UNCOVERING THE GENETIC ARCHITECTURE AND METABOLIC BASIS OF AMINO ACID COMPOSITION IN MAIZE KERNELS USING MULTI-OMICS INTEGRATION

A Dissertation presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

VIVEK SHRESTHA

Dr. Ruthie Angelovici, Dissertation Supervisor

DECEMBER 2020

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

UNCOVERING THE GENETIC ARCHITECTURE AND METABOLIC BASIS OF AMINO ACID COMPOSITION IN MAIZE KERNELS USING MULTI-OMICS INTEGRATION

presented by Vivek Shrestha,

a candidate for the degree of Doctor of Philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

_____

Professor Ruthie Angelovici

_____

Professor Sherry Flint-Garcia

_____

Professor Elizabeth King

_____

Professor Jason Gillman

DEDICATION

Dedicated to my late Father and late Mother who have always been my inspiration and my strength. I hope to fulfill and follow your dreams.

ACKNOWLEDGMENTS

First and foremost, I express my sincere gratitude to my advisor Dr. Ruthie Angelovici for providing me the opportunity to work on such interesting and broader impact projects. I still remember the day when I first met you in the Maize Genetics Conference in Jacksonville Florida in 2016 and we began talking our research interest and I feel fortunate myself to be where I am today. You have been an awesome advisor, mentor, and a guardian throughout my PhD period. Thank you, Ruthie, for providing me an unquantifiable amount of guidance, motivation, love, and support and immense knowledge to grow both professionally and personally. I would say my journey in these five years have been challenging but also very rewarding at the end. Thank you for providing me opportunities to work with multiple collaborators that help me to boost my confidence and build my critical thinking ability. I am truly grateful to have you as my advisor.

I also express my profound gratitude to Drs. Sherry Flint-Garcia, Elizabeth King and Jason Gillman for serving in my committee and for providing me with constructive and meaningful comments and suggestions. Thank you very much, Sherry, for helping me from the very start of the project. Thank you for providing me those invaluable maize germplasms, helping in the field experimental trials, and supervising me throughout my PhD. Also, thank you for being with us as a faculty advisor for the Corteva Agri science Symposium Series at MU. Thank you, Libby, for sharing your invaluable statistical knowledge. You have been a wonderful mentor and an incredible teacher. I appreciate your teaching style- being so friendly to students and creating an amicable environment in the class; I really enjoyed and learnt a lot from your statistical class. Thank you, Jason, for always been there for me and providing suggestion and comments and helping me in providing guidance especially during the transition of my PhD towards the professional career. Thank you all for writing several recommendation letters on my behalf to

different scholarship committees, organizations, and job applications. I am truly indebted for that.

My sincere thanks to Dr. Abou Yobi for his contribution and assistance in performing analytical jobs for my experimental samples, supervising my research, providing critical comments and suggestion and helping me with improving and sharping my writing skills without whom I could not be able to succeed. Thank you very much, Abou. I appreciate your immense help and support. I would like to thank my fellow graduate students, lab mate and most importantly my friend Marianne Slaten, who have provided me immense help in my projects, provided ideas, inspiration and friendship and much help during the difficult times. You are such an amazing friend, Marianne, and a truly a kind being. Thank you very much. I cannot forget my other lab mate and friend Yen On Chan and former Angelovici lab members; Dr. Sarah Turner Hissong and Dr. Albert Batushansky; all of your contribution and collaboration in solving various problems throughout the project has helped me to move forward. Thank you for your immense effort in troubleshooting the scripting and the computing issues. I would like to thank my lab mates Sam Holden and Clement Bagaza and Jim Elder (USDA, MO) for providing immense help in propagating, managing and postharvest processing of maize diversity panel. A big shout out to all current and previous Angelovici Lab members including Gabi Akrap, Braden Zink, Carson Broeker, Edmond Rieffer, Isaiah DeShon and others for providing the positive vibes, right attitudes and all the fun filled environment in the lab and outside throughout my stay. I am thankful for your friendship, care and cooperation.

I am very much thankful to Dr. Alex Lipka from UIUC for all the help and guidance and supervision of my research work throughout my PhD. Although, you are not in my official committee member, Alex, I would not be wrong to say that your contribution is as equal to as my PhD committee member and I am thankful to you. I am

also indebted to my former MS advisor, Dr. Donald Auger, for being my role model to become a science lover and make me believe in hard work.

I would like to thank Missouri EPSCOR for funding my research. I appreciate it. My deepest gratitude goes to the Division of Biological Sciences (DBS) and the faculty members. I especially want to thank Dr. Paula McSteen for providing me the suggestion and guidance when I was very confused towards my PhD enrollment here at University of Missouri. Thank you for showing me the right path and thank you for your help, support, and the guidance. I would like to thank Dr. Jim Birchler for giving me the opportunity to rotate in his lab. and providing guidance to learn exciting techniques on maize cytogenetics. I am grateful to the lab. of Drs. David Braun, Paula McSteen and Sherry F. Garcia for providing immense help during the first year of the maize association panel plantation at the Genetics Farm, Mizzou. My first impression on DBS and Mizzou was amazing and the credit goes to these wonderful faculty members. My sincere thank you to Dr. Tim Beissinger for partly serving in my committee and supervising and helping me in the initial days of my research as well as being faculty advisor for the Corteva Agriscience Sympoisum Series at MU. I am also thankful to Corteva Agri science for the funding of the symposium at MU and the symposium series (MU Plant Research) committee members for the collaborative effort to lead the symposium. I am thankful to Dr. Felix Fritschi for providing me an opportunity to work in a nitrogen use efficiency project during my first year of the graduate school. Last but not the least, I am thankful to Dr. Tracie Gibson for giving me the opportunity to work with her as a teaching assistant in Cell Biology 2300 during my first few semesters.

I appreciate the love and support given to me by my family, my Dad and Moms, has been my source of courage and strength to me. You always motivated me to pursue this dream, and to be determined to make this dream come true. Although you are not physically present to witness this success, I know you are always there for me. My

grandparents who raised me and have had my back throughout my life. My brothers Kiran and Pukar, my sisters Sumu, Deepa and Darshu for your immense love and support and all my niece and nephews for their welcomed distractions during my studies. I am thankful to my in-law's family members; Ma, Abin and Mimi. I am much thankful to Dr. Avi Karn for his help and support throughout my studies.I am thankful to Singha, Abiskar, Prameela and Vishnu for your help, support and fun filled memories.

Lastly, my dearest love and appreciation goes to my wife, Mani, for her incredible help and support without whom I could be able to succeed. Thank you for always listening my research and providing feedbacks and suggestions. Thank you for the motivation, encouragement, love, and support and for always being with me. Thank you for believing in me when I doubted myself. To anybody that I failed to mention, I thank you all so much for your help.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Seeds are a major source of protein in human and livestock diets. Cereal grains are some of the most consumed seeds by both humans and livestock worldwide, with maize, wheat, and rice alone accounting for ~70% of the total cereal production. Maize is one of the major staple crops used for food, feed, and fuel. A mature maize kernel contains small embryo (10% of the volume) and a large endosperm (~90% of its volume). In terms of composition, majority of the kernel proportion contains around 90% of starch and around 8-10% of protein. Nine of the twenty amino acids cannot be synthesized by monogastric animals, including humans, and must be obtained through the diet and are considered essential amino acids (EAA): lysine, isoleucine, leucine, histidine, methionine, phenylalanine, threonine, tryptophan, and valine. The protein quality is poor in maize endosperm as the primary storage proteins are severely deficient in EAA such as lysine, tryptophan, and methionine. Such deficiencies can be detrimental since corn provides an important source of proteins for food in developing countries and for feed in developed countries such as the U.S. Failure to consume sufficient levels of EAA per day leads to severe malnutrition, even if one's calories requirements are met.

Many attempts to increase the EAA has demonstrated only limited success since seed can rebalance their amino acids composition even when major changes are introduced in their proteome. One possible approach to solve this applied problem is by seed EAA biofortification; however, many attempts at this task fall short and strongly indicates that even though we know most of the metabolic pathways of amino acids, we know very little about their regulation especially in seed. Therefore, the first step towards efficient

amino acid biofortification is to increase our fundamental understanding of their function, as well as the metabolic regulation and the biology of the plant seeds.

Despite the tight regulation within any given genotype seed amino acid composition display extensive natural variation which can be utilized to uncover the genetic basis and identify new targets for seed amino acids biofortification. Hence to uncover the genetic architecture of amino acids composition in maize kernels we used Goodman-Buckler maize association panel that consists of 282 diverse maize inbred lines including stiff stalk, non-stiff stalk, tropical and subtropical, sweetcorn and popcorn lines. I performed genome wide association study (GWAS) on both the protein bound amino acids (PBAA) and free amino acids (FAA). Although, GWAS is widely used to dissect the genetic architecture of complex traits, oftentimes the GWAS outputs the extensive list of genes particularly when using multiple phenotypic traits. To overcome this, I used an integrative multi-omics approach that combines GWAS and co-expression networks modules obtained from ten seed filling stages of B73 to uncover novel key regulatory genes, characterize biological process and prioritized the candidate genes that involved in shaping the natural variation of amino acid composition.

Chapter one of the dissertation is the general introduction and literature review on the seed amino acids. It briefly discuss the general introduction of PBAA and FAA, previous attempts done to improve seed PBAA and FAA composition, natural variation used to uncover the genetic architecture of complex traits including metabolic traits such as amino acids and finally discuss the multi-omics integration to uncover the genetic basis of complex traits.

Chapter two elaborates the comprehensive genetic basis of PBAA in maize kernels using integrative analysis of 76 PBAA GWAS with protein co-expression network modules. Previous studies have shown that manipulation of storage proteins and amino acid pathway genes have contributed in the improvement of quality protein maize however, my study strongly suggests that in addition to the manipulation of storage protein and amino acid metabolic genes, specific ribosomal genes along with other translation machinery could be the novel target for seed amino acids biofortification.

Chapter three discusses the genetic basis of FAA in maize kernels using integrative analysis of 109 FAA GWAS with protein co-expression network modules. I have presented here the comprehensive list of SNPs as well as the candidate genes and several biological processes including the translational machinery responsible for shaping the genetic architecture of FAA in seed.

Chapter four includes the conclusion and future works.

Maize is an important crop used for both food and feed and possesses great genotypic and phenotypic diversity. The results from my study has validated several previous characterized genes and identified novel key genes that regulate and shape the PBAA and FAA in maize kernels, which could be used further to target for amino acid biofortification.

CHAPTER 1: GENERAL INTRODUCTION AND LITERATURE REVIEW

1.1 Amino acid composition in seeds

Amino acids (AA) are the basic building blocks of proteins and are therefore essential for growth and development to sustain life. There are 20 proteogenic amino acid: Alanine (Ala), Arginine (Arg), Asparagine (Asn), Aspartate (Asp), Cysteine (Cys), Glutamine (Gln), Glutamate (Glu), Glycine (Gly), Histidine (His), Isoleucine (Ile), Leucine (Leu), Lysine (Lys), Methionine (Met), Phenylalanine (Phe), Proline (Pro), Serine (Ser), Threonine (Thr), Tryptophan (Trp), Tyrosine (Tyr) and Valine (Val). There are two functional pools of amino acid in seeds: the protein bound amino acid pool (PBAA), which account for 95% of seed total amino acids (TAA) and the free amino acid pool (FAA), which account for the remaining 5% (Amir, Galili, & Cohen, 2018; Muehlbauer, Gengenbach, Somers, & Donovan, 1994). FAAs serve as building blocks of protein synthesis, as main precursors for the synthesis of diverse group of primary and secondary metabolites, which includes organic acids, osmolytes, phytohormones, and secondary metabolites and as an alternative source of accessible energy (Amir et al., 2018; Angelovici, Fait, Fernie, & Galili, 2011; G. Galili & Hofgen, 2002), the majority of PBAA in seeds are deposited as seed storage proteins (SSP), which ensure proper desiccation and germination (Angelovici, Galili, Fernie, & Fait, 2010; Bewley, 1997).

In maize, out of 95% of the PBAA in seed, ~ 70% are in SSP. There are three main groups of SSP in maize categorized based on their solubility: water soluble albumins, salt-soluble globulins, and alcohol soluble prolamins (Peter R Shewry & Casey, 1999). Prolamins, also known as zeins, are the most abundant SSP (60-70% of the total storage proteins) in maize seeds (Larkins, 2017; Larkins & Hurkman, 1978) and

hence are the major determinants of seed amino acid composition in the maize kernel. Zein proteins are synthesized in the lumen of the rough endoplasmic reticulum where they form insoluble spherical accretions called protein bodies of generally 1-2 microns in diameter (Larkins, 2017; Lending & Larkins, 1989). Zeins lacks one or more essential amino acids such as lysine and tryptophan. Maize zeins are categorized into 4 types: α, β, γ and δ-zeins. Among the 4 types of zeins, the 19 and 22-kDa α zeins are the most abundant ones and mostly enriched in proline and glutamine (a non-essential amino acids), while β, γ and δ zeins, which are mostly enriched in cysteine, methionine, and other essential amino acids that are less abundant in maize endosperm. Therefore, zeins consists of plenty of non-essential amino acids but are deficient in several essential amino acids, thereby resulting in the poor seed amino acid nutritional value (Larkins, 2017; Larkins, Pedersen, Marks, & Wilson, 1984; Peter R Shewry, 2007).

1.2 Previous attempts to improve the amino acid levels and composition in seeds

Attempts to improve the seed storage proteins (PBAA) over the past two decades had uncovered a unique regulatory mechanism that governs amino acid composition and accumulation in seeds. In contrast to the expectation, knockdown of the highly abundant but EAA poor storage proteins or overexpression of high-quality proteins using transgenic approaches had very little effect on the overall relative composition of amino acids in seeds. In fact, seeds revealed their ability to rebalance protein composition to the original state even when severe perturbation to the proteome is introduced by transgenic measures (Larkins, 2017; M. Schmidt et al., 2011). In addition, some of the proteomic alterations that were introduced had negative effects on seed quality and led to brittle kernel texture, poor germination, poor longevity, as well as negative effects on plant

overall growth and yield. For example, the *opaque 2* (*o2*) mutant exhibits a substantial reduction in seed storage proteins, exhibits proteome rebalance with higher lysine content but comes with poor seed quality. *O2* encodes for a transcription factor belonging to the BZIP family; alteration of the gene results in a decrease of zein proteins and an increase in non-zein proteins that are rich in lysine.

Although, *o2* has double the amount of lysine compared to the wild type (Hunter et al., 2002), it comes with negative agronomic qualities such as soft chalky endosperm which is unfavorable for food processing and susceptible to many diseases and pests with poor germination and yield (Dizigan et al., 2007; Kodrzycki, Boston, & Larkins, 1989). Attempts were made using different opaque modifiers to improve the agronomic qualities of the *o2* mutant while maintaining the increase in lysine in the seed. The result was the development of quality protein maize (QPM) cultivars (Crow & Kermicle, 2002; Geetha, Lending, Lopes, Wallace, & Larkins, 1991; Vasal, Villegas, Bjarnason, Gelaw, & Goertz, 1980). Endosperm texture was improved to a greater extant in the QPM cultivars along with high lysine and tryptophan and better protein digestibility, however, the yield was still compromised (Sofi, Wani, Rather, & Wani, 2009). In addition, QPM breeding is a complex process involving selection of multiple, unlinked *o2* modifier loci, while maintaining the homozygous recessive *o2* background, which is both challenging and time consuming.

Interestingly, the proteome rebalancing and FAA reprogramming phenomena are followed by a major increase in free amino acids in maize seeds. The FAA reprogramming function is unclear since the total PBAA composition or levels are mostly unchanged. In general, the regulation of the FAA pool and its interplay with the PBAA

pool is unclear even though they are precursors and products. Several studies have focused on increasing FAA to achieve seed amino acid biofortification. Studies demonstrated that FAA are highly interconnected and perturbation of a single gene in the amino acid metabolic pathway often leads to the changes in the entire amino acid metabolic network along with adverse effects on the agronomic and yield traits (Dizigan et al., 2007; Gu, Jones, & Last, 2010). For instance, Dizigan et al (2007) expressed a Lys feedback insensitive biosynthesis gene in maize that overcame the natural regulation of free Lys and led to a substantially higher accumulation of free lysine in seeds (Dizigan et al., 2007), however, it also came with negative pleiotropic effects on growth and yield penalties in the maize cultivar. Similarly, Zhu et al. (2003) demonstrated a substantial increase in *Arabidopsis* seed free lysine by increasing its synthesis and blocking catabolism. Nevertheless, this manipulation results in severe fitness penalty including poor germination and growth, which was found to be caused by the deprivation of alternative energy source during germination (Angelovici et al., 2011; X. Zhu & Galili, 2003). Gu et al. (2010) reported that in *Arabidopsis* seeds, knock out of a leucine catabolic gene causes increase in 12 FAA. Interestingly, it does not alter the overall PBAA composition of seeds (Gu et al., 2010), which suggests that there is a disconnect between the PBAA from the availability of its precursors (FAA) during the dry seed developmental stage.

Neither the genetic basis of the rebalancing mechanism nor the genetic basis governing the interplay between FAA and PBAA were exposed using reverse genetic approaches, most likely since knock out mutants are either lethal or FAA and PBAA are complex traits that are governed by multiple genes. The natural variation of these traits

certainly supports the later. Recent studies have shown that there is extensive natural variation in seed amino acids (both FAA and PBAA absolute levels and relative composition) across inbred lines (Angelovici et al., 2016; Deng et al., 2017) despite their tight regulation in seeds. Hence, exploiting the natural variation of these traits to uncover the genetic basis of their regulation could prove very beneficial.

1.3 Using natural variation to uncover the genetic basis of amino acids

Genome wide association study (GWAS) is widely used to dissect the architecture of several complex traits including the primary and secondary metabolites traits (Angelovici et al., 2016; Angelovici et al., 2013; Cook et al., 2012; Deng et al., 2017; Wu et al., 2016). A recent review on GWAS and various GWAS working models can be found in Shrestha, Awale, & Karn (2019). The majority of amino acid GWAS have been done in seed FAA in *Arabidopsis*, while there are limited studies done on GWAS of seed FAA and PBAA in maize. Angelovici et al. (2013) performed GWAS on branched chain amino acid levels in *Arabidopsis* seeds using 360 ecotypes. The authors found the strong association of BCAA with branched chain amino acid transferases; *BCAT1* and *BCAT2*. Using linkage analysis and reverse genetic approaches, the authors found that the allelic variation of *BCAT2* is responsible for the natural variation of seed BCAAs as well as FAA homeostasis in seeds (Angelovici et al., 2013). A recent study by Slaten et al (2020) on FarmCPU, GWAS on free Glutamine (Gln) related traits in *Arabidopsis* seed identified several multiple candidate genes. Using molecular validation, the authors revealed an unexpected association between the aliphatic glucosinolates and the Gln related traits, which indicates that secondary metabolites might have key functions in primary seed metabolism and also play key roles in sharping the seed metabolic

homeostasis (Slaten, Yobi, et al., 2020a). Qin et. al (2019) studied the GWAS of 15 PBAA in 249 soybean association panel and found 14 annotated amino acid metabolism related genes including pyrroline-5-carboxylate reductase and B L-selenocystathione (Qin et al., 2019). A strong SNP-PBAA association was found by a study on PBAA GWAS in maize by Deng et. al (2017). The study was conducted using 513 diverse maize inbred lines from China as well as 3 recombinant inbred line (RIL) populations for linkage study. The authors found 247 and 281 significant loci using GWAS study at two different field locations respectively. They found some important candidate genes regulating PBAA natural variation in maize seeds that included *o2* and *o2* modifier genes including the 27kD γ zein (Deng et al., 2017).

1.4 Challenges in GWAS and the use of integrative approaches combined with GWAS to whittle down candidate genes

Although GWAS is widely used to dissect the genetic basis of complex trait, it is challenging to identify and validate the key regulatory genes from an extensive list of genes especially when GWAS is done using multiple traits. To overcome this challenge, integration of multiple omics (genomics, transcriptomics, and proteomics) techniques has been found to be effective in identifying key regulations of different metabolites in a wide variety of species that includes *Arabidopsis* (S. Wu et al., 2016), maize (Schaefer et al., 2018), soybean (Qi et al., 2018), rapeseed (Yao et al., 2020) and peanut (H. Zhang et al., 2019). For example, Wu et al. (2016) demonstrated that comparing two or more orthogonal genome scale data sets is a potential approach to identify the key regulatory elements with more confidence. The study showed that in *Arabidopsis*, comparing GWAS output with metabolite-transcript correlation network analysis (S. Wu et al.,

2016) helped identify both previously identified as well as novel key functional regulatory genes of both primary and secondary metabolites. Schaefer et al (2018) integrated results from maize ionome GWAS with gene expression networks and facilitated the prioritization of key candidate genes (Schaefer et al., 2018). The authors demonstrated that coexpression networks are a powerful tool in prioritizing candidate causal genes from GWAS loci but also suggest that the success of such strategies highly depends on the gene expression data context related to the phenotype studied. A meta-QTL analysis combined with weighted gene correlation network analysis (WGCNA) was used to reveal hub genes for soybean seed storage composition during seed development (Qi et al., 2018). Using WGCNA, the authors identified 47 modules and reported several hub gene that were involved in soybean oil and seed storage protein accumulation processes including vacuolar protein sorting 35 (VPS35), Zn-dependent exopeptidase superfamily protein, ABI3b, LEC1d and cupin family protein (Qi et al., 2018). The GWAS and co-expression network combination was also used to uncover the key regulatory genes of oleic acid in rapeseed and to develop functional haplotype markers for the improvement of oleic acid content in rapeseed (Yao et al., 2020). Similarly, A GWAS and co-expression network were used to uncover the ionomic variation in cultivated peanut. The authors used 13 mineral elements and performed GWAS using a 120-accession association panel and combined them with a gene coexpression network. They reported several QTLs and candidate genes for boron, copper, sodium, sulfur, and zinc (H. Zhang et al., 2019).

1.5 High throughput phenotyping of seed amino acids enables genomic analysis

In addition to the challenges in prioritizing GWAS candidate genes, finding an appropriate, high throughput, cost-effective, and accurate method to quantity seed amino acids is also a major challenge. To date one of the main challenges was the lack of high resolution, low cost high throughput phenotyping method that can facilitate phenotyping large populations and implementation of quantitative genetic approaches such as GWAS. Low resolution methods such as NIR are used, but predominantly for preliminary sample screening. In addition, FAA and PBAA cannot be separated using these methods. High resolution methods that are based on high performance liquid chromatography (HPLC) method may take up to two hours for a single analysis and are highly expensive. To overcome this challenge, one can implement a low cost, microscale (3-4 mg) high throughput (96 multiplex) detection method that is based on LC-MS/MS multiple reaction monitoring (MRM) approach, which combines both accuracy and affordability (Angelovici et al. 2013). Recently a similar method was also develped for PBAA (Yobi & Angelovici, 2018) that has enabled extensive genomic analysis of amino acids in seeds.

Therefore, in this dissertation, I, first quantified 16 PBAA and 20 FAA absolute levels from the 282 Goodman-Buckler association panel using the methods described above. I, then, used these absolute levels, and calculated relative composition and biochemical family related ratios to perform GWAS. Altogether, 76 PBAA and 109 FAA GWAS were used. In order to prioritize the GWAS candidate gene list, an integrative multi-omics approach that integrates both PBAA and FAA GWAS with the protein co-expression modules that has strong association with the PBAA and FAA traits

respectively obtained from 10 seed filling stages of B73 was used. This study highlights and uncovers novel key regulatory genes, characterize biological processes, and prioritized candidate genes that are involved in shaping the natural variation of amino acid composition.

