). The validation procedure involved dividing the 435-line data set randomly into ten subsets, and then one subset (10% of the data) served as the validation set and the other 9 subsets (90% of the data) were combined to serve as the training set. This validation process was repeated ten times by changing the validation set with another 10% subset, and therefore every line participated once in the validation set. The model equation was built for the training set and then used for prediction in the validation set; the cross-validation predicted values are referred to as genomic estimated breeding values (GEBV) (Heffner et al., 2009). Pearson correlation coefficients of the GEBV with the observed values (least square means in ln(y+1) scale, as described earlier) and the root mean square error (RMSE), the square root of

average squared deviation of actual from predicted values, were calculated as measures of prediction accuracy.

RESULTS

Analyses of Variance and Correlations

There was significant variation among inbred lines for all carotenoids ($P < 0.01$) (data not shown). Pearson correlation coefficients among carotenoid concentrations were all significant ($P < 0.01$), except between β-cryptoxanthin and lutein (Table III-2). Strong correlations ($r \geq 0.6$) were observed between β-cryptoxanthin and zeaxanthin, and for total provitamin A with β-cryptoxanthin and β-carotene. Spearman rank correlation coefficients between the two year environments were highly significant for all carotenoids ($r \geq 0.76$, $P < 0.01$), indicating that the carotenoid phenotypes were generally consistent across years (Table III-2).

Population Structure

The population structure among the 435 lines was well described by the K-means clustering method followed by the discriminant analysis of principal components (DAPC) (Figure III-1), where BIC model selection on K-means clustering using the GBS data set indicated that four clusters were most likely for this population (Supplemental Figure III-2 A). The grouping of the lines based on the K-means and DAPC methods using this data set were highly associated with each other (Supplemental Figure III-2 B). Moreover, grouping based on the 55K data set corresponded to some extent with that based on the GBS, although there was some inconsistency in grouping of lines to groups 2 and 3 (Figure III-2 A). The GBS-based grouping provided satisfactory results in annotating the regular principal component analysis plot

based on the 55K+GBS data (Figure III-2 C), reflecting the fact that the combined data were dominated by the GBS SNPs and/or GBS SNPs being free of ascertainment bias better captured diversity. Proportion of phenotypic variation due to population structure were 5-7% for zeaxanthin and lutein:zeaxanthin ratio, 17-18% for lutein and β-cryptoxanthin, and 26-28% for β-carotene and total provitamin A concentrations.

Association Mapping

The 55K platform had larger proportion of SNPs having low minor allele frequency (MAF<0.2 = 35%) than the GBS (57%) and the 55K+GBS (51%) data sets (Figure III-3). The 436 lines had average heterozygosity rate of 0.04, reflecting that most of these lines were in advanced inbreeding generations.

Eight SNP markers were significantly associated with zeaxanthin, seven each with lutein:zeaxanthin ratio and β-cyptoxanthin, 18 with β-carotene, and four with total provitamin A carotenoid concentrations (FDR-adjusted P-value < 0.005) (Table III-3 and Figure III-4). Association mapping identified the zeaxanthin epoxidase (*ZEP*) gene, located on chromosome 2 at 44,440,299–44,449,237 bp, where the most significant SNP marker (located inside the gene, at chromosome 2: 44,448,432, P FDR < 0.001) had an effect of 4.67 μg g^{-1} and explained 14% of the variance for zeaxanthin concentration. A SNP located 1.8Mb from the lycopene-epsilon-cyclase (*LcyE*) gene (chromosome 8: 138,882,594-138,889,812) (Table III-1) was significantly associated with lutein:zeaxanthin ratio (chromosome 8: 137,047,779) (Table III-3). No SNP was significantly associated with lutein concentration.

The most significant SNP marker for β-carotene (chromosome 10: 135,911,707, P FDR < 0.001, R^2 = 10%) was located near the β-carotene hydroxylase 1 gene, *CrtRB1* (chromosome 10:

136,057,214–136,060,219) (Table III-3). This SNP explained 10% of the variation for β-carotene concentration and had the effect size of 2.63 μg g^{-1}. Another significant SNP (chromosome 10: 135,170,838) that also located in relatively close proximity to the *CrtRB1* gene explained less of the phenotypic variation (6%), but had larger effect (4.86 μg g^{-1}) than the previously mentioned SNP. While most of the significant SNPs were located on chromosome 10, a few others were located on chromosomes 1, 2, 3 and 9, each explaining 6-7% of the variation for β-carotene concentration, and with effect sizes ranging from 0.10 to 0.89 μg g^{-1} (Table III-3). Interestingly, one significant SNP for β-cryptoxanthin (chromosome 9: 154,545,873, P FDR = 0.004) was located relatively close to the β-carotene hydroxylase 5 (*hyd5*) gene, suggesting that the *hyd5* gene has a role in increasing β-cryptoxanthin concentration in the grain.

Several significant SNPs are located in close proximity with the genes controlling putative uncharacterized proteins (www.maizesequence.org). For instance, there are three SNPs with favorable rare alleles, each located within a gene on chromosomes 2, 10, and 10 (si605047e03, LOC100279968, and LOC100194220) controlling putative uncharacterized proteins affecting zeaxanthin, β-cryptoxanthin, and β-carotene concentrations, respectively. Highly significant SNPs on chromosome 10 and chromosome 8 are likely associated with the β-carotene hydroxylase 1 (*CrtRB1*) and the lycopene epsilon-cyclase (*LcyE*) genes, respectively. Moreover, the results suggest that the β-carotene hydroxylase *hdy5* gene has a role on controlling β-cryptoxanthin concentrations in maize grain.

Correcting for population structure through PCA was important as illustrated in Figure III-5, where the model that accounted for genotype and population structure (Model 2) was

visibly superior to the model involving only genotypes (Model 1). This highlights the fact that some of the groups were genetically similar or distinct from each other and making adjustments for population structure in the model was necessary.

Genomic Predictions

Additive linear models using selected SNP markers successfully predicted carotenoid concentrations, as indicated by coefficients of determination (R^2) greater than 0.6 between observed and ten-fold cross-validation predicted values (GEBVs) ($r \geq 0.8$, $P < 0.01$) (Table III-4 and Figure III-6). The smaller subsets of significant SNPs (82-102) reported herein for various carotenoid traits could be of potential utility in designing marker based prediction/selection strategies. The best accuracies were obtained for predictions of β-carotene and total provitamin A concentrations, with observed-predicted correlations of 0.83 and 0.80, respectively, and root mean square errors (RMSE) of 1.01 and 2.21 μg g^{-1} (Table III-4). The SNP subset modeling total provitamin A concentration may be a useful selection tool in breeding programs.

DISCUSSION

Genotype by environment interaction effects can influence QTL mapping results, requiring the effect of QTL to be estimated for each environment (for example, see Zhang et al., 2008; Tétard-Jones et al., 2012). Although the significance of genotype by environment interaction effects could not be directly assessed (because only a single replication was grown at each environment), the Spearman rank correlation coefficients among years were large and significant for all traits ($r > 0.7$, $P < 0.01$), indicating that the genotype by year interaction, if

significant, was not of an extreme crossover type; therefore, all analyses used data combined across years.

The phenotypic correlation coefficients among carotenoids were generally as expected based on their known relationships in the carotenoid biosynthetic pathway (described in Yan et al., 2010). Lutein, which is on the alpha-branch, was only slightly or not significantly associated with carotenoids on the beta-branch of the pathway (β-carotene, β-cryptoxanthin, zeaxanthin). By contrast, there was a strong relationship between β-cryptoxanthin and zeaxanthin ($r = 0.65$, $P > 0.01$), as zeaxanthin is located downstream of β-cryptoxanthin on the β-branch of the pathway.

The most significant SNP from the association mapping for zeaxanthin resides inside the zeaxanthin epoxidase (*ZEP*) gene region. Vallabhaneni and Wurtzel (2009) suggested that the *ZEP* gene has a role in the conversion of zeaxanthin to violaxanthin (a precursor of abscisic acid in maize endosperm) and has two transcripts (*ZEP1* and *ZEP2*) that are negatively correlated with carotenoid accumulation. The effect size of the SNP (the difference of phenotypic value of the homozygous-major-allele relative to the homozygous-minor-allele genotypes) was +4.67 μg g^{-1}, indicating that the favorable allele was common in the population. With the minor allele frequency (MAF) of 0.24, the favorable allele of this SNP was possessed in about 76% of lines.

In the case of β-carotene, that the most significant marker (chromosome 10: 135,911,707) is assumed to be related to the β-carotene hyrdroxylase 1 (*CrtRB1*) gene (chromosome 10: 136,057,214-136,060,219) because polymorphisms of this gene have been confirmed to have large effects on β-carotene in a wide range of germplasm (Babu et al., 2012). Additionally, the negative effect size of this SNP (-2.63 μg g^{-1}) indicated that the favorable allele of the SNP was rare in the population; only possessed by around 11% of the lines (MAF=0.11).

For lutein:zeaxanthin ratio, one of the significant markers (chromosome 8: 137,047,779) may be related to the lycopene epsilon-cyclase (*LcyE*) gene (chromosome 8: 138,882,594-138,889,812), although there are a few other genes between this marker and the gene. Supporting this hypothesis is that the *LcyE* gene has been reported to have relatively large effect on relative carotenoid concentrations in the α- versus β-branches of the carotenoid pathway (Harjes et al., 2008) which is related to lutein:zeaxanthin ratio. The fact that this SNP is somewhat distant from the *LcyE* gene, is analogous to findings from another study that the *Y1* gene controlling grain color was strongly detected even with markers located around 3.7Mb away (see Chapter IV).

Presence or absence allelic state is not one of the possible states in the imputation methodused herein; therefore, the imputation procedure may have dismissed important variation. Re-assessing the association mapping results using methods that detect presence-absence variation (PAV) would merit future research.

The stepwise selection method had two advantages compared to ridge regression for predicting total provitamin A concentration using SNP markers, i.e. much fewer markers were required (98 versus 38,421), and more accurate prediction (correlation between observed and predicted values equal to 0.80 versus 0.35) was achieved. However, the R^2 values presented herein likely overestimate the predictive value of the model because the 200 candidate SNPs were selected based on association mapping tests involving all inbred lines (436) instead of only using the 90% training set. This finding from stepwise regression complements those of other studies in which ridge regression resulted in good accuracy for genomic selection for maturity (days to anthesis, days to silking, and anthesis to silking interval) in maize (Guo et al., 2012) and

barley (Iwata and Jannink, 2011). Results described by Hu et al. (2011) for genomic prediction of somatic embryo number in soybean suggest that using a model that includes epistasis could be significantly better than an additive model for predicting breeding values, which is a hypothesis that could be investigated in future research.

The ability to predict carotenoid concentrations using relatively few SNP markers opens the opportunity to enhance provitamin A breeding programs using efficient and rapid genomic selection instead of lower-through-put and costlier carotenoid phenotyping. This molecular approach would be useful for high-throughput screening and selection of genotypes with favorable alleles in segregating early generations, when the number of families and progenies are typically large. Although the cost of phenotyping carotenoids has decreased by using an ultra-high performance liquid chromatography (UPLC) platform in place of high performance liquid chromatography (HPLC), the cost of biochemical phenotyping remains expensive (N. Palacios, CIMMYT, pers. comm.). Therefore, although further validations and perhaps model-refining are needed, it is expected that these findings on SNP-based prediction can be useful in provitamin A carotenoid biofortification breeding programs.

CONCLUSIONS

Association mapping successfully identified the previously known zeaxanthin epoxidase gene (*ZEP*) on chromosome 2, and three other genes with favorable rare alleles on chromosomes 2, 10, and 10 (si605047e03, LOC100279968, and LOC100194220, respectively), controlling putative uncharacterized proteins affecting zeaxanthin, β-cryptoxanthin, and β-carotene concentrations, respectively. Highly significant SNPs on chromosome 10 and chromosome 8 are likely associated with the β-carotene hydroxylase 1 (*CrtRB1*) and the lycopene epsilon-cyclase

(*LcyE*) genes, respectively. Moreover, the results suggest that the β-carotene hydroxylase *hdy5* gene has a role on controlling β-cryptoxanthin concentrations in maize grain. A linear model of 98 SNP markers accurately predicted total provitamin A concentrations and could therefore be a valuable tool for maize provitamin A biofortification programs.

ACKNOWLEDGMENTS

Support for this work was provided by HarvestPlus (www.harvestplus.org), an international program that develops micronutrient-rich staple food crops to reduce hidden hunger among malnourished populations. Thanks go to Thanda Dhliwayo, Germán Mingramm, and José Luis Coyac for their support of all field activities, the field staff at CIMMYT Tlaltizapan research station that hosted the trials, and the staff of CIMMYT's maize quality laboratory for conducting the carotenoid analyses. Funding support provided by the Directorate General of Higher Education of Indonesia for this PhD study is highly appreciated.

REFERENCES

Abdurakhmonov, I.Y., and A. Abdukarimov. 2008. Application of association mapping to understanding the genetic diversity of plant germplasm resources. International journal of plant genomics 2008: 574927. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2423417&tool=pmcentrez&rendertype=abstract (verified 9 August 2012).

Abdurakhmonov, I.Y., R.J. Kohel, J.Z. Yu, a E. Pepper, a a Abdullaev, F.N. Kushanov, I.B. Salakhutdinov, Z.T. Buriev, S. Saha, B.E. Scheffler, J.N. Jenkins, and a Abdukarimov. 2008. Molecular diversity and association mapping of fiber quality traits in exotic G. hirsutum L. germplasm. Genomics 92(6): 478–87. Available at http://www.ncbi.nlm.nih.gov/pubmed/18801424 (verified 1 August 2012).

Agrama, H. a., G.C. Eizenga, and W. Yan. 2007. Association mapping of yield and its components in rice cultivars. Molecular Breeding 19(4): 341–356. Available at http://www.springerlink.com/index/10.1007/s11032-006-9066-6 (verified 31 August 2012).

Babu, R., N.P. Rojas, S. Gao, J. Yan, and K. Pixley. 2012. Validation of the effects of molecular marker polymorphisms in LcyE and CrtRB1 on provitamin A concentrations for 26 tropical maize populations. Theoretical and Applied Genetics. Available at http://www.springerlink.com/index/10.1007/s00122-012-1987-3 (verified 12 October 2012).

Bernardo, R. 2002. Breeding for Quantitative Traits in Plants. Stemma Press, MN.

Brachi, B., N. Faure, M. Horton, E. Flahauw, A. Vazquez, M. Nordborg, J. Bergelson, J. Cuguen, and F. Roux. 2010. Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS genetics 6(5): e1000940. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2865524&tool=pmcentrez&rendertype=abstract (verified 18 July 2012).

Buckler, E.S., J.B. Holland, P.J. Bradbury, C.B. Acharya, P.J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J.C. Glaubitz, M.M. Goodman, C. Harjes, K. Guill, D.E. Kroon, S. Larsson, N.K. Lepak, H. Li, S.E. Mitchell, G. Pressoir, J. a Peiffer, M.O. Rosas, T.R. Rocheford, M.C. Romay, S. Romero, S. Salvo, H. Sanchez Villeda, H.S. da Silva, Q. Sun, F. Tian, N. Upadyayula, D. Ware, H. Yates, J. Yu, Z. Zhang, S. Kresovich, and M.D. McMullen. 2009. The genetic architecture of maize flowering time. Science (New York, N.Y.) 325(5941): 714–8. Available at http://www.ncbi.nlm.nih.gov/pubmed/19661422 (verified 17 July 2012).

Ducrocq, S., D. Madur, J.-B. Veyrieras, L. Camus-Kulandaivelu, M. Kloiber-Maitz, T. Presterl, M. Ouzunova, D. Manicacci, and A. Charcosset. 2008. Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical

information. Genetics 178(4): 2433–7. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2323828&tool=pmcentrez&rendertype=abstract (verified 2 August 2012).

D'hoop, B.B., M.J. Paulo, R. a. Mank, H.J. Eck, and F. a. Eeuwijk. 2007. Association mapping of quality traits in potato (Solanum tuberosum L.). Euphytica 161(1-2): 47–60. Available at http://www.springerlink.com/index/10.1007/s10681-007-9565-5 (verified 31 August 2012).

Elshire, R.J., J.C. Glaubitz, Q. Sun, J. a Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PloS one 6(5): e19379. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087801&tool=pmcentrez&rendertype=abstract (verified 18 July 2011).

Galicia, L., E. Nurit, A. Rosales, and N. Palacios-Rojas (Eds). 2008. Maize nutrition quality and plant tissue analysis laboratory.

Graham, R.D., R.M. Welch, and H.E. Bouis. 2001. Addresing micronutrient malnutrition through enhancing the nutritional quality of staple foods: principles, perspectives, and knowledge gaps. 70.

Guo, Z., D.M. Tucker, J. Lu, V. Kishore, and G. Gay. 2012. Evaluation of genome-wide selection efficiency in maize nested association mapping populations. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 124(2): 261–75. Available at http://www.ncbi.nlm.nih.gov/pubmed/21938474 (verified 30 July 2012).

Harjes, C.E., T.R. Rocheford, L. Bai, T.P. Brutnell, C.B. Kandianis, S.G. Sowinski, A.E. Stapleton, R. Vallabhaneni, M. Williams, E.T. Wurtzel, J. Yan, and E.S. Buckler. 2008. Natural genetic variation in lycopene epsilon-cyclase tapped for maize biofortification. Science (New York, N.Y.) 319(5861): 330–3. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2933658&tool=pmcentrez&rendertype=abstract (verified 1 August 2011).

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic Selection for Crop Improvement. Crop Science 49(1): 1. Available at https://www.crops.org/publications/cs/abstracts/49/1/1 (verified 13 October 2012).

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. Crop Science 52(1): 146. Available at https://www.crops.org/publications/cs/abstracts/52/1/146 (verified 3 October 2012).

Hu, Z., Y. Li, X. Song, Y. Han, X. Cai, S. Xu, and W. Li. 2011. Genomic value prediction for quantitative traits under the epistatic model. BMC genetics 12(1): 15. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3038975&tool=pmcentrez&rendertype=abstract (verified 31 August 2012).

Iwata, H., and J.-L. Jannink. 2011. Accuracy of Genomic Selection Prediction in Barley Breeding Programs: A Simulation Study Based On the Real Single Nucleotide Polymorphism Data of Barley Breeding Lines. Crop Science 51(5): 1915. Available at https://www.crops.org/publications/cs/abstracts/51/5/1915 (verified 3 August 2012).

Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC genetics 11(1): 94. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2973851&tool=pmcentrez&rendertype=abstract (verified 13 July 2012).

McMullen, M.D., S. Kresovich, H.S. Villeda, P. Bradbury, H. Li, Q. Sun, S. Flint-Garcia, J. Thornsberry, C. Acharya, C. Bottoms, P. Brown, C. Browne, M. Eller, K. Guill, C. Harjes, D. Kroon, N. Lepak, S.E. Mitchell, B. Peterson, G. Pressoir, S. Romero, M. Oropeza Rosas, S. Salvo, H. Yates, M. Hanson, E. Jones, S. Smith, J.C. Glaubitz, M. Goodman, D. Ware, J.B. Holland, and E.S. Buckler. 2009. Genetic properties of the maize nested association mapping population. Science (New York, N.Y.) 325(5941): 737–40. Available at http://www.ncbi.nlm.nih.gov/pubmed/19661427 (verified 12 July 2012).

Nuss, E.T., and S. a. Tanumihardjo. 2010. Maize: A Paramount Staple Crop in the Context of Global Nutrition. Comprehensive Reviews in Food Science and Food Safety 9(4): 417–436. Available at http://doi.wiley.com/10.1111/j.1541-4337.2010.00117.x (verified 16 July 2012).

Pfeiffer, W.H., and B. McClafferty. 2007. HarvestPlus: Breeding Crops for Better Nutrition. Crop Science 47(Supplement_3): S–88. Available at https://www.crops.org/publications/cs/abstracts/47/Supplement_3/S-88 (verified 8 August 2011).

Prasanna, B.M., K. Pixley, M.L. Warburton, and C.-X. Xie. 2010. Molecular marker-assisted breeding options for maize improvement in Asia. Molecular Breeding 26(2): 339–356. Available at http://www.springerlink.com/index/10.1007/s11032-009-9387-3 (verified 19 September 2012).

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. Nature genetics 38(8): 904–9. Available at http://www.ncbi.nlm.nih.gov/pubmed/16862161 (verified 26 October 2012).

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155(2): 945–59. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461096&tool=pmcentrez&rendertype=abstract.

SAS Institute. 2009. SAS/STAT 9.2 User's Guide. Second ed. SAS Institute, Inc., NC.

Skøt, L., M.O. Humphreys, I. Armstead, S. Heywood, K.P. Skøt, R. Sanderson, I.D. Thomas, K.H. Chorlton, and N.R.S. Hamilton. 2005. An association mapping approach to identify flowering time genes in natural populations of Lolium perenne (L.). Molecular Breeding 15(3): 233–245. Available at http://www.springerlink.com/index/10.1007/s11032-004-4824-9 (verified 7 August 2012).

Storey, J.D. 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64(3): 479–498. Available at http://doi.wiley.com/10.1111/1467-9868.00346.

Stracke, S., G. Haseneyer, J.-B. Veyrieras, H.H. Geiger, S. Sauer, A. Graner, and H.-P. Piepho. 2009. Association mapping reveals gene action and interactions in the determination of flowering time in barley. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 118(2): 259–73. Available at http://www.ncbi.nlm.nih.gov/pubmed/18830577 (verified 12 July 2012).

R Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Tétard-Jones, C., M. a Kertesz, and R.F. Preziosi. 2012. Identification of plant quantitative trait loci modulating a rhizobacteria-aphid indirect effect. PloS one 7(7): e41524. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3406024&tool=pmcentrez&rendertype=abstract (verified 12 October 2012).

Weir, B.S. 2008. Linkage disequilibrium and association mapping. Annual review of genomics and human genetics 9: 129–42. Available at http://www.ncbi.nlm.nih.gov/pubmed/18505378 (verified 4 August 2012).

Yan, J., C.B. Kandianis, C.E. Harjes, L. Bai, E.-H. Kim, X. Yang, D.J. Skinner, Z. Fu, S. Mitchell, Q. Li, M.G.S. Fernandez, M. Zaharieva, R. Babu, Y. Fu, N. Palacios, J. Li, D. Dellapenna, T. Brutnell, E.S. Buckler, M.L. Warburton, and T. Rocheford. 2010. Rare genetic variation at Zea mays crtRB1 increases beta-carotene in maize grain. Nature genetics 42(4): 322–7. Available at http://www.ncbi.nlm.nih.gov/pubmed/20305664 (verified 28 July 2011).

Yu, J., and E.S. Buckler. 2006. Genetic association mapping and genome organization of maize. Current opinion in biotechnology 17(2): 155–60. Available at http://www.ncbi.nlm.nih.gov/pubmed/16504497 (verified 17 July 2012).

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38(2): 203–8. Available at http://www.ncbi.nlm.nih.gov/pubmed/16380716 (verified 12 July 2012).

Zhang, K., J. Tian, L. Zhao, and S. Wang. 2008. Mapping QTLs with epistatic effects and QTL x environment interactions for plant height using a doubled haploid population in cultivated wheat. Journal of genetics and genomics = Yi chuan xue bao 35(2): 119–27. Available at http://www.ncbi.nlm.nih.gov/pubmed/18407059.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H.P. Maurer, F.H. Longin, N. Ranc, and J.C. Reif. 2012. Accuracy of genomic selection in European maize elite breeding populations. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 124(4): 769–76. Available at http://www.ncbi.nlm.nih.gov/pubmed/22075809 (verified 22 August 2012).

Table III-1. Candidate genes of the carotenoid pathway

Chr	Position	Gene name	Abbreviation
1	17,660,941-17,667,054	Phytoene desaturase	*pds1*
2	15,865,938-15,868,219	β-carotene hydroxylase	*hyd1*
2	44,440,299-44,449,237	Zeaxanthin epoxidase	*ZEP*
6	55,671,246-55,674,458	Phytoene synthase	*psy*
7	17,470,585-17,479,020	ζ-carotene desaturase	*zds1*
8	138,882,594-138,889,812	lycopene epsilon-cyclase	*LcyE*
8	168,273,042-168,276,092	Phytoene synthase 2	*psy2*
9	153,692,212-153,694,576	β-carotene hydroxylase	*hyd5*
10	136,057,214-136,060,219	β-carotene hydroxylase	*CrtRB1*
10	4,705,086-4,705,639	Phytoene synthase 3	*psy3*

Adapted mainly from www.maizesequence.org. Chr=chromosome.

Table III-2. Pearson phenotypic correlation coefficients among carotenoids (from least square means, N=435, below diagonal) and Spearman rank correlation coefficients for each carotenoid evaluated in the two years' environments (N=316, diagonal).

Trait	Lutein	Zeaxanthin	β-Crypto-xanthin	β-Carotene	Total Provitamin A
Lutein	0.76**				
Zeaxanthin	0.18**	0.83**			
β-Cryptoxanthin	-0.04	0.65**	0.78**		
β-Carotene	-0.14**	0.20**	0.39**	0.82**	
Total Provitamin A	-0.17**	0.35**	0.63**	0.95**	0.80**

All phenotypic values were in ln(y+1) scale.
** Significant at 0.01 level.

Table III-3. Significant SNP markers (FDR-adjusted P-value < 0.005) from the 55K+GBS combined association mapping data sets

SNP	Chr	Position	FDR P[#]	R^2	MAF	Effect (μg g^{-1})[##]	Nearby gene[+]
Zeaxanthin							
S2_44448432	2	44,448,432	< 0.001	0.14	0.24	4.67	*ZEP*
S2_232361461	2	232,361,461	< 0.001	0.08	0.41	-3.39	**
S5_36547775	5	36,547,775	< 0.001	0.08	0.08	4.29	
S3_7937918	3	7,937,918	< 0.001	0.08	0.21	3.97	
S2_202278793	2	202,278,793	0.001	0.07	0.15	-5.59	
S9_141933919	9	141,933,919	0.002	0.06	0.06	4.65	
S5_16428141	5	16,428,141	0.003	0.06	0.06	4.46	
S10_30784997	10	30,784,997	0.003	0.06	0.36	3.41	
Lutein:zeaxanthin ratio							
S5_36547775	5	36,547,775	< 0.001	0.10	0.08	-0.33	
S3_15864036	3	15,864,036	< 0.001	0.08	0.10	-0.34	
S8_161903119	8	161,903,119	0.001	0.08	0.07	-0.31	
S2_196377782	2	196,377,782	0.001	0.07	0.06	-0.27	
S8_20831735	8	20,831,735	0.003	0.07	0.07	-0.29	
S8_137047779	8	137,047,779	0.003	0.07	0.11	-0.24	*LcyE*
S7_156112715	7	156,112,715	0.003	0.06	0.06	-0.32	
β-cryptoxanthin							
S2_234259494	2	234,259,494	0.004	0.07	0.19	-1.13	**
S5_11691569	5	11,691,569	0.004	0.07	0.17	1.19	
S4_87550096	4	87,550,096	0.004	0.06	0.06	-1.44	
S3_20642693	3	20,642,693	0.004	0.07	0.08	-0.92	
S1_211006905	1	211,006,905	0.004	0.06	0.15	1.35	
S9_20462571	9	20,462,571	0.004	0.06	0.18	1.31	
S9_154545873	9	154,545,873	0.004	0.07	0.12	1.20	*hyd5*
Total provitamin A							
S2_109492015	2	109,492,015	0.003	0.07	0.08	1.72	
S1_2693878	1	2,693,878	0.003	0.06	0.18	0.45	
S10_135911707	10	135,911,707	0.004	0.07	0.11	-3.14	*CrtRB1*
S10_133272617	10	133,272,617	0.004	0.07	0.09	-3.45	
S3_173371841	3	173,371,841	0.004	0.06	0.08	1.29	

SNP	Chr	Position	FDR P[#]	R^2	MAF	Effect ($\mu g\ g^{-1}$)[##]	Nearby gene[+]
β-carotene							
S10_135911707	10	135,911,707	< 0.001	0.10	0.11	-2.63	*CrtRB1*
S10_133272618	10	133,272,618	< 0.001	0.08	0.09	-2.62	
S10_69430367	10	69,430,367	0.001	0.07	0.07	-2.68	
S10_76506479	10	76,506,479	0.001	0.07	0.05	-6.45	**
S10_97726883	10	97,726,883	0.001	0.07	0.06	-3.12	
S10_125952485	10	125,952,485	0.001	0.07	0.10	-1.52	
S10_85009641	10	85,009,641	0.001	0.07	0.07	-2.63	
S9_124760961	9	124,760,961	0.001	0.07	0.26	0.62	
S2_109492013	2	109,492,013	0.001	0.07	0.08	0.89	
S10_19070007	10	19,070,007	0.001	0.06	0.05	-4.44	
S10_95917855	10	95,917,855	0.002	0.06	0.19	-0.89	
S2_117049318	2	117,049,318	0.003	0.06	0.23	-0.54	
S10_135170838	10	135,170,838	0.003	0.06	0.06	-4.86	
S3_173371841	3	173,371,841	0.004	0.06	0.08	0.62	
S10_119126746	10	119,126,746	0.004	0.06	0.37	0.62	
S1_2693878	1	2,693,878	0.004	0.06	0.18	0.10	
S10_80045974	10	80,045,974	0.004	0.06	0.10	-2.27	
S10_74807759	10	74,807,759	0.004	0.06	0.06	-5.24	

Chr = chromosome, FDR = false discovery rate, MAF = minor allele frequency

[#] FDR-adjusted P-value

[##] Effect = average phenotype of homozygous-major-allele – average phenotype of homozygous-minor-allele at the marker.

[+] The nearest previously identified gene according to Table III-1 (the furthest distance to the marker was around 1.8mb).

** SNP located inside a gene controlling a putative uncharacterized protein (www.maizesequence.org)

Table III-4. Prediction of carotenoids concentrations using additive linear models of SNP markers with 10-fold cross-validations[+]

Phenotype	No. of selected SNPs[#]	r yy'[##]	RMSE[##] (μg g^{-1})	Observed mean (μg g^{-1})	Predicted mean (μg g^{-1})
Lutein	93	0.84**	1.71	4.43	4.32
Zeaxanthin	95	0.84**	3.83	9.63	9.39
β-Cryptoxanthin	82	0.80**	1.37	3.89	3.79
β-Carotene	102	0.83**	1.01	2.32	2.25
Total provitamin A	98	0.80**	2.21	4.95	4.81

[+] Data were analyzed using the ln(y+1) transformation on the observed phenotypic least square means (y), but the results are presented in original scale (μg g^{-1}).

[#] The SNPs were selected using stepwise procedure from a candidate model contained 200 most significant SNPs (uncorrected p-value < 0.005) for each phenotype.

[##] r yy' = Pearson correlation coefficient among observed and cross-validation predicted values (GEBV). RMSE = root mean square error.

** Significant at P<0.01.

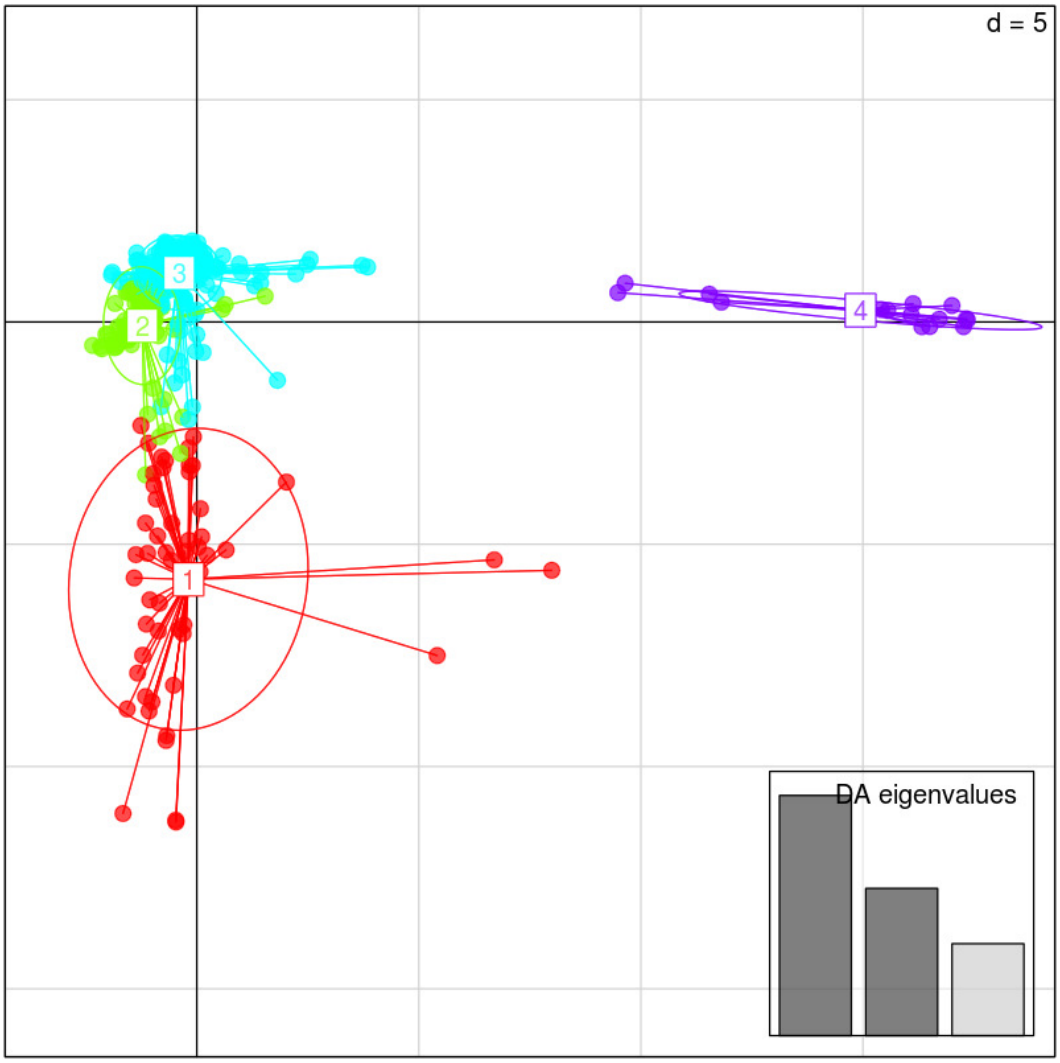
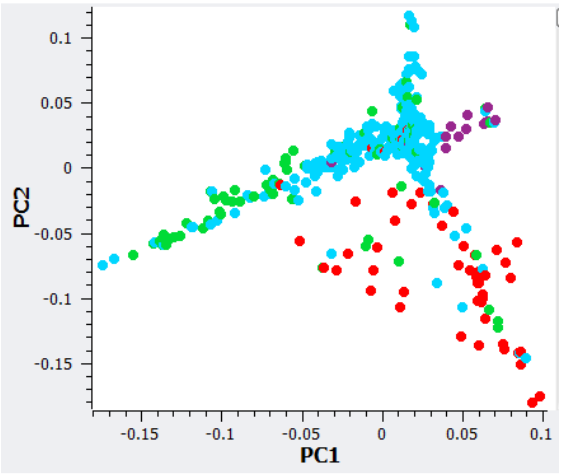
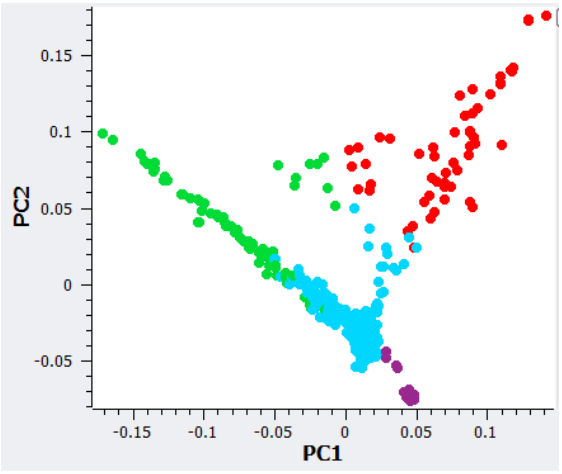


Figure III-1. DAPC plot based on the GBS data using 26 principal components and three linear discriminants under the dominant genetic model. The X and Y axes are the first and the second axes of linear discriminants, respectively.

(A)



(B)



(C)

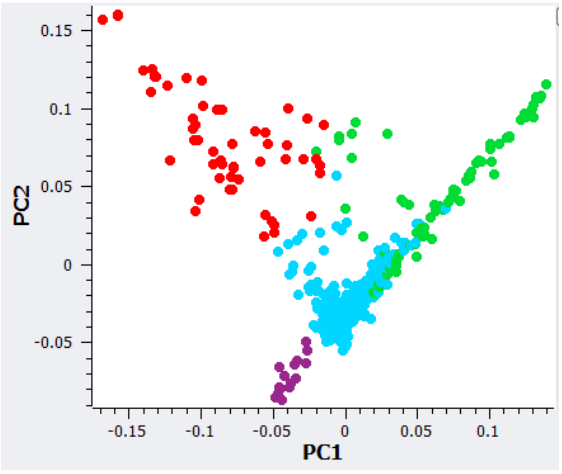
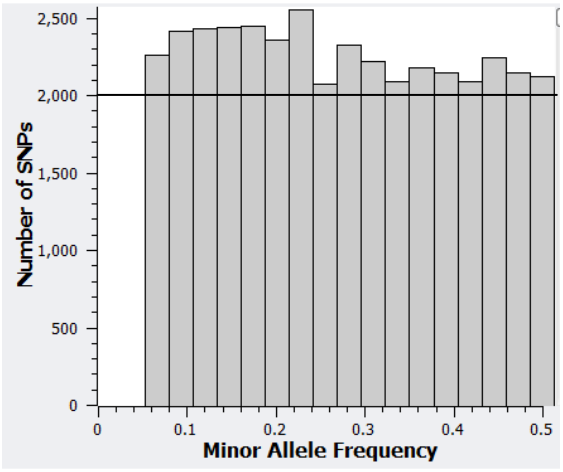
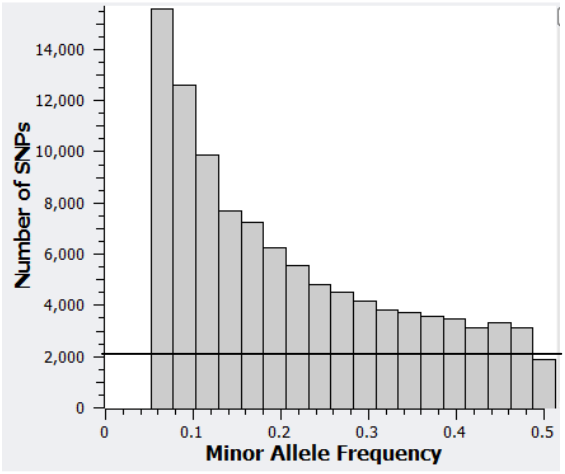


Figure III-2. Plots of the first and the second principal components computed from the (A) 55K, (B) GBS, and (C) 55K+GBS genotype data. Red, green, cyan, and purple colors represent GBS' DAPC group 1, 2, 3, and 4, respectively.

(A)



(B)



(C)

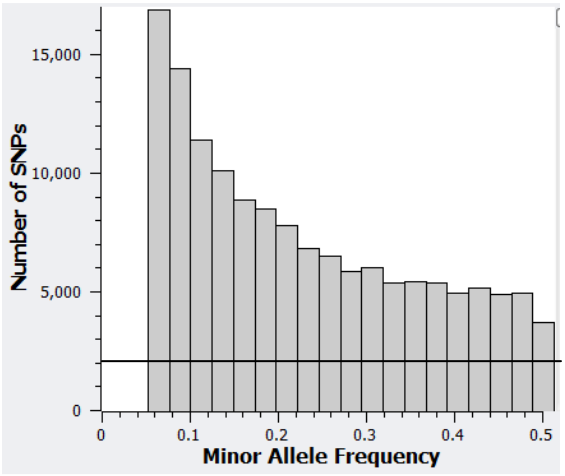


Figure III-3. Distribution of minor allele frequency of the (A) 55K, (B) GBS, and (C) 55K+GBS data sets.

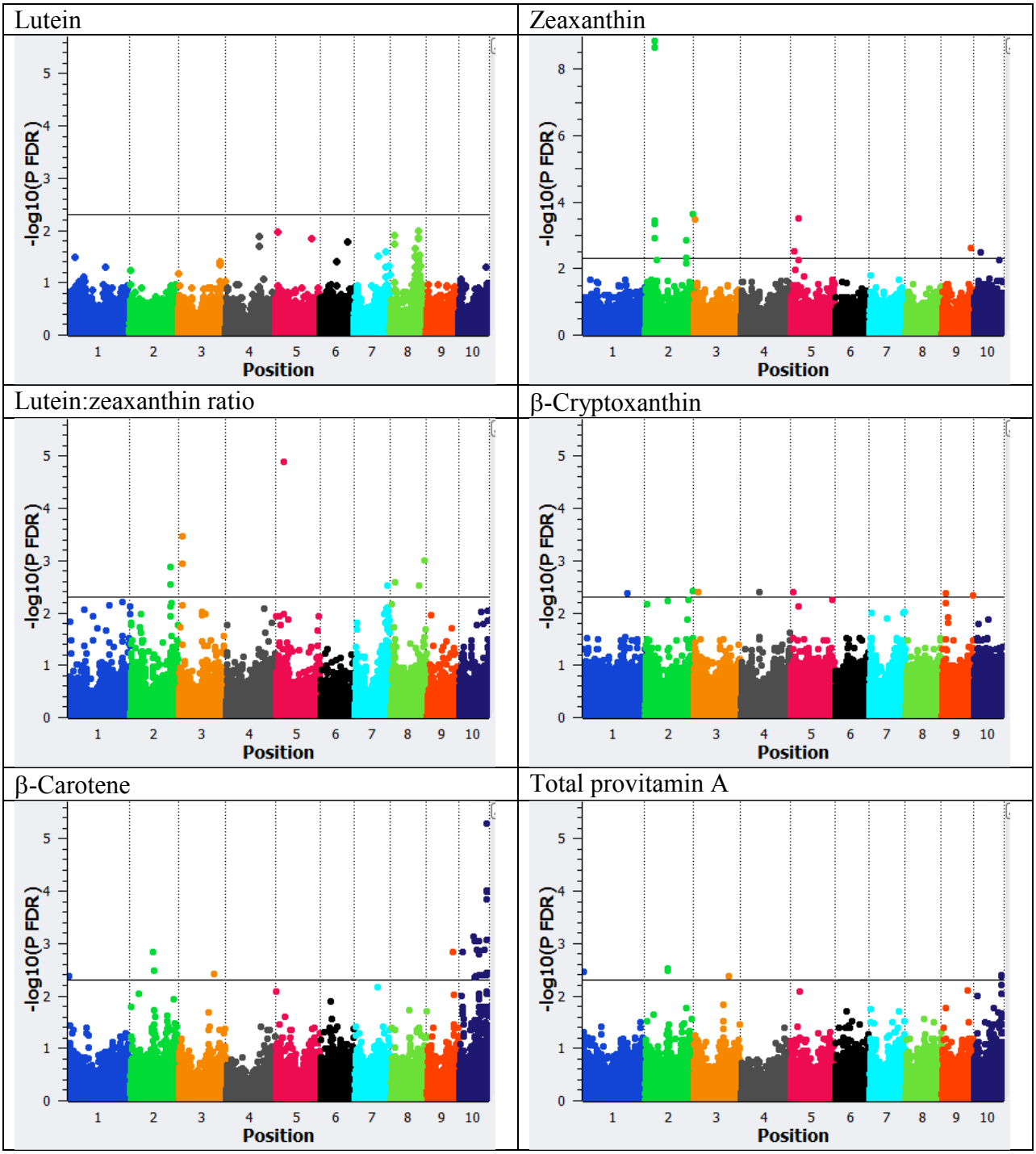
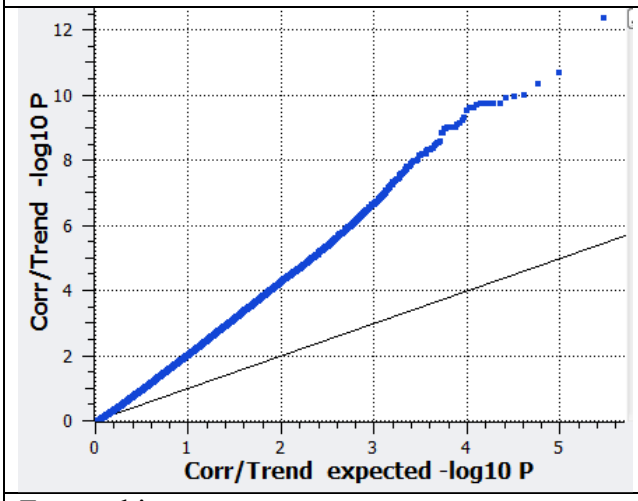
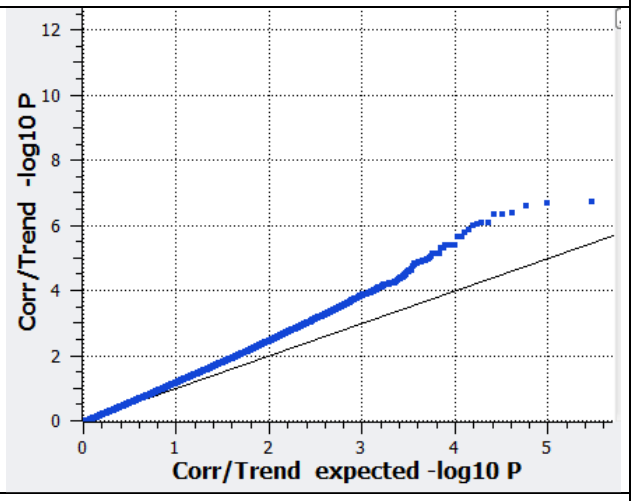
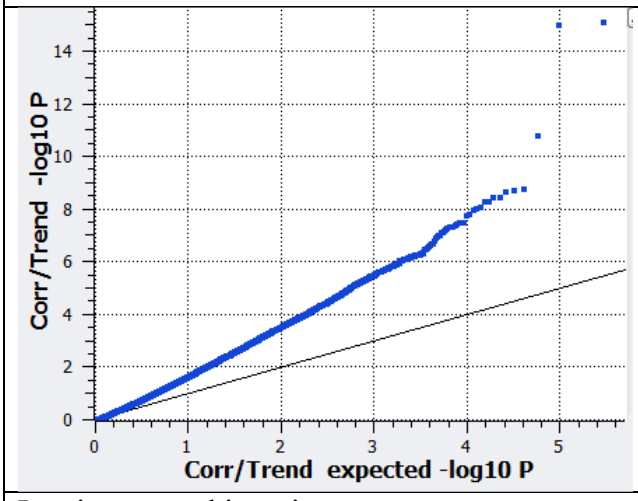
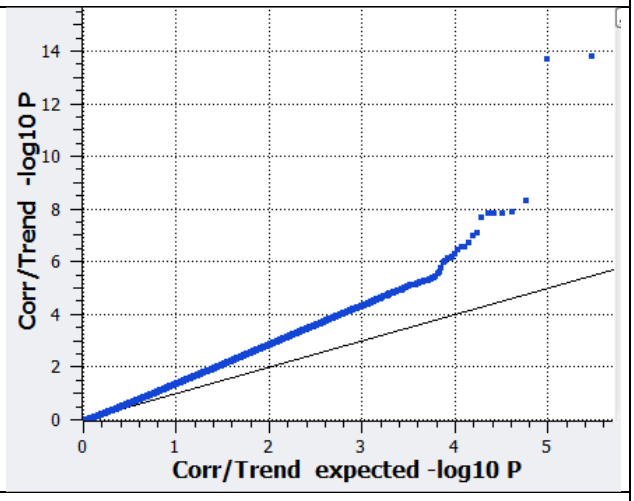
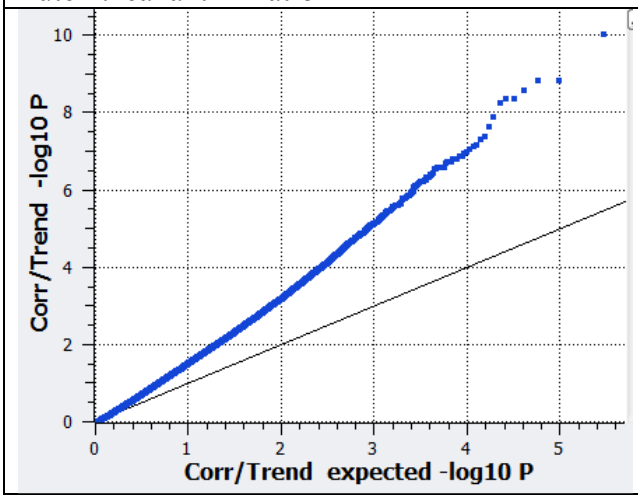
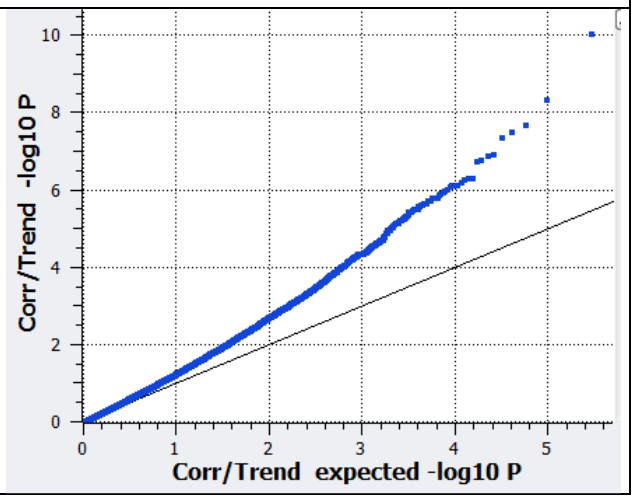


Figure III-4. Manhattan plots from the association mapping results of lutein, zeaxanthin, lutein:zeaxanthin ratio, β-cryptoxanthin, β-carotene, and total provitamin A concentrations using the Model 2 (with principal components) on the 55K+GBS combined data sets. All phenotypic values (y) were transformed to ln(y+1) prior to analyses. The black horizontal line is the 0.005 FDR-adjusted P-value significant thresholds.