1.6 References:

Amir R, Galili G, Cohen H (2018) The metabolic roles of free amino acids during seed development. Plant Science

Angelovici R, Batushansky A, Deason N, Gonzalez-Jorge S, Gore MA, Fait A, DellaPenna D (2016) Network-guided GWAS improves identification of genes affecting free amino acids. Plant physiology:pp. 01287.02016

Angelovici R, Fait A, Fernie AR, Galili G (2011) A seed high-lysine trait is negatively associated with the TCA cycle and slows down Arabidopsis seed germination. New Phytologist 189 (1):148-159

Angelovici R, Galili G, Fernie AR, Fait A (2010) Seed desiccation: a bridge between maturation and germination. Trends in plant science 15 (4):211-218

Angelovici R, Lipka AE, Deason N, Gonzalez-Jorge S, Lin H, Cepela J, Buell R, Gore MA, DellaPenna D (2013) Genome-wide analysis of branched-chain amino acid levels in Arabidopsis seeds. The Plant cell 25 (12):4827-4843

Bewley JD (1997) Seed germination and dormancy. The Plant cell 9 (7):1055-1066

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia SA (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant physiology 158 (2):824-834

Crow JF, Kermicle J (2002) Oliver Nelson and quality protein maize. Genetics 160 (3):819-821

Deng M, Li D, Luo J, Xiao Y, Liu H, Pan Q, Zhang X, Jin M, Zhao M, Yan J (2017) The genetic architecture of amino acids dissection by association and linkage analysis in maize. Plant biotechnology journal 15 (10):1250-1263

Dizigan MA, Kelly RA, Voyles DA, Luethy MH, Malvar TM, Malloy KP (2007) High lysine maize compositions and event LY038 maize plants. Google Patents,

Galili G, Hofgen R (2002) Metabolic engineering of amino acids and storage proteins in plants. Metab Eng 4 (1):3-11. doi:10.1006/mben.2001.0203

Geetha K, Lending CR, Lopes MA, Wallace JC, Larkins BA (1991) opaque-2 modifiers increase gamma-zein synthesis and alter its spatial distribution in maize endosperm. The Plant cell 3 (11):1207-1219

Gu L, Jones AD, Last RL (2010) Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant. The Plant Journal 61 (4):579-590

Hunter BG, Beatty MK, Singletary GW, Hamaker BR, Dilkes BP, Larkins BA, Jung R (2002) Maize opaque endosperm mutations create extensive changes in patterns of gene expression. The Plant cell 14 (10):2591-2612

Kodrzycki R, Boston RS, Larkins BA (1989) The opaque-2 mutation of maize differentially reduces zein gene transcription. The Plant cell 1 (1):105-114

Larkins BA (2017) Maize Kernel Development. CABI,

Larkins BA, Hurkman WJ (1978) Synthesis and deposition of zein in protein bodies of maize endosperm. Plant Physiology 62 (2):256-263

Larkins BA, Pedersen K, Marks MD, Wilson DR (1984) The zein proteins of maize endosperm. Trends in Biochemical Sciences 9 (7):306-308

Lending CR, Larkins BA (1989) Changes in the zein composition of protein bodies during maize endosperm development. The Plant cell 1 (10):1011-1023

Muehlbauer G, Gengenbach B, Somers D, Donovan C (1994) Genetic and amino-acid analysis of two maize threonine-overproducing, lysine-insensitive aspartate kinase mutants. Theoretical and applied genetics 89 (6):767-774

Qi Z, Zhang Z, Wang Z, Yu J, Qin H, Mao X, Jiang H, Xin D, Yin Z, Zhu R (2018) Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development. Plant, cell & environment 41 (9):2109-2127

Qin J, Shi A, Song Q, Li S, Wang F, Cao Y, Ravelombola W, Song Q, Yang C, Zhang M (2019) Genome Wide Association Study and Genomic Selection of Amino Acid Concentrations in Soybean Seeds. Frontiers in Plant Science 10:1445

Schaefer RJ, Michno J-M, Jeffers J, Hoekenga O, Dilkes B, Baxter I, Myers CL (2018) Integrating coexpression networks with GWAS to prioritize causal genes in maize. The Plant cell 30 (12):2922-2942

Schmidt M, Barbazuk WB, Sandford M, May GD, Song Z, Zhou W, Nikolau BJ, Herman EM (2011) Silencing of soybean seed storage proteins results in a rebalanced protein composition preserving seed protein content without major collateral changes in the metabolome and transcriptome. Plant Physiology:pp. 111.173807

Shewry PR (2007) Improving the protein content and composition of cereal grain. Journal of cereal science 46 (3):239-250

Shewry PR, Casey R (1999) Seed proteins. In: Seed proteins. Springer, pp 1-10

Shewry PR, Halford NG (2002) Cereal seed storage proteins: structures, properties and role in grain utilization. Journal of experimental botany 53 (370):947-958

Shrestha V, Awale M, Karn A (2019) Genome Wide Association Study (GWAS) on Disease Resistance in Maize. In: Disease Resistance in Crop Plants. Springer, pp 113-130

Slaten ML, Yobi A, Bagaza C, Chan YO, Shrestha V, Holden S, Katz E, Kanstrup C, Lipka AE, Kliebenstein DJ (2020) mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis. Plant Physiology

Sofi P, Wani SA, Rather A, Wani SH (2009) Quality protein maize (QPM): genetic manipulation for the nutritional fortification of maize. Journal of Plant Breeding and Crop Science 1 (6):244-253

Vasal SK, Villegas E, Bjarnason M, Gelaw B, Goertz P (1980) Genetic modifiers and breeding strategies in developing hard endosperm opaque-2 materials. Genetic modifiers and breeding strategies in developing hard endosperm opaque-2 materials:37-73

Wu S, Alseekh S, Cuadros-Inostroza Á, Fusari CM, Mutwil M, Kooke R, Keurentjes JB, Fernie AR, Willmitzer L, Brotman Y (2016) Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in Arabidopsis thaliana. PLoS genetics 12 (10):e1006363

Yao M, Guan M, Zhang Z, Zhang Q, Cui Y, Chen H, Liu W, Jan HU, Voss-Fels KP, Werner CR (2020) GWAS and co-expression network combination uncovers multigenes with close linkage effects on the oleic acid content accumulation in Brassica napus. BMC genomics 21:1-12

Yobi A, Angelovici R (2018) A High-Throughput Absolute-Level Quantification of Protein-Bound Amino Acids in Seeds. Current protocols in plant biology 3 (4):e20084

Zhang H, Wang ML, Schaefer R, Dang P, Jiang T, Chen C (2019) GWAS and coexpression network reveal ionomic variation in cultivated peanut. Journal of agricultural and food chemistry 67 (43):12026-12036

Zhu X, Galili G (2003) Increased lysine synthesis coupled with a knockout of its catabolism synergistically boosts lysine content and also transregulates the metabolism of other amino acids in Arabidopsis seeds. The Plant cell 15 (4):845-853

CHAPTER 2: MULTI-OMICS APPROACH REVEALS THE INVOLVEMENT OF TRANSLATIONAL REPROGRAMMING IN SHAPING AMINO ACID COMPOSITION DURING KERNEL MATURATION IN MAIZE.

ABSTRACT

Maize seeds are a critical source of protein for humans and livestock worldwide, despite being deficient in several essential amino acids. They are therefore an important target for biofortification efforts. Studies centered on eliminating the highly abundant but poorly balanced seed storage proteins have revealed that the regulation of seed amino acids is complex and does not rely on a handful of proteins. In this study, we used two complementary omics-based approaches to shed new light on the genes and biological processes that underlie the regulation of seed amino acid composition. I first conducted a genome wide association study to identify a list of candidate genes involved in the natural variation of 76 seed protein bound amino acid related traits. I then used a weighted gene co-expression network analysis to associate protein expression with seed amino acid composition dynamics during kernel maturation. I found that in contrast to the rising seed storage proteins almost half of the proteome, was significantly reduced during kernel maturation. That proteome is enriched in the translational machinery components, such as ribosomal proteins, which strongly suggests a translational reprogramming is occurring. The reduction was significantly associated with decreases in several amino acids, including lysine and methionine, pointing to its role in shaping the seed amino acid composition. I compared the candidate gene lists generated with both approaches and found a nonrandom overlap of 80 genes. A functional analysis of these genes showed a tight interconnected cluster dominated by translational machinery genes, especially

ribosomal proteins, lending further support for a role of translation dynamics in shaping seed amino acid composition. The findings strongly suggest that seed biofortification strategies that target the translation machinery dynamics should be considered and explored.

1. Introduction

Cereal grains are an important food source for humans and livestock worldwide. Maize, wheat, and rice account for approximately 70% of total cereal production (Peter R Shewry, 2007; P. R. Shewry & Halford, 2002). Of these three cereals, maize (*Zea mays ssp. mays*) is the most productive crop in terms of yield per acreage; the United States alone produced around 400 million tons in 2018, for example (FAOSTAT, 2018). Maize seeds, or kernels, consist of a large endosperm (~90% of the total dry seed weight) and a small embryo (~10% of the total dry seed weight) (Flint-Garcia, Bodnar, & Scott, 2009; Watson, 2003). Although maize kernels are dominated by carbohydrates—roughly 70% of their composition is starch and about 10% is protein (Flint-Garcia et al., 2009; Watson, 2003; Y. Wu & Messing, 2014). Both humans and livestock, especially in developing countries, rely heavily on maize as a protein source (B. Shen & Roesler, 2017; Shiferaw, Prasanna, Hellin, & Bänziger, 2011). This dietary reliance is problematic since maize kernels are deficient in several essential amino acids (EAA); that is, those amino acids that humans and livestock cannot synthesize in their bodies and must obtain from their diet.

Previous studies have posited that this deficiency in essential amino acids in seed is due to the abundance of seed storage proteins (SSPs). The reason is that SSPs can contribute to more than 60% of a seed's total amino acid composition but are very poor in

several EAAs (Larkins, 2017; Larkins et al., 1984; J. Messing, 1983; Peter R Shewry, 2007; P. R. Shewry & Halford, 2002). The most abundant SSPs in maize kernels are the zeins, which are members of the prolamin group (Boston & Larkins, 2009; Larkins, 2017; Larkins & Hurkman, 1978). During seed development, zeins are deposited in the endosperm in specific protein bodies (Lending & Larkins, 1989). While rich in proline (Pro) and glutamine (Gln), zeins lack several essential amino acids, including lysine (Lys), methionine (Met), and tryptophan (Trp) (Larkins, 2017; Peter R Shewry, 2007). Attempts have been made to boost the EAA composition of kernels by knocking down or out the zeins, but these efforts have not yielded any significant improvements. To the contrary, studies of various zein mutants have revealed that seeds largely maintain their protein levels and protein bound amino acid (PBAA) composition in response to large perturbations to their proteome (Morton, Jia, Zhang, & Holding, 2015; M. Schmidt et al., 2011; Y. Wu, Wang, & Messing, 2012). This natural phenomenon, which is termed proteomic rebalancing, is a highly conserved mechanism in seeds, having been reported not only in maize (Morton et al., 2015) but also in soybean (M. Schmidt et al., 2011), *Arabidopsis* (Withana-Gamage et al., 2013), camelina (M. A. Schmidt & Pendarvis, 2017), and wheat (Altenbach, Tanaka, & Allen, 2014). Neither the underlying molecular mechanism nor the natural function of proteomic rebalancing is understood. Nonetheless, the phenomenon strongly implies a greater complexity to the metabolic regulation of a seed's PBAA levels and composition than just the seed storage protein levels.

Quantitative genetic approaches have proven to be efficacious for interrogating the genetic architecture of complex traits, including PBAAs, and their regulatory genetic mechanisms. Genome wide association studies (GWAS) have proven to be a powerful

tool for uncovering metabolic QTLs (mQTL) that underlie the natural variation of metabolic traits (Angelovici et al., 2013; Deng et al., 2017; Slaten, Yobi, et al., 2020a; S. Wu et al., 2016) and, consequently, facilitating the identification of key regulatory genes as well as targets for marker assisted selection for breeding. A GWAS of seed PBAAs measured from 79 wild soybean accessions, for example, elucidated the genetic basis that underlies the deficiency of essential sulfur amino acids and led to identification of several QTLs involved in the natural variation of aspartic acid (Asp) and Gln (La et al., 2019). Likewise, a large meta-analysis of soybean seed compositional QTLs identified 156 QTLs related to PBAAs, which included a number of potential candidate genes related to their metabolic pathways (Van & McHale, 2017). Surprisingly, very few GWAS of PBAAs measured from maize kernels were conducted. Deng et al. (2017) conducted an extensive association and linkage mapping study of maize PBAAs measured from 513 lines of a Chinese diversity panel that led to the identification of 247 and 281 significant QTLs in two different environments (Deng et al., 2017). The authors subsequently identified additional QTLs via a linkage mapping of three RIL populations, but an in-depth analysis was performed on only three candidate genes.

The large number of candidate genes can be one of the drawbacks of GWAS. Often, it is impractical to validate or prioritize the candidate genes using classical genetic approaches. Further, the candidate genes often are involved in the natural variation of traits in a very specific combination of environments and timepoints and/or developmental stages. While some candidate genes identified by GWAS are key regulatory genes that belong to a specific relevant biological process, it can be hard to

differentiate these genes from other candidate genes relevant only to the specific conditions under which the plants were grown.

An integrative multi-omics approach can help overcome this challenge. Several recent studies have combined GWAS with a co-expression network analysis as a means to efficiently whittle down candidate gene lists to a few novel key regulatory genes involved in shaping the natural variation of a phenotype of interest (Qi et al., 2018; Schaefer et al., 2018; S. Wu et al., 2016). A study in maize, for example, demonstrated the power of integrating co-expression networks with GWAS to prioritize casual genes from a GWAS candidate gene list (Schaefer et al., 2018). The authors cautioned, however, that the success of similar strategies depends heavily on the relevance of the gene expression data to the phenotypes studied. A study in *Arabidopsis*, for example, found numerous functional associations between genes and primary metabolites by combining quantitative genetic mapping with correlation networks of transcripts and metabolites taken from a stress time course study (S. Wu et al., 2016). With this method, the authors were able to identify ~90 candidate associations between structural genes and primary metabolites, some of which had been previously characterized, reinforcing the effectiveness of their strategy. Similarly, a meta-analysis study in soybean integrated metabolic and transcriptomic data, GWAS, and a mapping association to identify genes responsible for seed compositional traits, including amino acids (Qi et al., 2018). In this example, the authors used a weighted gene co-expression network analysis (WGCNA) to find an association between co-expression modules from RNA seq data taken during seed development and a seed compositional trait. The highly connected genes from the associated co-expression modules (i.e., hub genes) were then compared to previously

identified mQTLs for the traits. Several of the hub genes matched the identified mQTLs, supporting their key role in the compositional trait of interest.

It stands to reason that coupling GWAS with a proteomic expression network analysis would be biologically more relevant than gene expression for a trait such as seed PBAA composition, which ultimately reflects proteome dynamics. Such an approach is rarely used, however, since proteomic analyses can be expensive and hard to generate. Here, I took advantage of a relatively affordable, high-throughput shotgun proteomic methodology to overcome this limitation. Thus, in this study, I performed GWAS on PBAAs measured from dry maize kernels taken from a 282-association panel and generated a candidate gene list. I then performed an association analysis between proteomic expression data and PBAA composition quantified from a developmental series of B73 plants to generate an orthogonal candidate gene list. From a comparison of these two gene lists, I identified 80 high confident candidate genes (HCCG) that were both strongly associated with seed PBAA composition during maturation and involved in the natural variation of these traits in dry kernels. A downstream functional analysis and comprehensive ranking of these 80 genes showed that our approach was very efficient in identifying both characterized and, more importantly, previously uncharacterized biological processes involved in seed PBAA composition. This analysis revealed surprising insights about the role that the translational machinery genes, especially ribosomal proteins, play in the seed amino acid composition. I suggest these proteins may be new avenues to explore for seed PBAA biofortification.

2. Materials and Methods

2.1 Germplasm

I obtained phenotypic data from 279 lines of the Goodman-Buckler maize association panel (Flint-Garcia et al., 2005). These lines were grown over the summers of 2017 and 2018 with two replications, each using a randomized complete block design, at the Genetics Farm near Columbia, MO. Briefly, 13 kernels per inbred line were grown in 10 feet rows. Self-pollination was done for every standing plant in a row, and measures were taken to avoid cross contamination. Pollinated ears were harvested at maturity, and cob husks removed. Ears were dried, shelled, and bulked to form representative composite grain samples for each inbred line. Twenty-five random seeds from each inbred bulked sample were ground into fine powder for PBAA quantification.

2.2 PBAA quantification

Seed PBAAs were extracted and detected using a high-throughput absolute quantification protocol for 16 protein-bound amino acids. The protocol combines a microscale protein hydrolysis step and an absolute quantification step using multiple reaction monitoring—based liquid chromatography–tandem mass spectrometry detection, as described in (Yobi & Angelovici, 2018). Due to acid hydrolysis, Gln and Asn were converted to Glu and Asp, respectively; therefore, Glx denotes Gln and Glu, and Asx denotes Asn and Asp. Trp and Cys were destroyed, and Gly was poorly detected. In total, 15 PBAAs were analyzed which represent 17 PBAAs.

2.3 Phenotypic data analysis and GWAS

I evaluated a total of 76 PBAA absolute, relative composition, and biochemical traits. All traits were treated independently. Metabolic ratio traits were derived prior to

calculation of the best linear unbiased predictions (BLUP) to minimize noise. For each trait, the outlier removal, optimal transformation, and BLUP calculation were performed as previously described (Slaten, Chan, Shrestha, Lipka, & Angelovici, 2020). The analysis was performed in R version 3.4.3. Variance component estimates from a mixed linear model, where taxa, replication, and year are fitted as random effects, were used to estimate broad sense heritability on a line mean basis as previously described (Holland, Nyquist, & Cervantes-Martínez, 2003). Ser/Total heritability was 0 – because the BLUP of Ser/Total did not converge to the model – and was therefore removed from the downstream GWAS analysis.

The association panel was previously genotyped with the Illumina MaizeSNP50 BeadChip (Cook et al., 2012) and with genotyping-by-sequencing (Elshire et al., 2011) as described in (Lipka et al., 2013). Single-nucleotide polymorphisms (SNPs) were filtered using a minor allele frequency greater than 0.05, and a total of 422,161 SNPs were used for the GWAS analysis. I used GAPIT (Lipka et al., 2012) mixed linear model (MLM) and FarmCPU (X. Liu, Huang, Fan, Buckler, & Zhang, 2016) to conduct the univariate GWAS. False discovery rate (FDR) (Benjamini & Hochberg, 1995) was used to correct the multiple hypothesis testing problem at 5%. Candidate gene lists were obtained using a 200 kb window size (100 kb on either side) of each significant SNP. The physical locations and annotations of genes were based on Maize AGP_V2 (http://ftp.maizesequence.org/release-5b/filtered-set/).

2.4 Gene functional categorization using MapMan

MapMan version 3.6; (Lohse et al., 2014) was used for the functional categorization of the candidate genes generated by GWAS. The genes were mapped to

corresponding Bins using the *Zea mays* mapping database *Zm_B73_5b_FGS_cds_2012.m02* obtained from https://mapman.gabipd.org/mapmanstore. A total of 35 functional gene categories were in the mapping database.

2.5 PBAA and seed proteome quantification of B73 inbred lines across ten seed filling stages

B73 inbred lines were grown during the summer of 2018 at the Genetics Farm near Columbia, MO. Samples from 10 different seed filling stages were collected, starting at 10 days after pollination (DAP) and then every four days until 46 DAP. The 10 time points collected were 10, 14, 18, 22, 26, 30, 34, 38, 42, and 46 DAP. Three biological replicates were collected from each seed developmental stage (3 Rep * 10 time points =30 samples). The whole ear from each sample was then harvested and frozen with liquid nitrogen, and 15 random developing seeds also were collected and stored immediately in liquid nitrogen. The seeds were lyophilized for 5 days and then finely ground for PBAA quantification.

2.6 Proteomic analysis

Protein extraction was performed based on the (Hurkman & Tanaka, 1986) method as described in (Yobi et al., 2020). Briefly, 5 mg finely ground seed powder were weighed and extracted with Tris-HCl buffered phenol and an SDS extraction buffer. After trypsin digest and purification, the peptides were analyzed on a Bruker timsTOF PRO using the PASEF (1) method. The acquired data were submitted to the PEAKS DB search engine (version 8.5, Bioinformatics Solutions Inc.) for peak picking. Protein

identification was completed using the MaizeGDB database (Lawrence, Dong, Polacco, Seigfried, & Brendel, 2004). Proteins with spectral counts ≥ 4 were retained for analysis.

2.7 WGCNA analysis

I performed a weighted protein correlation network analysis (Langfelder & Horvath, 2007) using the R package WGCNA (Langfelder & Horvath, 2008). I chose the soft threshold power β=12 to construct the co-expression network as the R^2 reached around 80% at ensuring the network was close to the scale-free network. I used Pearson correlation, blockwiseModules function, and the dynamic tree cut algorithm (Langfelder, Zhang, & Horvath, 2007) with a height of 0.20 and a minimum module size of 30. Modules are defined as the branches of the dendrogram.

2.8 GO enrichment analysis

GO enrichment analysis was performed using AgriGO_V2 (Tian et al., 2017). The following parameters were used to determine the GO biological process, cellular component, and molecular function terms that were overrepresented ($p<0.05$): a hypergeometric test with a 5% FDR correction, a custom reference that consisted of 2648 proteins detected in my proteomics study, *Zea mays* as the select organism, and GO full plant ontologies.

2.9 eQTL analysis of the 80 HCCG

I used the gene expression dataset from (Kremling et al., 2018). The authors collected 3' RNA-seq data from seven different tissues (including seeds at 350 growing degree days) from 255 inbred lines of the Goodman-Buckler association panel. I limited my eQTL analysis by using expression datasets specifically from the seed and to 80 high confidence candidate genes (HCCG). I first choose 80 HCCG gene expression levels

from the seed specific dataset and associated them with the corresponding GWAS tag SNPs by using the student t-test followed by an FDR correction at a 0.05 significance level. However, we defined an eQTL as significant only if it explained the variation of the corresponding PBAA related trait. Hence, we performed a Pearson correlation test between the normalized gene expression level and the corresponding PBAA trait from the GWAS.

2.10 Protein-protein interaction of the 80 HCCG using STRING

I constructed and visualized the protein-protein interaction (PPI) network associated with the 80 HCCG using the Search Tool for the Retrieval of Interacting Genes/Proteins database STRING V11.0(Szklarczyk et al., 2019). Active interaction sources, including high-throughput lab experiments, gene co-expression and previous knowledge from curated databases specific to *Zea mays*, were used to construct the PPI network at medium confidence (> 0.4) and high confidence (> 0.7) levels (Szklarczyk et al., 2019). I used the MCL clustering algorithm within STRING (Szklarczyk et al., 2019) to further investigate strong interactions among nodes in the PPI network.

3. Results

3.1 Seed PBAA absolute levels and relative composition displayed different relationships and heritability.

The PBAA content and composition of dry maize kernels was measured from 279 inbred lines that belonged to a 282 line maize diversity panel (Flint-Garcia et al., 2005). Two replicates of this panel were grown in 2017 and again in 2018. Fifteen PBAAs were quantified from the dry kernels. Due to the extraction method, the 15 PBAAs represent 17 PBAAs, since Asn and Gln hydrolyze to Asp and Glu and are denoted as Asx and Glx,

respectively (see Materials and Methods and Yobi et. al 2018) **(Supplemental Data S2.1).**

I found considerable natural variation in most PBAAs, both in terms of absolute levels and in relative composition **(Table S2.1a, b; Figure 2.1a, b).** A relative composition trait is defined as the ratio of an individual amino acid to the sum of the 15 measured amino acids (e.g., Ala/Total, Ile/Total). I categorized the natural variation of these traits into three groups: Group 1 were those with PBAA levels > 10%, Group 2 were those with PBAA levels between 10 and 2%, and Group 3 were with those with PBAA levels < 2% **(Figure 2.1a, b and Table S2.1a, b).** In general, the broad sense heritability of the PBAA absolute traits were moderate to high; the exceptions were Arg and Glx (Glu + Gln), which each showed low heritability (0.05 and 0.33, respectively) **(Table S2.1a).** Interestingly, the broad sense heritability values of many PBAA relative compositions were substantially lower than their absolute levels **(Table S2.1a, b**).

I performed a pairwise Pearson correlation analysis to evaluate the relationship between PBAA absolute levels and relative composition. For absolute PBAA, all pairwise correlations were significant at a qFDR-values <0.05 and were exclusively positive **(Figure 2.1c and Table S2.2a, b)**. The strongest correlation was between Ala and Leu (r = 0.96) absolute levels, whereas the weakest correlation was between Lys and Leu (r = 0.27) absolute levels **(Figure 2.1c and Table S2.2a)**. Notably, both Met, and Lys demonstrated relatively larger number of weaker pairwise correlations with the other PBAAs **(Figure 2.1c and Table S2.2a)**. The pairwise correlation analysis showed a very different pattern for PBAA relative compositions. For this trait, we found both positive and negative correlations and numerous non-significant correlations among the relative

PBAAs **(Figure 2.1d and Table S2.2c, d)**. The strongest positive correlation was between Lys/Total and Arg/Total (r = 0.74), and the strongest negative correlation was between Lys/Total and Leu/Total (r = -0.83) **(Figure 2.1d and Table S2.2c).** In general, only a few PBAA relative compositions, including those of Leu, Lys, Arg, His, and Val, had multiple moderate to strong correlations. Overall, we found that PBAA absolute levels were largely heritable and had strong positive correlations and that PBAA relative compositions were less heritable and had low to moderate negative and positive correlations.

3.2 GWAS of PBAA related traits resulted in 1399 potential candidate genes

To capture the breadth of the genetic architecture underlying this natural variation in PBAAs, I performed a GWAS on 76 PBAA-related traits. These traits included: (1) the absolute levels in nmole/mg, (2) the relative composition of each trait represented as the ratio of each PBAA to the total sum of the PBAAs measured (i.e. Lys/ Total), and (3) metabolic ratios based on the potential relationship within each amino acid metabolic family (e.g., Lys/the Asp family pathway: Ile + Met + Thr + Asx + Lys). A full list of the 76 traits is in **(Table S2.3)**, and the calculated values are elaborated in **(Supplemental Data S2.1)**. I describe a similar use of amino acid derived traits for GWAS in (Angelovici et al., 2016; Deng et al., 2017). For simplicity, I use the one letter code to describe the ratio-related traits; these abbreviations are in **(Table S2.3).**

GWAS were performed using the FarmCPU model (X. Liu et al., 2016) and the GAPIT MLM model (Lipka et al., 2013). Results from the latter model were not significant; therefore, we present only the results from FarmCPU. We used FarmCPU previously to successfully identify key genes involved in free amino acid metabolism in

Arabidopsis (Slaten, Yobi, et al., 2020a). Overall, the model yielded 40 traits (out of the 76) with significant SNP-trait associations at 5% FDR correction (Benjamini & Hochberg, 1995) and 277 unique (i.e., non-redundant) SNPs **(Supplemental Data S2.2).** **(Figure 2.2a)** shows the partitioning of the 277 unique SNPs across the ten maize chromosomes. A visualization of the distribution of the significant SNP-trait associations using a 1 Mb window span showed several potential hotspots on chromosomes 1, 2, 6, 7, and 9 (**Figure S2.1a)**. A PBAA trait categorical summary of the GWAS results by amino acid family is in (**Table 2.1)**.

I extracted 1399 unique candidate genes from 200 kb intervals centered around the significant SNPs identified for the 40 traits (100 kb upstream and 100 kb downstream) **(Supplemental Data S2.2)** as was described in previous studies and to compensate for a low marker coverage (Ching et al., 2002; Flint-Garcia, Thornsberry, & Buckler IV, 2003; J. Yan et al., 2009). I detected six candidate genes that were associated with the highest number of traits and that were also the most significant associations. These genes were extracted from the same SNP and included two SSP proteins, GRMZM2G138689 (50 kD γ-zein), and GRMZM2G138727 (27 kD γ-zein) **(Supplemental Data S2.2)**.

I next asked whether any specific biological processes or pathways were enriched across the 1399 candidate genes. An enrichment analysis using AgriGo (Tian et al., 2017) found no enrichments. I next used MapMan 3.6 (Lohse et al., 2014) to assess the functional categorization of the genes **(Figure 2.2b)** and found that the four top functional categories were protein (13.6%), RNA (10.5%), signaling (4.6%), and transport (4.0%) **(Figure 2.2b)**.

In sum, the GWAS identified 1399 candidate genes. To whittle this list down, I chose to generate an orthogonal candidate gene list with which to compare these genes.

3.3 The dynamics of individual PBAA relative compositions during kernel maturation highlighted opposing patterns.

I generated an orthogonal candidate gene list associated with seed PBAA regulation and homeostasis by performing an association analysis of PBAA composition with protein co-expression during seed maturation. I collected 15 developing B73 kernels from three biological replicas at ten time points; kernels were collected every 4 days, starting 10 days after pollination (DAP) through desiccation (46 DAP-dry kernels). At 10 DAP, kernels transition to maturation and the storage compounds, especially SSPs, accumulate (Larkins, 2017; Sabelli & Larkins, 2009).