Model 1 (without principal components)	Model 2 (with principal components)
Lutein	
	
Zeaxanthin	
	
Lutein:zeaxanthin ratio	
	

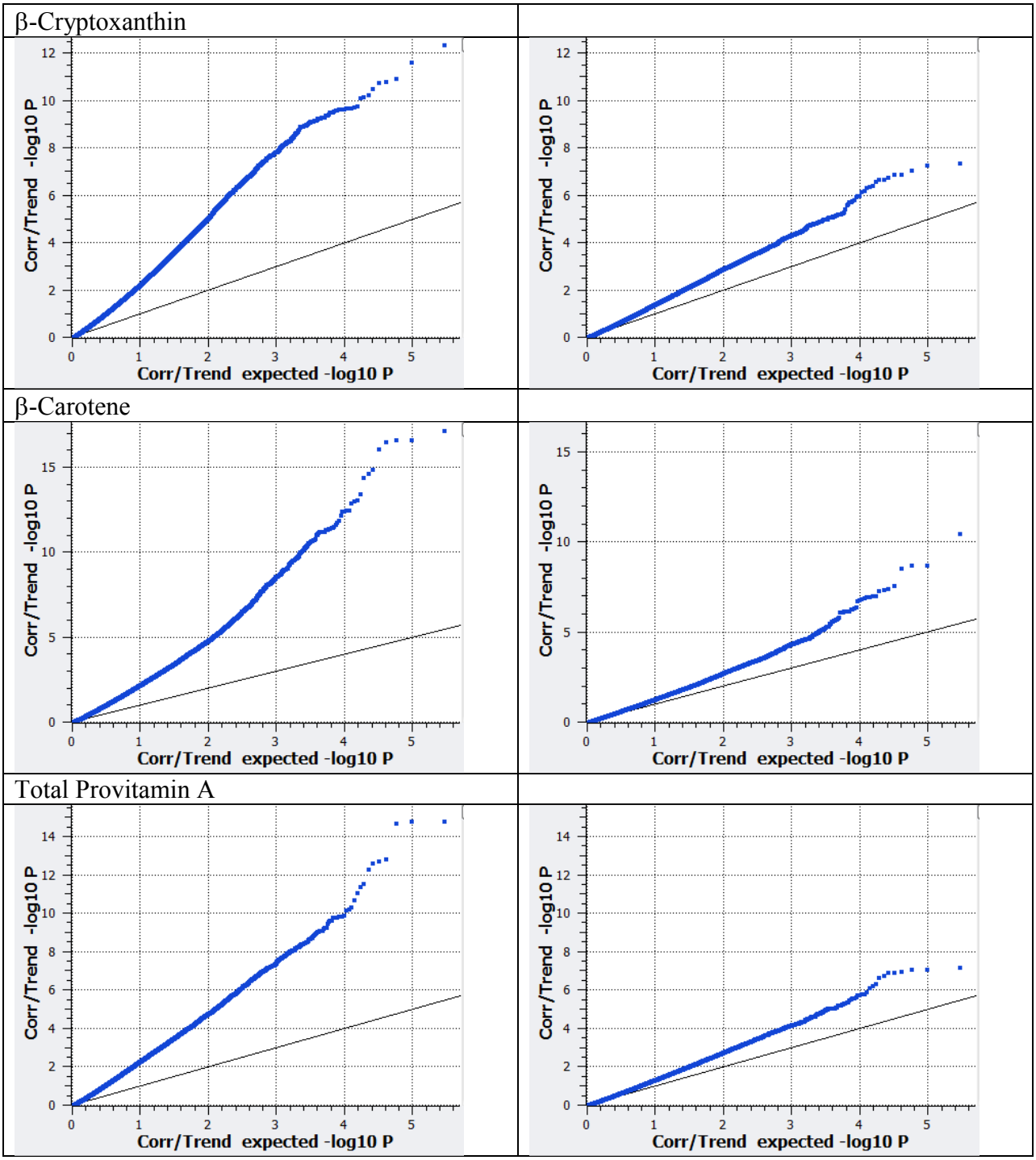


Figure III-5. Plot of observed versus expected -$\log_{10}$(P-values) plots for lutein, zeaxanthin, lutein:zeaxanthin ratio, β-cryptoxanthin, β-carotene, and total provitamin A concentrations evaluating two association mapping models in the 55K+GBS combined data sets. G = genotype, Q = ten principal components. All phenotypic values (y) were transformed to ln(y+1) prior to analyses.

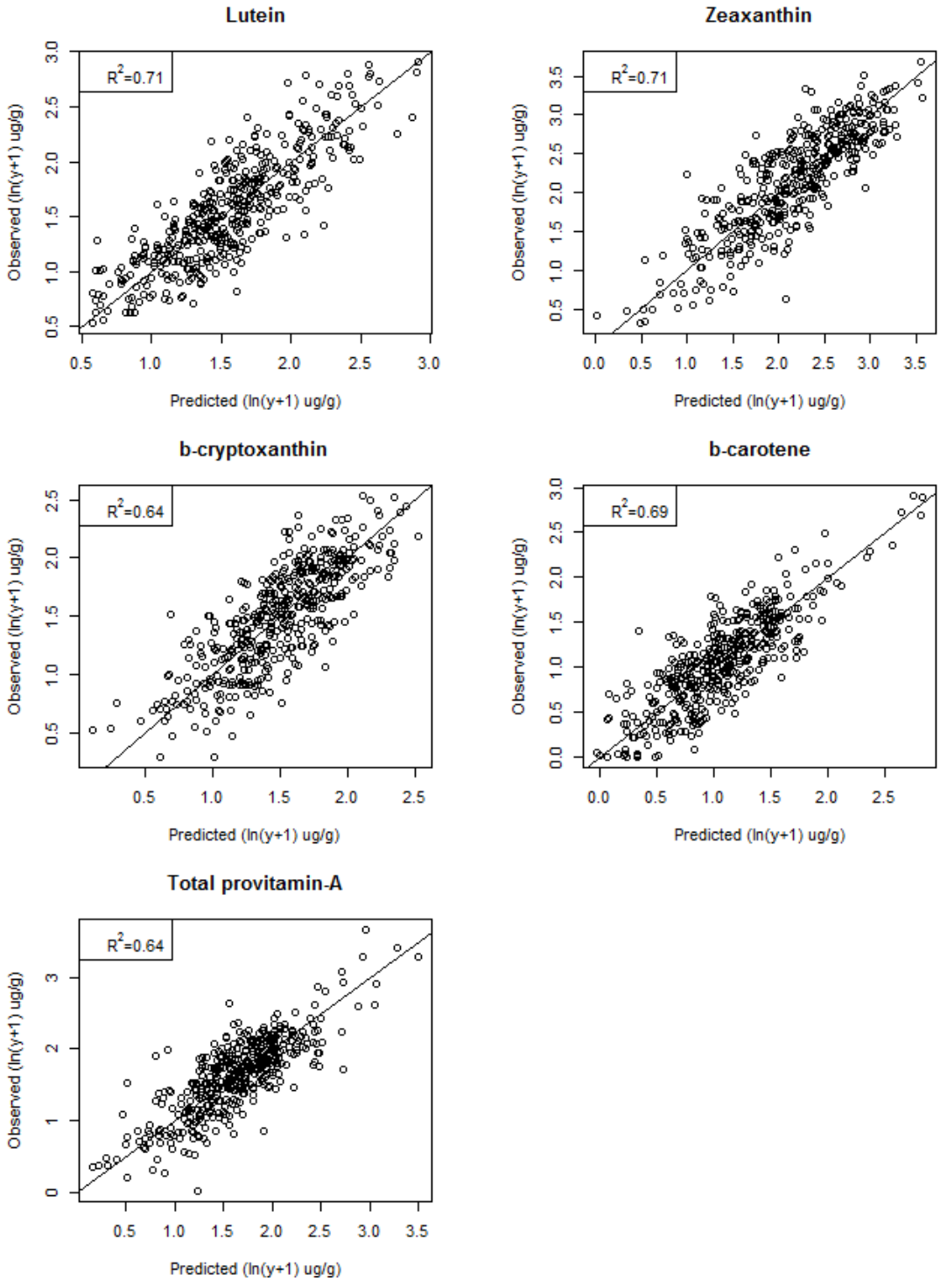
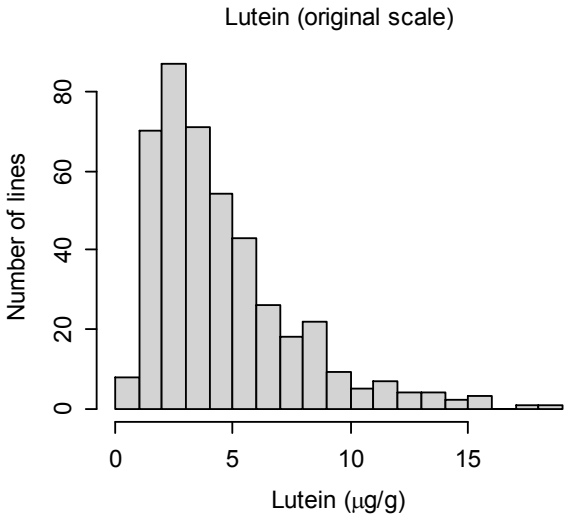
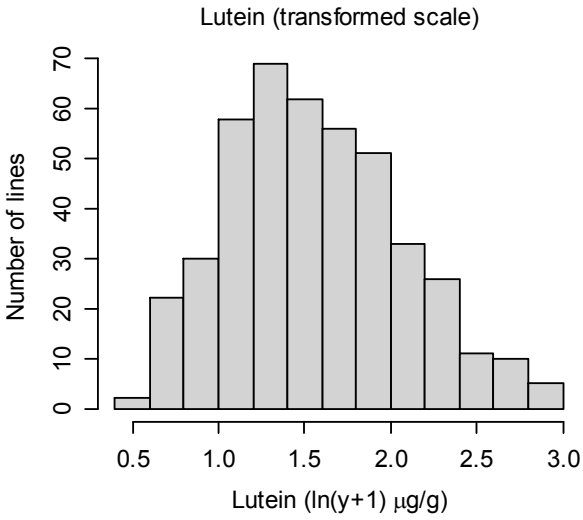
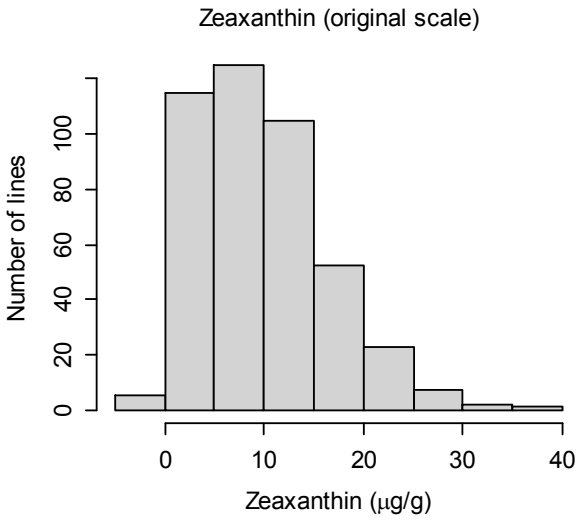
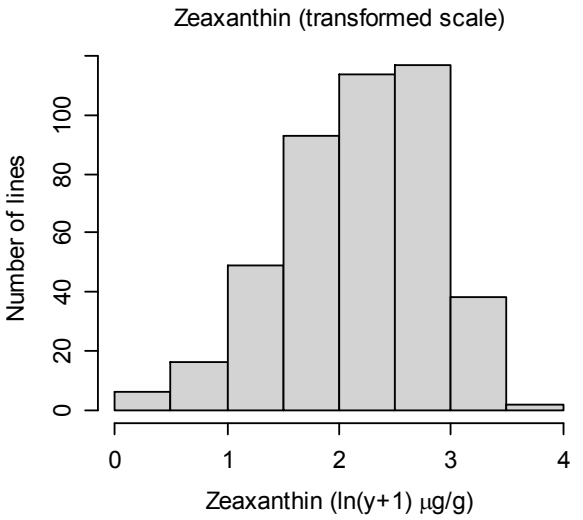


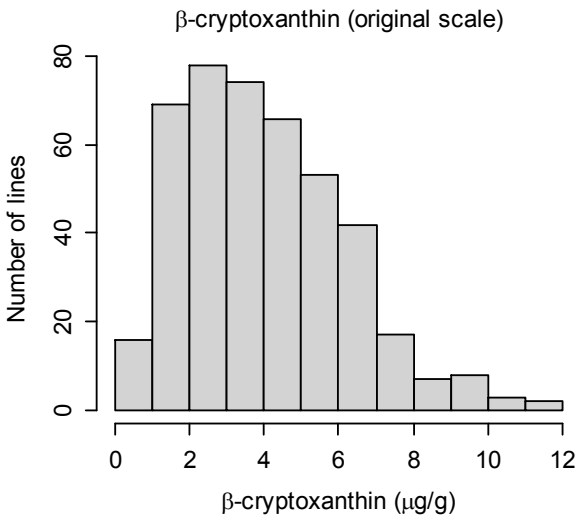
Figure III-6. Scatter plots of observed versus 10-fold cross-validation predicted values (GEBV) from the additive linear models in ln(y+1) scale of lutein, zeaxanthin, β-cryptoanthin, β-carotene, and total provitamin A carotenoids concentrations.

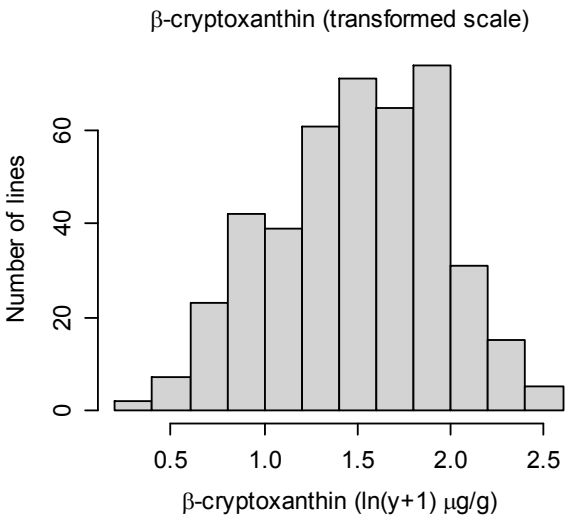

Lutein (original scale)
Number of lines
0
20
40
60
80
0
5
10
15
Lutein (μg/g)

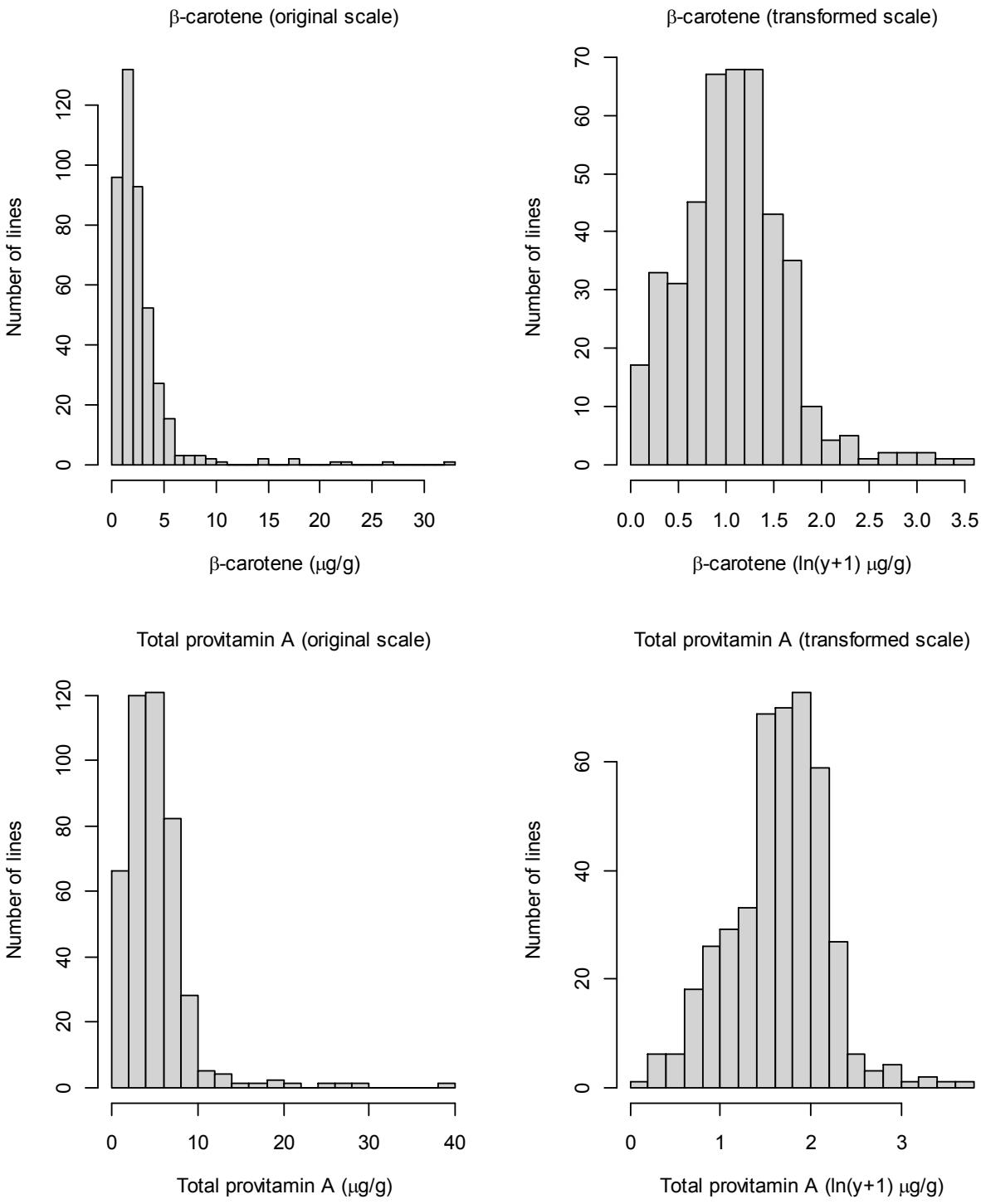

Lutein (transformed scale)
Number of lines
0
10
20
30
40
50
60
70
0.5
1.0
1.5
2.0
2.5
3.0
Lutein (ln(y+1) μg/g)


Zeaxanthin (original scale)
Number of lines
0
20
40
60
80
100
0
10
20
30
40
Zeaxanthin (μg/g)


Zeaxanthin (transformed scale)
Number of lines
0
20
40
60
80
100
0
1
2
3
4
Zeaxanthin (ln(y+1) μg/g)

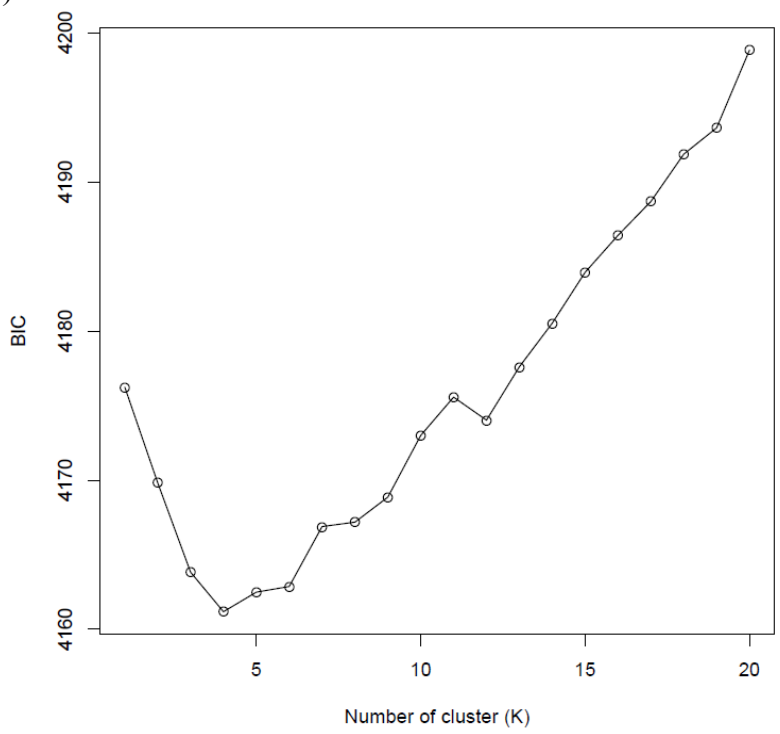

β-cryptoxanthin (original scale)
Number of lines
0
20
40
60
80
0
2
4
6
8
10
12
β-cryptoxanthin (μg/g)


β-cryptoxanthin (transformed scale)
Number of lines
0
20
40
60
0.5
1.0
1.5
2.0
2.5
β-cryptoxanthin (ln(y+1) μg/g)

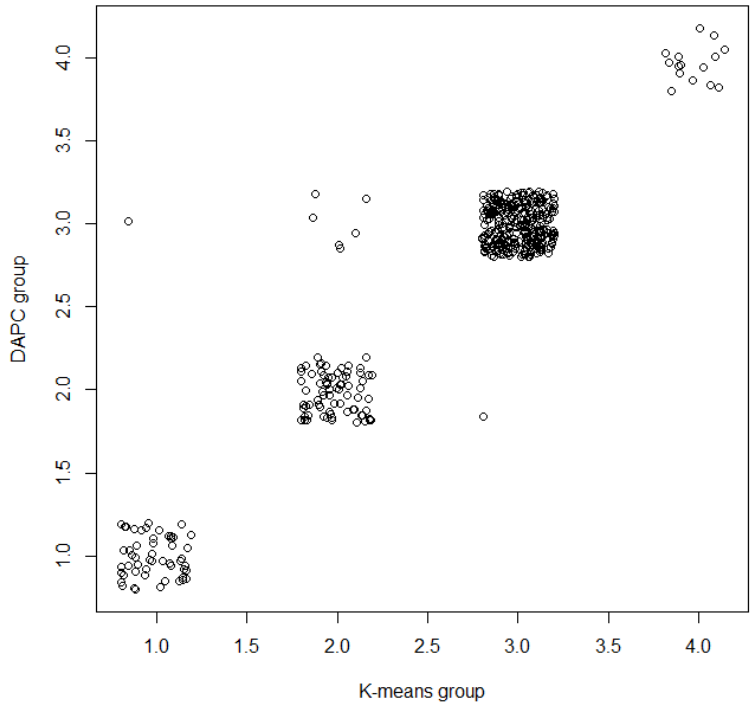


Supplemental Figure III-1. Distributions of phenotypic values (y) of each trait in original scale ($\mu g\ g^{-1}$) and after ln(y+1) transformation. Some negative values were due to the REML estimation and should be interpreted as zero.

(A)



(B)



Supplemental Figure III-2. (A) K-means clustering model comparisons using Bayesian Information Criterion (BIC) values for 1 to 20 clusters and (B) plot of line membership in the DAPC groups versus the K-means groups (B) from the analyses using the GBS data set.

CHAPTER IV

MOLECULAR DIVERSITY AND GENOMIC ASSESSMENT OF FOUR CIMMYT ASSOCIATION MAPPING PANELS BASED ON HIGH DENSITY SNP DATA

ABSTRACT

Genome-wide association studies have been extensively used for mapping simple traits and dissecting complex traits in maize. In preparation for using CIMMYT's association mapping panels to study complex traits, this research tested and validated association mapping models for these panels using simple traits for which the controlling genes have been identified. The objectives of this study were to (1) assess the ability of four CIMMYT association mapping panels to identify SNP markers associated with grain color and QPM phenotypes, (2) understand the macro population structure in each of the panels and its influence on association results based on high density marker data, and (3) identify specific markers that may contribute significantly to the population structure differentiation. The association mapping panels identified the *psy* gene on chromosome 6 controlling grain color and the *o2* gene on chromosome 7 responsible for lysine and tryptophan content, thereby validating the appropriateness of linear model with principal components as covariates in association mapping analyses and the potential for these panels to identify allelic diversity for additional, more complex traits. Several SNP markers contributed importantly to population structure, and a few of them located in close proximity to previously identified genes. Moderate to large population structure (F_{ST}>0.05) within each panel confirmed the need to control for population structure in subsequent association mapping studies.