For each sample, I quantified the PBAA levels and calculated the relative composition using the same methods described above **(Supplemental Dataset S2.3)**. Hierarchical clustering was used to cluster the trends of the relative compositions across the ten developmental stages **(Figure 2.3)**. I choose to only analyze the trends of relative composition as I reasoned they are the most relevant to compare/associate with protein expression that are measured per equal protein content.

The patterns of PBAA relative composition aligned well with the known seed storage protein composition and accumulation trends, which resulted in elevation of the branched chain amino acid (BCAA) content and reductions in Lys and Met (Gorissen et al., 2018). My analysis of PBAA relative composition trends resulted in four clusters. Cluster I (all the BCAAs and Tyr) was characterized by a gradual but continual elevation in relative composition with a peak at seed desiccation. Cluster II (Ser, Pro, Thr) was

characterized by an initial reduction in relative composition followed by an increase after 26 DAP. Cluster III (Met, Ala, Asx, Glx) was characterized by a continual but gradual decrease in relative composition. And, finally, cluster IV (Arg, Phe, Lys, Gly, His) showed an increase initially (until 18 DAP) followed by a decrease in relative composition **(Figure 2.3a, b)**.

3.4 Five protein co-expression modules were highly associated with PBAA compositional dynamic patterns

In addition to PBAA quantification, I analyzed protein expression levels from each sample using a shotgun proteomic approach (see Materials and Methods). I identified 6361 proteins and then removed those with low spectral counts and poor reproducibility (see Materials and Methods), leaving 2648 good quality proteins **(Supplemental Data S2.4)**. I performed a weighted gene co-expression network analysis (WGCNA) (Langfelder & Horvath, 2008) on these filtered proteins and then constructed an undirected and weighted protein co-expression network using the optimum soft threshold **(Figure S2.2)**. Notably, our digestion and detection method filtered out the highly abundant zeins, allowing us both to get a deeper coverage of the seed proteome and to skip the exclusion of these proteins using protein fractionation. The 2648 proteins were assigned to eight modules: blue, turquoise, brown, green, yellow, black, red, and gray **(Figure 2.4a)**. A list of proteins in each module with their respective annotations is in **(Supplemental Data S2.5).** The turquoise module had the most proteins (853 proteins;), and the black module had the least (66 proteins) **(Figure 2.4a and Supplemental Data S2.5).** Visualization of the expression pattern of eigen protein of a particular module indicated that, as the kernel matured, proteins from the blue and

turquoise modules decreased (45% of all proteins detected) and proteins from the brown and green modules increased **(Figure 2.4b)**. Thus, almost half of all proteins analyzed declined during seed maturation (**Figure 2.4b**). I calculated the Eigengene-based module connectivity, or module membership (kME), for each particular protein within a given module in order to investigate its potentiality to be a candidate for hub gene (kME>0.7). A full list of proteins and their kME within a respective module is in **(Supplemental Data S2.5)**.

Next, I assessed the functional relationships between the seed PBAA relative composition trends and the protein expression patterns across the seed developmental stages. Here, I used WGCNA to perform a correlation between the relative PBAA composition traits and the protein expression modules. To create a modules-PBAA composition association, I calculated the module Eigen protein (ME), the first principal component of a given module. The correlation between the ME of each module and the PBAA composition trait are shown in **(Figure 2.4a).** Five modules (blue, turquoise, brown, green, and black) correlated highly (r_{abs} =~0.7-0.94) with multiple PBAA compositional traits. The gray, yellow, and red modules had mostly low and non-significant correlations with the PBAA traits **(Figure 2.4a).**

I used AgriGO_V2 to carry out a GO enrichment analysis on the proteins identified from the five modules (Tian et al., 2017). Enrichment for the blue, turquoise, brown, and black module proteins is in **(Table S2.4).** I did not find enrichment for proteins in the green module. The most significant enrichments were found for proteins in the turquoise module, and the terms included structural molecule activity, structural constituent of ribosome, ribosomal subunit, and translation (corrected p-values 1.80E-28

- 2.30E-19). The blue module was enriched for the term purine ribonucleoside metabolic process (corrected p-value 0.013) **(Table S2.4).** The most enriched categories for the brown and black modules were response to light intensity (corrected p-value 3.90E-05) and carbohydrate biosynthetic process (corrected p-value 0.00017), respectively (**Table S2.4).** Altogether, I identified 1583 candidate proteins that were highly correlated with the PBAA compositional dynamics during kernel maturation. I converted these candidate proteins into their respective gene ID, providing us with the orthogonal gene list with which to compare to the gene list generated by GWAS.

3.5 Eighty high confidence candidate genes (HCCG) were identified by intersecting the GWAS and WGCNA candidate gene lists

I compared my two orthogonal candidate gene lists and searched for genes that were highly associated with my traits in both approaches. The logic of this approach is as follows: if overlap in the lists is not random, then genes that are both highly associated with PBAA across development and are also part of the genetic architecture of PBAA in the dry kernel are key regulatory genes. A comparison between the GWAS and WGCNA candidate gene lists yielded 80 overlapping genes **(Figure 2.5).** I refer to these as high confidence candidate genes (HCCG) **(Table S2.5).**

I tested my assumption that these genes resulted from a non-random overlap between the GWAS and WGCNA analyses. I used GeneOverlap (L. Shen, 2014) to perform 10,000 overlap simulations. I extracted a random subset of genes without replacement with the same number of genes as the GWAS candidate gene list (i.e., 1399 genes) and overlapped them with the WGCNA gene list. I used the AGP_V2 maize genome annotation, which consists of 39,656 genes. This analysis showed that 80 genes

is a larger number than what would be expected (p=7e-04) using a Fisher's Exact test, which supports my assumption that the overlap was not random.

3.6 A protein-protein interaction network analysis of the 80 HCCG showed a large, tightly interconnected functional cluster containing mostly protein synthesis related genes

I was interested in the potential functional interaction between the 80 HCCG proteins. I constructed and visualized a protein-protein interaction (PPI) network associated with the 80 HCCG by using the Search Tool for the Retrieval of Interacting Genes/Proteins database STRING (V11.0; (Szklarczyk et al., 2019). The PPI network generated by STRING consisted of 80 nodes (42 connected at least to one other protein, and 38 were unconnected; not shown in the figure) and 91 edges with an average node degree of 2.27 **(Figure 2.6a)**. Each node represents a HCCG, and each edge represents the interaction between nodes/HCCG. The number of edges is larger than is expected for a random network of the same size (p-value 6.58e-05), indicating that these interactions are not random. To further investigate the functional interaction among the nodes, I used the MCL clustering algorithm in STRING, which resulted in 11 clusters **(Figure 2.6a)**. The genes and clusters are summarized in (**Table S2.5).** Cluster 1, which contained 19 genes, was the largest and most interconnected functional cluster. The other clusters had only two to three genes. Using MapMan (Lohse et al., 2014) to assess the functional categorization of the 11 clusters **(Figure 2.6a, b and Table S2.5),** we found that the majority of genes in cluster 1 belong to protein synthesis, degradation, and folding. Two genes are categorized as stress genes but could be involved in protein folding as they are chaperons. Strikingly, of these 19 genes in cluster 1, nine were ribosomal gene subunits.

Cluster 2 (green yellow) included three zein storage proteins; cluster 3 (green) included TCA and electron transport chain genes; and the remaining clusters included genes related to carbohydrate, amino acid, lipid metabolism, cell wall degradation, and cell vesical transport **(Figure 2.6b and Table S2.5).**

I wanted to investigate protein and gene expressions from cluster 1, so I performed expression visualization of both **(Figure 2.7a, b).** The proteomic data were the same data used for the WGCNA **(Figure 2.7a).** The gene expression data **(Figure 2.7c)** were extracted from a public dataset (Chen et al., 2014). The majority of cluster 1 proteins were highly expressed at 10 DAP and then gradually but continually decreased throughout the remainder of seed development **(Figure 2.7a, b).** The exceptions were two heat shock proteins that demonstrated the opposite trend **(Figure 2.7a).** I also visualized the transcript expression patterns of the same genes using a public dataset (Chen et al., 2014) that contains the maturity timepoints up until 38 DAP **(Figure 2.7c).** Interestingly, the expression patterns of most proteins diverged from the gene expression patterns one, toward late maturation and desiccation. While proteins showed a consistent decrease, gene expression was elevated toward these stages **(Figure 2.7c, d)**. The latter observation could be used to infer that transcript elevation is not manifested in translation of these proteins **(Figure 2.7e).**

To further investigate expression levels of the cluster 1 genes, I used a heatmap to visualize the gene expression dataset from (Stelpflug et al., 2016) at various developmental stages, including whole seed, endosperm, embryo, ear, tassel, internode, and leaf **(Figure S2.3).** Overall, I found gene expression was higher in seeds and reproductive tissues as compared to the vegetative tissues. I also found high gene

expression in the embryo (in most of the stages), endosperm (particularly in the early stages from 12 to 14 DAP), whole seed, ear, and tassel. Gene expression was relatively low in the different leaf stages; however, expression in the internode was higher relative to the leaf **(Figure S2.3)**. The latter pattern indicates that these ribosomal proteins are not housekeeping genes and are extensively differentially expressed across the various tissues, most especially in seeds.

3.7 eQTL-mQTL analysis indicate that only few HCCG are driven by expression

I evaluated whether my identified mQTLs are driven, at least in part, by significantly different expression levels in the seeds. I used 3' RNA-sequencing data from maize developing seeds (350 growing degree DAP) collected from 255 inbred lines of the Goodman-Buckler association panel, as previously published in (Kremling et al., 2018). Out of the 80 HCCG, 18 (22.5%) had significant associations between their expression levels and their corresponding GWAS tag SNP locus and, hence, were potential eQTL candidates **(Table S2.6a).** However, I was interested in whether these eQTLs explained the variation in the corresponding PBAA-related traits. Hence, I did a Pearson correlation test between the normalized gene expression levels of those 18 genes and their corresponding PBAA traits from the GWAS **(Table S2.6b).** Out of these 18, only four genes showed a significant correlation at a 5% level of significance **(Table S2.6b)**. These four eQTLs/mQTLs were significantly correlated with 11 PBAA traits **(Table S2.6b).** Interestingly, one gene, Glutelin 2 (GRMZM2G138727), associated with eight PBAA traits (i.e., H/M, H/Z, Z/ZHPR, H/Total, L/IVL, V/A, V/LAV, V/Total; **Table S2.6b)**. Notably, most correlations, although significant, were low **(Table S2.6b)**, which we infer is a consequence of using data coming from two independents

experiments or from additional biological factors at play. In sum, the eQTL/mQTL analysis suggests that most of the mQTLs I identified are not driven by expression variation.

3.8 Ranking the 80 HCCG by comprehensively integrating all performed analyses

Finally, I prioritized and ranked the 80 HCCG by integrating data from all analyses performed. I interrogated each of the 80 genes using the following five criteria: 1) Does it have a significant association (based on the GWAS analysis) with multiple traits? 2) Does it have a high WGCNA-kME, which represents the connectivity of a given protein with a module eigengene (kME> 0.7) 3) Does it have an mQTL that is driven by gene expression (mQTL/eQTL)? 4) Does it have a STRING analysis connectivity > 0.4, which represents the inclusion in the PPI network; and 5) Does it have a STRING analysis connectivity > 0.7, which represents high connectivity within the PPI network. Each gene was given a score, from 1 to 5, that reflected how many criteria it fulfilled. I then ranked the 80 genes according to their score (high to low). Genes with the same score were further ranked by the number of traits associated with it in the GWAS (the more traits, the higher the rank) (**Supplemental Data S2.6).** I used the following logic as the basis for the five criteria: genes that are associated with multiple traits in the GWAS and/or are highly connected in either the co-expression (WGCNA) analysis or the functional analysis (STRING) and/or showing eQTL/mQTL are key genes that are involved in the PBAA composition in seeds.

The top 15 genes are summarized in **(Table 2.2).** GRMZM2G138727 (27kD γ-zein) met all five criteria and was the top ranked gene, followed by GRMZM2G058760 (ferredoxin NADP reductase1-fnr1) and GRMZM2G138689 (50kD γ-zein). **(Table 2.2)**.

27kD γ-zein and 50kD γ-zein are the two important genes that were previously reported to be involved in PBAA composition in seed (Deng et al., 2017; Guo et al., 2013), confirming the effectiveness of my ranking approach. Four ribosomal genes were also ranked among the top 15 HCCG as was elongation initiation factor 3 (eIF3) and 26S protease regulatory subunits **(Table 2.2),** indicating that protein metabolism synthesis and degradation machinery, especially the translational machinery, must play a key role in shaping and regulating PBAAs in maize kernels. Three genes, which included eIF 3, Zein 2 (16 kD zein), and a 40S ribosomal subunit, were responsible for Met-related traits. While a different ribosomal protein was associated with Lys related trait. Both Lys and Met are two amino acids that are deficient in maize. Many of the top-ranked genes were associated with Arg, Met, His, and Lys (minor amino acids) or BCAA (major amino acids) related traits.

4. Discussion

The development of effective seed amino acid biofortification strategies requires a fundamental understanding of the mechanism that underlies PBAA composition. I have learned from multiple studies that have targeted SSPs in various plant species that the genetic and metabolic bases of seed composition are interconnected and complex and do not rely merely on the expression of specific proteins (Larkins & Hurkman, 1978; Joachim Messing, 1983; Morton et al., 2015). In this study, I used and integrated two omics approaches to analyze the genetic basis of PBAA natural variation and its association with the compositional dynamics during seed maturation. My results provide new and strong evidence that the structural components of the translational machinery play a key role in seed PBAA regulation.

4.1 Functional categorization of candidate genes resulting from GWAS of PBAA-related traits highlights proteins and RNA metabolism.

The phenomenon of proteomic rebalancing makes altering seed PBAA composition via mutation difficult (Morton et al., 2015; Y. Wu & Messing, 2014). Nevertheless, we know that seed PBAA composition is a complex trait that varies across natural and artificial populations of maize and other crops (Deng et al., 2017; La et al., 2019). Indeed, my study found that both PBAA absolute levels (nmol/mg) and PBAA relative composition (PBAA/TPBAA) show significant natural variation **(Figure 2.1a, b).** Interestingly, a correlation analysis of both these traits across a large association panel showed a strong and significant positive correlation between PBAA absolute levels, but significant correlations, both positive and negative, for only a handful of PBAA relative composition levels **(Figure 2.1b).** I interpret this result to mean that the high coordination observed in the overall amino acid absolute levels results from a general change in the overall protein levels in the different genotypes as opposed to the relative composition, which itself more likely reflects the natural variation of the proteomic composition in the seeds. For example, the relative compositions of Lys and Leu demonstrate a strong negative correlation, which very likely reflects variation in the abundance of SSPs, which are high in Leu but poor in Lys (Gorissen et al., 2018). My analysis also found a high positive correlation between the natural variation of Lys and Asx, which aligns with a previous study where proteomic perturbations that led to high levels of Lys also led to high levels of Asx (Hunter et al., 2002).

These findings support the notion that different PBAA-related ratios represent different aspects of the relationship between seed PBAAs. Therefore, we included in my

GWAS all relevant metabolic ratios of PBAA based on their biochemical affiliations and interpreted the resultant candidate genes to be part of the comprehensive genetic architecture of PBAAs. A similar approach has proven to be very effective in other metabolic studies in maize and *Arabidopsis* (Angelovici et al., 2013; Deng et al., 2017; Slaten, Yobi, et al., 2020b). My GWAS and candidate gene extraction approach yielded a relatively large number of unique candidate genes for all PBAA-related traits, most likely due to the relatively permissive statistical correction used for my analysis. However, since I intended to intersect the GWAS candidate gene list with an orthogonal candidate gene list, the length of the list did not pose a concern. Nevertheless, I conducted a functional analysis of these candidate genes and found that both "proteins" and "RNA" were the largest identified categories **(Figure 2.2b)**. This finding suggests that protein metabolism and gene expression may lie at the heart of seed PBAA.

4.2 An increase in several PBAA relative composition during seed maturation is negatively correlated with multiple translational machinery components.

I intersected my GWAS with proteomic expression data. This combination is rarely implemented because of the cost and difficulty associated with proteomic analyses. Similar multi-omics studies have used transcriptomic expression data. However, I reasoned that the association between overall seed PBAA composition dynamics during maturation and proteome expression patterns would yield a more direct and biologically relevant findings. From my analysis, I learned that the relative composition levels of Leu, Ile, Val, Tyr, and Pro were elevated toward seed desiccation **(Figure 2.3)** and were positively associated with three proteomic co-expression modules (brown, green, and black) that also gradual increased with kernel maturation **(Figure 2.4a, b)**. In contrast,

Ala, Asx, Glx, Gly, His, and Lys relative compositions decreased during kernel maturation and were strongly associated with two proteins co-expression modules (blue and turquoise) that also gradually decreased **(Figure 2.4a, b).** These PBAA composition dynamics are consistent with the elevation of SSPs that are poor in Lys and rich in Leu and Pro (Gorissen et al., 2018; Hunter et al., 2002). My analysis further revealed that 45% of the detected proteins were reduced during kernel maturation, while only 14% increased (**Figure 2.4a, b)**. This large reduction in proteins undoubtedly contributed to the PBAA composition in the dry seed. The effect of this reduction on PBAA composition might be on par with the elevation of SSPs; however, since my proteomic analysis was semi-quantitative, a future study is needed to confirm this claim. Nonetheless, this finding should be of great interest to seed amino acid biofortification efforts since it suggests that the proteins to target may be the reduced ones, and not the abundant SSPs.

The large reduction of proteins during kernel development and maturation coincided with two major processes. The first is the accumulation of SSPs, which starts around 10 DAP and increases with seed maturity (Larkins, 2017; Sabelli & Larkins, 2009). The second is a programmed cell death (PCD) that is initiated around 12-16 DAP and expands to engulf the entire starchy endosperm by late development (Young, Gallie, & DeMason, 1997), which lead to the degradation of many proteins in the endosperm. It is interesting then that the proteins included in the two protein expression modules (turquoise and blue module) that showed reductions toward desiccation were significantly enriched highly in biological processes related to translation and gene expression **(Table S2.4)**. It stands to reason that many gene expression related proteins are downregulated

during maturation as a result of PCD and in preparation for desiccation and dormancy. The large reduction in proteins related to translation is, nonetheless, surprising since this period is characterized by massive SSP synthesis and accumulation (Larkins, 2017; Prioul et al., 2008; Sabelli & Larkins, 2009). Even more striking is that a large portion of these proteins are ribosomal proteins which are associated with protein synthesis **(Supplemental Data S2.5; turquoise module).** These results could indicate that specific types of ribosomes are associated with the translation of SSPs and that this reduction may be part of a global translational reprogramming or some sort of translational switch. Global translational control has been reported for stresses that produce energy deficits, including hypoxia (Branco-Price, Kawaguchi, Ferreira, & Bailey-Serres, 2005; Branco-Price, Kaiser, Jang, Larive, & Bailey-Serres, 2008), heat (Yangueez, Castro-Sanz, Fernandez-Bautista, Oliveros, & Castellano, 2013), and drought (Kawaguchi & Bailey-Serres, 2002; Kawaguchi, Girke, Bray, & Bailey-Serres, 2004). Selective translational regulation also has been associated with dark/light transition (Bailey-Serres & Juntawong, 2012), photomorphogenesis (M.-J. Liu et al., 2013), daily clock cycle (Missra et al., 2015), and symbiosis with nitrogen fixing bacteria (Reynoso, Blanco, Bailey-Serres, Crespi, & Zanetti, 2013). Consistently, a study by Shamimuzzaman & Vodkin (2014) that found an increase of 64 ribosomal proteins during the transition of soybean cotyledons from storage organs to photosynthetic organs (4 days after seed imbibition and germination), supporting an involvement of specific ribosomal changes in developmental transitions (Shamimuzzaman & Vodkin, 2014). In sum, I infer from the results of my association analysis of proteomic expression and seed PBAA that a

translational reprogramming is likely to occurs during seed maturation which play a key role in shaping the PBAA composition.

4.3 An analysis of HCCG relationships reveals a tightly connected cluster dominated by translational machinery components.

I used two orthogonal approaches to pinpoint the genetic basis of PBAA composition in seed. I reasoned that (1) genes that are associated in both analyses are key regulatory genes determining PBAA composition in seeds; and (2) that the overlap in genes between the two approaches is not random. My multi-omics approach yielded 80 high confidence candidate genes (HCCG), which my statistical analyses supports my non-random assumption.

A protein-protein interaction analysis of the 80 HCCG found several biological processes previously implicated in seed PBAA composition as well as biological processes not previously reported. The previously characterized processes were storage proteins, transport and amino acids, lipid, and energy and carbohydrate metabolism. The novel processes were related to translational machinery, especially ribosomal proteins **(Figure 2.6a).** Lipid, carbohydrate, and energy metabolism affect levels of PBAA in crop seeds (Deng et al., 2017; M. Jia et al., 2013; Miclaus, Wu, Xu, Dooner, & Messing, 2011), while cell transport, select amino acid metabolic pathways and storage proteins affect PBAA composition (Hunter et al., 2002; Morton et al., 2015; Pandurangan et al., 2012; Wang & Larkins, 2001). Interestingly, my GWAS identified only three γ-zeins **(Supplemental Data S2.2),** notably the same three zeins also identified as part of my top ranked HCCG list **(Table 2.2).** Among these three zeins, Glutelin2 ranked highest. This ranking was consistent with its previous detection in a GWAS of PBAAs measured from

a different maize association panel (Deng et al., 2017), highlighting its general importance in the natural variation of seed PBAA. This protein is also involved in the initiation of SSP protein bodies (Guo et al., 2013). Together, these results strongly support the validity of my multi-omics approach as well as the integrative scoring method I used to rank my HCCG list.

Surprisingly, the functional analysis of the 80 HCCG highlighted the role that core protein metabolism, especially ribosomal proteins, plays in seed PBAA composition. The PPI analysis showed that the 80 HCCG included only one large tightly connected group dominated by genes related to protein metabolism, where more than half were components of the translational machinery, with nine ribosomal proteins and one translational initiation factor EIF3. The remaining proteins included four protein chaperons, two involved in protein degradation, and one vesicle–fusing ATPase **(Table S2.5).** Four of the ribosomal genes and EIF3 are in my top 15 HCCG list **(Table 2.2).** An analysis of publicly available transcriptional data of the genes in cluster 1 indicated that these genes are differentially expressed and have relatively high expression levels in reproductive tissues, mainly endosperm and embryo **(Figure S2.3).** Hence, I reason that a key factor shaping seed PBAA is specific translational attenuation, most probably driven by alteration to specific components of the translational machinery, namely the ribosomal proteins. Studies over the past two decades have shed light on the adaptive nature of the ribosomal proteins and their potential selective regulation of translation. Multiple ribosomal protein mutants, for example, have revealed a functional role for these proteins in developmental and stress related processes (Ma & Dooner, 2004; Tzafrir et al., 2004; H. Yan et al., 2016; J. Zhang et al., 2016). Many ribosomal proteins are transcriptionally

regulated during stress, such as in sucrose feeding (Gamm et al., 2014), cold, heat, and UV-B (Sáez-Vásquez & Delseny, 2019; Sormani, Masclaux-Daubresse, Daniele-Vedele, & Chardon, 2011) as reviewed in (Martinez-Seidel, Beine-Golovchuk, Hsieh, & Kopka, 2020). We also know from ribosomal profiling of stem cells that ribosomal heterogeneity at the level of core ribosomal proteins facilitates preferential translation of specific mRNAs (Shi et al., 2017). Most recently, it was proposed that plant evolution directing high ribosomal proteins paralogue divergence toward functional heterogeneity (Martinez-Seidel et al., 2020) .

Previous studies have established that proteomic reprogramming and rebalancing are due, in large part, to translational regulation rather than transcriptional regulation (Morton et al., 2015; M. Schmidt et al., 2011). For example, proteomic characterization of the RDM4 mutant, which exhibits proteomic reprogramming and rebalancing, found enrichment for protein biosynthesis, especially ribosomal biogenesis, in the endosperm for proteins that increased (S. Jia et al., 2020). These previous observations align with a key role of ribosomal proteins in seed proteome and seed PBAA composition. An additional finding that supports this hypothesis is the identification of eIF3 as a HCCG. eIF3, which consists of 12 subunits, is the most complex and largest initiation factor, participating in nearly all major steps of translation initiation (Browning & Bailey-Serres, 2015; Merchante, Stepanova, & Alonso, 2017). Alterations to eIFs can have drastic effects on translation, either by global repression or upregulation (Merchante et al., 2017). Mutants of several subunits of the eIF3 complex in *Arabidopsis* have been shown to affect the translation of genes in specific processes (Merchante et al., 2017). Notably, eIF protein levels, especially eIF2, are enhanced in several maize opaque mutants that

undergo rebalancing but still display elevated levels of Lys (Habben, Kirleis, & Larkins, 1993; Habben, Moro, Hunter, Hamaker, & Larkins, 1995; M. Jia et al., 2013)

5. Conclusion

This study provides further support for the effectiveness of using an integrative multi-omics approach to shed new light on the biological processes involved in regulation of complex metabolic traits in plants, here seed PBAA composition in maize. With this approach and by using publicly available datasets, I uncovered two uncharacterized factors associated with seed PBAA: a translational reprogramming during seed maturation and the complex dynamic and heterogeneity of the translational machinery, especially ribosomal proteins. I propose that the role of ribosomal proteins and the translational dynamic in seed PBAA composition represent new and exciting avenues for seed amino acid biofortification that might be the key for overcoming proteomic reprogramming and rebalancing.

6. References

Altenbach, S. B., Tanaka, C. K., & Allen, P. V. (2014). Quantitative proteomic analysis of wheat grain proteins reveals differential effects of silencing of omega-5 gliadin genes in transgenic lines. *Journal of cereal science, 59*(2), 118-125.

Amir, R., Galili, G., & Cohen, H. (2018). The metabolic roles of free amino acids during seed development. *Plant science*.

Angelovici, R., Batushansky, A., Deason, N., Gonzalez-Jorge, S., Gore, M. A., Fait, A., & DellaPenna, D. (2016). Network-guided GWAS improves identification of genes affecting free amino acids. *Plant physiology*, pp. 01287.02016.

Angelovici, R., Fait, A., Fernie, A. R., & Galili, G. (2011). A seed high-lysine trait is negatively associated with the TCA cycle and slows down Arabidopsis seed germination. *New Phytologist, 189*(1), 148-159.