Keywords: population structure, association mapping, grain color, quality protein maize

INTRODUCTION

Maize is an important cereal in the world and accounts for 30-50% of total caloric intake for many people in sub-Saharan Africa (Smale et al., 2011). While agriculture accounts for 20 to 40% of its gross domestic product, there is a large yield gap between potential and achieved maize yields in this region (Godfray et al., 2010), partly because of frequent droughts and pervasive low-input farming practices. Additionally, many people in this region suffer from malnutrition, including widespread vitamin A deficiency (WHO, 2009; Nuss and Tanumihardjo, 2010). The development and commercialization of high-yielding maize varieties, tolerant to drought, adapted to low input environments such as possessing improved nitrogen use efficiency, and having increased levels of provitamin A carotenoid concentrations in the grain can be a valuable component of an integrated strategy to overcome these problems.

The International Maize and Wheat Improvement Center (CIMMYT) has been working for more than 40 years to enhance global food security and alleviate poverty and hidden hunger worldwide through developing and disseminating sustainable agricultural technologies, including improved varieties. Several research projects targeting African environments are ongoing, including biofortification of maize with provitamin A (in collaboration with HarvestPlus), the drought tolerant maize for Africa (DTMA) initiative, and the improved maize for African soils (IMAS) project (www.cimmyt.org). Conventional breeding methods, including molecular approaches, have been used complementarily and extensively for developing hybrids and open pollinated maize varieties. While some traits of interest are controlled by few genes, such as b-carotene (Harjes et al., 2008; Yan et al., 2010; Babu et al., 2012), most are complex, including yield, drought and low-soil-nitrogen tolerance.

Association mapping has been extensively used to dissect complex traits in maize (Yan et al., 2011), complementing the linkage mapping approach. While strong relatedness structure effects are generally not an issue in human populations (Rosenberg et al., 2002), they are a major challenge for association mapping in maize where population structure and complex familial relationship often exists in mapping populations (Yu and Buckler, 2006). Failure to control population structure might result in false positive associations that confound true associations (Pritchard et al., 2000), complicating the identification of functional polymorphisms controlling traits of interest. Some models have been developed to correct for population structure in mapping populations, including the linear model which includes population structure information in the model, and the mixed linear model which includes both population and familial structure (Yu et al., 2006).

CIMMYT has developed four association mapping panels to identify polymorphisms associated with carotenoid concentration in grain, drought tolerance, nitrogen use efficiency, and protein quality; these panels are known as the carotenoid association mapping (CAM), DTMA, IMAS, and quality protein maize (QPM) panels, respectively. In preparation for using these association mapping panels to study complex traits (e.g. drought and low-N tolerance), this research tested and validated association mapping models for these panels using simple traits for which the controlling genes have been identified. The objectives of this study were to (1) assess the ability of four CIMMYT association mapping panels to identify SNPs associated with grain color and QPM phenotypes, (2) understand the macro population structure in each of the panels and its influence on association results based on high density marker data, and (3) identify specific markers that may contribute significantly to the population structure differentiation.

MATERIAL AND METHODS

Source of Germplasm and Genotype Data

CIMMYT's carotenoid association mapping, CAM (436 lines), drought tolerant maize for Africa, DTMA (277 lines), improved maize for African soils, IMAS (376 lines), and quality protein maize, QPM (248 lines) association mapping panels were used for this research. The four panels were combined to form a meta panel consisting of 1,337 lines.

Two single nucleotide polymorphism (SNP) genotyping platforms, the 55,000 (55K) and genotyping-by-sequencing (GBS), were used to genotype the lines. A combined genotype data set consisting 455,086 SNP markers was formed by including 40,033 SNPs from the 55K and 415,053 SNPs from GBS with call rate (proportion of non-missing genotypes at a marker) and minor allele frequency of at least 0.3 and 0.01, respectively. The 55K data were imputed before combining, but the GBS data were not. The 55K genotyping used the MaizeSNP50 Genotyping BeadChip from Illumina (catalog is available at www.illumina.com) and was done at Syngenta facility, Slater, IA, and the GBS genotyping was conducted at the Institute for Genomic Diversity, Cornell University, Ithaca, NY.

Population Structure

Analysis of population structure was performed using the K-means clustering method followed by discriminant analysis of principal components (DAPC) (Jombart et al., 2010). The DAPC used synthetic variables (principal components) with large variance among groups and small variance within groups, which is useful to elucidate relative distance between groups. Additionally, the DAPC was used to identify SNPs with different allele frequencies among groups, which may be associated with traits that contribute to population structure. The

identified SNPs were then used as queries against the Maize Sequence database (www.maizesequence.org). The DAPC analyses for the DTMA, IMAS, CAM, and QPM panels were conducted using 40,033 SNPs from the 55K platform to reduce computing time. Additionally, the principal component analysis (PCA) was performed on the meta panel (1337 lines) using the combined genotype data (455,086 SNPs) and the results were annotated with the pedigree group of the lines.

The genetic differentiations among the DAPC groups in each panel were measured using the fixation index (F statistic comparing genetic diversity among subpopulation relative to total population, F_{ST}), which can range from 0 to 1. An F_{ST} range of 0.00 – 0.05 suggests little genetic differentiation, 0.05 – 0.15 indicates moderate, 0.15 – 0.25 indicates large, and above 0.25 suggests very large genetic diversity among subpopulations (Wright, 1978). The F_{ST} values were computed using the SVS software (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com). Euclidean distances among the centers of DAPC groups and average Euclidean distances among individuals within DAPC groups in each panel were calculated using R software (R Core Team, 2012). Correlation analyses between F_{ST} and Euclidean distances among groups in each panel, and between average and standard deviation of within-group distances and group sizes combining all panels, were also performed using the R software.

Association Mapping

To assess the performance of the four CIMMYT association mapping panels for identifying SNPs associated with simple traits, association mapping analyses were performed for grain color (yellow versus white) and protein quality (presence versus absence of functional opaque-2 along with associated modifiers that influence kernel hardness and tryptophan content)

using the correlation/trend test (Weir, 2008). These analyses were performed for all panels except for grain color in the CAM panel, because most of the lines in this panel are yellow; however, this panel was included in the combined meta panel analysis for grain color. Number of lines in each phenotype class is shown in Table IV-1. Both grain color and QPM traits were considered to have binary phenotypes, with code of zero for white-grain and non-QPM lines, and code one assigned to yellow-grain and QPM lines.

The meta panel of 455,086 SNPs was used for association mapping. The average call rate ranged from 0.74 to 0.88 for the CAM, DTMA, and IMAS panels (Table IV-2). For the QPM panel, 79% of its lines (196 of 248) were genotyped in the 55K platform but not in the GBS, and therefore their GBS genotype data were treated as missing. Fifty-two lines in the QPM panel that were genotyped in both platforms have an average call rate of 0.90 (Table IV-2).

The analyses were performed using the SVS software (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com). Two association mapping models were used:

$$Y = SNP*\beta + \varepsilon \text{ (Model 1)}$$

$$Y = SNP*\beta + PC*\alpha + \varepsilon \text{ (Model 2)}$$

where Y = response variable (least square means of carotenoids phenotype), SNP = SNP marker, and PC = principal component coordinate from the PCA. The twenty most-significant SNPs (Bonferroni-adjusted P-value < 0.001) based on Model 2 were selected from each association mapping panel for grain color and QPM phenotypes. These significant SNPs were put in the 500kb windows, and the most significant SNPs in each window were kept. The SNPs were then sorted by their physical position to identify significant SNPs that occur in more than one panel.

RESULTS

Population Structure

The K-means clustering grouped lines into four clusters for the CAM, DTMA, and QPM panels, and five clusters for the IMAS panel. The Bayesian information criterion (BIC)-based model selection suggested that the combined meta panel has 16 clusters; however, for better understanding and interpretation, the grouping of this panel was simplified into five clusters. This number of clusters was determined based on visual observation of the principal components plot, and the lowest BIC values for number of groups fewer than ten (data not shown). Original clusters from the meta panel (16 clusters) did not correspond with clusters from individual panels (r=0.24, Supplemental Figure IV-1).

The average minor allele frequency across 455,036 SNPs in all panels ranged from 0.14 to 0.15 (Table IV-2 and Figure IV-1), indicating that most SNPs had sufficient minor allele frequency for association mapping (> 0.05) (Abdurakhmonov and Abdukarimov, 2008). On average across all markers, 18% of the lines possessed at least one minor allele, and the QPM panel had the highest average (25%). Additionally, the average rate of heterozygosity was relatively low for all panels (0.01-0.04), as expected for inbred line populations (Yang et al., 2011).

Strong population structures were observed among some groups of lines in the four panels and in the combined meta panel, but group sizes within each panel were quite diverse. In the CAM panel, group 3 was genetically very different from group 4 (F_{ST}=0.33) and group 1 (F_{ST}=0.32) (Table IV-3). There was also a very strong genetic differentiation among groups 1 and 2 in the DTMA (F_{ST}=0.38), 2 and 3 in the IMAS (F_{ST}=0.33), and 1 and 3 in the QPM panels

(F_{ST}=0.34). In the meta panel, although the differentiation among groups was not as large as in the individual panels, there were large genetic differences between groups 1 and 3 (F_{ST}=0.19), 2 and 3 (F_{ST}=0.18) and 1 and 2 (F_{ST}=0.16) (Table IV-3).

Average Euclidean distances among centers of DAPC groups were significantly higher than average distances among individuals within each group for all panels ($P < 0.01$). Lines' memberships in K-means groups corresponded closely with that in DAPC groups, as indicated by correlation coefficients greater than 0.9 for all panels (Table IV-4). Moreover, the DAPC plots (Figure IV-2 to IV-6) generally represented the F_{ST} measures very well; for example, in the CAM panel, groups 1-3 and 1-4 (F_{ST}=0.33 and 0.32), were depicted further apart than groups 1-2 (F_{ST}=0.09) (Figure IV-2). Other examples include groups 1 and 2 in the DTMA, 2 and 3 in the IMAS, and 1 and 3 in the QPM panels, all of which had F_{ST}>0.30 and were plotted with larger distances among them, than between them and other groups (Figure IV-3 to IV-5). Large and significant correlation coefficients (r=0.83-0.95, P<0.05) between F_{ST} and Euclidean distances among centers of DAPC groups in all association mapping panels support the finding that these two measurements of genetic diversity among groups are associated (Table IV-5).

The grouping of lines based on PCA and DAPC corresponded with known pedigree relatedness, although these associations were imperfect and not quantifiable (Figure IV-7, Figure IV-8 and Supplemental Table IV-1). Lines derived from La Posta Sequia, a group of drought tolerant lines, corresponded mainly with DAPC group 1, South African lines with group 2, Mexico and Colombia lines with group 4, and provitamin A and Thailand lines with group 5 (Figure IV-8). Kenya and Zimbabwe lines mostly clustered in DAPC group 4 along with the Mexico and Colombia lines, reflecting the fact that these pedigree groups are genetically similar

with each other; indeed, the Kenya, Zimbabwe and Colombia lines in these panels have mostly been developed by CIMMYT breeders over 2-3 decades, relying heavily initially on Mexican germplasm.

The DAPC plots for each panel help visualize the separation between groups along discriminant analysis (DA) axis 1 (the X axis) or DA axis 2 (the Y axis). Markers contributing to separation of groups were then identified, and found that most of the SNPs with largest contribution (coefficients or loadings) to DA axes are located inside or nearby genes annotated as putative uncharacterized proteins (Table IV-6). However, some interesting genes were revealed contributing to population structure, such as plastocyanin (chromosome 6; close to the PZE-106022533 marker) for the CAM panel, acetolactate synthase (chromosome 4, by the PZE-104068430) and remorin (chromosome 1, near the SYN27560) for the DTMA panel, and acyltransferase putative uncharacterized protein (chromosome 1, in close proximity to PZE-101188060) for the QPM panel.

Association Mapping

Three SNPs on chromosome 6 (PZE-106028491, PZE-106033993, and PZE-106035079) and four SNPs on chromosome 8 (PZE-108040088, PZE-108041027, PZE-108041477, SYN3892) were highly significantly associated with grain color in more than one of the DTMA, IMAS, QPM, and meta association mapping panels (Table IV-7 and Figure IV-9). Five additional SNPs (PZE-106033961, S6_79469507, SYN2634, S6_82015505, and S6_82019628) were also significantly associated with yellow grain color in one of the panels. The S6_82019628 marker (chromosome 6: 82,019,628) is located inside the phytoene synthase (*psy*) gene (chromosome 6: 82,017,148-82,021,007), explained a relatively large proportion (27%) of

yellow versus white grain color variation (Table IV-7). Larger phenotypic variance (73-76% in three panels) was explained by the PZE-106035079 marker (chromosome 6: 80,875,894). For most SNPs in the DTMA, IMAS, and QPM panels, the difference between number of lines having homozygous-predominant-allele (A_1A_1) and those with homozygous-rare-allele (A_2A_2) was positive for the white phenotype and negative for the yellow (Table IV-7 and Figure IV-10), reflecting the fact that the panels contain more white than yellow lines. Type I error rate was reduced in Model 2 (with population structure information) relative to Model 1 (without population structure information), as indicated visually in the Q-Q plots (Figure IV-11).

For the QPM binary phenotype, five SNPs on chromosome 7 (SYNGENTA6482, SYN36579, S7_10550478, SYN6900, S7_11335555) were highly significantly associated in more than one panel and explained 26-73% of QPM versus non-QPM variation (Table IV-8 and Figure IV-12). Two among these, S7_10550478 (chromosome 7: 10,550,478) and SYN6900 (chromosome 7: 10,582,734) are located in close proximity to the opaque-2 (*o2*) gene (chromosome 7: 10,793,452-10,796,233) (Table IV-8). The difference between number of lines having homozygous-predominant-allele (A_1A_1) and lines with homozygous-rare-allele (A_2A_2) was positive for non-QPM lines and negative for QPM lines at the majority of SNPs (Table IV-8 and Figure IV-13), reflecting the fact that most of the lines in these panels are not QPM. In agreement with the results from grain color association mapping, the probability-probability (P-P) plots (Figure IV-14) visually indicate that using principal components as covariates in the linear model could reduce false positives relative to the model with SNP markers only.

DISCUSSION

The four association mapping panels and the combined meta panel successfully identified the phytoene synthase (*psy*) gene on chromosome 6 controlling yellow versus white grain color. This gene follows Mendelian inheritance, with yellow endosperm resulting from the dominant allele (*Y1*) and white (lack of color) endosperm resulting from the recessive allele (*y1*) (Egesel et al., 2003). All panels were also able to identify the simply-inherited *o2* gene on chromosome 7, for which the recessive allele (*o2*) results in increased lysine and tryptophan – two essential amino acids – content in maize endosperm (Crow and Kermicle, 2002; Atlin et al., 2010). The ability of all association mapping panels to identify simple traits suggests that the models and methods are appropriate for near-future association studies for more complex traits.

The power of association mapping relies on diversity of lines for traits of interest, population size, and marker density and quality (Yu et al., 2008). For simple traits such as grain color, however, association mapping was able to identify the controlling gene despite having a large number of lines with missing phenotypic and genotype data such as in the QPM panel (23% missing phenotype, 75% missing genotype) (see Table IV-1 and Materials and Methods). The power to detect the *psy* locus of the QPM panel was similar to that of the DTMA panel, as indicated visually in their Manhattan plots (Figure IV-12), even though the QPM panel had fewer lines having phenotypic data (198 lines) and more unbalanced yellow:white phenotype ratio (0.22) than the DTMA panel (274 lines and 0.49, respectively). One explanation for this similarity is that both panels shared in common a few significant SNP markers in the *psy* gene region with similar effect pattern, that is, similar difference in number of lines with A_1A_1 and A_2A_2 genotypes at the markers. For the most significant marker in both panels (PZE-10603579)

from the simple model analyses, for instance, the difference between these homozygous classes in the DTMA panel was -62 and 147 for yellow and white phenotypes, while in the QPM panel it was -36 and 140, resulting in high marker significance for the DTMA ($-\log_{10}$(Bonferroni-adjusted P)=12.1) and the QPM (26.3) panels.

In the association mapping study for the QPM phenotype, most non-QPM lines had the predominant allele (A_1) at most significant SNPs across four panels and the combined meta panel (Figure IV-13), reflecting the fact that non-QPM lines outnumbered QPM lines in these panels. The predominant allele was also possessed by most white lines at most significant markers for grain color mapping across the DTMA, IMAS and QPM panels (Figure IV-10).

Genetic diversity among groups in each association mapping panel generally ranged from moderate to large (F_{ST}>0.05, Table IV-3). The discriminant analysis of principal component (DAPC) analysis provided satisfactory results for clustering the lines based on genotypic data, as indicated by its agreement with fixation index (F_{ST}) measures and its ability to reveal a few known genes contributing to population structure. However, because these association mapping panels consist of diverse sets of inbred lines and such panels typically have complex familial relationships (Yu and Buckler, 2006), large-effect polymorphisms affecting the separation of groups were not identified as, for example, was reported by (Jombart et al., 2010) for mutations of seasonal influenza virus strains. Furthermore, some association between SNPs-based grouping and pedigree relatedness were found, suggesting that high-density SNP markers could be useful in grouping diverse tropical maize germplasm. Incidentally, no obvious separation of lines in the meta panel according to their classification by breeders into CIMMYT heterotic

groups was found (result not shown), supporting previous findings that greater genetic diversity exists within than among CIMMYT heterotic groups (Wen et al., 2011).

The DAPC group sizes were unbalanced in all panels, with the largest group in each panel consisting of 44-77% of all individuals in the respective group (Supplemental Table IV-2). Combining all panels, there was a significant negative correlation coefficient between group sizes and average Euclidean distances among individuals within groups (r=-0.43, P=0.04, n=22) and between group sizes and within-group standard deviations (r=-0.48, P=0.02, n=22). These findings indicated that the average and variance of distances among individuals in the large groups were smaller than for small groups, meaning that even though some groups were large, the individuals within them could be genetically similar to each other.

Among the genes contributing to population structure in the QPM panel, the acyltransferase putative gene on chromosome 1 was hypothesized to have a related function with one of the existing opaque genes (*o2*, *fl2*, and *o7*) or with a putative new gene affecting high lysine phenotype. In support of this hypothesis, the *o7* gene was recently reported as an acyl-CoA synthetase-like (*ACS*) gene on chromosome 10, with its recessive allele (*o7-ref*), which results in opaque phenotype, having a 12-bp deletion in the second exon of *ACS* (Miclaus et al., 2011). Furthermore, one of the SNPs contributing to differentiation of group 2 in the DTMA panel is located near the acetolactate synthase (*ALS*) gene, which is expressed in maize embryo and endosperm and is responsible for producing the first enzymes for synthesizing the amino acids leucine, valine, and isoleucines.

CONCLUSIONS

The CIMMYT association mapping panels successfully identified genes controlling grain color (*psy* gene on chromosome 6) and QPM phenotypes (*o2* gene on chromosome 7), thereby validating the appropriateness of Model 2 (linear model with principal components as covariates) in association mapping analyses and the potential for these panels to identify allelic diversity for additional, more complex traits. Several SNP markers contributed importantly to population structure, and a few of them located in close proximity to previously identified genes. Moderate to large population structure (F_{ST}>0.05) within each panel confirmed the need to control for population structure in subsequent association mapping studies and suggested that the K-means clustering followed by discriminant analysis of principal components can be used to identify groups of germplasm.

REFERENCES

Abdurakhmonov, I.Y., and A. Abdukarimov. 2008. Application of association mapping to understanding the genetic diversity of plant germplasm resources. International journal of plant genomics 2008: 574927. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2423417&tool=pmcentrez&rendertype=abstract (verified 9 August 2012).

Atlin, G.N., N. Palacios, R. Babu, B. Das, S. Twumasi-Afriyie, D.K. Friesen, H. De Groote, B. Vivek, and K.V. Pixley. 2010. Quality Protein Maize: Progress and Prospects. *In* Plant Breeding Reviews.

Babu, R., N.P. Rojas, S. Gao, J. Yan, and K. Pixley. 2012. Validation of the effects of molecular marker polymorphisms in LcyE and CrtRB1 on provitamin A concentrations for 26 tropical maize populations. Theoretical and Applied Genetics. Available at http://www.springerlink.com/index/10.1007/s00122-012-1987-3 (verified 12 October 2012).

Crow, J.F., and J. Kermicle. 2002. Oliver Nelson and Quality Protein Maize. 821(March): 819–821.

Egesel, C.O., J.C. Wong, R.J. Lambert, and T.R. Rocheford. 2003. Gene dosage effects on carotenoid concentration in maize grain. Maydica 48: 183–190.

Godfray, H.C.J., J.R. Beddington, I.R. Crute, L. Haddad, D. Lawrence, J.F. Muir, J. Pretty, S. Robinson, S.M. Thomas, and C. Toulmin. 2010. Food security: the challenge of feeding 9 billion people. Science (New York, N.Y.) 327(5967): 812–8. Available at http://www.ncbi.nlm.nih.gov/pubmed/20110467 (verified 5 October 2012).

Harjes, C.E., T.R. Rocheford, L. Bai, T.P. Brutnell, C.B. Kandianis, S.G. Sowinski, A.E. Stapleton, R. Vallabhaneni, M. Williams, E.T. Wurtzel, J. Yan, and E.S. Buckler. 2008. Natural genetic variation in lycopene epsilon-cyclase tapped for maize biofortification. Science (New York, N.Y.) 319(5861): 330–3. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2933658&tool=pmcentrez&rendertype=abstract (verified 1 August 2011).

Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC genetics 11(1): 94. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2973851&tool=pmcentrez&rendertype=abstract (verified 13 July 2012).

Miclaus, M., Y. Wu, J.-H. Xu, H.K. Dooner, and J. Messing. 2011. The maize high-lysine mutant opaque7 is defective in an acyl-CoA synthetase-like protein. Genetics 189(4): 1271–80. Available at http://www.ncbi.nlm.nih.gov/pubmed/21926304 (verified 29 October 2012).

Nuss, E.T., and S. a. Tanumihardjo. 2010. Maize: A Paramount Staple Crop in the Context of Global Nutrition. Comprehensive Reviews in Food Science and Food Safety 9(4): 417–436. Available at http://doi.wiley.com/10.1111/j.1541-4337.2010.00117.x (verified 16 July 2012).

Pritchard, J.K., M. Stephens, N. a Rosenberg, and P. Donnelly. 2000. Association mapping in structured populations. American journal of human genetics 67(1): 170–81. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1287075&tool=pmcentrez&rendertype=abstract.

Rosenberg, N. a, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L. a Zhivotovsky, and M.W. Feldman. 2002. Genetic structure of human populations. Science (New York, N.Y.) 298(5602): 2381–5. Available at http://www.ncbi.nlm.nih.gov/pubmed/12493913 (verified 8 October 2012).

Smale, M., D. Byerlee, and T. Jayne. 2011. Maize Revolutions in Sub-Saharan Africa. Policy Research Working Paper 5659. The World Bank.

R Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Weir, B.S. 2008. Linkage disequilibrium and association mapping. Annual review of genomics and human genetics 9: 129–42. Available at http://www.ncbi.nlm.nih.gov/pubmed/18505378 (verified 4 August 2012).

WHO. 2009. Global prevalence of vitamin A deficiency in populations at risk 1995–2005: WHO Global Database on Vitamin A Deficiency. DatabaseAvailable at http://www.who.int/nutrition/publications/micronutrients/.

Wen, W., J.L. Araus, T. Shah, J. Cairns, G. Mahuku, M. Bänziger, J.L. Torres, C. Sánchez, and J. Yan. 2011. Molecular Characterization of a Diverse Maize Inbred Line Collection and its Potential Utilization for Stress Tolerance Improvement. Crop Science 51(6): 2569. Available at https://www.crops.org/publications/cs/abstracts/51/6/2569 (verified 5 March 2012).

Wright, S. 1978. Evolution and the Genetics of Population, Volume 4. The University of Chicago Press, Chicago.

Yan, J., C.B. Kandianis, C.E. Harjes, L. Bai, E.-H. Kim, X. Yang, D.J. Skinner, Z. Fu, S. Mitchell, Q. Li, M.G.S. Fernandez, M. Zaharieva, R. Babu, Y. Fu, N. Palacios, J. Li, D. Dellapenna, T. Brutnell, E.S. Buckler, M.L. Warburton, and T. Rocheford. 2010. Rare genetic variation at Zea mays crtRB1 increases beta-carotene in maize grain. Nature genetics 42(4): 322–7. Available at http://www.ncbi.nlm.nih.gov/pubmed/20305664 (verified 28 July 2011).

Yan, J., M. Warburton, and J. Crouch. 2011. Association Mapping for Enhancing Maize (L.) Genetic Improvement. Crop Science 51(2): 433. Available at https://www.crops.org/publications/cs/abstracts/51/2/433 (verified 18 July 2012).