Angelovici, R., Fait, A., Zhu, X., Szymanski, J., Feldmesser, E., Fernie, A. R., & Galili, G. (2009). Deciphering transcriptional and metabolic networks associated with lysine metabolism during Arabidopsis seed development. *Plant physiology, 151*(4), 2058-2072.

Angelovici, R., Galili, G., Fernie, A. R., & Fait, A. (2010). Seed desiccation: a bridge between maturation and germination. *Trends Plant Sci, 15*(4), 211-218.

Angelovici, R., Lipka, A. E., Deason, N., Gonzalez-Jorge, S., Lin, H., Cepela, J., . . . DellaPenna, D. (2013). Genome-wide analysis of branched-chain amino acid levels in Arabidopsis seeds. *Plant Cell, 25*(12), 4827-4843.

Bailey-Serres, J., & Juntawong, P. (2012). Dynamic light regulation of translation status in Arabidopsis thaliana. *Frontiers in plant science, 3*, 66.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bewley, J. D. (1997). Seed germination and dormancy. *Plant Cell, 9*(7), 1055-1066.

Binder, S. (2010). Branched-chain amino acid metabolism in Arabidopsis thaliana. *The Arabidopsis Book/American Society of Plant Biologists, 8*.

Boston, R. S., & Larkins, B. A. (2009). The genetics and biochemistry of maize zein storage proteins. In *Handbook of Maize* (pp. 715-730): Springer.

Branco-Price, C., Kawaguchi, R., Ferreira, R. B., & Bailey-Serres, J. (2005). Genome-wide analysis of transcript abundance and translation in Arabidopsis seedlings subjected to oxygen deprivation. *Annals of Botany, 96*(4), 647-660.

Branco-Price, C., Kaiser, K. A., Jang, C. J., Larive, C. K., & Bailey-Serres, J. (2008). Selective mRNA translation coordinates energetic and metabolic adjustments to cellular oxygen deprivation and reoxygenation in Arabidopsis thaliana. *The Plant Journal, 56*(5), 743-755.

Browning, K. S., & Bailey-Serres, J. (2015). Mechanism of cytoplasmic mRNA translation. *The Arabidopsis Book/American Society of Plant Biologists, 13*.

Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A., & Lai, J. (2014). Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant physiology, 166*(1), 252-264.

Ching, A., Caldwell, K. S., Jung, M., Dolan, M., Smith, O. S. H., Tingey, S., . . . Rafalski, A. J. (2002). SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC genetics, 3*(1), 19.

Cook, J. P., McMullen, M. D., Holland, J. B., Tian, F., Bradbury, P., Ross-Ibarra, J., . . . Flint-Garcia, S. A. (2012). Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant physiology, 158*(2), 824-834.

Crow, J. F., & Kermicle, J. (2002). Oliver Nelson and quality protein maize. *Genetics, 160*(3), 819-821.

Deng, M., Li, D., Luo, J., Xiao, Y., Liu, H., Pan, Q., . . . Yan, J. (2017). The genetic architecture of amino acids dissection by association and linkage analysis in maize. *Plant biotechnology journal, 15*(10), 1250-1263.

Dizigan, M. A., Kelly, R. A., Voyles, D. A., Luethy, M. H., Malvar, T. M., & Malloy, K. P. (2007). High lysine maize compositions and event LY038 maize plants. In: Google Patents.

Dong, J., & Horvath, S. (2007). Understanding network concepts in modules. *BMC systems biology, 1*(1), 24.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One, 6*(5), e19379.

Fait, A., Angelovici, R., Less, H., Ohad, I., Urbanczyk-Wochniak, E., Fernie, A. R., & Galili, G. (2006). Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. *Plant physiology, 142*(3), 839-854.

Fait, A., Fromm, H., Walter, D., Galili, G., & Fernie, A. R. (2008). Highway or byway: the metabolic role of the GABA shunt in plants. *Trends Plant Sci, 13*(1), 14-19.

Fait, A., Nesi, A. N., Angelovici, R., Lehmann, M., Pham, P. A., Song, L., . . . Fernie, A. R. (2011). Targeted enhancement of glutamate-to-γ-aminobutyrate conversion in Arabidopsis seeds affects carbon-nitrogen balance and storage reserves in a development-dependent manner. *Plant physiology, 157*(3), 1026-1042.

FAOSTAT. (2018). Retrieved from http://www.fao.org/faostat/en/#rankings/countries_by_commodity

Flint-Garcia, S. A., Bodnar, A. L., & Scott, M. P. (2009). Wide variability in kernel composition, seed characteristics, and zein profiles among diverse maize inbreds, landraces, and teosinte. *Theoretical and Applied Genetics, 119*(6), 1129-1142.

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler IV, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual review of plant biology, 54*(1), 357-374.

Flint-Garcia, S. A., Thuillet, A. C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., . . . Buckler, E. S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal, 44*(6), 1054-1064.

Galili, G., & Amir, R. (2013). Fortifying plants with the essential amino acids lysine and methionine to improve nutritional quality. *Plant biotechnology journal, 11*(2), 211-222.

Galili, G., Avin-Wittenberg, T., Angelovici, R., & Fernie, A. R. (2014). The role of photosynthesis and amino acid metabolism in the energy status during seed development. *Frontiers in plant science, 5*, 447.

Galili, G., & Hofgen, R. (2002). Metabolic engineering of amino acids and storage proteins in plants. *Metab Eng, 4*(1), 3-11. doi:10.1006/mben.2001.0203

Gamm, M., Peviani, A., Honsel, A., Snel, B., Smeekens, S., & Hanson, J. (2014). Increased sucrose levels mediate selective mRNA translation in Arabidopsis. *BMC plant biology, 14*(1), 306.

Gebauer, F., & Hentze, M. W. (2004). Molecular mechanisms of translational control. *Nature reviews Molecular cell biology, 5*(10), 827-835.

Geetha, K., Lending, C. R., Lopes, M. A., Wallace, J. C., & Larkins, B. A. (1991). opaque-2 modifiers increase gamma-zein synthesis and alter its spatial distribution in maize endosperm. *Plant Cell, 3*(11), 1207-1219.

Gorissen, S. H., Crombag, J. J., Senden, J. M., Waterval, W. H., Bierau, J., Verdijk, L. B., & van Loon, L. J. (2018). Protein content and amino acid composition of commercially available plant-based protein isolates. *Amino acids, 50*(12), 1685-1695.

Gu, L., Jones, A. D., & Last, R. L. (2010). Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant. *The Plant Journal, 61*(4), 579-590.

Guo, X., Yuan, L., Chen, H., Sato, S. J., Clemente, T. E., & Holding, D. R. (2013). Nonredundant function of zeins and their correct stoichiometric ratio drive protein body formation in maize endosperm. *Plant physiology, 162*(3), 1359-1369.

Habben, J. E., Kirleis, A. W., & Larkins, B. A. (1993). The origin of lysine-containing proteins in opaque-2 maize endosperm. *Plant molecular biology, 23*(4), 825-838.

Habben, J. E., Moro, G. L., Hunter, B. G., Hamaker, B. R., & Larkins, B. A. (1995). Elongation factor 1 alpha concentration is highly correlated with the lysine content of maize endosperm. *Proceedings of the National Academy of Sciences, 92*(19), 8640-8644.

Hernandez-Sebastia, C., Marsolais, F., Saravitz, C., Israel, D., Dewey, R. E., & Huber, S. C. (2005). Free amino acid profiles suggest a possible role for asparagine in the control of

storage-product accumulation in developing seeds of low- and high-protein soybean lines. *J Exp Bot, 56*(417), 1951-1963. doi:10.1093/jxb/eri191

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

Hunter, B. G., Beatty, M. K., Singletary, G. W., Hamaker, B. R., Dilkes, B. P., Larkins, B. A., & Jung, R. (2002). Maize opaque endosperm mutations create extensive changes in patterns of gene expression. *Plant Cell, 14*(10), 2591-2612.

Hurkman, W. J., & Tanaka, C. K. (1986). Solubilization of plant membrane proteins for analysis by two-dimensional gel electrophoresis. *Plant Physiol, 81*(3), 802-806.

Jia, M., Wu, H., Clay, K. L., Jung, R., Larkins, B. A., & Gibbon, B. C. (2013). Identification and characterization of lysine-rich proteins and starch biosynthesis genes in the opaque2mutant by transcriptional and proteomic analysis. *BMC plant biology, 13*(1), 60.

Jia, S., Yobi, A., Naldrett, M. J., Alvarez, S., Angelovici, R., Zhang, C., & Holding, D. R. (2020). Deletion of maize RDM4 suggests a role in endosperm maturation as well as vegetative and stress-responsive growth. *Journal of Experimental Botany*.

Kawaguchi, R., & Bailey-Serres, J. (2002). Regulation of translational initiation in plants. *Current opinion in plant biology, 5*(5), 460-465.

Kawaguchi, R., Girke, T., Bray, E. A., & Bailey-Serres, J. (2004). Differential mRNA translation contributes to gene regulation under non-stress and dehydration stress conditions in Arabidopsis thaliana. *The Plant Journal, 38*(5), 823-839.

Kim, H., Hirai, M. Y., Hayashi, H., Chino, M., Naito, S., & Fujiwara, T. (1999). Role of O-acetyl-L-serine in the coordinated regulation of the expression of a soybean seed storage-protein gene by sulfur and nitrogen nutrition. *Planta, 209*(3), 282-289.

Kim, W.-S., Chronis, D., Juergens, M., Schroeder, A. C., Hyun, S. W., Jez, J. M., & Krishnan, H. B. (2012). Transgenic soybean plants overexpressing O-acetylserine sulfhydrylase accumulate enhanced levels of cysteine and Bowman–Birk protease inhibitor in seeds. *Planta, 235*(1), 13-23.

Kodrzycki, R., Boston, R. S., & Larkins, B. A. (1989). The opaque-2 mutation of maize differentially reduces zein gene transcription. *Plant Cell, 1*(1), 105-114.

Kremling, K. A., Chen, S.-Y., Su, M.-H., Lepak, N. K., Romay, M. C., Swarts, K. L., . . . Buckler, E. S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature, 555*(7697), 520.

La, T., Large, E., Taliercio, E., Song, Q., Gillman, J. D., Xu, D., . . . Scaboo, A. (2019). Characterization of select wild soybean accessions in the USDA germplasm collection for seed composition and agronomic traits. *Crop Science, 59*(1), 233-251.

Langfelder, P., & Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology, 1*(1), 54.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics, 9*(1), 559.

Langfelder, P., Zhang, B., & Horvath, S. (2007). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics, 24*(5), 719-720.

Larkins, B. A. (2017). *Maize Kernel Development*: CABI.

Larkins, B. A., & Hurkman, W. J. (1978). Synthesis and deposition of zein in protein bodies of maize endosperm. *Plant physiology, 62*(2), 256-263.

Larkins, B. A., Pedersen, K., Marks, M. D., & Wilson, D. R. (1984). The zein proteins of maize endosperm. *Trends in Biochemical Sciences, 9*(7), 306-308.

Lawrence, C. J., Dong, Q., Polacco, M. L., Seigfried, T. E., & Brendel, V. (2004). MaizeGDB, the community database for maize genetics and genomics. *Nucleic acids research, 32*(suppl_1), D393-D397.

Lea, P. J., Sodek, L., Parry, M. A., Shewry, P. R., & Halford, N. G. (2007). Asparagine in plants. *Annals of Applied Biology, 150*(1), 1-26.

Lending, C. R., & Larkins, B. A. (1989). Changes in the zein composition of protein bodies during maize endosperm development. *Plant Cell, 1*(10), 1011-1023.

Lipka, A. E., Gore, M. A., Magallanes-Lundback, M., Mesberg, A., Lin, H., Tiede, T., . . . Rocheford, T. (2013). Genome-wide association study and pathway level analysis of tocochromanol levels in maize grain. *G3: Genes, Genomes, Genetics*, g3. 113.006148.

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

Liu, M.-J., Wu, S.-H., Wu, J.-F., Lin, W.-D., Wu, Y.-C., Tsai, T.-Y., . . . Wu, S.-H. (2013). Translational landscape of photomorphogenic Arabidopsis. *Plant Cell, 25*(10), 3699-3710.

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS genetics, 12*(2), e1005767.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., . . . Usadel, B. (2014). M ercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, cell & environment, 37*(5), 1250-1258.

Ma, Z., & Dooner, H. K. (2004). A mutation in the nuclear-encoded plastid ribosomal protein S9 leads to early embryo lethality in maize. *The Plant Journal, 37*(1), 92-103.

Martinez-Seidel, F., Beine-Golovchuk, O., Hsieh, Y.-C., & Kopka, J. (2020). Systematic review of plant ribosome heterogeneity and specialization. *Frontiers in plant science, 11*, 948.

Merchante, C., Hu, Q., Heber, S., Alonso, J., & Stepanova, A. N. (2016). A ribosome footprinting protocol for plants. *Bio Protoc, 6*, e1985.

Merchante, C., Stepanova, A. N., & Alonso, J. M. (2017). Translation regulation in plants: an interesting past, an exciting present and a promising future. *The Plant Journal, 90*(4), 628-653.

Messing, J. (1983). The manipulation of zein genes to improve the nutritional value of corn. *Trends Biotechnol, 1*, 54-59.

Messing, J. (1983). The manipulation of zein genes to improve the nutritional value of corn. *Trends in biotechnology, 1*(2), 54-59.

Miclaus, M., Wu, Y., Xu, J.-H., Dooner, H. K., & Messing, J. (2011). The maize high-lysine mutant opaque7 is defective in an acyl-CoA synthetase-like protein. *Genetics, 189*(4), 1271-1280.

Missra, A., Ernest, B., Lohoff, T., Jia, Q., Satterlee, J., Ke, K., & von Arnim, A. G. (2015). The circadian clock modulates global daily cycles of mRNA ribosome loading. *Plant Cell, 27*(9), 2582-2599.

Morton, K. J., Jia, S., Zhang, C., & Holding, D. R. (2015). Proteomic profiling of maize opaque endosperm mutants reveals selective accumulation of lysine-enriched proteins. *Journal of Experimental Botany, 67*(5), 1381-1396.

Muehlbauer, G., Gengenbach, B., Somers, D., & Donovan, C. (1994). Genetic and amino-acid analysis of two maize threonine-overproducing, lysine-insensitive aspartate kinase mutants. *Theoretical and Applied Genetics, 89*(6), 767-774.

Pandurangan, S., Pajak, A., Molnar, S. J., Cober, E. R., Dhaubhadel, S., Hernández-Sebastià, C., . . . Marsolais, F. (2012). Relationship between asparagine metabolism and protein concentration in soybean seed. *Journal of Experimental Botany, 63*(8), 3173-3184.

Pitchers, W. R., Nye, J., Márquez, E. J., Kowalski, A., Dworkin, I., & Houle, D. (2017). The power of a multivariate approach to genome-wide association studies: an example with Drosophila melanogaster wing shape. *bioRxiv*, 108308.

Prioul, J. L., Méchin, V., Lessard, P., Thévenot, C., Grimmer, M., Chateau-Joubert, S., . . . Murigneux, A. (2008). A joint transcriptomic, proteomic and metabolic analysis of

maize endosperm development and starch filling. *Plant biotechnology journal, 6*(9), 855-869.

Qi, Z., Zhang, Z., Wang, Z., Yu, J., Qin, H., Mao, X., . . . Zhu, R. (2018). Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development. *Plant, cell & environment, 41*(9), 2109-2127.

Qin, J., Shi, A., Song, Q., Li, S., Wang, F., Cao, Y., . . . Zhang, M. (2019). Genome Wide Association Study and Genomic Selection of Amino Acid Concentrations in Soybean Seeds. *Frontiers in plant science, 10*, 1445.

Reynoso, M. A., Blanco, F. A., Bailey-Serres, J., Crespi, M., & Zanetti, M. E. (2013). Selective recruitment of m RNA s and mi RNA s to polyribosomes in response to rhizobia infection in M edicago truncatula. *The Plant Journal, 73*(2), 289-301.

Sabelli, P. A., & Larkins, B. A. (2009). The development of endosperm in grasses. *Plant physiology, 149*(1), 14-26.

Sáez-Vásquez, J., & Delseny, M. (2019). Ribosome biogenesis in plants: from functional 45S ribosomal DNA organization to ribosome assembly factors. *Plant Cell, 31*(9), 1945-1967.

Schaefer, R. J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., & Myers, C. L. (2018). Integrating coexpression networks with GWAS to prioritize causal genes in maize. *Plant Cell, 30*(12), 2922-2942.

Schmidt, M., Barbazuk, W. B., Sandford, M., May, G. D., Song, Z., Zhou, W., . . . Herman, E. M. (2011). Silencing of soybean seed storage proteins results in a rebalanced

protein composition preserving seed protein content without major collateral changes in the metabolome and transcriptome. *Plant physiology*, pp. 111.173807.

Schmidt, M. A., & Pendarvis, K. (2017). Proteome rebalancing in transgenic Camelina occurs within the enlarged proteome induced by β-carotene accumulation and storage protein suppression. *Transgenic research, 26*(2), 171-186.

Shamimuzzaman, M., & Vodkin, L. (2014). Transcription factors and glyoxylate cycle genes prominent in the transition of soybean cotyledons to the first functional leaves of the seedling. *Functional & integrative genomics, 14*(4), 683-696.

Shen, B., & Roesler, K. (2017). Maize kernel oil content. *Maize kernel development*, 160-174.

Shen, L. (2014). GeneOverlap: An R package to test and visualize gene overlaps. *R Package*.

Shewry, P. R. (2007). Improving the protein content and composition of cereal grain. *Journal of cereal science, 46*(3), 239-250.

Shewry, P. R., & Casey, R. (1999). Seed proteins. In *Seed proteins* (pp. 1-10): Springer.

Shewry, P. R., & Halford, N. G. (2002). Cereal seed storage proteins: structures, properties and role in grain utilization. *Journal of Experimental Botany, 53*(370), 947-958. doi:DOI 10.1093/jexbot/53.370.947

Shi, Z., Fujii, K., Kovary, K. M., Genuth, N. R., Röst, H. L., Teruel, M. N., & Barna, M. (2017). Heterogeneous ribosomes preferentially translate distinct subpools of mRNAs genome-wide. *Molecular cell, 67*(1), 71-83. e77.

Shiferaw, B., Prasanna, B. M., Hellin, J., & Bänziger, M. (2011). Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food security, 3*(3), 307.

Shrestha, V., Awale, M., & Karn, A. (2019). Genome Wide Association Study (GWAS) on Disease Resistance in Maize. In *Disease Resistance in Crop Plants* (pp. 113-130): Springer.

Singh, B. K. (1998). *Plant amino acids: biochemistry and biotechnology*: CRC Press.

Slaten, M. L., Chan, Y. O., Shrestha, V., Lipka, A. E., & Angelovici, R. (2020). HAPPI GWAS: Holistic Analysis with Pre and Post Integration GWAS. *Bioinformatics*. doi:10.1093/bioinformatics/btaa589

Slaten, M. L., Yobi, A., Bagaza, C., Chan, Y. O., Shrestha, V., Holden, S., . . . Kliebenstein, D. J. (2020). mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis. *Plant physiology*.

Sofi, P., Wani, S. A., Rather, A., & Wani, S. H. (2009). Quality protein maize (QPM): genetic manipulation for the nutritional fortification of maize. *Journal of Plant Breeding and Crop Science, 1*(6), 244-253.

Sormani, R., Masclaux-Daubresse, C., Daniele-Vedele, F., & Chardon, F. (2011). Transcriptional regulation of ribosome components are determined by stress according to cellular compartments in Arabidopsis thaliana. *PLoS One, 6*(12), e28070.

Stelpflug, S. C., Sekhon, R. S., Vaillancourt, B., Hirsch, C. N., Buell, C. R., de Leon, N., & Kaeppler, S. M. (2016). An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *The plant genome, 9*(1), 1-16.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., . . . Bork, P. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research, 47*(D1), D607-D613.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., . . . Su, Z. (2017). agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic acids research, 45*(W1), W122-W129.

Tuin, L. G., & Shelp, B. J. (1994). In situ [14C] glutamate metabolism by developing soybean cotyledons I. Metabolic routes. *Journal of Plant Physiology, 143*(1), 1-7.

Turner-Hissong, S. D., Bird, K. A., Lipka, A. E., King, E. G., Beissinger, T. M., & Angelovici, R. (2020). Genomic prediction informed by biological processes expands our understanding of the genetic architecture underlying free amino acid traits in dry Arabidopsis seeds. *G3: Genes, Genomes, Genetics*.

Tzafrir, I., Pena-Muralla, R., Dickerman, A., Berg, M., Rogers, R., Hutchens, S., . . . Patton, D. (2004). Identification of genes required for embryo development in Arabidopsis. *Plant physiology, 135*(3), 1206-1220.

Van, K., & McHale, L. K. (2017). Meta-analyses of QTLs associated with protein and oil contents and compositions in soybean [Glycine max (L.) Merr.] seed. *International journal of molecular sciences, 18*(6), 1180.

Vasal, S. K., Villegas, E., Bjarnason, M., Gelaw, B., & Goertz, P. (1980). Genetic modifiers and breeding strategies in developing hard endosperm opaque-2 materials. *Genetic modifiers and breeding strategies in developing hard endosperm opaque-2 materials.*, 37-73.

Wang, X., & Larkins, B. A. (2001). Genetic Analysis of Amino Acid Accumulation inopaque-2 Maize Endosperm. *Plant physiology*, *125*(4), 1766-1777.

Watson, S. (2003). Description, development, structure and composition of the corn kernel. *Corn: chemistry and technology*, *2*, 69-106.

Withana-Gamage, T. S., Hegedus, D. D., Qiu, X., Yu, P., May, T., Lydiate, D., & Wanasundara, J. P. (2013). Characterization of Arabidopsis thaliana lines with altered seed storage protein profiles using synchrotron-powered FT-IR spectromicroscopy. *Journal of Agricultural and Food Chemistry*, *61*(4), 901-912.

Wu, S., Alseekh, S., Cuadros-Inostroza, Á., Fusari, C. M., Mutwil, M., Kooke, R., . . . Brotman, Y. (2016). Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in Arabidopsis thaliana. *PLoS genetics*, *12*(10), e1006363.

Wu, Y., & Messing, J. (2014). Proteome balancing of the maize seed for higher nutritional value. *Frontiers in plant science*, *5*, 240.

Wu, Y., Wang, W., & Messing, J. (2012). Balancing of sulfur storage in maize seed. *BMC plant biology*, *12*(1), 77.

Xia, Z., Wang, M., & Xu, Z. (2018). The maize sulfite reductase is involved in cold and oxidative stress responses. *Frontiers in plant science, 9,* 1680.

Yan, H., Chen, D., Wang, Y., Sun, Y., Zhao, J., Sun, M., & Peng, X. (2016). Ribosomal protein L18aB is required for both male gametophyte function and embryo development in Arabidopsis. *Scientific reports, 6*, 31195.

Yan, J., Shah, T., Warburton, M. L., Buckler, E. S., McMullen, M. D., & Crouch, J. (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS One, 4*(12), e8451.

Yangueez, E., Castro-Sanz, A. B., Fernandez-Bautista, N., Oliveros, J. C., & Castellano, M. M. (2013). Analysis of genome-wide changes in the translatome of Arabidopsis seedlings subjected to heat stress. *PLoS One, 8*(8), e71425.

Yao, M., Guan, M., Zhang, Z., Zhang, Q., Cui, Y., Chen, H., . . . Werner, C. R. (2020). GWAS and co-expression network combination uncovers multigenes with close linkage effects on the oleic acid content accumulation in Brassica napus. *BMC Genomics, 21*, 1-12.

Yobi, A., & Angelovici, R. (2018). A High-Throughput Absolute-Level Quantification of Protein-Bound Amino Acids in Seeds. *Current protocols in plant biology, 3*(4), e20084.

Yobi, A., Bagaza, C., Batushansky, A., Shrestha, V., Emery, M. L., Holden, S., . . . Angelovici, R. (2020). The complex response of free and bound amino acids to water stress during the seed setting stage in Arabidopsis. *The Plant Journal, 102*(4), 838-855.

Young, T. E., Gallie, D. R., & DeMason, D. A. (1997). Ethylene-mediated programmed cell death during maize endosperm development of wild-type and shrunken2 genotypes. *Plant physiology, 115*(2), 737-751.

Zhang, H., Wang, M. L., Schaefer, R., Dang, P., Jiang, T., & Chen, C. (2019). GWAS and coexpression network reveal ionomic variation in cultivated peanut. *Journal of Agricultural and Food Chemistry, 67*(43), 12026-12036.

Zhang, J., Yuan, H., Yang, Y., Fish, T., Lyi, S. M., Thannhauser, T. W., . . . Li, L. (2016). Plastid ribosomal protein S5 is involved in photosynthesis, plant development, and cold stress tolerance in Arabidopsis. *Journal of Experimental Botany, 67*(9), 2731-2744.

Zhu, W., & Zhang, H. (2009). Why do we test multiple traits in genetic association studies? *Journal of the Korean Statistical Society, 38*(1), 1-10.

Zhu, X., & Galili, G. (2003). Increased lysine synthesis coupled with a knockout of its catabolism synergistically boosts lysine content and also transregulates the metabolism of other amino acids in Arabidopsis seeds. *Plant Cell, 15*(4), 845-853.

7. Figures

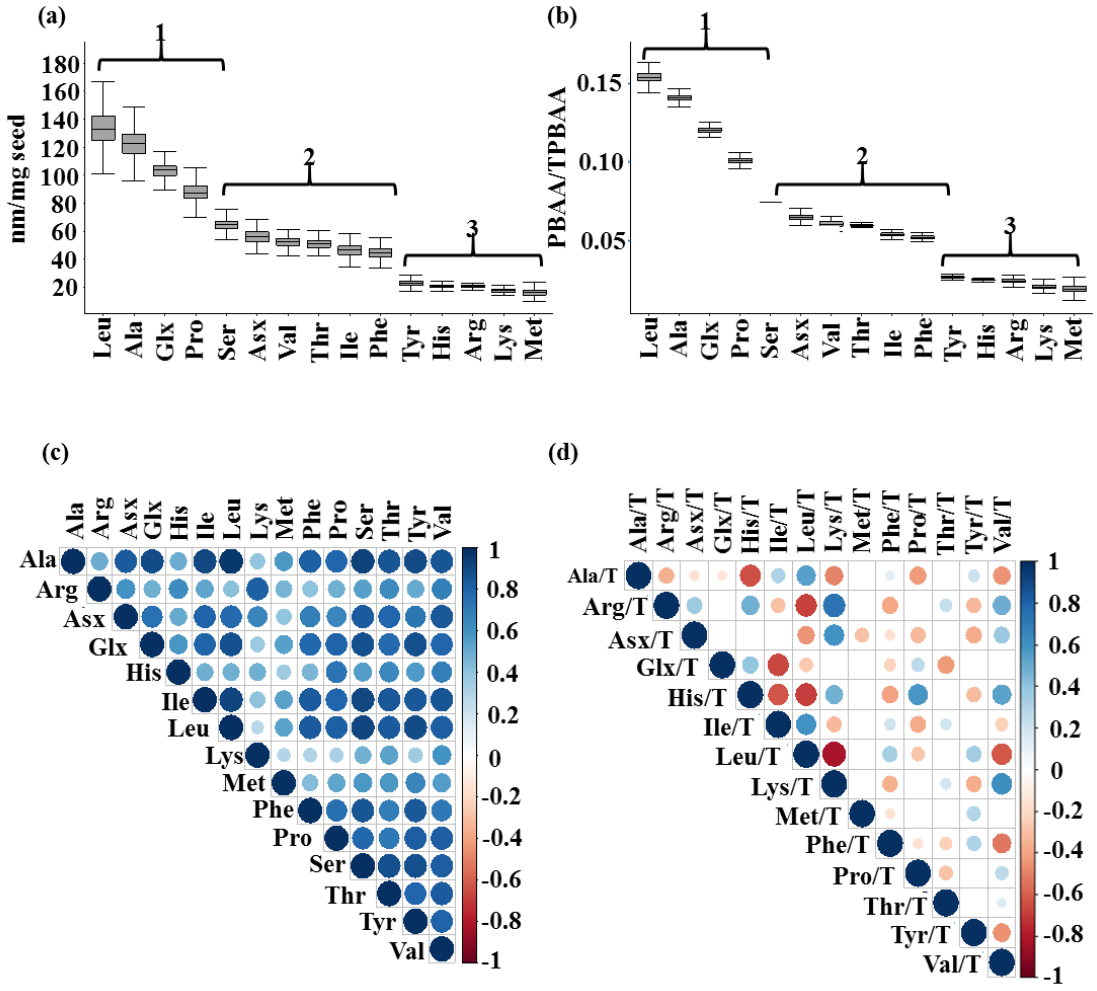


Figure 2. 1: **The natural variation and relationships of PBAA measured from the diversity panel.** Boxplot showing the PBAA (a) absolute levels and (b) relative compositional distribution in the 279 taxa from the Goodman-Buckler maize association panel. Pairwise Pearson correlation analysis between the absolute PBAA levels (c) and relative composition (d) using back transformed BLUPs of 279 taxa from Goodman-Buckler maize association panel. The correlation matrix was visualized in R v.3.4.3 (R Core Team). Each dot represents a significant correlation coefficient (r) at qFDR values <0.05. Blue dots indicate positive correlation, and red dots indicate negative correlations. Asx denotes Asn + Asp. Glx denotes Gln + Glu. Numbers on (a) and (b) represent groups based on PBAA absolute levels (a) and PBAA/TPBAA ratios (b) where 1 is for PBAA levels > 10%; 2 is for PBAA levels between 10 and 2%; and 3 is for PBAA levels < 2%.

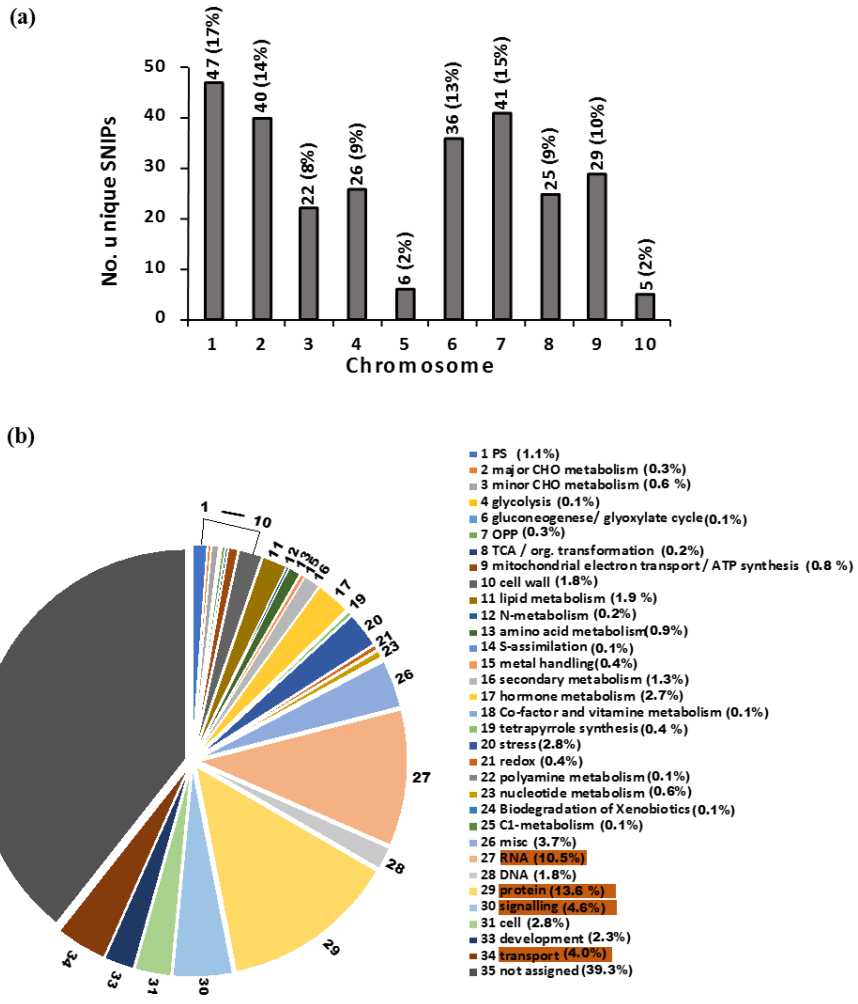


Figure 2.2**: The genomic distribution of the significant unique SNPs found in GWAS and the functional categorization of the extracted candidate. (a)** The partition of the significant unique SNPs across the 10 chromosomes in maize. (**b**) Pie chart representing the functional categorization of the 1*399* GWAS candidate genes using MapMan version 3.6. The percentage in parenthesis represents the proportion of genes that falls into a functional category. The top four categories are highlighted in dark orange and include protein, RNA, signaling, and transport.

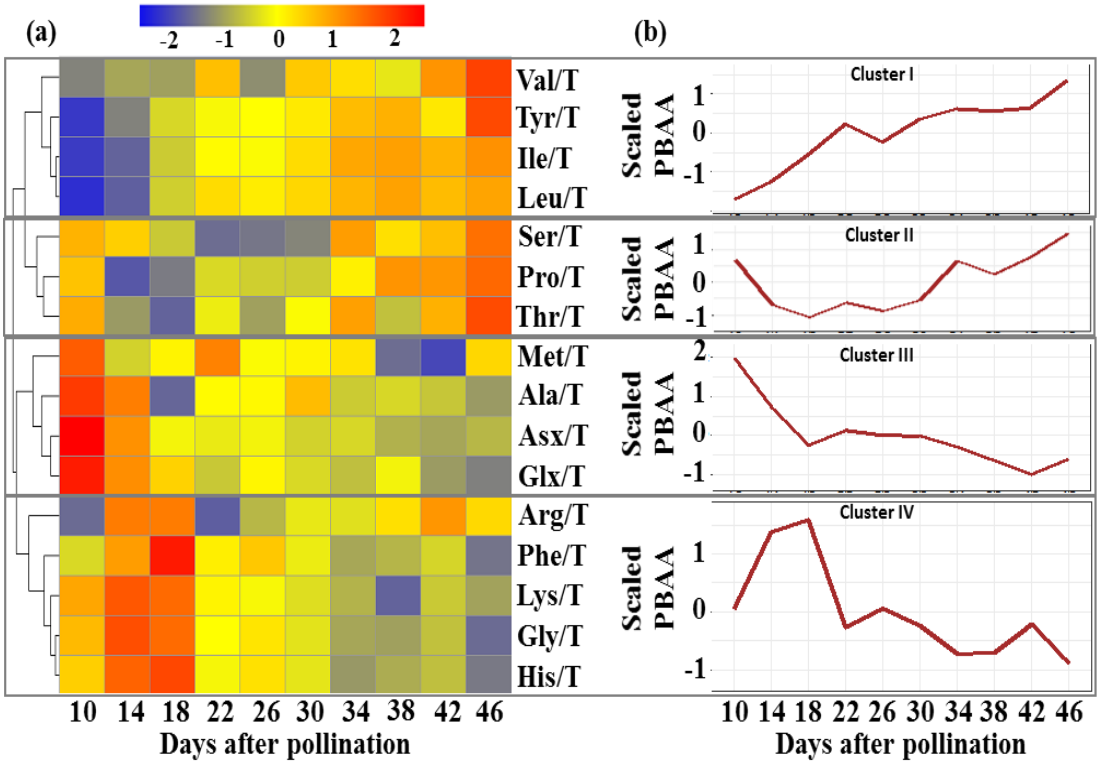


Figure 2. 3: **Seed PBAA composition dynamics during maturation.**
(a) Heatmap and (b) hierarchical clustering trends of the PBAA relative compositions across ten seed filling time points of maize inbred B73. The average values of three biological replicates from each time point were scaled and used to create the heatmap (*n* =3). Blue indicates low values for PBAA accumulation, and red indicate high accumulations. The red line in (b) indicates the average expression pattern of individual PBAA accumulation within a cluster (*n* = X) and was created using geom line "mean" function in ggplot2 package in R v.3.4.3.

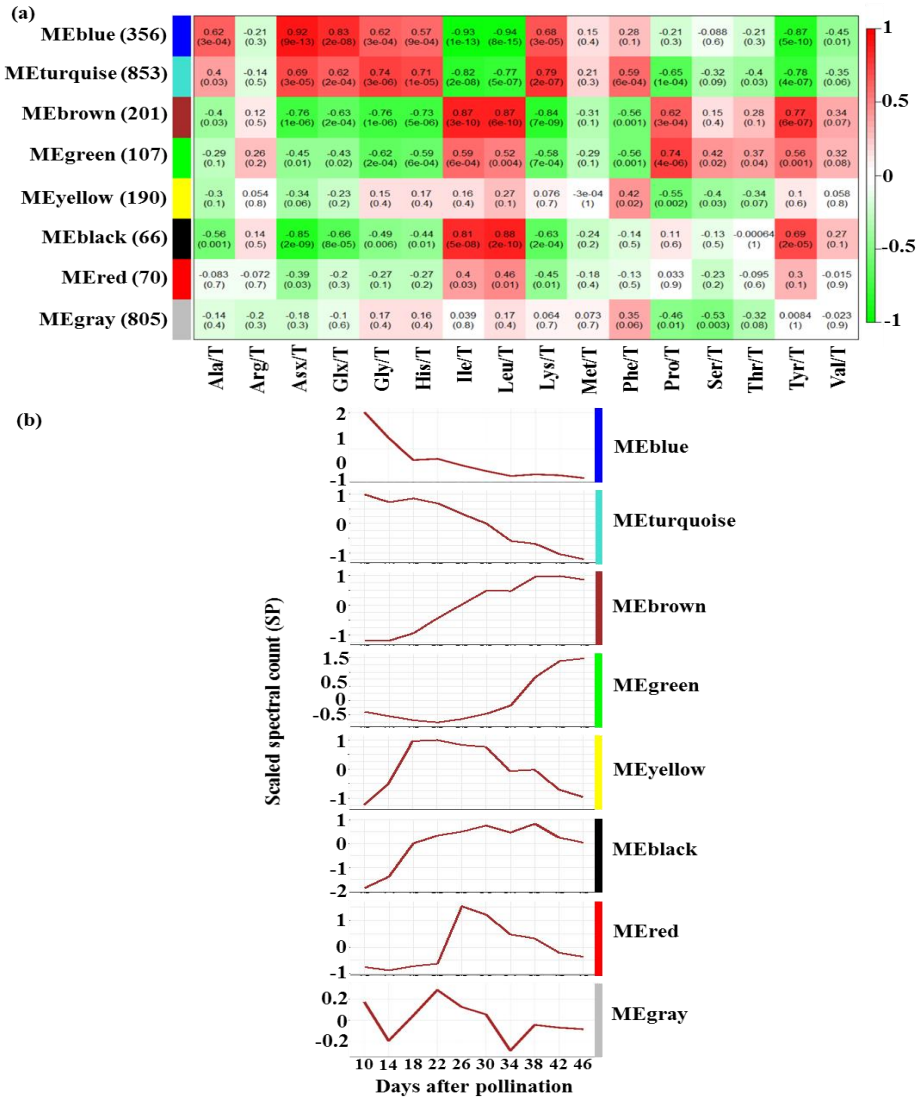


Figure 2. 4: **Relationships among protein co-expression modules and PBAA compositional dynamics during seed maturation.** (a) Module-trait relationships from the WGCNA analysis. Module names are displayed as rows on the left y-axis (e.g., MEblue denotes modules eigen protein for blue module). The relative PBAA composition traits (e.g., Ala/T, which is the ratio of Ala/Sum total of all PBAAs levels) are displayed in columns on the x-axis. The total number of proteins for each respective module is appears in parentheses along the y-axis. Each cell shows the correlation coefficients between modules Eigen protein (ME)-PBAA traits (top number) and the corresponding p-value (bottom number in parenthesis). The module-trait relationships are colored based on their correlation: red is a strong positive correlation, and green is a strong negative correlation. (b) The expression trend of Eigen protein found in the corresponding modules across the seed development time points. The x-axis is the 10 timepoints, in days after pollination. The y-axis is the expression of module eigen protein using the scaled spectral count of protein (scaled SP).

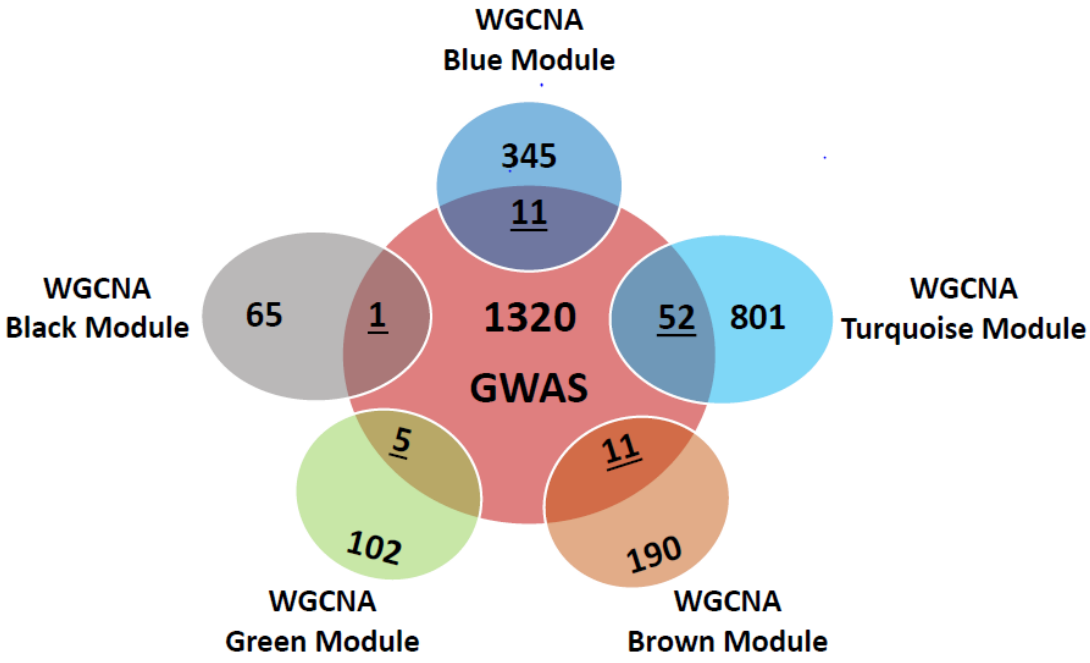
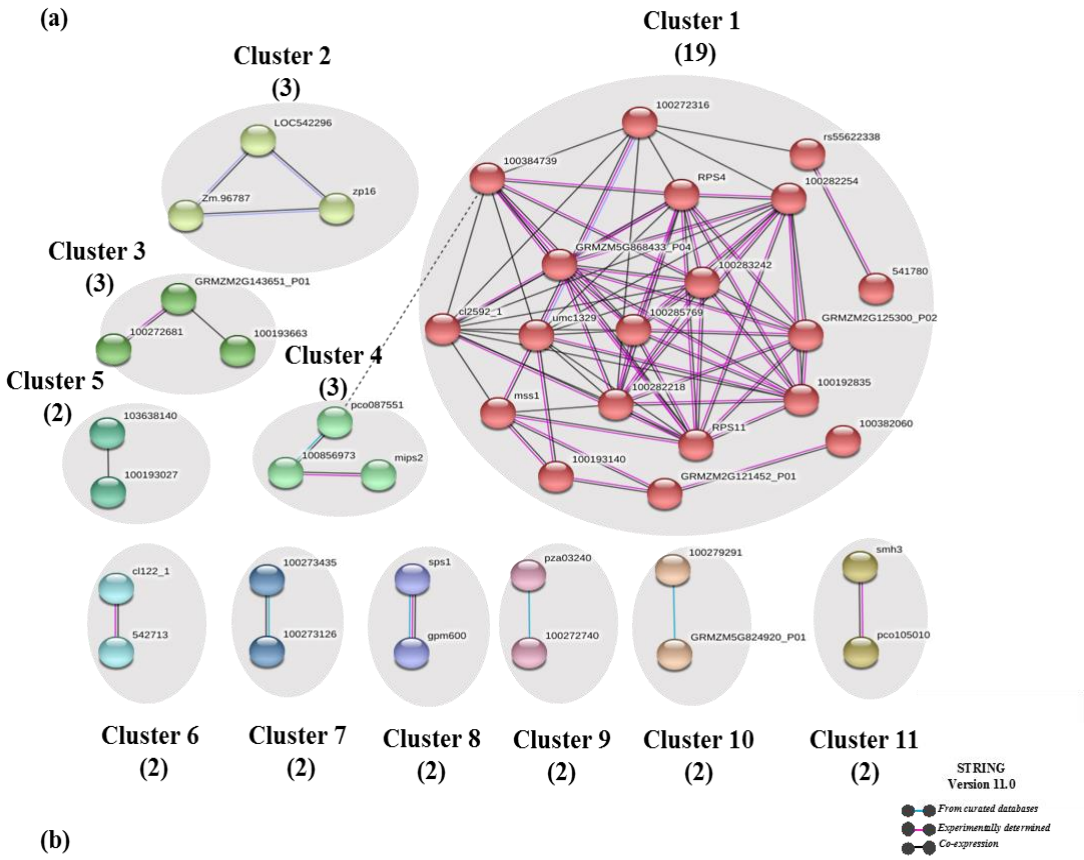


Figure 2. 5: **Comparison between the candidate genes lists of WGCNA and GWAS.** Van diagram depicting the 80 genes that overlap with both the GWAS candidate gene list and the relevant proteomic co-expression modules (turquoise, brown, green, black, and blue modules). I refer to them as high confidence candidate genes (HCCG).

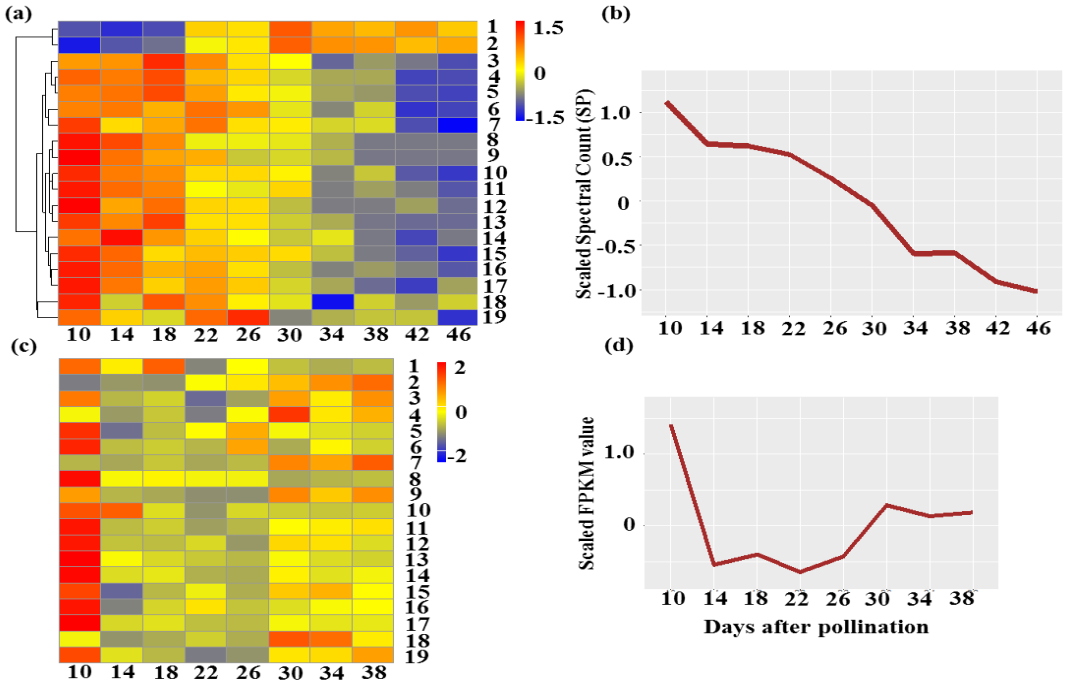
(a)



(b)

String			MapMan category
Cluster number	Cluster color	Gene count	Bin name
1	Red	19	Protein synthesis, degradation, and folding; stress
2	Green yellow	3	Zein storage protein
3	Green	3	TCA and electron transport chain
4	Light green	3	Minor CHO metabolism and oxidative pyrophosphate (OPP)
5	Medium see green	2	Cell vesicle transport and protein post translation modification (PPTM)
6	Cyan	2	Sulfur assimilation and OPP
7	Dark cyan	2	Lysine biosynthesis
8	Purple	2	Major and minor CHO metabolism
9	Pink	2	Amino acid metabolism and protein degradation
10	Sandy brown	2	Cell wall degradation
11	Brown	2	Lipid metabolism and PPTM

Figure 2. 6: **Protein-protein interaction (PPI) of the 80 HCCG list**. (a) A PPI of the 80 HCCG was created using STRING V11.0. HCCG are indicated by nodes labeled with the encoding protein symbol from STRING. Interaction between nodes are indicated by edges. Smooth line edges indicate intra-cluster interactions, and dotted edges indicate inter-cluster interaction. Cluster analysis using MCL algorithm resulted in 11 distinct clusters. (b) Table representation of cluster numbers, color, gene count within each cluster using STRING, and the Bin name is the functional category of the clusters using MapMan version 3.6.0.



(e)

Gene Number	Gene ID	Description
1	GRMZM2G070863	Hsp70-Hsp90 organizing protein 3
2	GRMZM2G360681	Chaperone protein ClpB1; Heat-shock protein 101 ; Belongs to the ClpA/ClpB family
3	GRMZM2G044137	60S ribosomal protein L30; Uncharacterized protein
4	GRMZM2G102779	Eukaryotic translation initiation factor 3 subunit D
5	GRMZM2G110185	Uncharacterized protein
6	GRMZM2G121452	26S proteasome non-ATPase regulatory subunit 8
7	GRMZM2G122767	T-complex protein 1 subunit delta; Molecular chaperone
8	GRMZM2G110626	T-complex protein 1 subunit alpha; Molecular chaperone
9	GRMZM2G377797	40S ribosomal protein S16
10	GRMZM2G125300	40S ribosomal protein S21
11	GRMZM2G145280	60S ribosomal protein L13
12	GRMZM2G044800	40S ribosomal protein S11
13	GRMZM2G140116	60S ribosomal protein L14-1
14	GRMZM2G125271	40S ribosomal protein S4
15	GRMZM2G587327	Vesicle-fusing ATPase
16	GRMZM2G124576	annotation not available
17	GRMZM5G868433	60S ribosomal protein L7-2
18	GRMZM2G005080	Proteasome subunit alpha type
19	GRMZM2G024354	Ribosomal protein L15

Figure 2. 7: **Protein and gene expression from cluster 1.** Heatmap of 19 genes in cluster 1 (red) obtained from the 80 HCCG protein-protein interaction in Figure 6a. (a) Protein expression pattern across the ten seed developmental stages of B73 obtained from shotgun proteomic sequencing. (b) Hierarchical clustering was used to cluster the proteins using the scaled data for spectral counts of proteins (scaled SP). (c) Gene expression pattern of the same 19 genes across eight seed developmental stages of B73 obtained from Chen et. al (2014) in the same order as (a). (d) Hierarchical clustering was used to cluster the gene using the scaled data from the FPKM gene values. Red indicates high expression, and blue indicates low expression. (e) The names and annotations of the 19 proteins/genes in cluster 1.