Yang, X., S. Gao, S. Xu, Z. Zhang, B.M. Prasanna, L. Li, J. Li, and J. Yan. 2011. Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. Molecular Breeding 28(4): 511–526. Available at http://www.springerlink.com/index/10.1007/s11032-010-9500-7 (verified 12 November 2012).

Yu, J., and E.S. Buckler. 2006. Genetic association mapping and genome organization of maize. Current opinion in biotechnology 17(2): 155–60. Available at http://www.ncbi.nlm.nih.gov/pubmed/16504497 (verified 17 July 2012).

Yu, J., J.B. Holland, M.D. McMullen, and E.S. Buckler. 2008. Genetic design and statistical power of nested association mapping in maize. Genetics 178(1): 539–51. Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2206100&tool=pmcentrez&rendertype=abstract (verified 12 July 2012).

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38(2): 203–8. Available at http://www.ncbi.nlm.nih.gov/pubmed/16380716 (verified 12 July 2012).

Table IV-1. Number of lines in each class for grain color and quality protein maize (QPM) phenotypes for four association mapping panels and the combined meta panel.

Panel[#]	Grain color			QPM		
	Yellow	White	Missing	QPM	Non-QPM	Missing
CAM	419	15	2	19	417	0
DTMA	90	184	3	10	267	0
IMAS	79	290	7	17	359	0
QPM	36	162	50	122	104	22
Meta	624	651	62	168	1147	22

[#] CAM=carotenoid association mapping, DTMA=drought tolerant maize for Africa, IMAS= improved maize for African soils.

Table IV-2. Average minor allele frequency (MAF), call rate, heterozygosity rate, and fraction of genotypes with minor allele for the combined 455,086 SNPs in four association mapping panels and the meta panel.

Panel[#]	MAF	Call rate	Heterozygosity rate	Fraction with minor allele[##]
CAM	0.14	0.78	0.02	0.16
DTMA	0.14	0.88	0.01	0.14
IMAS	0.15	0.74	0.04	0.17
QPM	0.14	0.90 ^	0.04	0.25
Meta	0.15	0.69	0.03	0.18

[#] CAM=carotenoid association mapping, DTMA=drought tolerant maize for Africa, IMAS= improved maize for African soils.
[##] Proportion of genotypes having at least one minor allele relative to total number of genotypes in the respective panel.
^ Calculated from 52 lines which were genotyped in both 55K and GBS; the remaining 196 lines were genotyped in 55K but not in GBS.

Table IV-3. Pairwise fixation index (F_{ST}) and Euclidean distances among centers of DAPC groups (below and above diagonal, respectively), and average Euclidean distances among individuals within groups (diagonal) in the four association mapping panels.

CAM

Group	1	2	3	4
1	2.49	5.38	15.57	8.13
2	0.09	1.35	14.24	6.74
3	0.32	0.20	3.29	14.49
4	0.22	0.11	0.33	3.38

DTMA

Group	1	2	3	4
1	2.93	13.94	12.69	6.54
2	0.38	3.67	15.64	11.99
3	0.20	0.39	1.93	10.83
4	0.07	0.24	0.11	1.57

IMAS

Group	1	2	3	4	5
1	2.63	13.10	13.34	7.32	9.04
2	0.16	3.44	18.57	14.65	16.33
3	0.19	0.33	4.31	10.78	10.91
4	0.08	0.18	0.12	2.72	4.89
5	0.11	0.20	0.12	0.03	1.55

QPM

Group	1	2	3	4
1	3.38	14.58	16.87	16.09
2	0.17	1.59	8.14	6.08
3	0.34	0.13	2.55	9.88
4	0.28	0.08	0.24	2.48

Meta

Group	1	2	3	4	5
1	2.60	14.64	15.28	11.64	13.08
2	0.16	2.30	13.08	8.30	9.93
3	0.19	0.18	4.32	10.08	11.30
4	0.08	0.07	0.09	2.15	5.61
5	0.12	0.10	0.13	0.03	2.49

Table IV-4. Pearson correlation coefficients (r) between line memberships in the K-means grouping and that in the DAPC grouping.

Panel	r	P-value	n
CAM	0.97	< 0.01	436
DTMA	0.92	< 0.01	277
IMAS	1.00	< 0.01	376
QPM	0.98	< 0.01	248
Meta	0.99	< 0.01	1337

n=number of pairs of groups.

Table IV-5. Pearson correlation coefficients (r) between fixation index (F_{ST}) and Euclidian distances among centers of DAPC groups in each association mapping panel.

Panel	r	P-value	n
CAM	0.86	0.03	6
DTMA	0.90	0.02	6
IMAS	0.95	< 0.01	10
QPM	0.83	0.04	6
Meta	0.91	< 0.01	10

n=number of pairs of groups.

Table IV-6. SNP markers contributing to population structure based on DAPC

SNP	Chr	Position	Nearby gene	Description	Loading[#] (10^{-4})
CAM (Axis 1)					
PZE-109054925	9	92,514,795	GRMZM2G173358	Transcribed locus, moderately similar to XP_002438749.1 hypothetical protein SORBIDRAFT_10g025475 [Sorghum bicolor]	2.5
PZE-107080376	7	129,794,525	AC208031.3_FG002	Putative uncharacterized protein	2.4
PZE-106022533	6	54,467,828	GRMZM2G071450	Plastocyanin	2.4
PZE-103053199	3	59,688,226	GRMZM2G118641	NA	2.4
PZE-104045526	4	75,026,384	GRMZM2G146518	NA	2.3
SYN19272	7	120,756,804	GRMZM2G099049	TSA: Zea mays contig51361, mRNA sequence	2.3
DTMA (Axis 2)					
PZE-104015488	4	14,819,684	GRMZM2G139621	NA	1.7
PZE-104068430	4	135,176,829	GRMZM2G407044	Acetolactate synthase/ amino acid binding protein	1.7
PZE-105077348	5	86,129,643	GRMZM2G082683	Putative uncharacterized protein	1.7
SYN19749	9	8,796,333	GRMZM2G142072	TSA: Zea mays contig14675, mRNA sequence	1.6
SYN537	5	2,799,579	GRMZM2G108677	Putative uncharacterized protein	1.6
SYN27560	1	290,807,970	GRMZM2G004511	Remorin	1.6
IMAS (Axis 1)					
PZE-110025098	10	39,823,394	GRMZM2G360615	Putative uncharacterized protein	3.4
PZE-110009748	10	7,484,892	GRMZM2G156506	Putative uncharacterized protein	3.0
PZE-104126595	4	210,331,391	GRMZM2G040720	Putative uncharacterized protein	2.8
PZE-103001192	3	1,620,622	GRMZM2G309152)	Putative uncharacterized protein	2.6
PZE-101016496	1	9,301,495	GRMZM5G823004	Putative uncharacterized protein	2.6
PZE-110022209	10	31,000,121	GRMZM2G117028)	Putative uncharacterized protein	2.6
Meta (Axis 1)					
PZE-109050317	9	84,557,567	GRMZM2G163641	TSA: Zea mays contig08153, mRNA sequence Source: UniGene Zm.22597	4.2
PZE-102111518	2	141,857,315	GRMZM2G130131	NA	4.0
PZE-109093862	9	135,732,763	GRMZM2G312481	NA	3.6
PZE-105094031	5	136,421,074	GRMZM2G346133)	Dynein light chain LC6, flagellar outer arm	3.5
PZE-106124630	6	164,936,617	GRMZM2G101515	Putative uncharacterized protein	3.5
PZA00587.4	10	84,518,416	GRMZM2G022793	TSA: Zea mays contig16448, mRNA sequence	3.4
QPM (Axis 1)					
PZE-101188060	1	232,746,246	GRMZM2G167438	Acyltransferase Putative uncharacterized protein	2.4
PZE-101196856	1	244,311,782	GRMZM2G320085	NA	2.2
SYN21695	6	133,484,372	GRMZM2G062333	Hypothetical protein	2.2
PZE-106075758	6	131,399,716	GRMZM2G126594	Putative uncharacterized protein	2.2
PZE-110066840	10	122,973,637	GRMZM2G361569	Plastid-specific 30S ribosomal protein 3 Putative uncharacterized protein	2.2
PZE-106062728	6	113,869,079	GRMZM2G165969	Putative uncharacterized protein	2.1

[#] Marker coefficient in the DAPC model

Table IV-7. List of highly significant SNPs for grain color phenotype in seven 500kb regions occurring in more than one association mapping panels.

Panel	Marker	Chr	Position	No. of nearby SNPs[#]	Bonferroni adjusted P-value	R^2	MAF	Effects[##]	
								Yellow lines	White lines
Meta	PZE-106028491	6	66,662,688	0	2.79E-07	0.28	0.17	-2	143
QPM	PZE-106028491	6	66,662,688	0	2.79E-07	0.28	0.12	-2	143
IMAS	PZE-106033961	6	78,310,770	9	1.57E-18	0.29	0.12	9	273
Meta	PZE-106033993	6	78,389,265	5	3.11E-10	0.35	0.26	-14	126
QPM	PZE-106033993	6	78,389,265	5	3.11E-10	0.35	0.20	-14	126
DTMA	S6_79469507	6	79,469,507	2	2.73E-04	0.15	0.14	21	166
IMAS	SYN2634	6	79,816,341	3	3.24E-16	0.26	0.11	10	273
Meta	PZE-106035079	6	80,875,894	4	5.31E-27	0.76	0.49	36	-140
DTMA	PZE-106035079	6	80,875,894	3	8.67E-13	0.29	0.35	-62	147
IMAS	PZE-106035079	6	80,875,894	5	7.69E-53	0.73	0.25	-70	254
QPM	PZE-106035079	6	80,875,894	4	5.31E-27	0.76	0.24	-36	140
DTMA	S6_82015505	6	82,015,505	3	2.24E-11	0.31	0.46	-78	97
IMAS	S6_82019628	6	82,019,628	0	6.32E-13	0.27	0.40	-55	117
Meta	PZE-108040088	8	63,389,630	2	5.99E-09	0.32	0.10	9	155
QPM	PZE-108040088	8	63,389,630	2	5.99E-09	0.32	0.09	9	155
Meta	PZE-108041027	8	64,951,311	0	1.38E-06	0.26	0.12	11	142
QPM	PZE-108041027	8	64,951,311	0	1.38E-06	0.26	0.11	11	142
Meta	PZE-108041477	8	65,998,702	2	6.64E-12	0.39	0.13	1	150
QPM	PZE-108041477	8	65,998,702	2	6.64E-12	0.39	0.12	1	150
Meta	SYN3892	8	139,128,233	0	6.80E-08	0.29	0.37	-20	107
QPM	SYN3892	8	139,128,233	0	6.80E-08	0.29	0.27	-20	107

MAF= minor allele frequency.

[#] Number of significant SNPs in the respective panel within 500kb region.

[##] The difference between numbers of homozygous-predominant-allele genotypes (A_1A_1) relative to homozygous-rare-allele genotypes (A_2A_2) at a particular locus for color binary phenotype: yellow (coded as 1) and white (0).

Table IV-8. List of highly significant SNPs for QPM phenotype in four 500kb regions occurring in more than one association mapping panels (all are on chromosome 7).

Panel	Marker	Position	No. of nearby SNPs[#]	Bonferroni adjusted P-value	R^2	MAF	Effects[##]	
							QPM lines	Non-QPM lines
CAM	SYNGENTA6482	8,587,447	0	1.82E-20	0.26	0.08	-10	375
QPM	SYNGENTA6482	8,587,447	2	1.50E-22	0.56	0.40	-63	102
Meta	SYNGENTA6482	8,587,447	0	5.60E-93	0.34	0.18	-76	928
CAM	SYN36579	9,216,124	0	1.03E-23	0.30	0.07	-9	385
IMAS	S7_9260251	9,260,251	0	2.29E-10	0.28	0.08	-7	216
QPM	SYN36579	9,216,124	2	1.33E-29	0.71	0.47	-92	98
Meta	SYN36579	9,216,124	2	6.66E-126	0.46	0.18	-108	963
CAM	SYN6900	10,582,734	0	8.56E-23	0.29	0.10	-15	364
DTMA	S7_10550478	10,550,478	0	1.90E-13	0.34	0.07	-6	220
IMAS	S7_10550478	10,550,478	0	1.78E-13	0.29	0.13	-12	224
QPM	SYN6900	10,582,734	5	2.22E-30	0.73	0.48	106	-84
Meta	SYN6900	10,582,734	3	3.65E-123	0.45	0.21	-132	901
DTMA	S7_11335555	11,335,555	0	6.07E-14	0.32	0.06	-5	238
IMAS	S7_11335555	11,335,555	3	6.32E-16	0.27	0.11	-11	281
QPM	SYN36048	11,356,756	3	1.58E-15	0.41	0.47	81	-58
DTMA	S7_13349992	13,349,992	2	2.84E-19	0.44	0.04	-3	234
IMAS	S7_13352214	13,352,214	8	2.07E-12	0.22	0.11	-10	293

MAF= minor allele frequency.

[#] Number of significant SNPs in the respective panel within 500kb region.

[##] The difference between numbers of homozygous-predominant-allele genotypes (A_1A_1) relative to homozygous-rare-allele genotypes (A_2A_2) at a particular locus for QPM binary phenotype: QPM line (coded as 1) and non-QPM line (0).

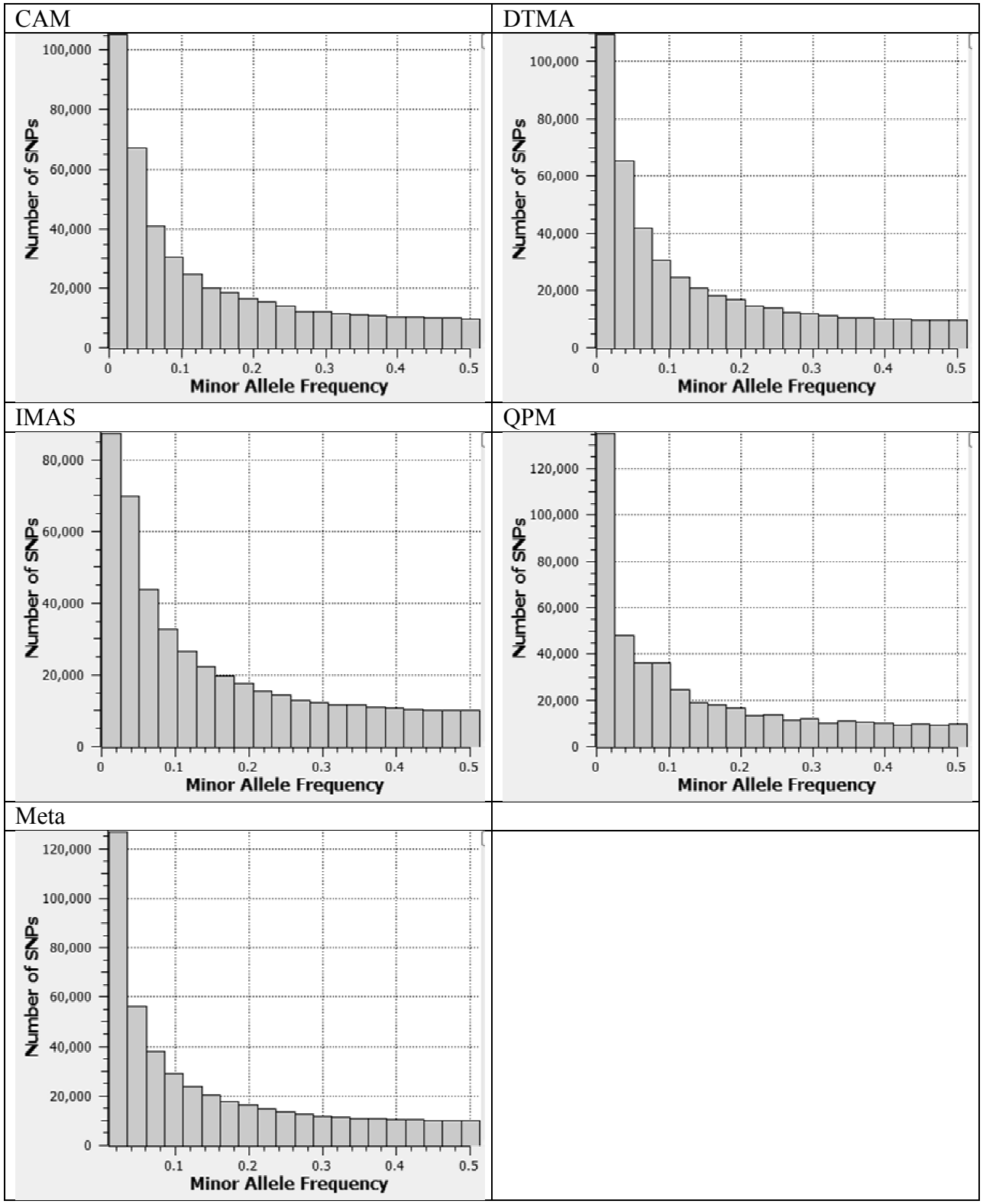


Figure IV-1. Distribution of minor allele frequency in five association mapping panels.

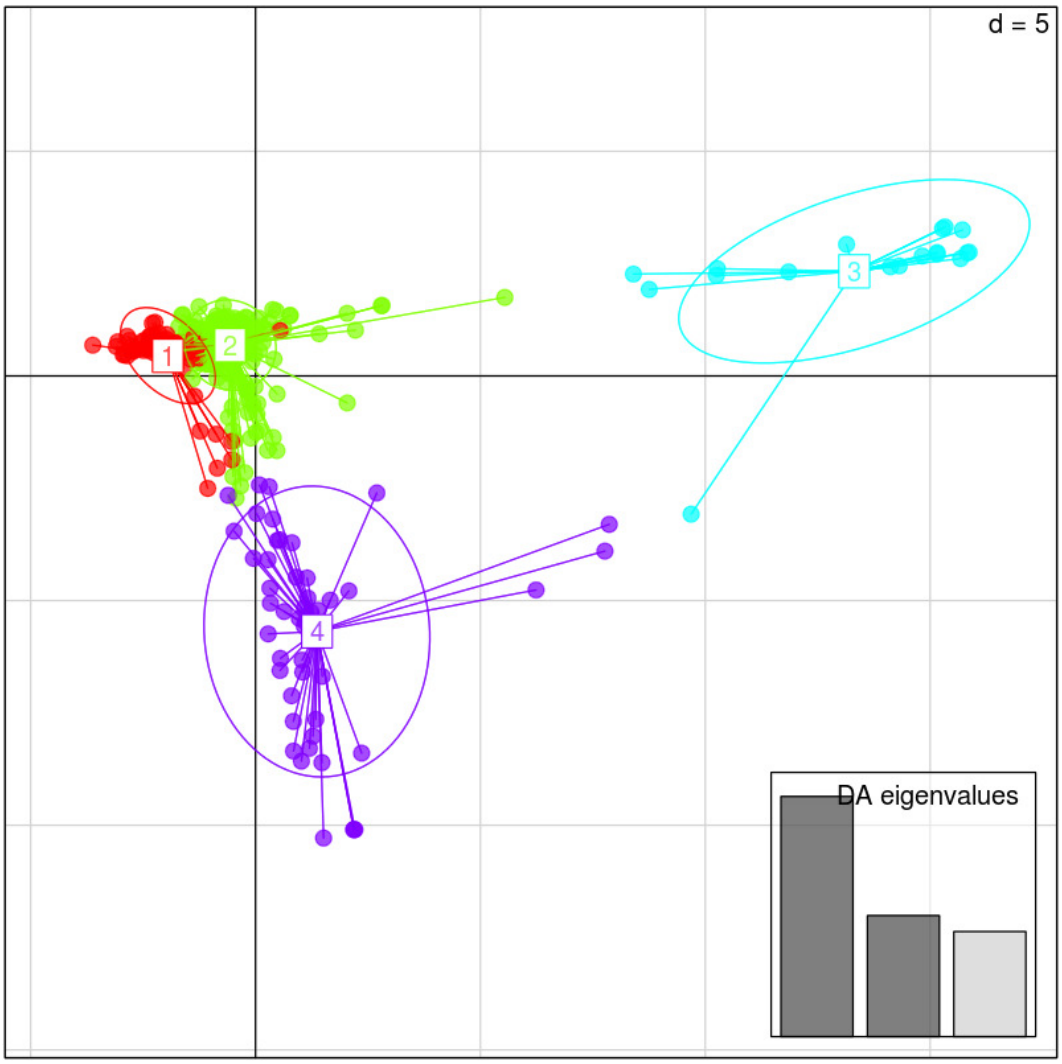


Figure IV-2. DAPC plot of the first and the second discriminant functions (X and Y axes, respectively) for the CAM panel.

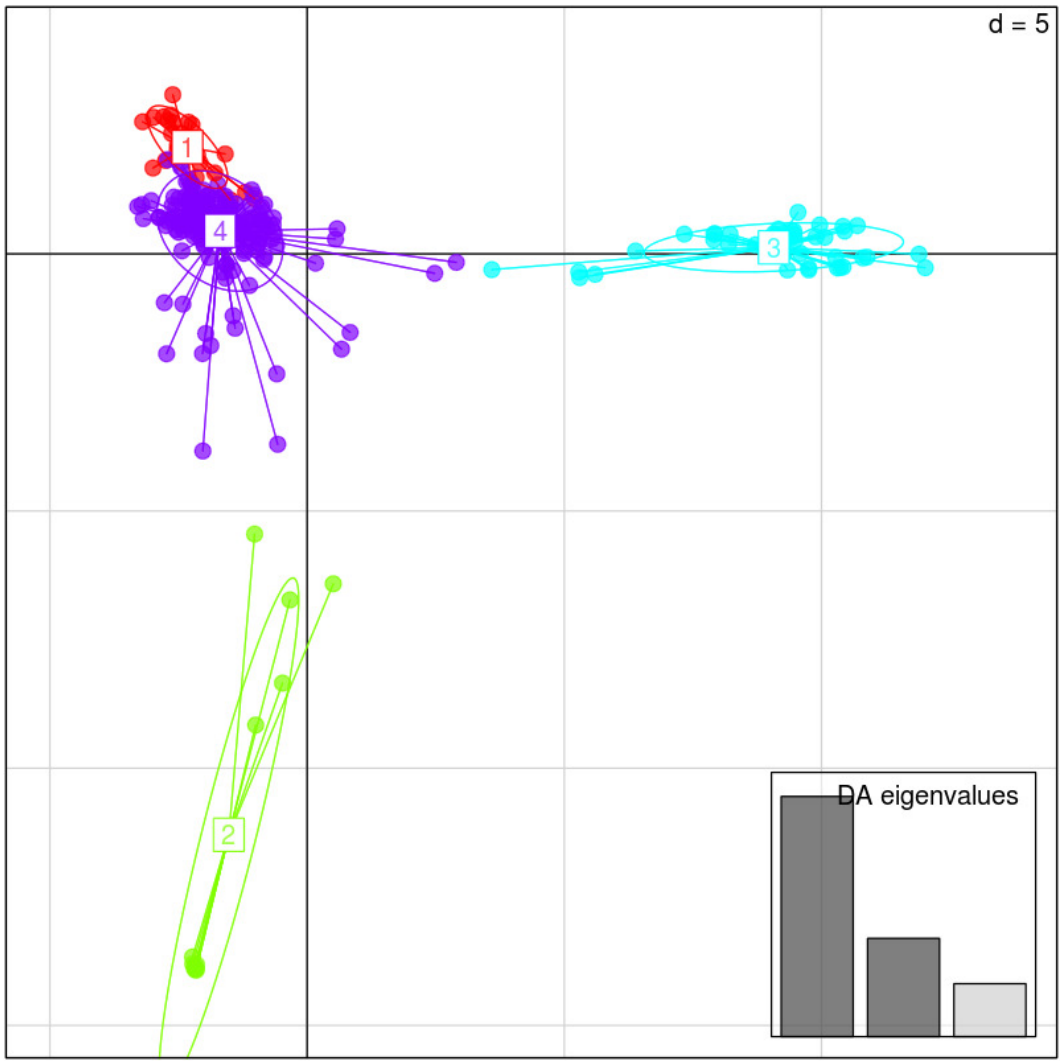


Figure IV-3. DAPC plot of the first and the second discriminant functions (X and Y axes, respectively) for the DTMA panel.

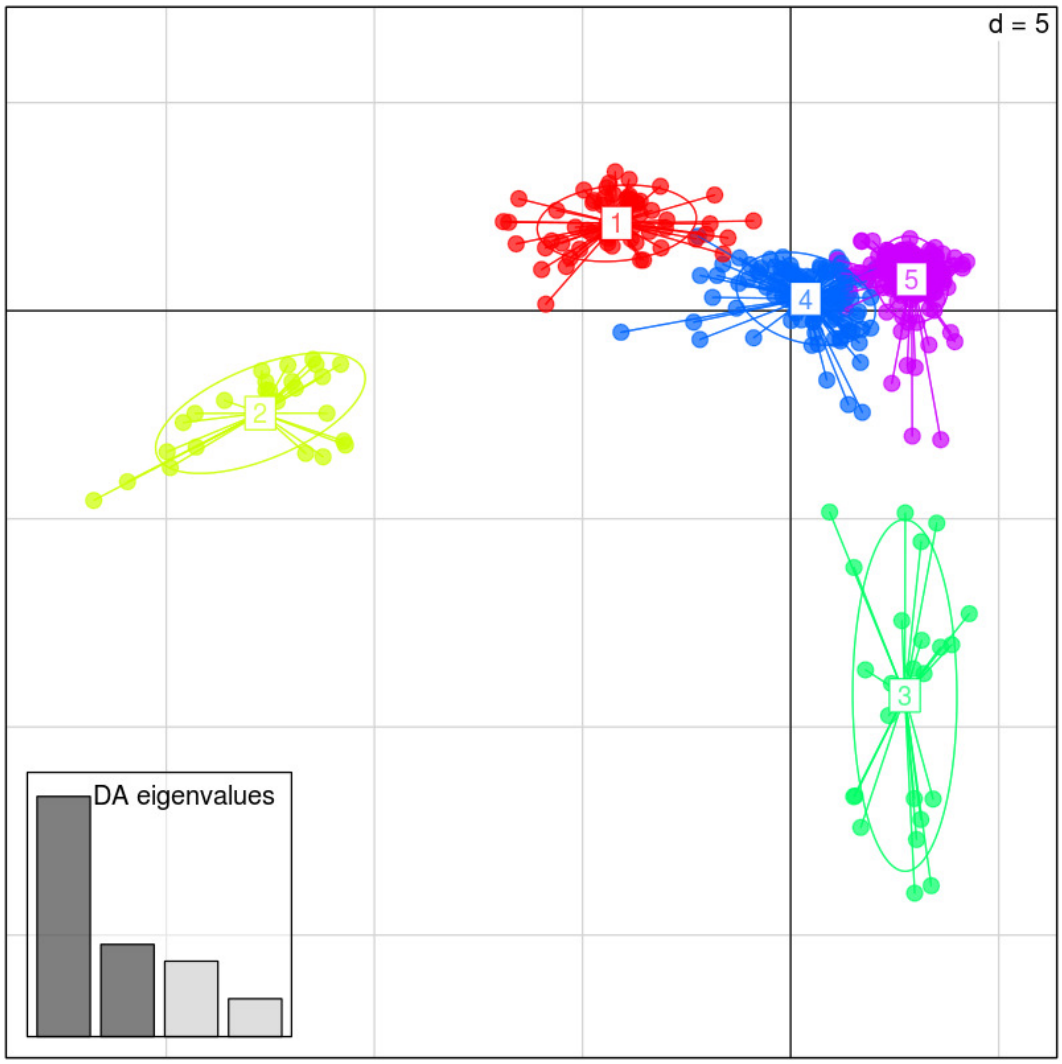


Figure IV-4. DAPC plot of the first and the second discriminant functions (X and Y axes, respectively) for the IMAS panel.

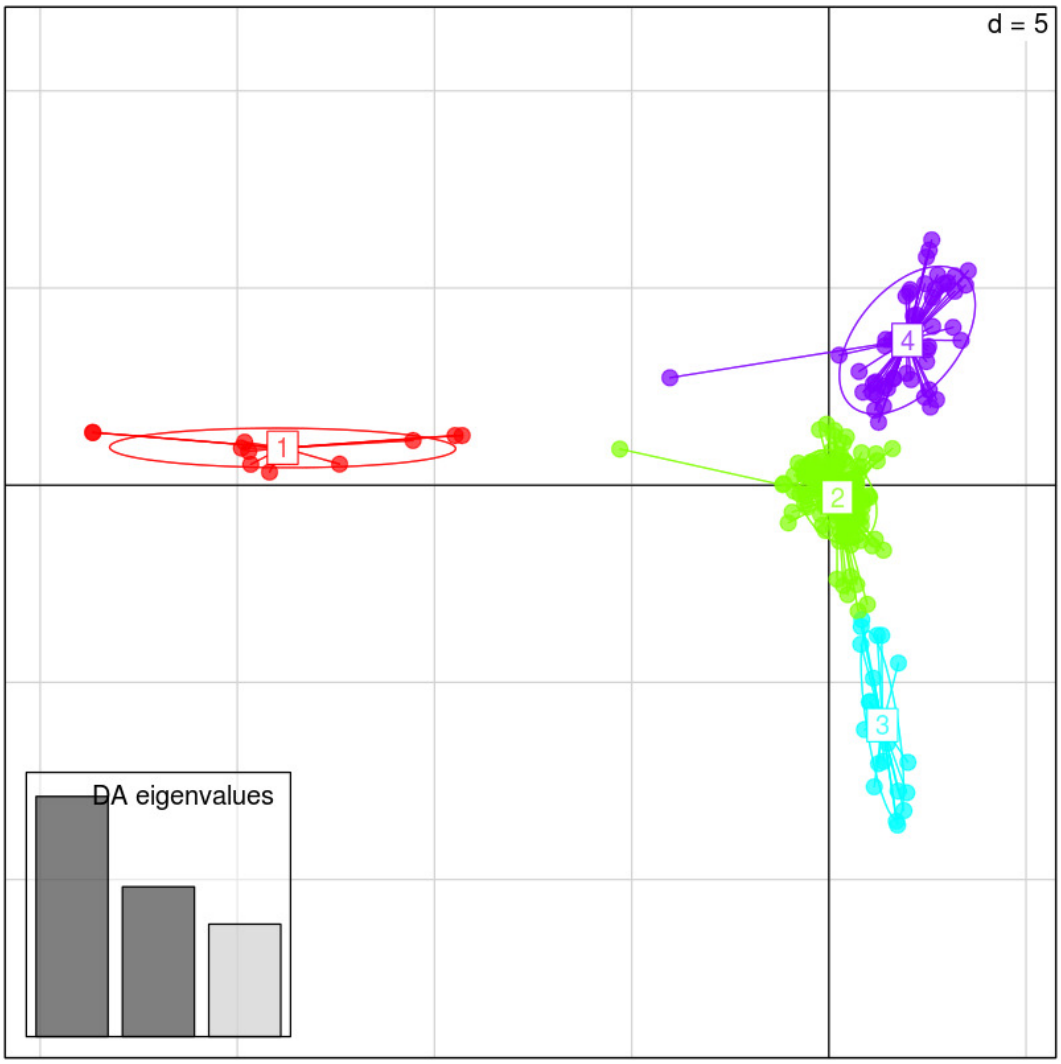


Figure IV-5. DAPC plot of the first and the second discriminant functions (X and Y axes, respectively) for the QPM panel.

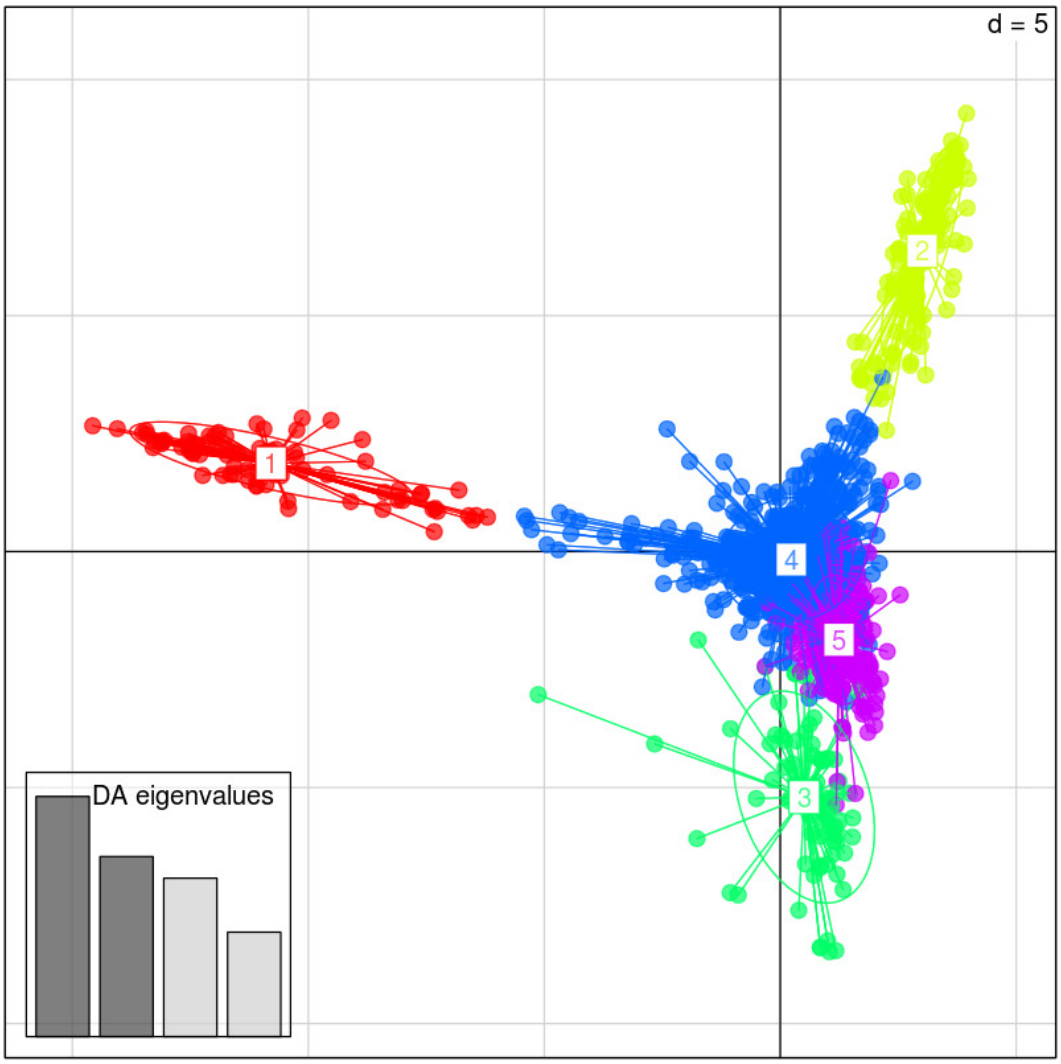
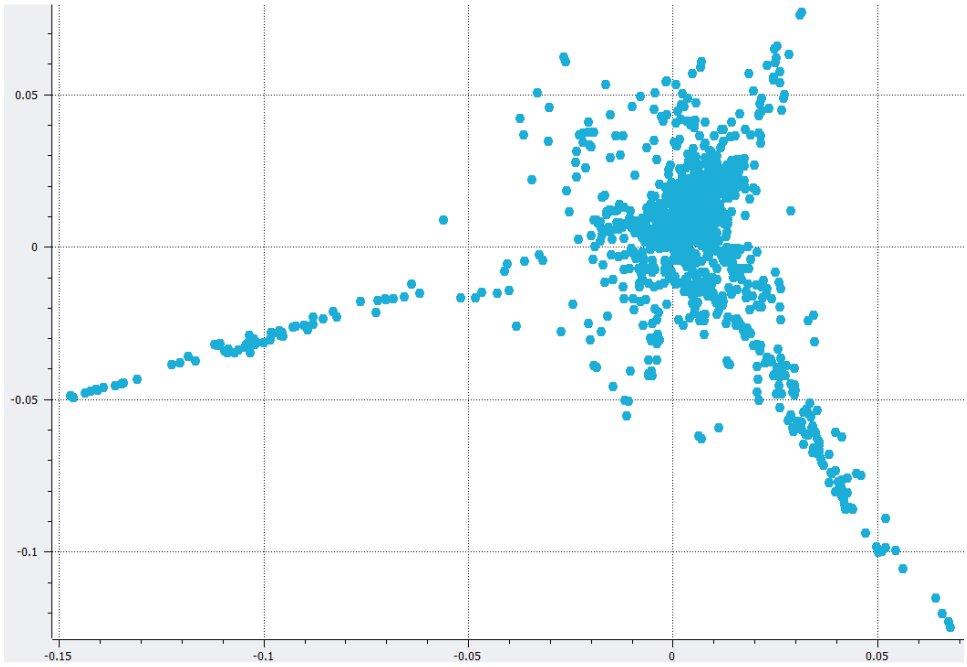


Figure IV-6. DAPC plot of the first and the second discriminant functions (X and Y axes, respectively) for the meta panel.

(A)



(B)

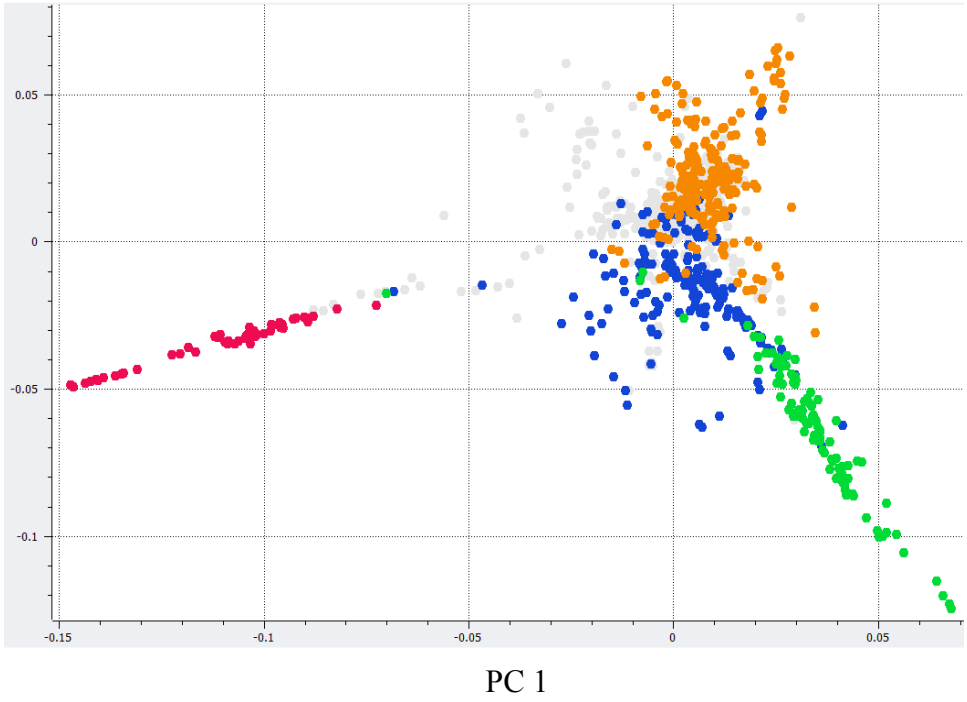


Figure IV-7. Plot of the first and the second PC from PCA analysis for the meta panel: (A) all lines, (B) differentiated by pedigree group: provitamin A and Thailand (orange), Kenya and Zimbabwe (blue), South Africa (green), La Posta Sequia (red), Mexico and Columbia (silver).

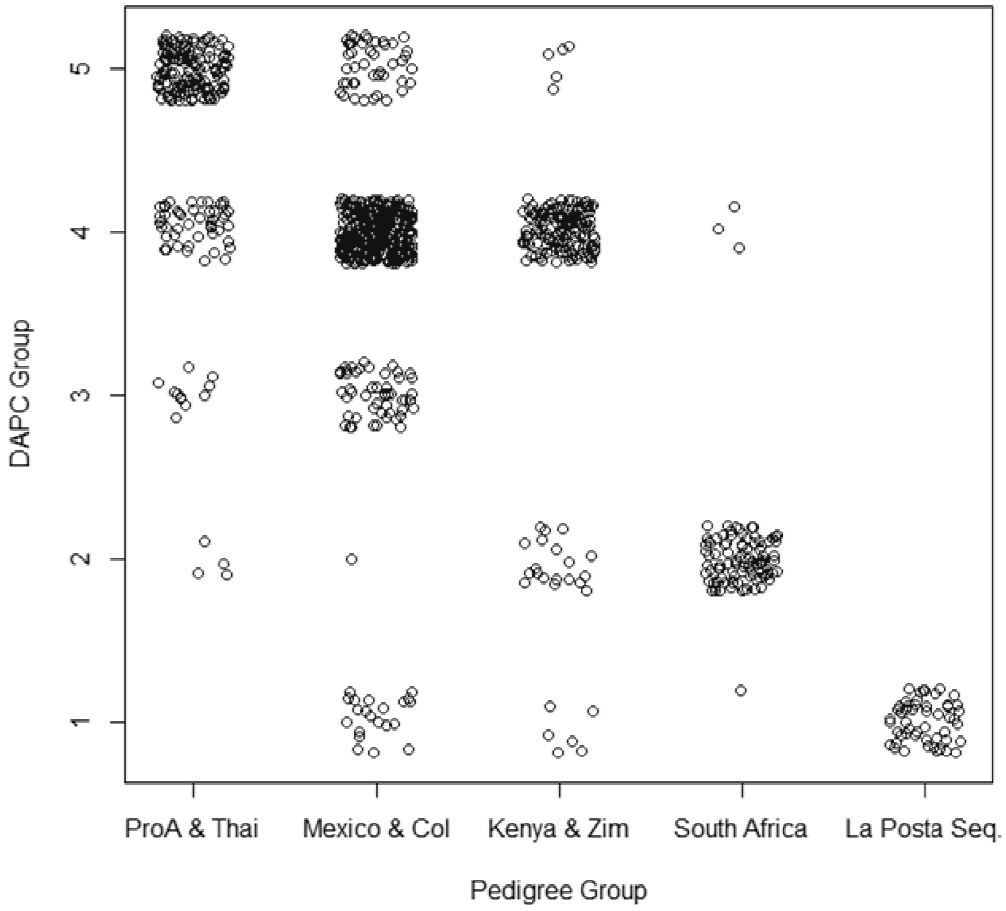
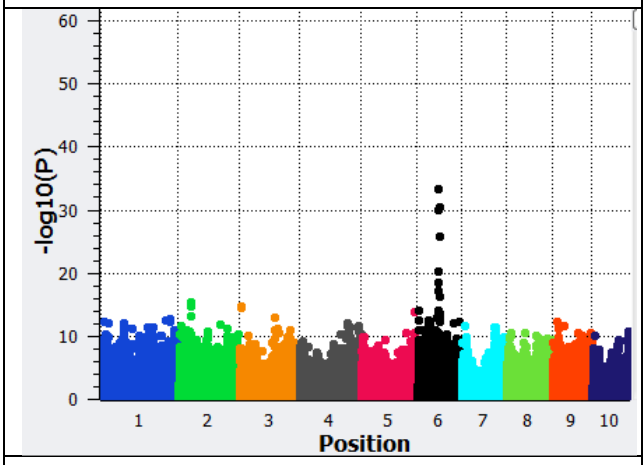
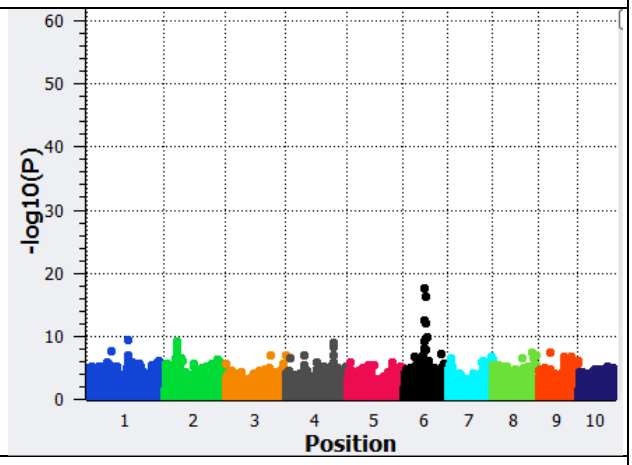
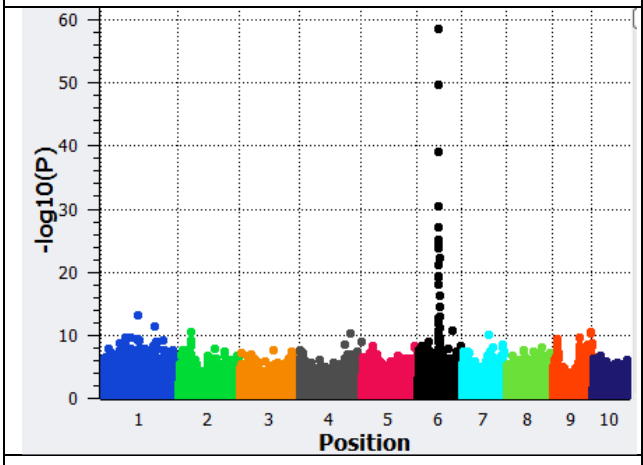
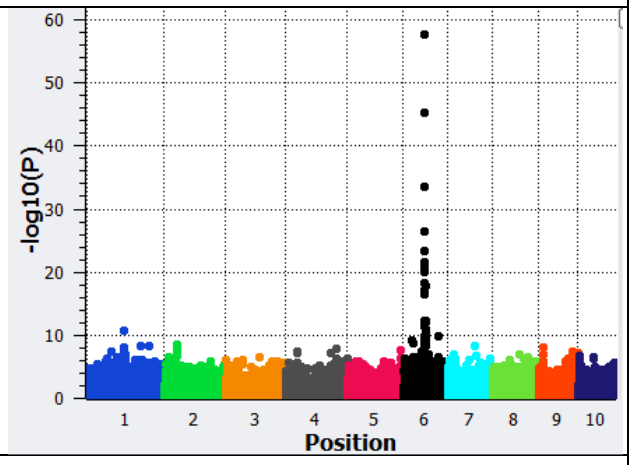
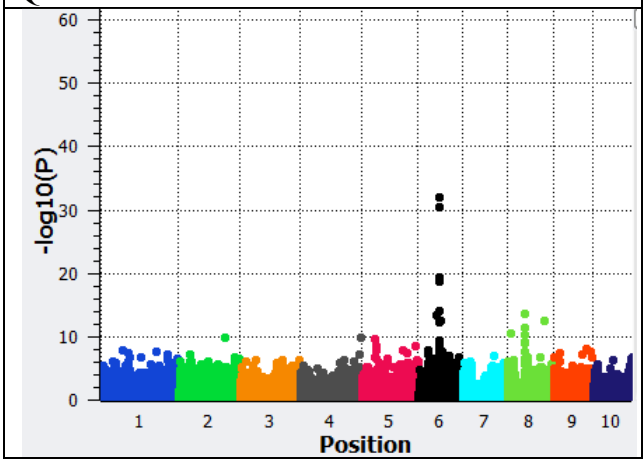
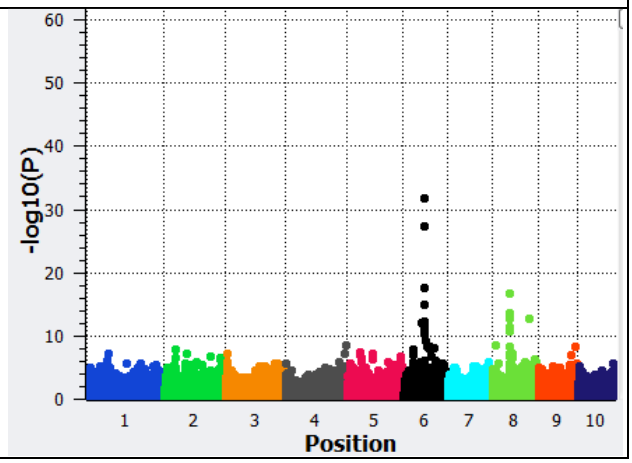


Figure IV-8. Relationship among pedigree groups and DAPC groups. ProA=provitamin A lines, Thai=Thailand, Col=Colombia, Zim=Zimbabwe.

Model 1 (without principal components)	Model 2 (with principal components)
DTMA	
	
IMAS	
	
QPM	
	

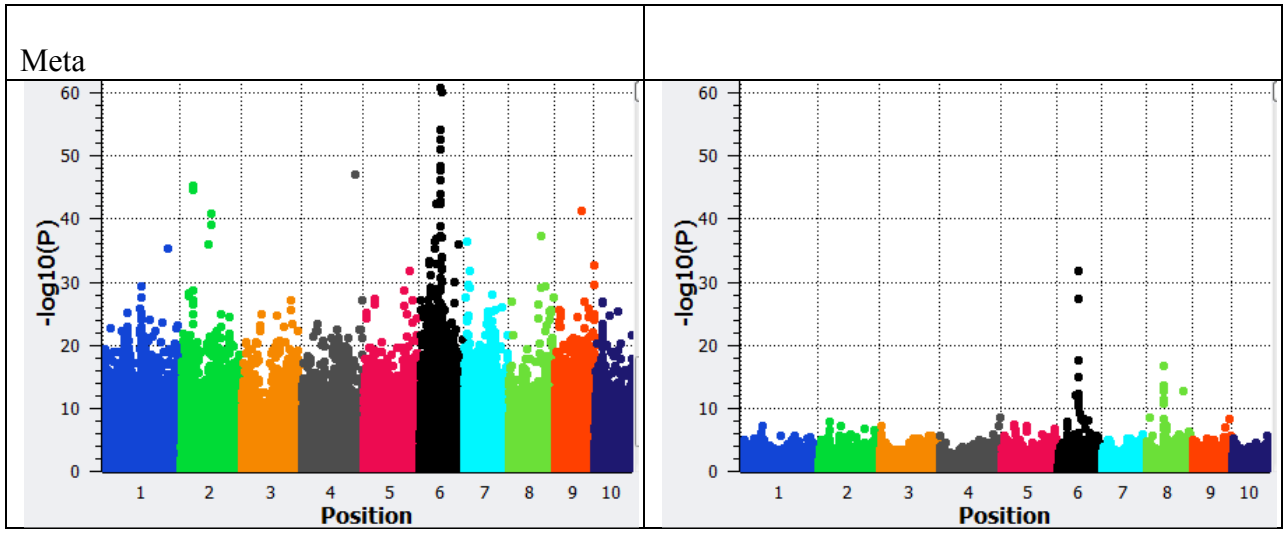


Figure IV-9. Association mapping of grain color binary phenotype on five association mapping panels using Model 1 (without principal components) and Model 2 (with principal components).