8. Supplementary Figures

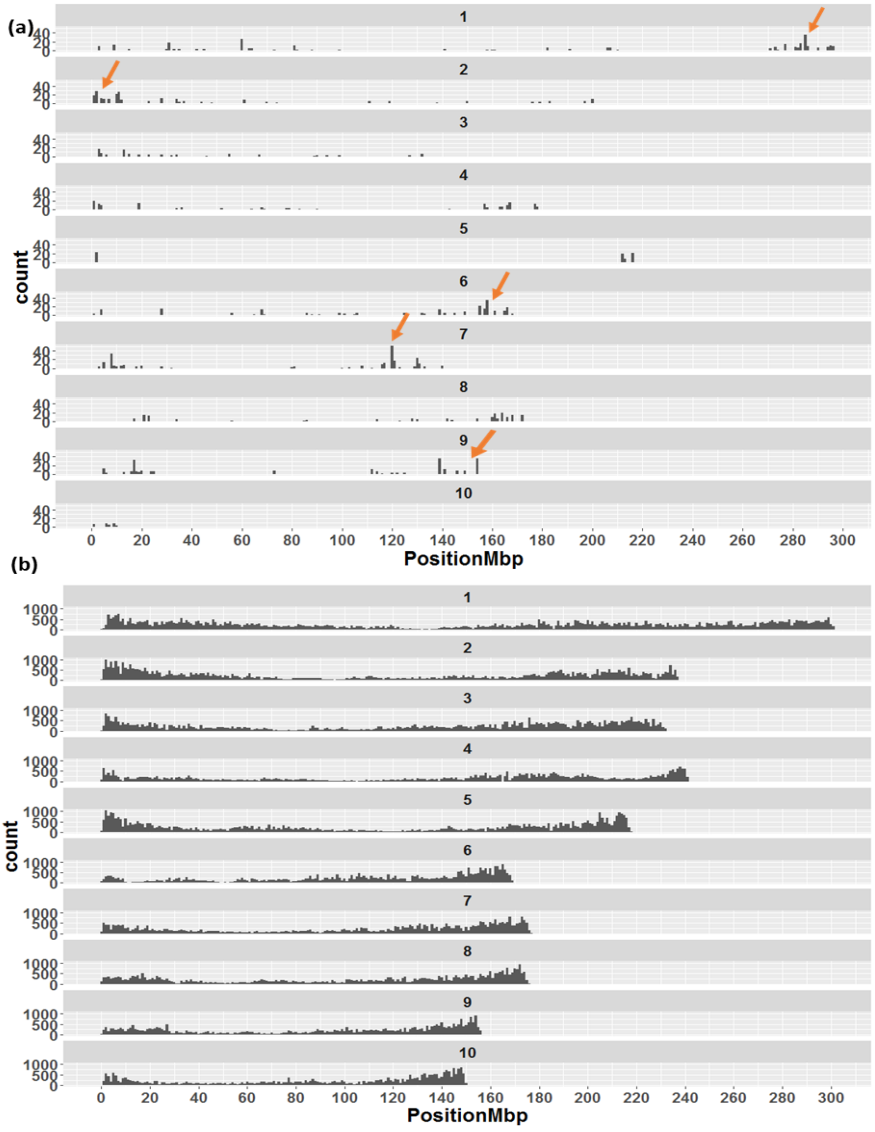


Figure S2.1 **Genomic distribution of significant and all SNPs across the genome (a)** Genome wide distribution of all the significant SNP found in our GWAS. **(b)** Genome wide SNP distribution of all the SNPs that were used for our analysis. The x-axis represents the position in Mbp across the genome while the y-axis represents the count in numbers of SNPs within the window size of 1 Mb. Several potential hotspot of SNP associations are detected (significant SNP count >20 for a given position) and marked by the orange arrows in chromosome 1, 2, 6, 7, and 9.

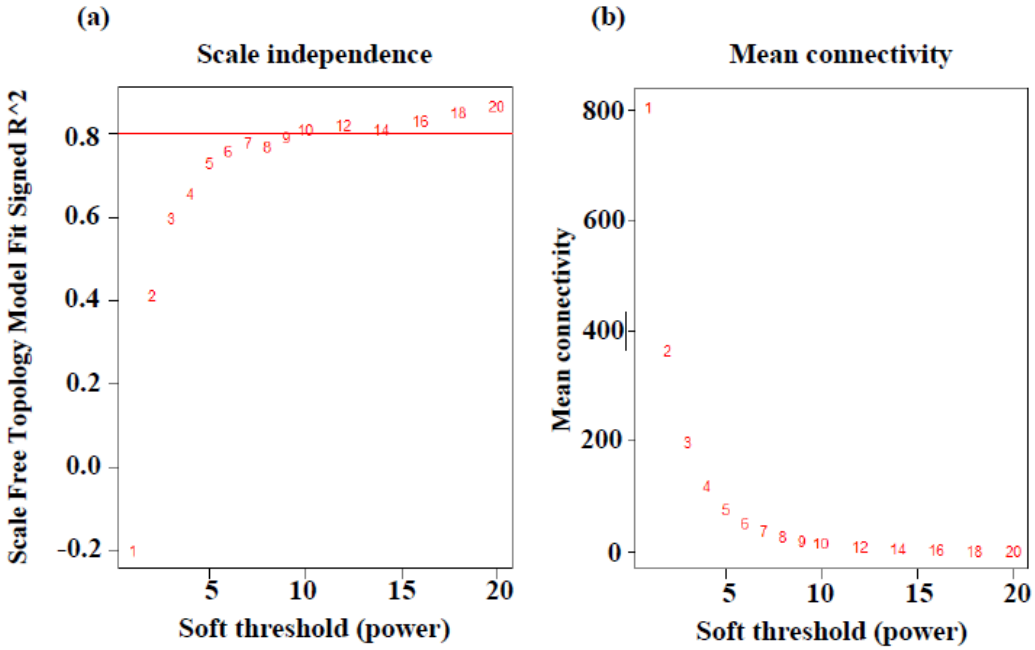
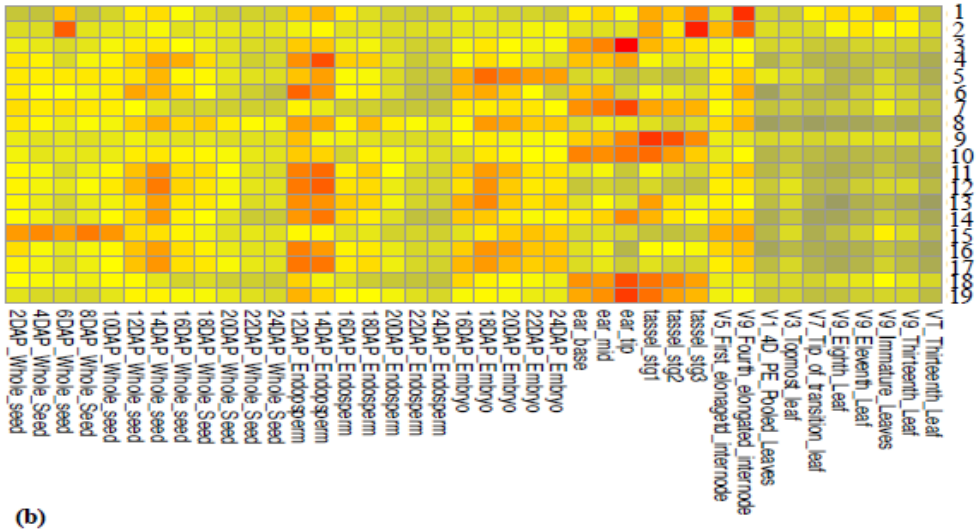


Figure S2.2 **Analysis of network topology for soft thresholding power for constructing the coexpression network**. (a) Shows the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). The red bar indicates scale free topology R^2 at 80%. (b) Displays the mean connectivity (degree, y-axis) as a function of the soft thresholding power (x-axis).

(a)



(b)

Gene Number	Gene ID	Description
1	GRMZM2G070863	Hsp70-Hsp90 organizing protein 3
2	GRMZM2G360681	Chaperone protein ClpB1; Heat-shock protein 101 ; Belongs to the ClpA/ClpB family
3	GRMZM2G044137	60S ribosomal protein L30; Uncharacterized protein
4	GRMZM2G102779	Eukaryotic translation initiation factor 3 subunit D
5	GRMZM2G110185	Uncharacterized protein
6	GRMZM2G121452	26S proteasome non-ATPase regulatory subunit 8
7	GRMZM2G122767	T-complex protein 1 subunit delta; Molecular chaperone
8	GRMZM2G110626	T-complex protein 1 subunit alpha; Molecular chaperone
9	GRMZM2G377797	40S ribosomal protein S16
10	GRMZM2G125300	40S ribosomal protein S21
11	GRMZM2G145280	60S ribosomal protein L13
12	GRMZM2G044800	40S ribosomal protein S11
13	GRMZM2G140116	60S ribosomal protein L14-1
14	GRMZM2G125271	40S ribosomal protein S4
15	GRMZM2G587327	Vesicle-fusing ATPase
16	GRMZM2G124576	annotation not available
17	GRMZM5G868433	60S ribosomal protein L7-2
18	GRMZM2G005080	Proteasome subunit alpha type
19	GRMZM2G024354	Ribosomal protein L15

Figure S2. 3 **Gene expression of cluster 1 members in the various maize tissues.** (a) Heatmap of tissue gene expression pattern of the cluster1 in Figure 6a; 19 genes across whole seed, endosperm, embryo, ear, tassel, internode, and leaves obtained from Stelpflug et. al (2016) and heatmap was created using the scaled data from the RPKM gene values. Red indicates higher expression while blue indicates lower expression. (b) The names and annotation of the proteins/genes within cluster 1.

9. Tables

Table 2.1 **A summary of 76 PBAA GWAS results.** Data are summarized by PBAA trait category: PBAA absolute levels, relative composition, aspartate family related ratios, BCAA family related ratios, glutamate family related ratios, and shikimate family related ratio traits. The table presents the absolute number (n) and the percentage (%) per family for the following trait parameters: total number of traits analyzed for the GWAS, number of significant traits in the GWAS at 5% FDR, number of unique (non-redundant) SNPs, average SNP per trait per family, and number of unique candidate genes from a 200 kb window of the peak SNP.

PBAA Traits Category	Total number of traits analyzed		No. of significant traits		No. of unique SNPs		Average SNP per trait	No. of unique Candidate genes	
	n	%	n	%	n	%		n	%
Absolute levels	15	20	5	12.5	36	12.9	7	219	16
Relative composition	14	18	8	20.0	60	21.6	8	317	23
Aspartate family related ratios	15	20	8	20.0	72	25.9	9	304	22
BCAA family related ratios	14	18	7	17.5	38	13.7	5	210	15
Glutamate family related ratios	13	17	9	22.5	61	22.3	7	296	21
Shikimate family related ratios	5	7	3	7.5	10	3.6	3	53	4
Total sum of the PBAA	76	100	40	100	277	100	39	1399	100

Table 2.2 **Ranking of the top 15 high confidence candidate gene (HCCG) list.** A data integration from GWAS, WGCNA, expression variation correlation with trait correlation (eQTL/mQTL), and STRING analyses were used for determination and a combined score and ranking were based on the cumulative number of criteria that each gene fulfilled. "1" means that the condition is satisfied, while "0" means that the condition is not satisfied. The criteria are [a]significant association based on GWAS analysis with multiple traits; [b]high WGCNA-kME, which represents the connectivity of a given protein with module eigengene (kME> 0.7); [c]mQTLs that is driven by mQTL/eQTL; [d]STRING analysis connectivity > 0.4, which represents the inclusion in the PPI network; and [e]STRING analysis connectivity > 0.7, which represents high connectivity within the PPI network- since STRING analysis at confidence score >0.7 and >0.4 is not independent, partial score (0.5) is given instead of full score (1) if the gene is included in the PPI network. [f]Combined score of all the five criteria. Sig. Traits (GWAS) represents traits that were significantly associated with that SNP in my GWAS.

			GWAS		**WGCNA**	**eQTL-mQTL**	**STRING**		
Rank	**Protein ID (HCCP)**	**Annotations**	**Significant Traits**	**Multiple associations[a]**	**HC gene[b]**	**Expression driven [c]**	**0.4[d]**	**0.7[e]**	**Combined score[f]**
1	GRMZM2G138727_P01	Glutelin-2 Precursor (Zein-gamma-27 kDa zein)	H/M, H/Z, Z/ZHPR, H/TOTAL, L/IVL, V/A, V/LAV, V/TOTAL	1	1	1	1	0.5	4.5
2	GRMZM2G058760_P01	ferredoxin NADP reductase1-fnr1	H/Z, H/ZHPR, H/TOTAL	1	1	1	1	0	4
3	GRMZM2G138689_P01	50kD gamma zein- gz50	Z/ZHPR, H/M, H/Z, H/TOTAL, L/IVL, V/A, V/LAV, V/TOTAL	1	1	0	1	0.5	3.5
4	GRMZM2G044800_P01	40S ribosomal protein S11	K/IMTXK, V/LAV	1	1	0	1	0.5	3.5
5	GRMZM2G440208_P01	6-phosphogluconate dehydrogenase	L/IVL, M/I	1	1	0	1	0	3
6	GRMZM2G090338_P01	sulfite reductase1-sir1	L/V, V/LAV	1	0	1	1	0	3
7	GRMZM2G176396_P02	proline iminopeptidase	L/LAV, V/IVL	1	0	0	1	0.5	2.5
8	GRMZM2G024354_P01	ribosomal protein l15	T/K, R/ZHPR	1	0	0	1	0.5	2.5
9	GRMZM2G125300_P02	40S ribosomal subunit protein S21	M/TOTAL	0	1	0	1	0.5	2.5
10	GRMZM2G360681_P01	heat-shock protein 101	V/A	0	1	0	1	0.5	2.5

11	GRMZM2G110185_P02	26s protease regulatory subunit 7	R/ZHPR	0	1	0	1	0.5	2.5
12	GRMZM2G060429_P01	Zein-beta Precursor (Zein-2)(16 kDa zein)(Zein Zc1)	M	0	1	0	1	0.5	2.5
13	GRMZM2G102779_P01	Eif3	M	0	1	0	1	0.5	2.5
14	GRMZM2G587327_P01	hypothetical protein LOC100382060	V/A	0	1	0	1	0.5	2.5
15	GRMZM2G140116_P01	60 ribosomal protein l14	H/ZHPR	0	1	0	1	0.5	2.5

10. Supplementary Tables

Table S2.1 **Statistical and heritability summary of the PBAA related traits.** The mean, standard deviation (SD), relative standard deviation (RSD), range and broad sense heritability (BSH) of 15 PBAAs absolute levels measured and calculated from the dry seeds of Goodman-Buckler maize association panel and the 15 calculated PBAA relative composition are described in (a) and (b) respectively. Statistics (mean, SD, RSD and range) were calculated using the dry seed PBAA measurement from back transformed BLUPs of the 279 taxa of Goodman-Buckler maize association panel while BSH was calculated using replicated data (raw data after outlier removal) from two years and two replica.

(a)		Back transformed BLUPs					
Trait	Trait Symbol	Mean	SD	Relative SD %	Min	Max	Broad Sense Heritability
Leu	L	133.61	15.21	11.39	83.72	186.82	0.78
Ala	A	122.57	12.25	9.99	84.87	167.88	0.89
Glx	Z	103.38	5.43	5.25	84.15	124.59	0.33
Pro	P	88.03	7.05	8.01	66.68	115.13	0.66
Ser	S	64.52	4.84	7.50	50.83	81.09	0.66
Asx	X	56.62	5.87	10.37	41.68	77.12	0.89
Val	V	52.51	4.18	7.97	41.65	68.19	0.79
Thr	T	51.23	4.17	8.13	40.59	69.05	0.66
Ile	I	46.66	5.22	11.19	31.76	66.74	0.86
Phe	F	44.51	4.74	10.65	30.37	59.83	0.56
Tyr	Y	23.07	2.53	10.95	15.54	33.94	0.90
His	H	20.71	1.33	6.41	17.27	25.65	0.71
Arg	R	20.67	1.10	5.33	17.02	24.03	0.05
Lys	K	17.61	1.68	9.53	14.20	24.40	0.81
Met	M	16.43	2.87	17.49	9.72	30.31	0.71

(b)		Back transformed BLUPs					
Trait	Trait Symbol	Mean	SD	Relative SD %	Min	Max	Broad Sense Heritability
Leu/T	L/T	0.15	0.01	3.51	0.12	0.17	0.50
Ala/T	A/T	0.14	0.00	1.91	0.12	0.15	0.70
Glx/T	Z/T	0.12	0.00	1.73	0.12	0.13	0.08
Pro/T	P/T	0.10	0.00	1.86	0.10	0.11	0.22
Ser/T	S/T	0.07	0.00	0.00	0.07	0.07	0.00
Asx/T	X/T	0.06	0.00	4.45	0.06	0.08	0.78
Val/T	V/T	0.06	0.00	3.55	0.06	0.07	0.61
Thr/T	T/T	0.06	0.00	0.94	0.06	0.06	0.10
Ile/T	I/T	0.05	0.00	2.42	0.05	0.06	0.72
Phe/T	F/T	0.05	0.00	2.39	0.05	0.06	0.25
Tyr/T	Y/T	0.03	0.00	2.81	0.02	0.03	0.45
Arg/T	R/T	0.02	0.00	2.70	0.02	0.03	0.02
His/T	H/T	0.02	0.00	7.47	0.02	0.03	0.85
Lys/T	K/T	0.02	0.00	11.05	0.02	0.04	0.83
Met/T	M/T	0.02	0.00	14.84	0.01	0.03	0.72

Table S2.2 **The relationship among the abs PBAA levels and among their relative composition.** (a) Pairwise Pearson correlation coefficients between the 15 absolute PBAA levels using back transformed BLUPs of 279 taxa from Goodman-Buckler maize association panel. (b) Significance of correlation analysis of absolute PBAA using corr.test function in R and the corresponding adjusted p-value using FDR correction at 5% level of significance. (c) Similarly, pairwise Pearson correlation coefficient between 14 relative (compositional) PBAA using back transformed BLUPs of 279 taxa from Goodman-Buckler maize association panel. Ser/Total trait did not converge for the back transformed BLUPs and hence excluded from the entire analysis. (d) Significance of relative PBAA correlation analysis using corr.test and the corresponding adjusted p-value using FDR correction at 5% level of significance. T stands for Total.

(a)	Ala	Arg	Asx	Glx	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Tyr	Val
Ala	1														
Arg	0.50	1													
Asx	0.83	0.60	1												
Glx	0.89	0.48	0.75	1											
His	0.49	0.62	0.50	0.59	1										
Ile	0.91	0.52	0.81	0.80	0.49	1									
Leu	0.96	0.41	0.78	0.87	0.49	0.92	1								
Lys	0.39	0.81	0.66	0.37	0.48	0.41	0.27	1							
Met	0.58	0.46	0.40	0.55	0.35	0.53	0.53	0.29	1						
Phe	0.82	0.40	0.69	0.79	0.46	0.84	0.84	0.31	0.43	1					
Pro	0.79	0.48	0.66	0.80	0.74	0.79	0.82	0.32	0.52	0.75	1				
Ser	0.92	0.54	0.83	0.87	0.55	0.90	0.91	0.48	0.58	0.85	0.78	1			
Thr	0.85	0.63	0.80	0.77	0.64	0.85	0.83	0.54	0.58	0.70	0.72	0.87	1		
Tyr	0.89	0.50	0.73	0.87	0.56	0.84	0.89	0.36	0.65	0.84	0.82	0.88	0.79	1	
Val	0.85	0.68	0.84	0.80	0.69	0.86	0.83	0.60	0.56	0.72	0.83	0.83	0.83	0.81	1

(b)	Ala	Arg	Asx	Glx	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Tyr	Val
Ala	0														
Arg	9E-19	0													
Asx	3E-72	5E-29	0												
Glx	1E-97	1E-17	1E-50	0											
His	2E-18	6E-31	2E-19	3E-27	0										
Ile	4E-104	6E-21	4E-65	3E-64	4E-18	0									
Leu	1E-154	8E-13	1E-57	9E-87	7E-18	8E-112	0								
Lys	1E-11	5E-67	5E-36	3E-10	1E-17	1E-12	3E-06	0							
Met	5E-26	3E-16	7E-12	5E-23	2E-09	7E-22	1E-21	6E-07	0						
Phe	4E-68	2E-12	2E-40	1E-59	9E-16	2E-73	6E-74	1E-07	5E-14	0					
Pro	8E-61	9E-18	2E-35	5E-64	1E-48	4E-61	9E-68	4E-08	1E-20	6E-52	0				
Ser	1E-115	1E-22	1E-72	2E-87	2E-23	4E-101	6E-109	2E-17	5E-26	3E-80	4E-58	0			
Thr	7E-79	8E-32	9E-62	8E-56	2E-33	3E-77	5E-73	2E-22	9E-26	1E-41	6E-45	9E-88	0		
Tyr	9E-97	3E-19	3E-47	1E-87	1E-24	4E-75	3E-96	8E-10	1E-34	4E-76	3E-69	3E-90	8E-61	0	
Val	3E-79	5E-39	5E-75	4E-62	6E-40	6E-81	5E-71	2E-28	6E-24	7E-45	2E-71	1E-70	1E-72	2E-65	0

(c)	Ala/T	Arg/T	Asx/T	Glx/T	His/T	Ile/T	Leu/T	Lys/T	Met/T	Phe/T	Pro/T	Thr/T	Tyr/T	Val/T
Ala/T	1.00													
Arg/T	-0.35	1.00												
Asx/T	-0.15	0.36	1.00											
Glx/T	-0.14	0.05	-0.12	1.00										
His/T	-0.64	0.49	0.05	0.40	1.00									
Ile/T	0.30	-0.30	0.03	-0.67	-0.62	1.00								
Leu/T	0.55	-0.69	-0.44	-0.27	-0.69	0.60	1.00							
Lys/T	-0.49	0.74	0.61	0.06	0.47	-0.32	-0.83	1.00						
Met/T	0.05	0.11	-0.29	-0.12	-0.09	-0.05	-0.06	-0.05	1.00					
Phe/T	0.13	-0.38	-0.17	-0.22	-0.41	0.21	0.34	-0.36	-0.15	1.00				
Pro/T	-0.42	0.04	-0.33	0.26	0.58	-0.37	-0.27	-0.05	0.00	-0.16	1.00			
Thr/T	-0.12	0.23	0.10	-0.42	0.09	0.19	-0.06	0.18	0.10	-0.24	-0.29	1.00		
Tyr/T	0.21	-0.32	-0.38	-0.06	-0.31	-0.07	0.34	-0.36	0.29	0.30	-0.08	-0.12	1.00	
Val/T	-0.45	0.49	0.37	0.04	0.54	-0.24	-0.61	0.61	-0.05	-0.53	0.27	0.13	-0.46	1.00

(d)	Ala/T	Arg/T	Asx/T	Glx/T	His/T	Ile/T	Leu/T	Lys/T	Met/T	Phe/T	Pro/T	Thr/T	Tyr/T	Val/T
Ala/T	0.00													
Arg/T	0.00	0.00												
Asx/T	0.01	0.00	0.00											
Glx/T	0.03	0.44	0.05	0.00										
His/T	0.00	0.00	0.44	0.00	0.00									
Ile/T	0.00	0.00	0.63	0.00	0.00	0.00								
Leu/T	0.00	0.00	0.00	0.00	0.00	0.00	0.00							
Lys/T	0.00	0.00	0.00	0.37	0.00	0.00	0.00	0.00						
Met/T	0.44	0.07	0.00	0.06	0.18	0.44	0.37	0.44	0.00					
Phe/T	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00				
Pro/T	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.44	0.96	0.01	0.00			
Thr/T	0.05	0.00	0.10	0.00	0.14	0.00	0.37	0.00	0.11	0.00	0.00	0.00		
Tyr/T	0.00	0.00	0.00	0.33	0.00	0.30	0.00	0.00	0.00	0.00	0.19	0.05	0.00	
Val/T	0.00	0.00	0.00	0.47	0.00	0.00	0.00	0.00	0.45	0.00	0.00	0.04	0.00	0.00

Table S2.3 **List of 76 seed PBAA traits used for GWAS**. These traits included from the quantification of 15 absolute levels and the calculation of their relative composition and known biochemical interactions (based on their affiliation with their respective amino acid families: Aspartate, Glutamate, BCAA and Shikimate).

Abs & relative composition to total AAs	**Biochemistry based metabolic ratios, grouped by AA families' affiliation**			
AA-Abs & AA/Total Total= Sum of 15 AA	**Asp Family =Ile, Met, Thr, Asx Lys (IMTXK)**	**BCAA Family= Ile, Val, Leu (IVL) Pyr Family=Leu, Ala, Val (LAV)**	**Bound amino acids one letter code**	
A	A/X	A/LAV	Ala	A
A/Total	I/IMTXK	I/IVL	Arg	R
F	IMTXK	I/L	Asx	X
F/Total	K/IMTXK	I/LAV	Glx	Z
H	M/I	I/V	His	H
H/Total	M/IMTXK	IVL	Ile	I
I	M/K	L/IVL	Leu	L
I/Total	T/I	L/LAV	Lys	K
K	T/IMTXK	L/V	Met	M
K/Total	T/K	LAV	Phe	F
L	T/Z	V/A	Pro	P
L/Total	X/IMTXK	V/IVL	Ser	S
M	X/K	V/LAV	Thr	T
M/Total	Z/I	Z/V	Tyr	Y
P	Z/K		Val	V
P/Total				
R	**Glu Family= Glx, His, Pro, Arg (ZHPR)**	**Shikimate Family = Phe,Tyr (FY)**		
R/Total	H/M	F/FY		
S	H/X	F/Z		
T	H/Z	FY		
T/Total	H/ZHPR	Y/F		
V	P/ZHPR	Y/FY		
V/Total	R/P			
X	R/Z			
X/Total	R/ZHPR			
Y	X/ZHPR			
Y/Total	Z/P			
Z	Z/R			
Z/Total	Z/ZHPR			
	ZHPR			

Table S2.4 **Enrichment analysis of the five protein co-expression modules.** GO enrichments of the proteins obtained from five WGCNA modules that are associated with PBAA composition dynamics across maturation (blue, turquoise, brown, green, and black) using AgriGo V2. P, biological process; F, molecular function F; C, cellular components. There was no significance GO enrichment terms for the green modules.