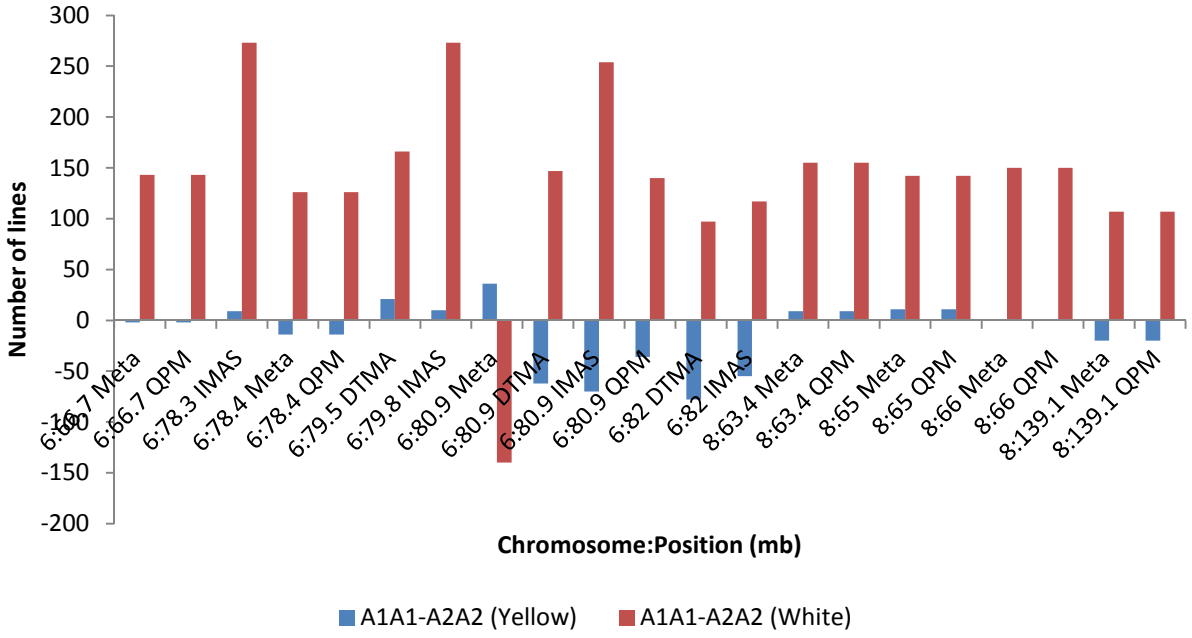
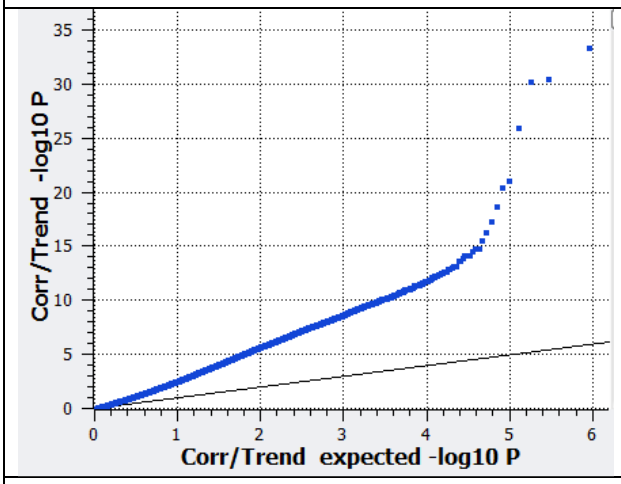
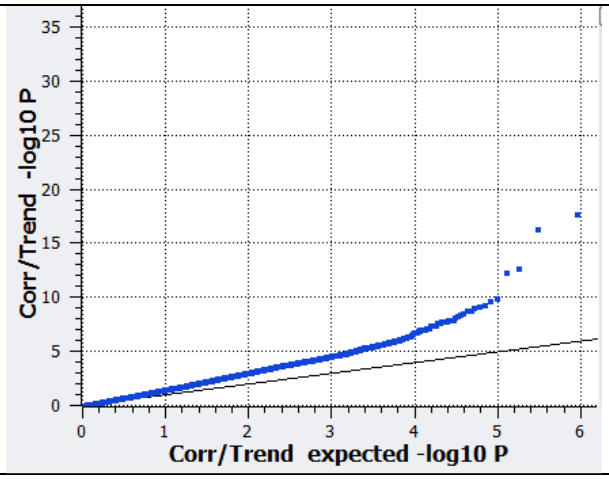
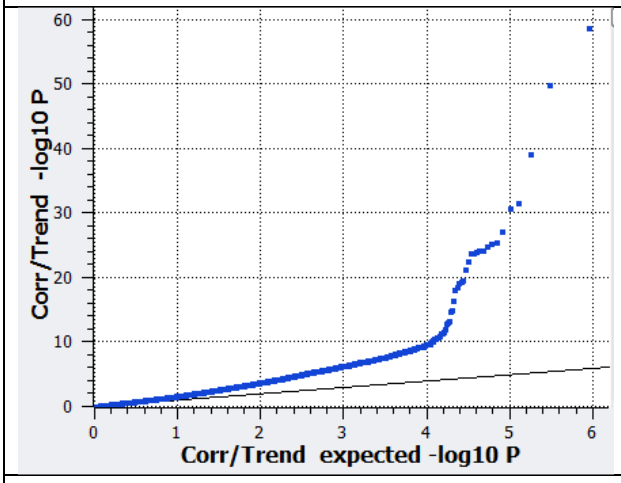
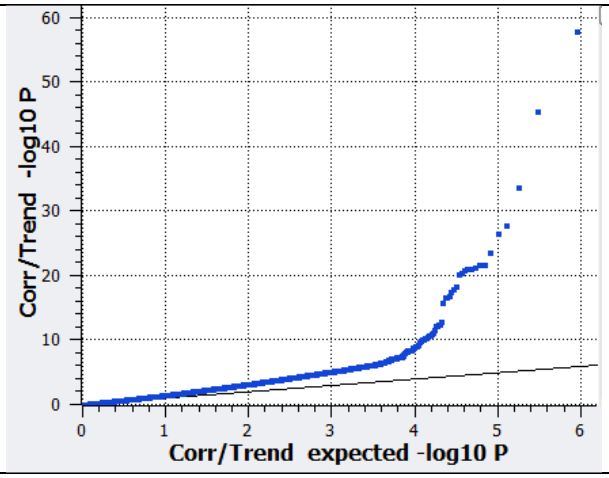
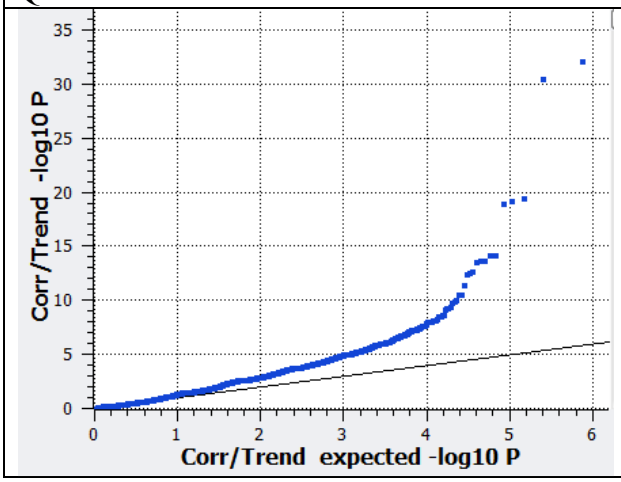
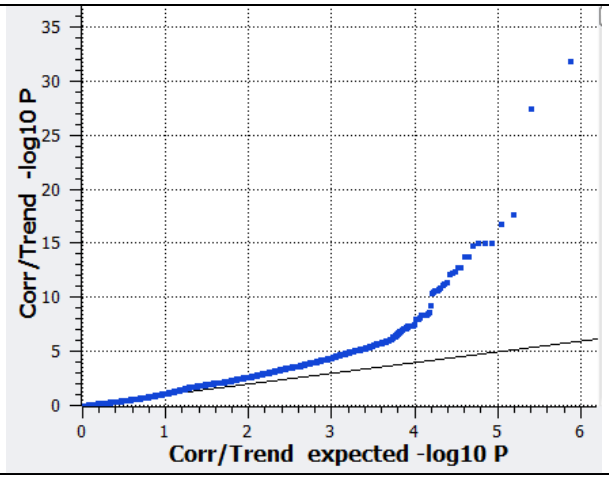


Figure IV-10. The effects of 21 most significant SNPs (12 unique) on grain color binary phenotype in five association mapping panels. A_1 and A_2 are predominant and rare alleles, respectively. The Y axis is the difference of number of lines having homozygous-predominant-allele (A_1A_1) relative to homozygous-rare-allele (A_2A_2) at the respective SNP for each phenotype class (yellow grain=blue and white grain=red). On the X axis, SNPs located on the same chromosome from left to right are in ascending order by physical position in bp.

Model 1 (without principal components)	Model 2 (with principal components)
DTMA	
	
IMAS	
	
QPM	
	

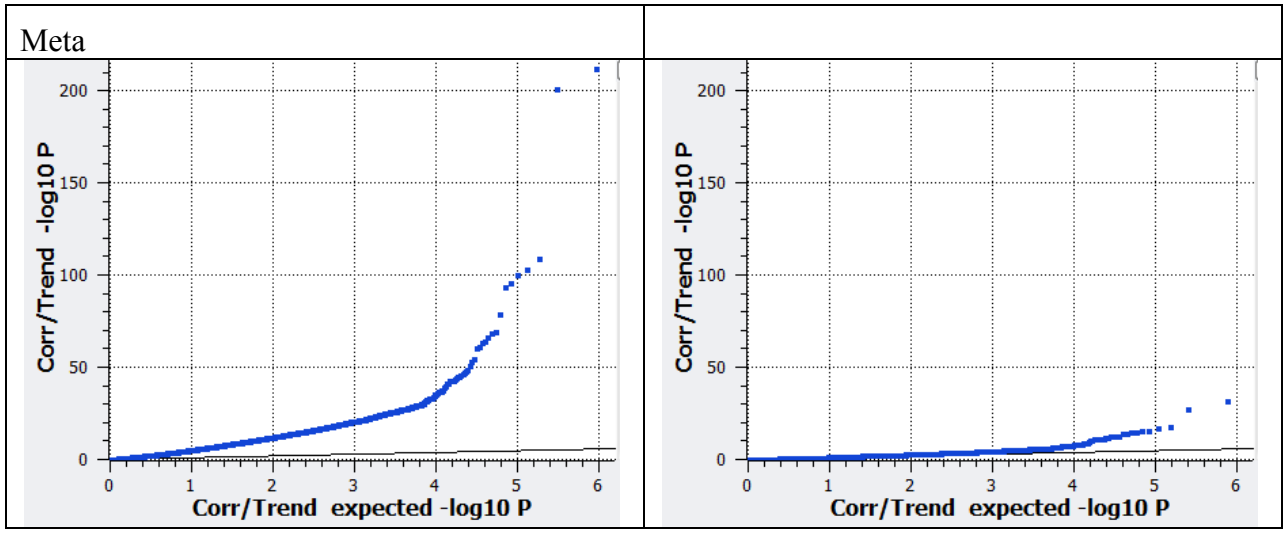
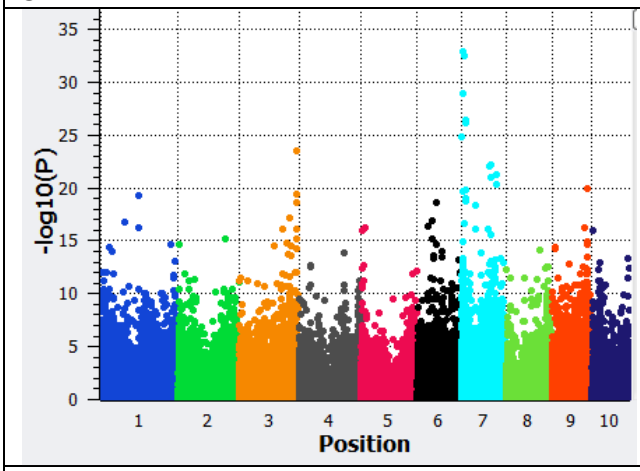
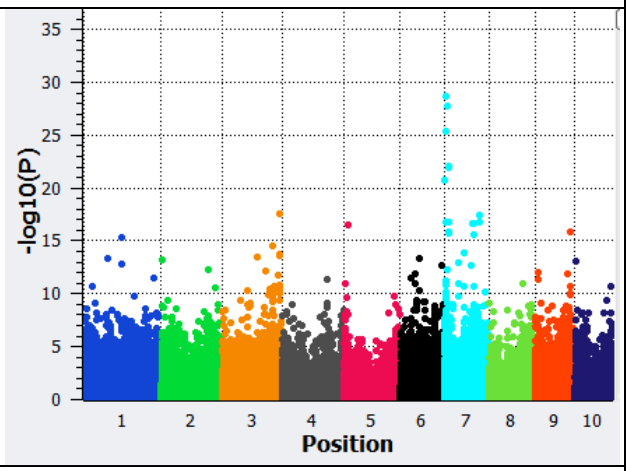
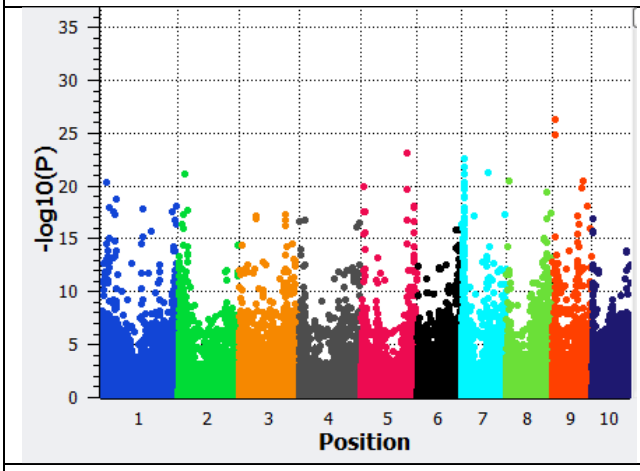
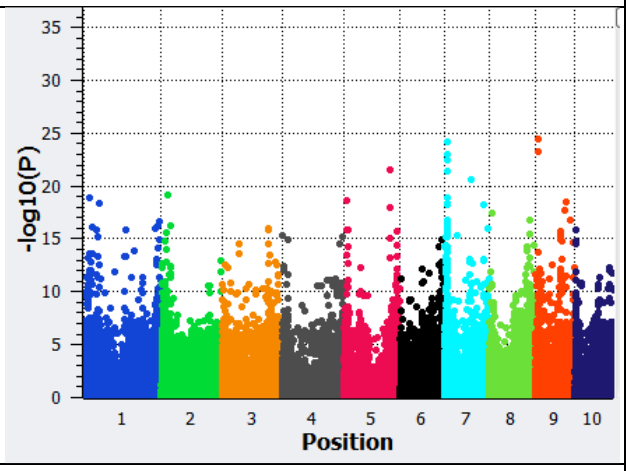
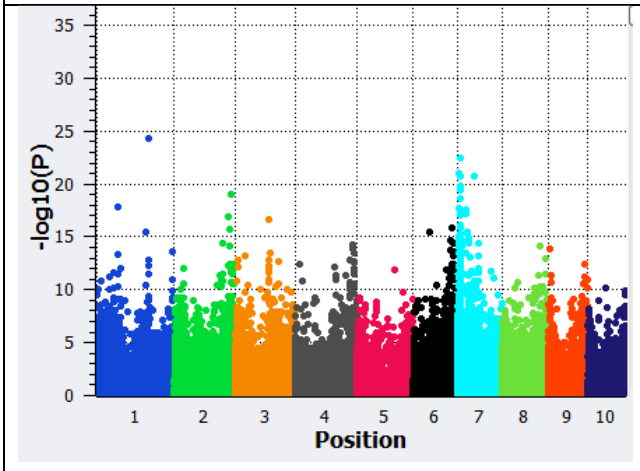
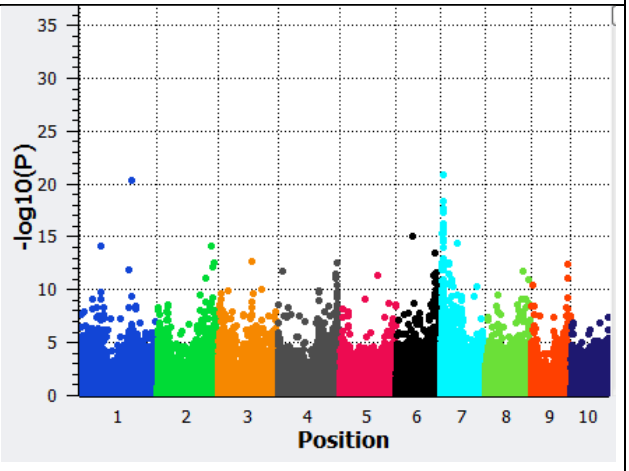


Figure IV-11. Probability-probability plots of P-values using Model 1 (without principal components) and Model 2 (with principal components) in association mapping of grain color binary phenotype.

Model 1 (without principal components)	Model 2 (with principal components)
CAM	
	
DTMA	
	
IMAS	
	

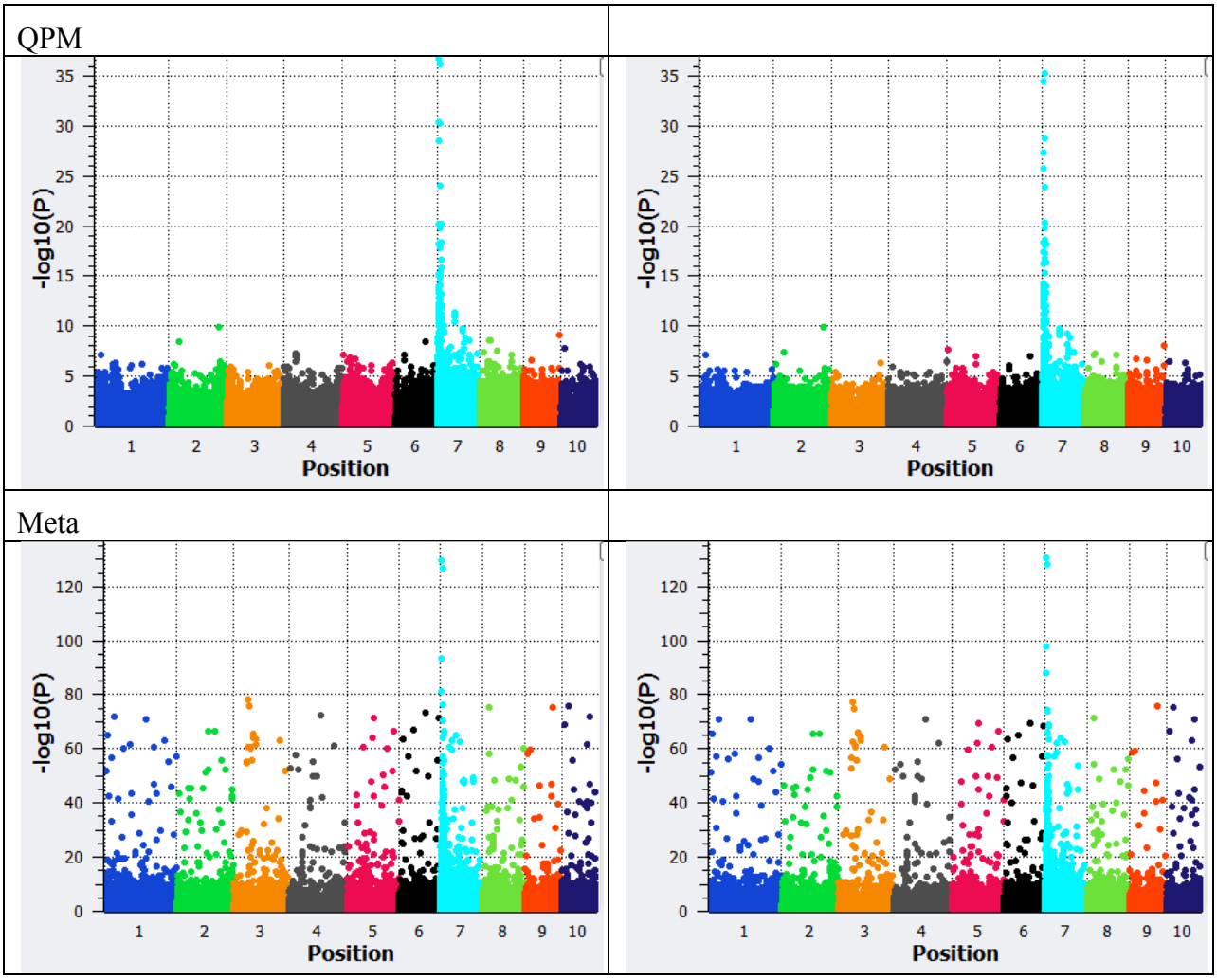


Figure IV-12. Association mapping of QPM binary phenotype on five association mapping panels using Model 1 (without principal components) and Model 2 (with principal components).