Blue Module

GO_acc	term_type	Term	query item	query total	ref item	ref total	p-value	FDR
GO:0046128	P	purine ribonucleoside metabolic process	7	332	8	2409	7.10E-06	0.013
GO:0042278	P	purine nucleoside metabolic process	7	332	9	2409	2.80E-05	0.026
GO:0046500	P	S-adenosylmethionine metabolic process	5	332	5	2409	5.20E-05	0.032
GO:0030244	P	cellulose biosynthetic process	7	332	10	2409	8.20E-05	0.038
GO:0070003	F	threonine-type peptidase activity	10	332	15	2409	4.30E-06	0.00076
GO:0004298	F	threonine-type endopeptidase activity	10	332	15	2409	4.30E-06	0.00076
GO:0050662	F	coenzyme binding	36	332	117	2409	3.50E-06	0.00076
GO:0005839	C	proteasome core complex	10	332	15	2409	4.30E-06	0.0018

Turquiose Module

GO_acc	term_type	Term	query item	query total	ref item	ref total	p-value	FDR
GO:0006412	P	translation	149	799	222	2409	7.30E-23	2.30E-19
GO:0010467	P	gene expression	170	799	283	2409	2.50E-18	4.00E-15
GO:0034645	P	cellular macromolecule biosynthetic process	172	799	300	2409	6.90E-16	7.30E-13
GO:0009059	P	macromolecule biosynthetic process	172	799	301	2409	1.00E-15	8.30E-13
GO:0044267	P	cellular protein metabolic process	225	799	435	2409	2.20E-13	1.40E-10
GO:0019538	P	protein metabolic process	251	799	519	2409	7.60E-11	4.00E-08
GO:0009166	P	nucleotide catabolic process	43	799	63	2409	2.10E-08	4.90E-06
GO:0009261	P	ribonucleotide catabolic process	43	799	63	2409	2.10E-08	4.90E-06
GO:0006195	P	purine nucleotide catabolic process	43	799	63	2409	2.10E-08	4.90E-06
GO:0009143	P	nucleoside triphosphate catabolic process	43	799	63	2409	2.10E-08	4.90E-06
GO:0009146	P	purine nucleoside triphosphate catabolic process	43	799	63	2409	2.10E-08	4.90E-06
GO:0009207	P	purine ribonucleoside triphosphate catabolic proce	43	799	63	2409	2.10E-08	4.90E-06
GO:0009203	P	ribonucleoside triphosphate catabolic process	43	799	63	2409	2.10E-08	4.90E-06
GO:0009154	P	purine ribonucleotide catabolic process	43	799	63	2409	2.10E-08	4.90E-06
GO:0044260	P	cellular macromolecule metabolic process	282	799	625	2409	2.50E-08	5.30E-06
GO:0046700	P	heterocycle catabolic process	46	799	73	2409	2.70E-07	5.50E-05
GO:0006184	P	GTP catabolic process	33	799	47	2409	3.10E-07	5.80E-05
GO:0009199	P	ribonucleoside triphosphate metabolic process	50	799	82	2409	3.80E-07	6.80E-05
GO:0009141	P	nucleoside triphosphate metabolic process	50	799	83	2409	6.40E-07	0.0001
GO:0009144	P	purine nucleoside triphosphate metabolic process	49	799	81	2409	6.90E-07	0.0001
GO:0009205	P	purine ribonucleoside triphosphate metabolic proce	49	799	81	2409	6.90E-07	0.0001
GO:0046039	P	GTP metabolic process	34	799	51	2409	1.40E-06	0.0002
GO:0043170	P	macromolecule metabolic process	312	799	729	2409	1.40E-06	0.0002
GO:0009150	P	purine ribonucleotide metabolic process	54	799	94	2409	1.90E-06	0.00026
GO:0006163	P	purine nucleotide metabolic process	56	799	99	2409	2.50E-06	0.00031
GO:0044249	P	cellular biosynthetic process	257	799	597	2409	4.60E-06	0.00057
GO:0043933	P	macromolecular complex subunit organization	43	799	72	2409	4.90E-06	0.00059
GO:0065003	P	macromolecular complex assembly	38	799	62	2409	7.20E-06	0.00082
GO:0034621	P	cellular macromolecular complex subunit organizat	39	799	65	2409	1.10E-05	0.0012
GO:0022607	P	cellular component assembly	42	799	72	2409	1.40E-05	0.0015
GO:0034622	P	cellular macromolecular complex assembly	34	799	55	2409	1.60E-05	0.0017
GO:0009259	P	ribonucleotide metabolic process	55	799	102	2409	2.00E-05	0.0019
GO:0009058	P	biosynthetic process	265	799	629	2409	2.00E-05	0.0019
GO:0051258	P	protein polymerization	13	799	15	2409	3.20E-05	0.003
GO:0016043	P	cellular component organization	87	799	180	2409	3.60E-05	0.0033
GO:0070727	P	cellular macromolecule localization	51	799	95	2409	4.40E-05	0.0039
GO:0006461	P	protein complex assembly	24	799	36	2409	4.70E-05	0.0039

GO:0070271	P	protein complex biogenesis	24	799	36	2409	4.70E-05	0.0039
GO:0034613	P	cellular protein localization	49	799	91	2409	5.50E-05	0.0045
GO:0006886	P	intracellular protein transport	48	799	90	2409	8.80E-05	0.007
GO:0009987	P	cellular process	576	799	1471	2409	9.20E-05	0.0072
GO:0043623	P	cellular protein complex assembly	20	799	29	2409	9.70E-05	0.0073
GO:0044085	P	cellular component biogenesis	47	799	88	2409	9.80E-05	0.0073
GO:0016192	P	vesicle-mediated transport	35	799	62	2409	0.00017	0.012
GO:0006753	P	nucleoside phosphate metabolic process	65	799	134	2409	0.00025	0.017
GO:0008104	P	protein localization	58	799	117	2409	0.00025	0.017
GO:0009117	P	nucleotide metabolic process	65	799	134	2409	0.00025	0.017
GO:0006139	P	nucleobase, nucleoside,	124	799	283	2409	0.00028	0.018
GO:0044237	P	cellular metabolic process	459	799	1175	2409	0.00031	0.02
GO:0045184	P	establishment of protein localization	57	799	116	2409	0.00037	0.023
GO:0015031	P	protein transport	57	799	116	2409	0.00037	0.023
GO:0022621	P	shoot system development	26	799	45	2409	0.00068	0.04
GO:0048367	P	shoot development	26	799	45	2409	0.00068	0.04
GO:0044248	P	cellular catabolic process	123	799	286	2409	0.00067	0.04
GO:0003002	P	regionalization	12	799	16	2409	0.00079	0.046
GO:0009887	P	organ morphogenesis	17	799	26	2409	0.00083	0.048
GO:0007389	P	pattern specification process	14	799	20	2409	0.00088	0.049
GO:0005198	F	structural molecule activity	124	799	154	2409	1.90E-31	1.80E-28
GO:0003735	F	structural constituent of ribosome	98	799	124	2409	2.50E-24	1.10E-21
GO:0017111	F	nucleoside-triphosphatase activity	90	799	146	2409	8.70E-12	2.50E-09
GO:0016462	F	pyrophosphatase activity	96	799	159	2409	1.10E-11	2.50E-09
GO:0016818	F	hydrolase activity	96	799	160	2409	1.80E-11	3.20E-09
GO:0016817	F	hydrolase activity, acting on acid anhydrides	99	799	167	2409	2.40E-11	3.60E-09
GO:0003924	F	GTPase activity	37	799	54	2409	1.70E-07	2.20E-05
GO:0003676	F	nucleic acid binding	129	799	265	2409	5.60E-07	6.40E-05
GO:0032555	F	purine ribonucleotide binding	181	799	404	2409	4.80E-06	0.0004
GO:0032553	F	ribonucleotide binding	181	799	404	2409	4.80E-06	0.0004
GO:0019001	F	guanyl nucleotide binding	47	799	81	2409	5.70E-06	0.0004
GO:0032561	F	guanyl ribonucleotide binding	47	799	81	2409	5.70E-06	0.0004
GO:0005525	F	GTP binding	47	799	81	2409	5.70E-06	0.0004
GO:0016887	F	ATPase activity	33	799	54	2409	3.00E-05	0.002
GO:0005200	F	structural constituent of cytoskeleton	11	799	12	2409	4.70E-05	0.0029
GO:0042623	F	ATPase activity, coupled	26	799	42	2409	0.00015	0.0084
GO:0017076	F	purine nucleotide binding	184	799	438	2409	0.00024	0.013
GO:0005524	F	ATP binding	138	799	326	2409	0.00076	0.039
GO:0032559	F	adenyl ribonucleotide binding	138	799	328	2409	0.001	0.048
GO:0033279	C	ribosomal subunit	85	799	104	2409	2.90E-23	2.50E-20
GO:0005840	C	ribosome	124	799	177	2409	5.60E-22	2.40E-19
GO:0043232	C	intracellular non-membrane-bounded organelle	199	799	332	2409	1.10E-20	2.50E-18
GO:0043228	C	non-membrane-bounded organelle	199	799	332	2409	1.10E-20	2.50E-18
GO:0022626	C	cytosolic ribosome	113	799	162	2409	4.70E-20	8.10E-18
GO:0030529	C	ribonucleoprotein complex	131	799	199	2409	2.00E-19	2.90E-17
GO:0044445	C	cytosolic part	82	799	111	2409	1.30E-17	1.70E-15
GO:0032991	C	macromolecular complex	261	799	488	2409	4.90E-17	5.20E-15
GO:0005730	C	nucleolus	118	799	194	2409	3.80E-14	3.70E-12
GO:0022625	C	cytosolic large ribosomal subunit	52	799	67	2409	2.60E-13	2.00E-11
GO:0031981	C	nuclear lumen	119	799	200	2409	2.50E-13	2.00E-11
GO:0015934	C	large ribosomal subunit	53	799	69	2409	3.30E-13	2.40E-11
GO:0044428	C	nuclear part	135	799	238	2409	1.20E-12	7.90E-11
GO:0015935	C	small ribosomal subunit	32	799	35	2409	1.40E-12	8.60E-11
GO:0022627	C	cytosolic small ribosomal subunit	26	799	29	2409	5.20E-10	3.00E-08
GO:0070013	C	intracellular organelle lumen	122	799	228	2409	1.40E-09	7.10E-08
GO:0043233	C	organelle lumen	122	799	228	2409	1.40E-09	7.10E-08
GO:0031974	C	membrane-enclosed lumen	122	799	230	2409	2.70E-09	1.30E-07
GO:0044446	C	intracellular organelle part	333	799	743	2409	6.40E-09	2.90E-07
GO:0044422	C	organelle part	333	799	744	2409	7.40E-09	3.20E-07
GO:0005634	C	nucleus	200	799	430	2409	9.70E-08	4.00E-06
GO:0005829	C	cytosol	230	799	512	2409	3.80E-07	1.50E-05
GO:0030662	C	coated vesicle membrane	18	799	21	2409	1.10E-06	4.10E-05

GO:0030120	C	vesicle coat	18	799	21	2409	1.10E-06	4.10E-05
GO:0016020	C	membrane	375	799	890	2409	1.30E-06	4.50E-05
GO:0044433	C	cytoplasmic vesicle part	18	799	22	2409	4.30E-06	0.00013
GO:0012506	C	vesicle membrane	18	799	22	2409	4.30E-06	0.00013
GO:0030659	C	cytoplasmic vesicle membrane	18	799	22	2409	4.30E-06	0.00013
GO:0048475	C	coated membrane	26	799	37	2409	5.30E-06	0.00015
GO:0030117	C	membrane coat	26	799	37	2409	5.30E-06	0.00015
GO:0030312	C	external encapsulating structure	115	799	242	2409	7.60E-06	0.00021
GO:0005618	C	cell wall	115	799	242	2409	7.60E-06	0.00021
GO:0030135	C	coated vesicle	19	799	25	2409	1.60E-05	0.00042
GO:0005886	C	plasma membrane	269	799	640	2409	2.10E-05	0.00054
GO:0030660	C	Golgi-associated vesicle membrane	13	799	15	2409	3.20E-05	0.00076
GO:0005798	C	Golgi-associated vesicle	13	799	15	2409	3.20E-05	0.00076
GO:0044425	C	membrane part	169	799	386	2409	3.80E-05	0.00088
GO:0015630	C	microtubule cytoskeleton	17	799	23	2409	8.30E-05	0.0019
GO:0043234	C	protein complex	126	799	282	2409	9.70E-05	0.0021
GO:0005874	C	microtubule	13	799	16	2409	0.00012	0.0025
GO:0005856	C	cytoskeleton	23	799	36	2409	0.00018	0.0037
GO:0016023	C	cytoplasmic membrane-bounded vesicle	20	799	31	2409	0.00038	0.007
GO:0030658	C	transport vesicle membrane	9	799	10	2409	0.00035	0.007
GO:0031982	C	vesicle	20	799	31	2409	0.00038	0.007
GO:0031988	C	membrane-bounded vesicle	20	799	31	2409	0.00038	0.007
GO:0031410	C	cytoplasmic vesicle	20	799	31	2409	0.00038	0.007
GO:0044459	C	plasma membrane part	94	799	208	2409	0.00037	0.007
GO:0000139	C	Golgi membrane	13	799	18	2409	0.00085	0.015
GO:0022624	C	proteasome accessory complex	9	799	11	2409	0.0014	0.021
GO:0005838	C	proteasome regulatory particle	9	799	11	2409	0.0014	0.021
GO:0030663	C	COPI coated vesicle membrane	9	799	11	2409	0.0014	0.021
GO:0005853	C	eukaryotic translation elongation factor 1 complex	6	799	6	2409	0.0014	0.021
GO:0030126	C	COPI vesicle coat	9	799	11	2409	0.0014	0.021
GO:0030133	C	transport vesicle	9	799	11	2409	0.0014	0.021
GO:0030137	C	COPI-coated vesicle	9	799	11	2409	0.0014	0.021
GO:0005911	C	cell-cell junction	84	799	190	2409	0.0015	0.022
GO:0055044	C	symplast	84	799	190	2409	0.0015	0.022
GO:0009506	C	plasmodesma	84	799	190	2409	0.0015	0.022
GO:0030054	C	cell junction	84	799	190	2409	0.0015	0.022
GO:0044430	C	cytoskeletal part	17	799	27	2409	0.0015	0.022
GO:0005737	C	cytoplasm	553	799	1462	2409	0.0018	0.026

Brown Module

GO_acc	term_type	Term	query item	query total	ref item	ref total	p-value	FDR
GO:0009642	P	response to light intensity	13	171	31	2409	1.00E-07	3.90E-05
GO:0009644	P	response to high light intensity	11	171	21	2409	5.60E-08	3.90E-05
GO:0042542	P	response to hydrogen peroxide	12	171	29	2409	3.80E-07	9.40E-05
GO:0000302	P	response to reactive oxygen species	12	171	34	2409	2.80E-06	0.00052
GO:0009408	P	response to heat	15	171	63	2409	3.90E-05	0.0058
GO:0009266	P	response to temperature stimulus	24	171	140	2409	9.70E-05	0.01
GO:0009314	P	response to radiation	16	171	76	2409	0.00011	0.01
GO:0009416	P	response to light stimulus	16	171	76	2409	0.00011	0.01
GO:0006979	P	response to oxidative stress	17	171	90	2409	0.00027	0.022

Black Module								
GO_acc	term_type	Term	query item	query total	ref item	ref total	p-value	FDR
GO:0016051	P	carbohydrate biosynthetic process	11	62	61	2409	9.00E-07	0.00017
GO:0034637	P	cellular carbohydrate biosynthetic process	11	62	58	2409	5.30E-07	0.00017
GO:0016137	P	glycoside metabolic process	6	62	20	2409	1.10E-05	0.001
GO:0019252	P	starch biosynthetic process	5	62	12	2409	9.50E-06	0.001
GO:0005982	P	starch metabolic process	6	62	22	2409	2.00E-05	0.0015
GO:0044262	P	cellular carbohydrate metabolic process	16	62	184	2409	7.40E-05	0.0045
GO:0009081	P	branched chain family amino acid metabolic process	5	62	26	2409	0.00057	0.03
GO:0006073	P	cellular glucan metabolic process	6	62	40	2409	0.00067	0.031
GO:0009250	P	glucan biosynthetic process	5	62	28	2409	0.00081	0.033
GO:0044042	P	glucan metabolic process	6	62	43	2409	0.00099	0.036

Table S2.5 **The 80 HCCG STRING and functional analysis.** String analysis of the 80 HCCG including cluster number, cluster color, Protein ID, STRING Protein name and protein description. MCL clustering method was used in STRING V11.0 resulting into 11 distinct cluster in the PPI network derived from 80 HCCG. "Unconnected" label in the cluster number indicates those proteins that are not connected in the given PPI analysis. MapMan functional categorization of the 80 HCCG and their respective bin, bin code, and bin name are also elaborated.

					MapMan Annotations		
Cluster number	**Cluster color**	**Protein ID**	**STRING protein name**	**Protein description**	**Bin**	**Bin code**	**Bin name**
1	**Red**	GRMZM2G070863_P01	rs55622338	Hsp70-Hsp90 organizing protein 3	20	20.2.1	stress abiotic heat
1	**Red**	GRMZM2G360681_P01	541780	Chaperone protein ClpB1; Heat-shock protein 101; Uncharacterized protein; Belongs to the ClpA/ClpB family	20	20.2.1	stress abiotic heat
1	**Red**	GRMZM2G587327_P01	100382060	Vesicle-fusing ATPase	29	29.5.11.20	protein degradation ubiquitin proteasome
1	**Red**	GRMZM2G044137_P01	100285769	60S ribosomal protein L30; Uncharacterized protein	29	29.2.1.2.2.30	protein synthesis ribosomal protein eukaryotic 60S subunit L30
1	**Red**	GRMZM2G140116_P01	100192835	60S ribosomal protein L14-1	29	29.2.1.2.2.14	protein synthesis ribosomal protein eukaryotic 60S subunit L14
1	**Red**	GRMZM2G024354_P01	100282218	Ribosomal protein L15; Belongs to the eukaryotic ribosomal protein eL15 family	29	29.2.1.2.2.15	protein synthesis ribosomal protein eukaryotic 60S subunit L15

1	**Red**	GRMZM2G102779_P01	cl2592_1	Eukaryotic translation initiation factor 3 subunit D; mRNA cap-binding component of the eukaryotic translation initiation factor 3 (eIF-3) complex, which is involved in protein synthesis of a specialized repertoire of mRNAs and, together with other initiation factors, stimulates binding of mRNA and methionyl-tRNAi to the 40S ribosome. The eIF-3 complex specifically targets and initiates translation of a subset of mRNAs involved in cell proliferation. In the eIF-3 complex, eif3d specifically recognizes and binds the 7-methylguanosine cap of a subset of mRNAs	29	29.2.3	protein synthesis initiation
1	**Red**	GRMZM2G377797_P01	100283242	40S ribosomal protein S16; Belongs to the universal ribosomal protein uS9 family	29	29.2.1.2.1.16	protein synthesis ribosomal protein eukaryotic 40S subunit S16
1	**Red**	GRMZM2G125300_P02	GRMZM2G125300_P02	40S ribosomal protein S21; Belongs to the eukaryotic ribosomal protein eS21 family	29	29.2.1.2.1.21	protein synthesis ribosomal protein eukaryotic 40S subunit S21
1	**Red**	GRMZM2G122767_P01	umc1329	T-complex protein 1 subunit delta; Molecular chaperone; assists the folding of proteins upon ATP hydrolysis	29	29.6	protein folding
1	**Red**	GRMZM2G121452_P01	GRMZM2G121452_P01	26S proteasome non-ATPase regulatory subunit 8; Uncharacterized protein	29	29.5.11.20	protein degradation ubiquitin proteasome

1	**Red**	GRMZM2G005080_P01	100193140	Proteasome subunit alpha type; The proteasome is a multi-catalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH; Belongs to the peptidase T1A family	29	29.5.11.20	protein degradation ubiquitin proteasome
1	**Red**	GRMZM5G868433_P04	GRMZM5G868433_P04	60S ribosomal protein L7-2; Uncharacterized protein	29	29.2.1.2.2.7	protein synthesis ribosomal protein eukaryotic 60S subunit L7
1	**Red**	GRMZM2G044800_P01	RPS11	40S ribosomal protein S11	29	29.2.1.2.1.11	protein synthesis ribosomal protein eukaryotic 40S subunit S11
1	**Red**	GRMZM2G145280_P01	100282254	60S ribosomal protein L13; Belongs to the eukaryotic ribosomal protein eL13 family	29	29.2.1.2.2.13	protein synthesis ribosomal protein eukaryotic 60S subunit L13
1	**Red**	GRMZM2G124576_P02	100384739	annotation not available	27	27.3.67	RNA regulation of transcription putative transcription regulator
1	**Red**	GRMZM2G125271_P01	RPS4	40S ribosomal protein S4	29	29.2.1.2.1.4	protein synthesis ribosomal protein eukaryotic 40S subunit S4
1	**Red**	GRMZM2G110185_P02	mss1	Uncharacterized protein	29	29.5.11.20	protein degradation ubiquitin proteasome

1	**Red**	GRMZM2G110626_P01	100272316	T-complex protein 1 subunit alpha; Molecular chaperone; assists the folding of proteins upon ATP hydrolysis	29	29.6	protein folding
2	**Green Yellow**	GRMZM2G138727_P01	LOC542296	Glutelin-2; Seed storage protein. It accounts for about 15% of the total endosperm protein content	35	35.2	not assigned unknown
2	**Green Yellow**	GRMZM2G138689_P01	Zm.96787	50 kDa gamma-zein; Zeins are major seed storage proteins	35	35.2	not assigned unknown
2	**Green Yellow**	GRMZM2G060429_P01	zp16	16 kDa gamma-zein; Zeins are major seed storage proteins	35	35.2	not assigned unknown
3	**Green**	GRMZM2G068455_P03	100193663	Malate dehydrogenase	8	8.1.9	TCA / org transformation TCA malate DH
3	**Green**	GRMZM2G097040_P01	100272681	NADH dehydrogenase [ubiquinone] flavoprotein 2 mitochondrial; NADH-ubiquinone oxidoreductase subunit	9	9.1.2	mitochondrial electron transport / ATP synthesis NADH-DH localization not clear
3	**Green**	GRMZM2G143651_P01	GRMZM2G143651_P01	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9 mitochondrial	9	9.1.2	mitochondrial electron transport / ATP synthesis NADH-DH localization not clear
4	**Light Green**	GRMZM2G004528_P03	mips2	Myo-inositol phosphate synthase; Putative inositol-3-phosphate synthase; Uncharacterized protein	3	3.4.3	minor CHO metabolism myo-inositol InsP Synthases

4	**Light Green**	GRMZM2G440208_P01	100856973	6-phosphogluconate dehydrogenase, decarboxylating; Catalyzes the oxidative decarboxylation of 6- phosphor-gluconate to ribulose 5-phosphate and CO(2), with concomitant reduction of NADP to NADPH; Belongs to the 6-phosphogluconate dehydrogenase family	7	7.1.3	OPP oxidative PP 6-phosphogluconate dehydrogenase
4	**Light Green**	GRMZM2G456086_P01	pco087551	Putative ribose-5-phosphate isomerase 3 chloroplastic; Ribose-5-phosphate isomerase; Uncharacterized protein	7	7.2.4	OPP non-reductive PP ribose 5-phosphate isomerase
5	**Medium Sea Green**	GRMZM2G030144_P01	100193027	AP-1 complex subunit gamma-1	31	31.4	cell vesicle transport
5	**Medium Sea Green**	GRMZM2G122135_P02	103638140	Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A beta isoform	29	29.4	protein posttranslational modification
6	**Cyan**	GRMZM2G058760_P01	542713	Ferredoxin--NADP reductase	7	7.3	OPP electron transfer
6	**Cyan**	GRMZM2G090338_P01	cl122_1	Sulfite reductase [ferredoxin], chloroplastic; Essential protein with sulfite reductase activity required in assimilatory sulfate reduction pathway during both primary and secondary metabolism and thus involved in development and growth	14	14.3	S-assimilation sulfite redox
7	**Dark Cyan**	GRMZM2G020446_P01	100273126	Diaminopimelate decarboxylase; Uncharacterized protein	13	13.1.3.5.5	amino acid metabolism synthesis aspartate family lysine diaminopimelate decarboxylase

7	**Dark Cyan**	AC182617.3_FGP001	100273435	Diaminopimelate epimerase chloroplastic	13	13.1.3.5.4	amino acid metabolism synthesis aspartate family lysine diaminopimelate epimerase
8	**Purple**	GRMZM5G875238_P01	sps1	Sucrose-phosphate synthase; Plays a role in photosynthetic sucrose synthesis by catalyzing the rate-limiting step of sucrose biosynthesis from UDP-glucose and fructose- 6-phosphate. Involved in the regulation of carbon partitioning in the leaves of plants. May regulate the synthesis of sucrose and therefore play a major role as a limiting factor in the export of photoassimilates out of the leaf. Plays a role for sucrose availability that is essential for plant growth and fiber elongation; Belongs to the glycosyltransferase 1 family	2	2.1.1.1	major CHO metabolism synthesis sucrose SPS
8	**Purple**	GRMZM2G122231_P01	gpm600	Trehalose-6-phosphate synthase	3	3.2.3	minor CHO metabolism trehalose potential TPS/TPP
9	**Pink**	GRMZM2G053720_P01	pza03240	Proline dehydrogenase; Converts proline to delta-1-pyrroline-5-carboxylate	13	13.2.2.2	amino acid metabolism degradation glutamate family proline
9	**Pink**	GRMZM2G176396_P02	100272740	Proline iminopeptidase	29	29.5	protein degradation
10	**Sandy Brown**	GRMZM5G824920_P01	GRMZM5G824920_P01	Glucan endo-13-beta-glucosidase 3; Putative O-Glycosyl hydrolase superfamily protein; Uncharacterized protein; Belongs to the glycosyl hydrolase 17 family	26	26.4.1	misc beta 1,3 glucan hydrolases glucan endo-1,3-beta-glucosidase

10	**Sandy Brown**	GRMZM2G181259_P01	100279291	Glycosyl hydrolase family protein; Belongs to the glycosyl hydrolase 3 family	10	10.6.1	cell wall degradation cellulases and beta -1,4-glucanases
11	**Brown**	GRMZM2G053803_P02	pco105010	Acyl-CoA binding protein	11	11.1.13	lipid metabolism FA synthesis and FA elongation acyl-CoA binding protein
11	**Brown**	GRMZM2G023667_P02	smh3	Glycylpeptide N-tetradecanoyltransferase; Adds a myristoyl group to the N-terminal glycine residue of certain cellular proteins	29	29.3.4.99	protein targeting secretory pathway unspecified
Unconnected	**NA**	GRMZM2G129451_P02	ss1	Starch synthase, chloroplastic/amyloplastic; Starch synthase I; Belongs to the glycosyltransferase 1 family. Bacterial/plant glycogen synthase subfamily	2	2.1.2.2	major CHO metabolism synthesis starch starch synthase
Unconnected	**NA**	GRMZM2G013002_P01	expB8	Beta-expansin 1a; Belongs to the expansin family	10	10.7	cell wall modification
Unconnected	**NA**	GRMZM2G135132_P01	103653801	Adenosine kinase 2	23	23.3.2.1	nucleotide metabolism salvage nucleoside kinases adenosine kinase
Unconnected	**NA**	GRMZM2G007791_P01	ago2a	Argonaute2a; Belongs to the argonaute family	27	27.3.36	RNA regulation of transcription Argonaute