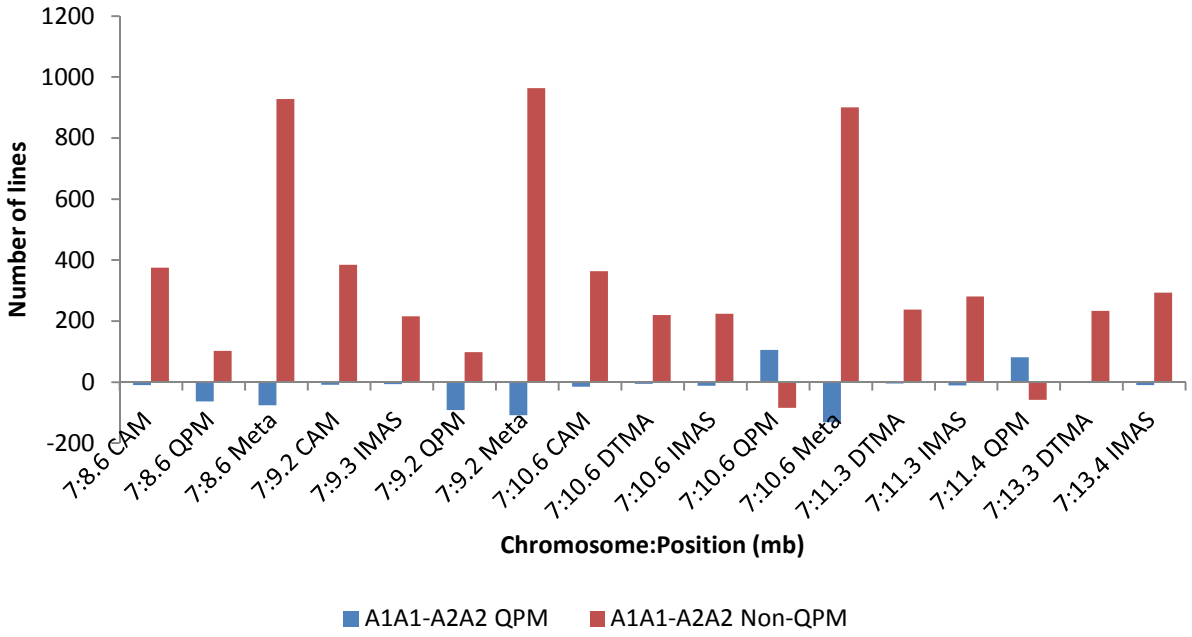
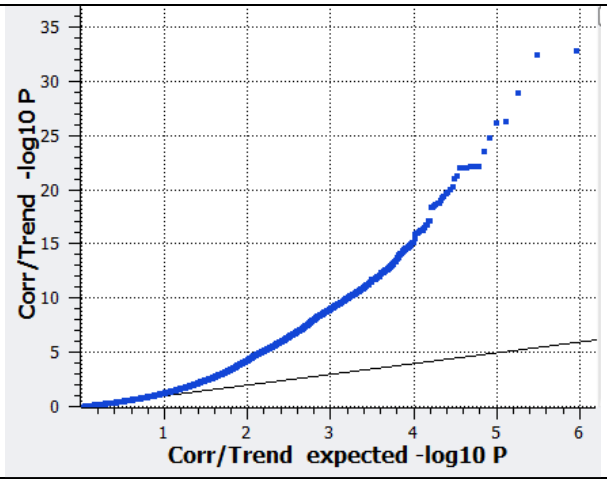
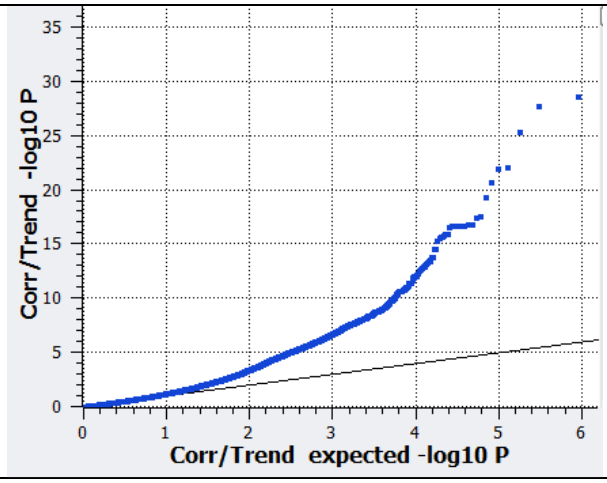
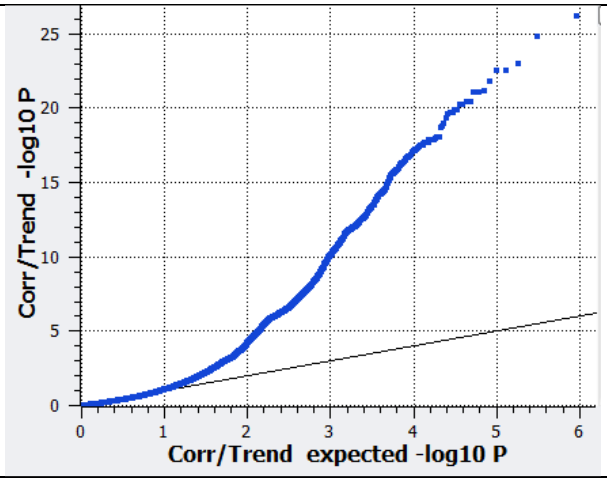
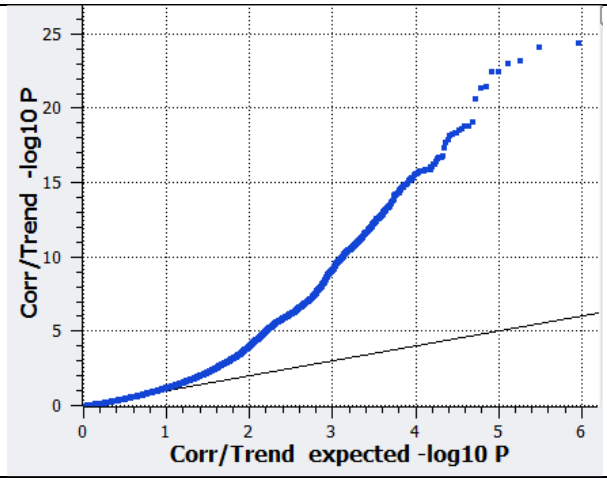
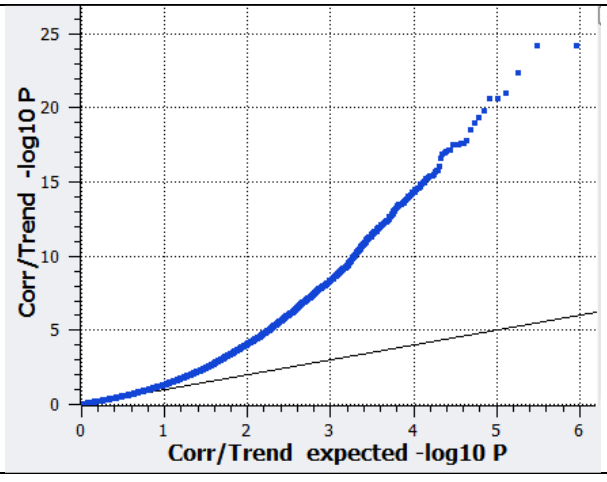
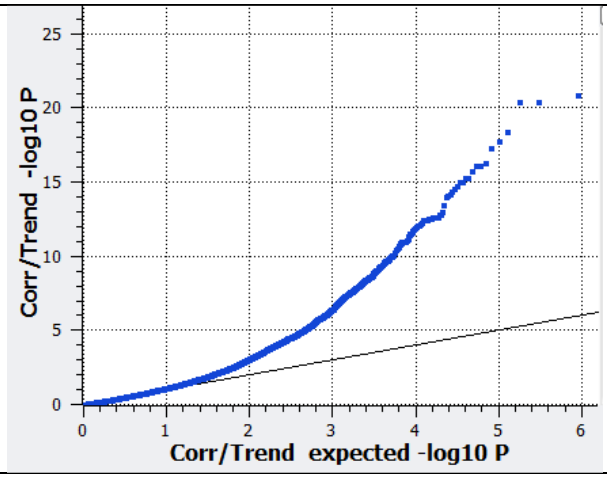


Figure IV-13. The effects of 17 most significant SNPs (11 unique) on QPM binary phenotype in five association mapping panels. A1 and A2 is predominant and rare alleles, respectively. The Y axis is the difference of number of lines having homozygous-predominant-allele (A_1A_1) relative to homozygous-rare-allele (A_2A_2) at the respective SNP for each phenotype class (QPM, blue and non-QPM, red). On the X axis, SNPs located on the same chromosome from left to right are ordered ascending by physical position in bp.

Model 1 (without principal components)	Model 2 (with principal components)
CAM	
	
DTMA	
	
IMAS	
	

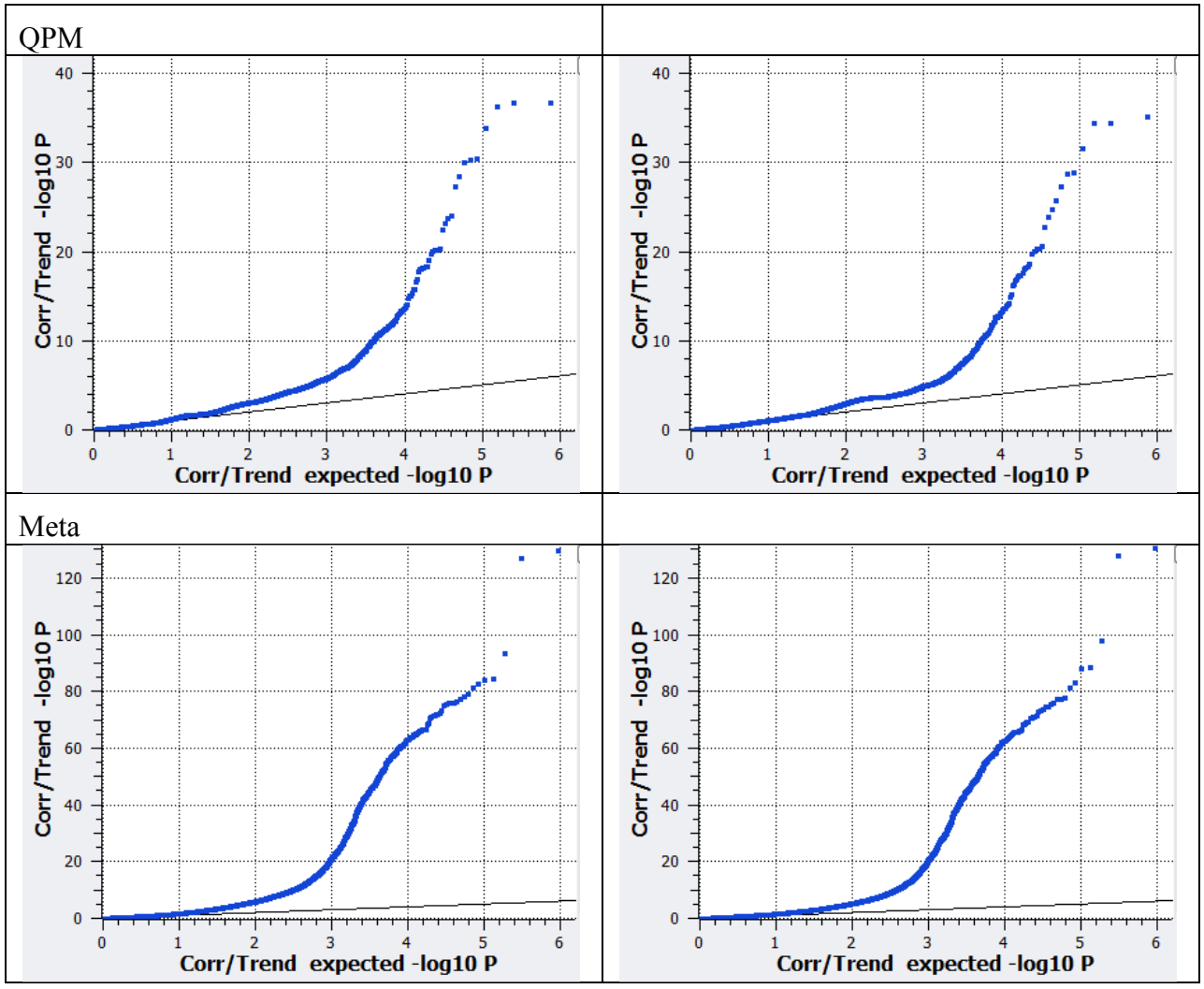


Figure IV-14. Probability-probability plots of observed versus expected P-values using the Model 1 (without principal components) and Model 2 (with principal components) in association mapping of QPM binary phenotype.

Supplemental Table IV-1. Pedigree group names and number of lines in each group

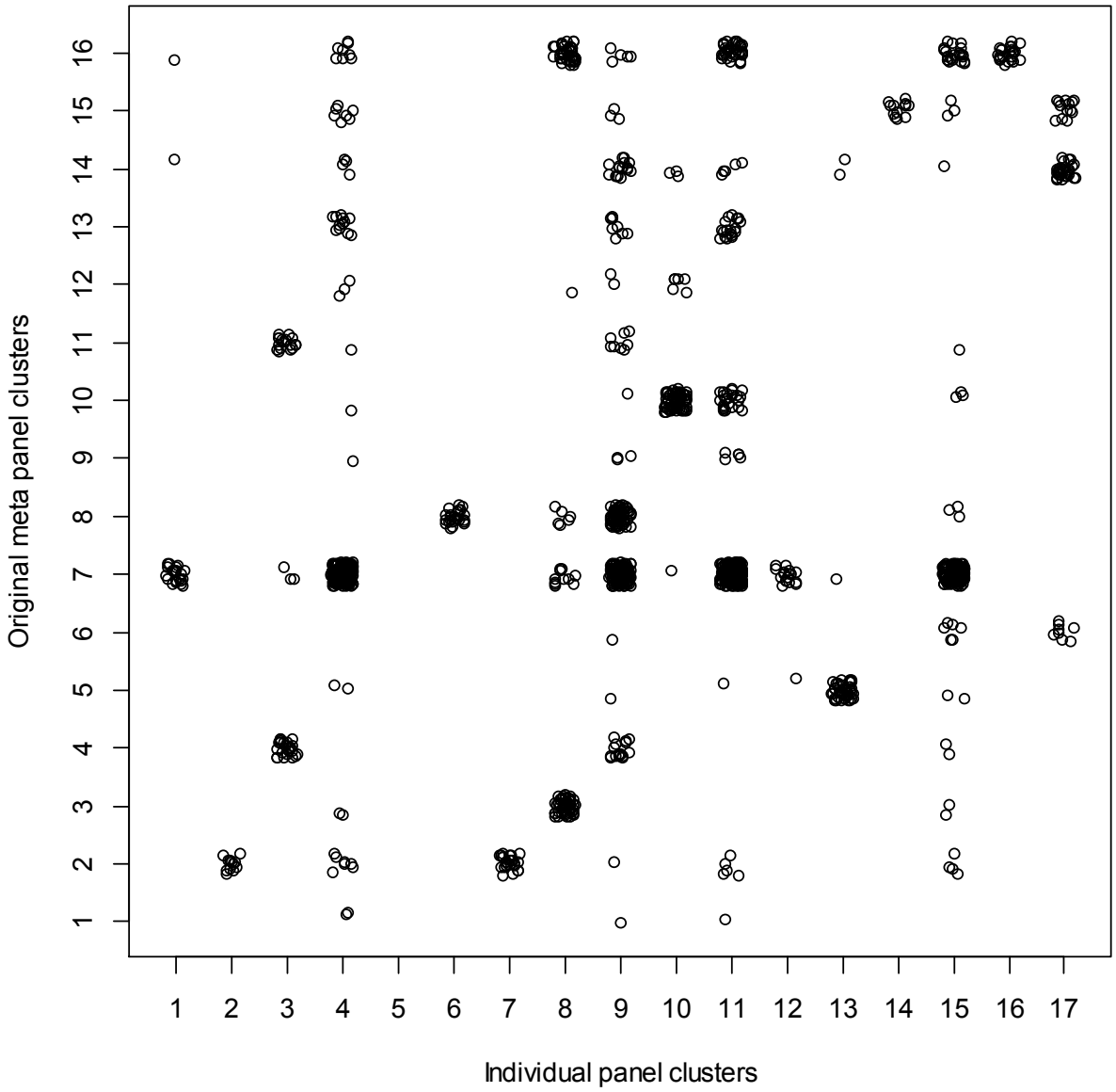
Group number	Group name	Number of lines
1	Thailand	50
1	Provitamin A	157
2	Mexico	219
2	Mexico-Drought	101
2	Mexico-Insect	27
2	Mexico-QPM	33
2	Colombia	83
3	Kenya	14
3	Kenya-Insect	77
3	Zimbabwe	93
4	South Africa	99
5	La Posta Sequia	53
NA	Other	331
	Total	1,337

NA=Not applicable; these lines would not be logically grouped.

Supplemental Table IV-2. Number of lines and each DAPC group for four association mapping panels and the meta panel.

Panel[#]	DAPC group	Number of lines
CAM (436 lines)	1	67
	2	305
	3	18
	4	46
DTMA (277 lines)	1	23
	2	12
	3	44
	4	198
IMAS (376 lines)	1	62
	2	25
	3	24
	4	100
	5	165
QPM (248 lines)	1	12
	2	164
	3	24
	4	48
Meta (1337 lines)	1	85
	2	131
	3	79
	4	820
	5	222

[#] CAM=carotenoid association mapping, DTMA=drought tolerant maize for Africa, IMAS= improved maize for African soils.



Supplemental Figure IV-1. Correspondence between original clusters from the meta panel (16 clusters) and clusters from individual panels (DTMA: cluster 1-4; IMAS: 5-9; CAM: 10-13; QPM: 14-17)

CHAPTER V

CONCLUSIONS AND FUTURE PERSPECTIVES

CONCLUSIONS

Significant general (GCA) and non-significant specific combining ability (SCA) effects indicated that provitamin A concentration was controlled primarily by additive gene action. Genotype x environment (GxE) effects on total provitamin A concentration, which could represent a challenge for developing cultivars with widespread impact on vitamin A malnutrition, were significant but were not of an extreme crossover-type, indicated by significant Spearman rank correlations among locations. Significant yield advantage of crosses among versus within putative heterotic groups formed by maximizing genetic distances, confirmed that molecular-marker-determined genetic distance, while not a panacea, can provide an effective starting point for further breeding work to develop useful heterotic groups.

Association mapping successfully identified the previously known zeaxanthin epoxidase gene (*ZEP*) on chromosome 2, and three other genes with favorable rare alleles on chromosomes 2, 10, and 10 (si605047e03, LOC100279968, and LOC100194220, respectively), controlling putative uncharacterized proteins affecting zeaxanthin, β-cryptoxanthin, and β-carotene concentrations, respectively. Highly significant SNPs on chromosome 10 and chromosome 8 are likely associated with the β-carotene hydroxylase 1 (*CrtRB1*) and the lycopene epsilon-cyclase (*LcyE*) genes, respectively. Additionally, results suggested that the β-carotene hydroxylase *hdy5* gene has a role in controlling β-cryptoxanthin concentrations in maize grain. Besides identifying previously reported genes in the pathway, association mapping identified two other loci,

affecting β-cryptoxanthin and β-carotene concentrations, which could be useful, albeit only after further validations in other populations, for marker-assisted selection for provitamin A carotenoids concentrations. Furthermore, a set of 98 SNP markers that accurately predicted total provitamin A concentrations was identified and is proposed for further validation and use as a marker-assisted selection tool in maize provitamin A biofortification programs.

A series of analyses using four CIMMYT association mapping panels (the Carotenoid Association Mapping (CAM), Drought Tolerant Maize for Africa (DTMA), the Improved Maize for African Soils (IMAS), and Quality Protein Maize (QPM) panels) as well as their combined meta panel successfully identified genes controlling grain color (*Y1* gene on chromosome 6) and QPM phenotypes (*o2* gene on chromosome 7). These results validated the appropriateness of the model with principal components in association mapping analyses and the potential for these panels to identify allelic diversity for additional, more complex traits. Moderate to large population structure (FST>0.05) within each panel confirmed the need to control for population structure in subsequent association mapping studies and suggested that the K-means clustering followed by discriminant analysis of principal components can be used to identify groups of germplasm. Several SNP markers contributed importantly to population structure, and a few of them located in close proximity to previously identified genes. One interest that could be investigated in further research is whether these genes have important effects on fitness or adaptation of lines to various environmental conditions, including – as suggested by the fact that one group identified herein was comprised primarily of drought tolerant lines – drought and low-nitrogen.

FUTURE PERSPECTIVES

CIMMYT's provitamin A maize biofortification breeding program has developed three-way hybrids with improved total provitamin A concentrations (7-8 μg g^{-1}) that have recently been released for commercialization in Zambia. These hybrids were developed prior to knowledge of allelic diversity and application of marker-assisted selection (MAS) for favorable alleles of genes in the carotenoid biosynthetic pathway. Second and third generation biofortified hybrids, with greatly enhanced concentrations of provitamin A (10-20 μg g^{-1}), have been developed using MAS for a favorable allele of *CrtRB1*, identified by association mapping (Yan et al., 2010), and are currently in advanced stages of testing at CIMMYT. This research investigated the prospect of identifying allelic diversity for additional genes in the carotenoid biosynthetic pathway for subsequent validation and potential use in MAS to further enhance the effectiveness of future provitamin A biofortification breeding efforts. Candidate loci were found for β-cryptoxanthin (LOC100279968) and for β-carotene (LOC100194220), as described earlier.

A second future prospect arising from this research is the development and refinement of heterotic groups to enhance the efficiency and effectiveness of biofortification breeding efforts. The process of developing a suitably-broad germplasm base to sustain breeding provitamin A biofortified maize at CIMMYT involved crossing and backcrossing numerous elite white-grained (0 μg g^{-1}), or yellow lines with average concentration of provitamin A (1-2 μg g^{-1}), with exotic (mainly temperate or Thai) sources with larger concentrations of provitamin A carotenoids (7-8 μg g^{-1}). The result was germplasm with increased provitamin A concentration, but also with blurred heterotic definition. The research described in Chapter II, using genetic distances to propose putative heterotic groups among advanced and elite CIMMYT provitamin A biofortified

lines, identified a potential starting point for further development of three heterotically complementary germplasm groups. It is generally recognized that heterotic groups do not exist in nature, but rather must be created via persistent selection (Tracy and Chandler, 2006), and results presented herein suggest a starting point for this work in CIMMYT's provitamin A biofortified breeding program. Furthermore, these results propose three heterotic groups, which will be particularly useful to this breeding program because it targets the development of three-way hybrids. While this study used 402 SNP markers, utilization of high density marker platforms such as 55K and/or GBS for developing putative heterotic groups and classifying inbred lines into the established groups, which might produce more accurate results, merits further research.

Recently developed inbred lines are greatly expanding the previously-known range of provitamin A concentration in maize. Adding these lines to existing association mapping panels, e.g. CIMMYT's carotenoid association mapping (CAM) panel, which was used in studies reported herein, will create broader and better-balanced phenotypic distribution from low (which is currently the majority of genotypes) to high provitamin A content. By adding greater genetic diversity to association mapping panels, particularly adding more lines with moderate to large provitamin A concentrations, their mapping power will be increased, and perhaps genes and useful alleles for genes with modifying effects will be identified. A considerable range has been reported for the effect of the favorable allele of *CrtRB1* on provitamin A concentration for different genetic backgrounds (Babu et al., 2012), for example, and suitably expanded association mapping panels might elucidate the genetic basis for this, opening the possibility of selecting for favorable genotypes in breeding programs.

Another future perspective suggested by the finding that size (291) of one group of lines in the CAM panel was much larger than the others, is that phenotyping and genotyping resources could be saved in association mapping projects by reducing the proportion of genetically similar lines from apparently over-represented group or groups. This hypothesis needs to be validated in further research, for example, by performing a series of association mapping analyses using reduced data sets in which a proportion of genetically similar lines has been randomly removed, and then comparing the power to detect genes in the carotenoid pathway with that of the full data set with all lines in the CAM panel.

Use of increased density of markers in association mapping may allow higher resolution mapping and lead to higher accuracy for identifying candidate genes. It was noteworthy that there were many missing data in the GBS dataset reported herein, and use of strict filtering criteria (minor allele frequency, MAF $\geq$ 0.05 and call rate, CR $\geq$ 0.8), necessary to ensure accuracy of missing-genotype data imputation, further reduced the GBS dataset from 680,000 to around 108,000 useful markers for association mapping (Chapter III). Further research should re-assess the CAM results using more relaxed filtering (e.g. MAF $\geq$ 0.01 and CR $\geq$ 0.3), which will yield about 390,000 GBS SNPs.

Marker assisted selection for the *CrtRB1* has been found useful for developing inbred lines having high (> 8 $\mu g\ g^{-1}$) provitamin A concentrations (Babu et al., 2012). Meanwhile, results reported in Chapter III indicated that genome-wide SNP markers can be useful for predicting breeding values of lines for total provitamin A concentrations. Further interesting research would be to build a more robust genomic prediction model by evaluating various methods (including stepwise regression, ridge regression, and Bayesian methods) using yellow-

grain lines from other CIMMYT association mapping panels (for example, the DTMA, IMAS, and QPM panels) as validation sets. It is likely that the lines in other CIMMYT association mapping panels encompass a narrower range of total provitamin A concentration (might be less than < 8 $\mu g\ g^{-1}$) than, and are genetically distinct (see Chapter IV) from those in the CAM panel. It would also interesting to evaluate prediction accuracies using differing numbers of candidate SNPs, considering that results reported herein indicated that use of fewer markers could give better prediction accuracy than use of maximum available number of markers (see Chapter II).

Results from Chapter III and IV demonstrated that the K-means clustering method and discriminant analysis of principal components (DAPC) were useful for classifying germplasm based on genetic similarity inferred from SNP markers. This result suggests that DAPC may be useful to predict the classification of lines into previously defined groups; for example, the CIMMYT breeding program could use DAPC to tentatively allocate new or exotic lines into its heterotic groups, thereby saving resources otherwise needed for yield trials with multiple testers to determine heterotic orientation of these lines.

REFERENCES

Babu, R., N.P. Rojas, S. Gao, J. Yan, and K. Pixley. 2012. Validation of the effects of molecular marker polymorphisms in LcyE and CrtRB1 on provitamin A concentrations for 26 tropical maize populations. Theoretical and Applied Genetics. Available at http://www.springerlink.com/index/10.1007/s00122-012-1987-3 (verified 12 October 2012).

Tracy, W.F. and M.A. Chandler. 2006. The historical and biological basis of the concept of heterotic patterns in corn bent dent maize. In: Lamkey, K.R. and Lee, M. Plant breeding: the Arnel R. Hallauer International Symposium. Blackwell Publishing.

Yan, J., C.B. Kandianis, C.E. Harjes, L. Bai, E.-H. Kim, X. Yang, D.J. Skinner, Z. Fu, S. Mitchell, Q. Li, M.G.S. Fernandez, M. Zaharieva, R. Babu, Y. Fu, N. Palacios, J. Li, D. Dellapenna, T. Brutnell, E.S. Buckler, M.L. Warburton, and T. Rocheford. 2010. Rare genetic variation at Zea mays crtRB1 increases beta-carotene in maize grain. Nature genetics 42(4): 322–7. Available at http://www.ncbi.nlm.nih.gov/pubmed/20305664 (verified 28 July 2011).