Unconnected	**NA**	GRMZM2G531230_P01	GRMZM2G531230_P01	annotation not available	13	13.2.3.1.1	amino acid metabolism degradation aspartate family asparagine L-asparaginase
Unconnected	**NA**	GRMZM2G044629_P03	100279878	UDP-N-acetylglucosamine diphosphorylase 2	10	10.1	cell wall precursor synthesis
Unconnected	**NA**	GRMZM2G082032_P01	GRMZM2G082032_P01	Putative isoaspartyl peptidase/L-asparaginase 2	13	13.2.3.1.1	amino acid metabolism degradation aspartate family asparagine L-asparaginase
Unconnected	**NA**	GRMZM2G036464_P01	gln5	Glutamine synthetase root isozyme 4	12	12.2.2	N-metabolism ammonia metabolism glutamine synthase
Unconnected	**NA**	GRMZM5G829778_P01	idh2	Isocitrate dehydrogenase [NADP]; Belongs to the isocitrate and isopropylmalate dehydrogenases family	8	8.1.4	TCA / org transformation TCA IDH
Unconnected	**NA**	GRMZM2G458549_P01	pco060326	Galactokinase	3	3.8.1	minor CHO metabolism galactose galactokinases
Unconnected	**NA**	GRMZM2G146754_P01	100302679	Betaine aldehyde dehydrogenase 2 mitochondrial	16	16.4.2	secondary metabolism N misc betaine
Unconnected	**NA**	GRMZM2G061969_P01	100280155	Phospholipase D; Hydrolyzes glycerol-phospholipids at the terminal phosphodiesteric bond	11	11.9.3	lipid metabolism lipid degradation lysophospholipases

Unconnected	**NA**	GRMZM2G013214_P01	GRMZM2G013214_P01	Uncharacterized protein	16	16.4.2	secondary metabolism N misc betaine
Unconnected	**NA**	GRMZM2G098167_P01	100191598	17.4 kDa class III heat shock protein; Belongs to the small heat shock protein (HSP20) family	20	20.2.1	stress abiotic heat
Unconnected	**NA**	GRMZM2G371375_P01	100280469	Embryonic protein DC-8	35	35.2	not assigned unknown
Unconnected	**NA**	GRMZM2G147882_P03	100192575	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein; Iron ion binding protein; Uncharacterized protein ; Belongs to the iron/ascorbate-dependent oxidoreductase family	17	17.5.1	hormone metabolism ethylene synthesis-degradation
Unconnected	**NA**	GRMZM2G051943_P01	cta1	Endochitinase A; Defense against chitin-containing fungal pathogens; Belongs to the glycosyl hydrolase 19 family. Chitinase class I subfamily	20	20.1	stress biotic
Unconnected	**NA**	GRMZM2G141473_P01	103644157	annotation not available	17	17.1.1	hormone metabolism abscisic acid synthesis-degradation
Unconnected	**NA**	GRMZM2G111143_P01	pco087970b	Glycosyl hydrolase superfamily protein; Belongs to the glycosyl hydrolase 17 family	26	26.4.1	misc beta 1,3 glucan hydrolases glucan endo-1,3-beta-glucosidase

Unconnected	**NA**	GRMZM2G124550_P02	100285555	annotation not available	29	29.4	protein posttranslational modification
Unconnected	**NA**	GRMZM2G052148_P01	rs55622871	annotation not available	29	29.3.1	protein targeting nucleus
Unconnected	**NA**	GRMZM2G045070_P01	GRMZM2G045070_P01	Topoisomerase-like protein	35	35.2	not assigned unknown
Unconnected	**NA**	GRMZM2G057576_P01	GRMZM2G057576_P01	Clathrin heavy chain; Clathrin is the major protein of the polyhedral coat of coated pits and vesicles	31	31.4	cell vesicle transport
Unconnected	**NA**	GRMZM2G162426_P01	103639290	Calcium pump3; Belongs to the cation transport ATPase (P-type) (TC 3.A.3) family	34	34.21	transport calcium
Unconnected	**NA**	GRMZM2G122810_P02	ESMT1	Methyltransferase; Cycloartenol-C-24-methyltransferase 1; Endosperm C-24 sterol methyltransferase; Uncharacterized protein; Belongs to the class I-like SAM-binding methyltransferase superfamily. Erg6/SMT family	17	17.3.1.2.1	hormone metabolism brassinosteroid synthesis-degradation sterols SMT1

Unconnected	**NA**	GRMZM2G328893_P01	GRMZM2G328893_P01	annotation not available	13	13.1.5.3.1	amino acid metabolism synthesis serine-glycine-cysteine group cysteine OASTL
Unconnected	**NA**	GRMZM2G574782_P01	IDP388	Probable bifunctional methylthioribulose-1-phosphate dehydratase/enolase-phosphatase E1 Methylthioribulose-1-phosphate dehydratase Enolase-phosphatase E1; In the C-terminal section; belongs to the HAD-like hydrolase superfamily. MasA/MtnC family	3	3.5	minor CHO metabolism others
Unconnected	**NA**	GRMZM2G053299_P01	umc2243	Actin-7; Belongs to the actin family	31	31.1	cell organization
Unconnected	**NA**	GRMZM2G126453_P01	103652718	AAA-type ATPase family protein	29	29.5.9	protein degradation AAA type
Unconnected	**NA**	GRMZM2G141818_P03	100383316	Argonaute104; Uncharacterized protein; Belongs to the argonaute family	27	27.3.36	RNA regulation of transcription Argonaute
Unconnected	**NA**	GRMZM5G823484_P01	GRMZM5G823484_P01	annotation not available	35	35.2	not assigned unknown

Unconnected	**NA**	GRMZM2G177928_P01	GRMZM2G177928_P01	Strictosidine synthase; Uncharacterized protein	16	16.4.1	secondary metabolism N misc alkaloid-like
Unconnected	**NA**	GRMZM2G052562_P02	znod1	Zea nodulation homolog1; Alpha-L-fucosidase 2; Alpha-L-fucosidase 2 isoform 1; Alpha-L-fucosidase 2 isoform 2; Uncharacterized protein	26	26.28	misc GDSL-motif lipase
Unconnected	**NA**	GRMZM2G126732_P01	acco20	Uncharacterized protein	17	17.5.1	hormone metabolism ethylene synthesis-degradation
Unconnected	**NA**	GRMZM2G051677_P02	frk2	Fructokinase-2; May play an important role in maintaining the flux of carbon towards starch formation. May also be involved in a sugar- sensing pathway	2	2.2.1.1	major CHO metabolism degradation sucrose fructokinase
Unconnected	**NA**	GRMZM2G081886_P01	103645840	annotation not available	13	13.1.5.1.1	amino acid metabolism synthesis serine-glycine-cysteine group serine phosphoglycerate dehydrogenase
Unconnected	**NA**	GRMZM2G019404_P01	mha2	Plasma membrane ATPase	34	34.1	transport p- and v-ATPases

Unconnected	**NA**	GRMZM2G051782_P01	GRMZM2G051782_P01	Tubulin alpha chain	31	31.1	cell organization

Table S2.6 (a) Summary of gene expression variation analysis using Student T-test. Gene and Tag SNP/locus from GWAS, chromosome number (Chr), position, alleles, p-values from Student T-test and adjusted p-values (FDR) were summarized. 18 gene expression variation were significant at adjusted p-values < 0.05 out of 80 tests performed. (b) Pearson correlation analysis between normalized gene expression and corresponding GWAS trait. Bold letters indicate that expression level of 4 genes out of 18 are significantly correlated with the PBAA levels and hence are the candidates for eQTL.

(a)

Significant Gene expression variation due to the GWAS tag SNP polymorphism							
Gene	**Annotation**	**Corresponding Tag SNP from GWAS**	**Chr**	**Position**	**Allele**	**P.value**	**FDR**
GRMZM5G823484	exocyst complex component 5	ss196445606	3	99274528	A/G	2.44E-09	5.74E-08
GRMZM2G126732	ACC oxidase20	S4_177597637	4	177597637	T/A	4.51E-06	0.0001084
GRMZM2G138727	Glutelin-2 Precursor (Zein-gamma)(27 kDa zein)	ss196479450	7	120236783	A/G	1.15E-05	0.0001798
GRMZM2G058760	Ferredoxin--NADP reductase	S1_285140826	1	285140826	G/C	4.34E-05	0.0005206
GRMZM2G177928	strictosidine synthase	S1_276683281	1	276683281	G/T	3.82E-05	0.0005206
GRMZM2G007791	protein argonaute 2	ss196433591	2	9984374	A/C	6.85E-05	0.000644
GRMZM2G456086	ribose-5-phosphate isomerase	ss196522048	7	9335049	A/G	6.00E-05	0.000644
GRMZM5G875238	Sucrose-phosphate synthase	S8_161897525	8	161897525	A/G	9.21E-05	0.0008035
GRMZM2G360681	heat-shock protein 101	S6_160624540	6	160624540	T/A	0.0001004	0.0008035
GRMZM2G531230	Putative isoaspartyl peptidase/L-asparaginase 2	S2_4988770	2	4988770	T/A	0.0024312	0.0122428
GRMZM2G122810	sterol methyltransferase1	S7_18139554	7	18139554	C/A	0.0023762	0.0122428
GRMZM2G181259	Glycosyl hydrolase family protein	S6_148954189	6	148954189	A/C	0.0029763	0.0122428
GRMZM2G019404	plasma-membrane H+ATPase2	S2_4291700	2	4291700	G/A	0.0027085	0.0122428
GRMZM2G371375	embryonic protein DC-8	S2_10778250	2	10778250	C/T	0.0061292	0.0226309
GRMZM2G090338	Sulfite reductase [ferredoxin], chloroplastic;	S6_157019045	6	157019045	G/C	0.0077057	0.0264195
GRMZM2G052148	nucleoporin interacting component	ss196424255	1	59588968	A/G	0.0071034	0.0417326
GRMZM2G125300	40S ribosomal subunit protein S21	ss196510015	5	1526576	A/G	0.0094171	0.0442606
GRMZM5G829778	Isocitrate dehydrogenase [NADP]	ss196524643	6	165506601	A/C	0.0086912	0.0442606

(b)

Correaltion between GWAS candidate gene expression and the trait variation			
Gene	**Trait**	**Correlation coefficient ®**	**P-value**
GRMZM5G823484	M/K	-2.98E-02	0.6827
GRMZM2G126732	H/Z	-6.17E-02	0.3988
GRMZM2G138727	**H/M, H/Z, Z/ZHPR, H/TOTAL, L/IVL, V/A, V/LAV, V/TOTAL**	0.1986084, 0.4979167, -0.474152, 0.3888341, -0.2535665, -0.3826112, -0.3321514, 0.3244373	**0.005145, 9.737e-14, 1.949e-12, 1.637e-08, 0.0003358, 2.88e-08, 1.974e-06, 3.301e-06**
GRMZM2G058760	H/Z, H/ZHPR, **H/TOTAL**	(-0.1747104), -0.1392256, -0.1943974	0.06792, 0.1469, **0.04185**
GRMZM2G177928	M/IMTXK, M, M/TOTAL	(-0.06727557), 0.01722733, -0.06246337	0.404, 0.831 , 0.4385
GRMZM2G007791	M	-6.77E-02	0.352
GRMZM2G456086	H/ZHPR	3.64E-02	0.6075
GRMZM5G875238	V/IVL	0.1349378	0.1068
GRMZM2G360681	V/A	(-0.0274067)	0.7037
GRMZM2G531230	H/Z	-8.25E-02	0.2552
GRMZM2G122810	H/Z	6.22E-02	0.4248
GRMZM2G181259	V/A	-6.46E-02	0.4398
GRMZM2G019404	A/X	6.61E-03	0.9499
GRMZM2G371375	M/K	0.1012778	0.1737
GRMZM2G090338	L/V, **V/LAV**	0.1323089, 0.1524204	0.08733, **0.04857**
GRMZM2G052148	H/X	-7.11E-02	0.323
GRMZM2G125300	M/TOTAL	-0.01657203	0.6037
GRMZM5G829778	**X**	-0.1473831	**3.68E-02**

CHAPTER 3: UNCOVERING THE GENETIC ARCHITECTURE OF FAA IN MAIZE KERNELS USING MULTI-OMICS INTEGRATION.

Abstract

Amino acids are the primary metabolites essential for protein synthesis and precursors for biosynthesis of numerous important compounds. The free amino acid (FAA) pool in seeds contributes to seed vigor, alternative energy source and contributes to seed desiccation and hence, its manipulation can potentially improve the nutritional status of crop. While genome wide association studies (GWAS) on FAAs have identified several regulatory genes mainly in *Arabidopsis*, the genetic regulation of seed FAA levels, composition, and biochemical regulation in maize kernels remains mostly unsolved. Hence, in this study, I first performed GWAS of FAAs as absolute levels, relative composition, and biochemical ratios that I measured from dry kernels of the 282-association panel and extracted a first list of candidate genes. Second, I performed an association analysis between proteomic expression data and the FAA biochemical traits I quantified from a developmental series of kernels from the inbred line B73 and similarly extracted a second list of candidate genes. Comparing the two candidate gene lists revealed 120 key candidate genes that were both strongly associated with seed FAA composition during maturation as well as involved in the natural variation of these traits in dry kernels. Functional analysis of these genes demonstrated a comprehensive genetic architecture of seed FAAs in maize that revealed several biological processes such as protein metabolism that includes translational machinery especially ribosomal proteins (RPs), amino acid metabolism, RNA transcription, sulfur assimilation, cell vesical transport, and TCA cycle. Understanding the coordinated working mechanisms of these

biological process in a system level would be beneficial to engineer the amino acids in seeds.

1. Introduction

Amino acids are primary metabolites that play a central role in plant growth and development as well as in providing key nutrients to humans and livestock. Out of the total amino acid pool, the free amino acid (FAA) pool contributes 1-10% in maize seeds (Amir et al., 2018; Muehlbauer et al., 1994). Despite their low levels in seeds, FAA are functionally diverse and essential for maintaining the overall seed development and seed amino acid homeostasis. In addition to being the building blocks of protein synthesis, FAA also serve as precursors of primary and secondary metabolites, which include organic acids, osmolytes, phytohormones, and secondary metabolites and can be utilized as an alternative source of accessible energy (Amir et al., 2018; Angelovici et al., 2011; G. Galili & Hofgen, 2002). The FAA pool in dry seeds has also been reported to ensure proper desiccation, longevity, germination, and seed vigor (Angelovici et al., 2011; Gad Galili, Avin-Wittenberg, Angelovici, & Fernie, 2014). Hence, due to their multi-faceted roles, the genetic regulation and the homeostasis of the FAA pool is complex, and more studies are needed to fully understand their genetic architecture and homeostasis, especially in seeds.

Despite the low proportion of the FAA pool in seeds, their manipulation can potentially improve the nutritional status of crop (Amir et al., 2018; Gad Galili & Amir, 2013). Several studies focused on increasing FAA to achieved seed amino acid biofortification have demonstrated that FAA are highly interconnected and perturbation of a single gene in the amino acid metabolic pathway can often lead to changes in the

entire amino acid metabolic network, along with adverse effects on agronomic and yield attributing traits (Dizigan et al., 2007; Gu et al., 2010). Hence, understanding more of the underlying genetic basis of the FAA metabolism and how they are intertwined with essential cellular processes in plants will facilitate manipulation of FAAs in seeds without strong deleterious and pleiotropic effects on the entire system.

Like other metabolic traits, FAA are complex metabolic traits and have shown extensive variability and high heritability across natural populations. Genome-wide association study (GWAS) have been used to uncover metabolic QTLs (mQTL) that underlie the natural variation of numerous metabolic traits (Angelovici et al., 2013; Deng et al., 2017; Slaten, Yobi, et al., 2020a; S. Wu et al., 2016), and has contributed to the identification of key regulatory genes or targets for maker assisted selection (MAS) for breeding. For instance, Angelovici et al. (2013) performed GWAS on branched chain amino acid levels in *Arabidopsis* seeds using 360 ecotypes. The authors found a strong association of BCAA with branched chain amino acid transferases: *BCAT1* and *BCAT2*. Using linkage analysis and reverse genetic approaches, the authors found that the allelic variation of *BCAT2* is responsible for the natural variation of seed BCAAs as well as free amino acid homeostasis in seeds (Angelovici et al., 2013). A recent study by Slaten et al (2020a) on FarmCPU GWAS on free glutamine (Gln) related traits in *Arabidopsis* seed identified several candidate genes and using molecular validation. The authors found an unexpected association between the aliphatic glucosinolates and Gln related traits, which indicates that FAA could be controlled by some secondary metabolites (Slaten, Yobi, et al., 2020a). While GWAS on FAAs have identified several regulatory genes mainly in

Arabidopsis, the GWAS study on regulation of seed FAA levels, composition, and biochemical regulation in Maize kernels is still far from being fully understood.

Although GWAS is widely used to dissect the genetic architecture of complex traits, one major limitation of using multiple traits GWAS is that the candidate gene list could be extensive. Previous studies indicated that an integrative multi-omics approach that combines GWAS and co-expression networks can help overcome this challenge and help whittle down the candidate gene list and identify of novel key regulatory genes that are involved in shaping the natural variation of the phenotype of interest (Qi et al., 2018; Schaefer et al., 2018; S. Wu et al., 2016).

Hence, in this study, I first performed GWAS studies of FAA absolute, relative composition, and biochemical ratios that I measured from mature, dry kernels of the 282-association panel. This analysis enabled the extraction of a candidate gene list. Second, I performed a weighted protein correlation network analysis between proteomic expression data and the FAA biochemical traits I quantified from a developmental series of a B73 line. This analysis enabled the extraction of a second candidate gene list. Comparing the candidate gene lists from the two approaches revealed 120 key candidate genes/proteins that were both strongly associated with seed FAA composition during maturation as well as involved in the natural variation of these traits in mature kernels. Functional analysis of these genes demonstrated that several biological processes such as protein metabolism that includes translational machinery especially RPs, amino acid metabolism, RNA transcription, sulfur assimilation, cell vesical transport, and TCA cycle are involved in shaping FAA composition in seeds. These results provide a comprehensive genetic

architecture of seed FAAs in maize and how these processes work in coordination to regulate FAA homeostasis in seeds.

2. Materials and Methods

2.1 Germplasm and GWAS Field Trial

I used the Goodman-Buckler maize association panel that consists of 282 diverse maize lines including tropical and sub-tropical, temperate, sweetcorn and popcorn lines (Flint-Garcia et al., 2005). This inbred panel was grown in the summers of 2017 and 2018 each with two replications each year, using a randomized complete block design at Genetics Farm, near Columbia, MO. Thirteen kernels per inbred line were grown in a 3 m row and each plant was self-pollinated to avoid any cross contamination. All the pollinated ears within a row were harvested at maturity and husk leaves were removed. The resulting ears were dried, shelled, and bulked to form a representative composite grain sample for each inbred line. Twenty-five random seeds from each inbred bulked sample were ground into fine powder for FAA quantification.

2.2 FAA quantification

To characterize the seed FAAs, freeze-dried seed tissue was ground to a fine powder and 7-8 mg were weighed, extracted, and analyzed for FAA quantification using UPLC-MS/MS (Waters Corporation, Milford, MA, USA) as described in (Angelovici et al., 2013). This method allowed us to accurately quantify all 20 proteogenic amino acids (FAA).

2.3 Phenotypic data analysis and GWAS

I evaluated a total of 109 FAAs traits - absolute, relative composition, and biochemical traits. All the traits were treated independently. Metabolic ratio traits were derived prior to the calculation of BLUPs to minimize noise. For each trait, outlier removal, optimal transformation, and BLUP calculation were performed as described in (Slaten, Chan, et al., 2020). Variance component estimates from mixed linear model where taxa, replication, and year were fitted as random effects were used to estimate the broad sense heritability on a line mean basis as described in Holland et al. (2003).

The association panel was previously genotyped with the Illumina MaizeSNP50 BeadChip (Cook et al., 2012) as well as with a genotyping-by-sequencing (GBS) method (Elshire et al., 2011) as described in Lipka et al. (2013). SNPs were filtered using minor allele frequency greater than 0.05 and a total of 407,120 SNPs were used for performing the GWAS analysis. I used GAPIT (Lipka et al., 2012) mixed linear model (MLM) and FarmCPU model (X. Liu et al., 2016) to conduct univariate GWAS. False discovery rate (FDR) (Benjamini & Hochberg, 1995) was used to correct the multiple hypothesis testing problem at 5%. A candidate gene list was obtained using 200 kb window size (100 kb on either side) of the significant SNPs. The physical locations and annotations of the genes were based on Maize AGP_V2; http://ftp.maizesequence.org/release-5b/filtered-set/ .

2.4 Gene functional categorization using MapMan

MapMan version 3.6 (Lohse et al., 2014) was used to assess the functional categorization of the candidate genes from GWAS. GWAS candidate genes were mapped to corresponding bins using the *Zea mays* mapping database

Zm_B73_5b_FGS_cds_2012.m02 obtained from https://mapman.gabipd.org/mapmanstore. A total of 35 functional gene categories were in the mapping database.

2.5 FAA and seed proteome quantification of B73 across ten seed filling stages

The B73 inbred line was grown in the summer of 2018 at Genetics Farm, near Columbia, MO. Samples from 10 different seed filling stages were collected starting at 10 days after pollination (DAP) and continued to sample every four days interval until 46 DAP. The 10 time points collected were 10, 14, 18, 22, 26, 30, 34, 38, 42 and 46 DAP. Three biological replicates were collected from each seed developmental stage. The whole ear from each sample was harvested and frozen into liquid nitrogen and 15 randomly chosen developing seeds were collected and stored immediately in liquid nitrogen. The seeds were then lyophilized for 5 days and then ground using a Perten Laboratory Mill 3310. These samples were then used for FAA quantification as detailed in Angelovici et al. (2013).

2.6 Proteomic analysis

Protein extraction was performed based on the Hurkman & Tanaka (1986) method as described in Yobi et al. (2020). Briefly, 5 mg fine powder were weighed and extracted with Tris-HCl buffered phenol and an SDS extraction buffer. After trypsin digest and purification, the peptides were analyzed on a Bruker timsTOF PRO using the PASEF (1) method. The acquired data were submitted to the PEAKS DB search engine (version 8.5, Bioinformatics Solutions Inc.) for peak picking and protein identification using the MaizeGDB database (Lawrence et al., 2004). Proteins were retained for analysis when their spectral counts were ≥ 4.

2.7 WGCNA analysis

I performed the weighted protein correlation network analysis (WGCNA) (Langfelder & Horvath, 2007) using the R package WGCNA (Langfelder & Horvath, 2008). I chose the soft threshold power $\beta=12$ to construct the co-expression network as the R^2 reached around 80% ensuring the network was close to the scale-free network. I used Pearson correlation to construct the co-expression network. I also constructed protein modules, which is a cluster of densely interconnected proteins with respect to co-expression. In order to construct protein modules from each network, I first used the topological overlap matrix derived from the resulting adjacency matrix. I then converted the topological overlap matrix into a dissimilarity measure by subtracting it from 1 as described in Dong & Horvath (2007) . Hence, the dissimilarity topological overlap measure was then used as input for the average linkage hierarchical clustering to create a dendrogram (clustering tree or modules). I use blockwiseModules function and the dynamic tree cut algorithm (Langfelder et al., 2007) with a height of 0.20 and a minimum module size of 30. Modules are defined as the branches of the dendrogram. Each module of a particular network is assigned to a unique color (e.g. turquoise, black, blue, etc.).

2.8 GO enrichment analysis

GO enrichment analysis was performed using AgriGO_V2 (Tian et al., 2017). The following parameters were used to determine the GO biological process, cellular component, and molecular function terms that were overrepresented ($p<0.05$); a hypergeometric test with a 5% FDR correction, a custom reference that consist of 2648 proteins detected in my proteomics study, *Zea mays* as the select organism, and GO full plant ontologies.

2.9 Protein-protein interaction (PPI) of 80 HCCG using STRING analysis

I constructed and visualized the PPI network associated with 120 HCCG using the Search Tool for the Retrieval of Interacting Genes/Proteins database STRING (STRING V11.0) (Szklarczyk et al., 2019). Active interaction sources including high-throughput lab experiments, gene coexpression as well as the previous knowledge in curated databases specific to species "*Zea mays*" were used to construct the PPI network both at medium confidence level > 0.4 and high confidence level > 0.7 (Szklarczyk et al., 2019). To further investigate the strong interaction among the nodes, I performed the clustering of the PPI network using MCL clustering algorithm within STRING (Szklarczyk et al., 2019).

3. Results

3.1 Both FAAs absolute and relative composition demonstrate high natural variation but only the abs levels show strong correlations

To evaluate the phenotypic variation of FAA absolute and relative composition, 20 proteogenic amino acids were quantified using water based extraction method and LC-MS/MS detection and quantification as described in Angelovici et al. (2013) **(Supplemental Data S3.1)** from 279 inbred lines of Goodman Buckler association panel (Flint-Garcia et al., 2005). Two replicates of this panel were grown in two years (2017, 2018). Relative composition traits were defined as the ratio of an individual amino acid to the sum of the 20 measured amino acids (e.g., Ala/Total).

My result demonstrated that there is an extensive natural variation in FAA both at the absolute levels and relative composition **(Table S3.1a, b and Figure 3.1a, b).** My data show that Pro, Asn, Glu, and Asp were the four most abundant FAAs in their

absolute levels and relative composition, whereas Ile, Trp, Met, and Cys were the four least abundant FAAs in both absolute level and relative composition **(Figure 3.1a, b, Table S3.1a, b)**. In general, the broad sense heritability of the FAA absolute traits was high, with the exception of Met and Thr that showed low heritability (0.48 and 0.42 respectively) (**Table S3.1a)**. The majority of the FAA relative composition also have high heritability with the exception of Cys/Total, Gly/Total and Thr/Total that showed moderate to low heritability (0.58, 0.04, and 0.11 respectively) **(Table S3.1b).** Interestingly, the relative composition broad sense heritability of Met/Total was substantially higher than its absolute level (0.94).

I performed Pearson correlation analysis to evaluate the relationship among the FAA absolute levels and relative composition. For absolute FAA absolute levels, all the pairwise correlations were significant at a qFDR-values <0.05 and were exclusively positive **(Figure 3.1c, Table S3.2a, b)**. The strongest correlation was between Ile and Leu as well as between Ile and Val (r = 0.88), whereas the weakest correlation was between Asn and Pro (r = 0.18) **(Figure 3.1c, Table S3.2a)**. Notably, Asn demonstrated relatively larger number of weaker pairwise correlations with the other FAAs **(Figure 3.1c, Table S3.2a)**. The pairwise correlation analysis of the FAA relative composition showed a very different pattern and consists of both positive and negative correlations and many of the correlations were not significant **(Figure 3.1d, Table S3.2c, d)**. The strongest positive correlation was found to be between the Ile/Total and Val/Total (r = 0.82) whereas the strongest negative correlation was found between Asn/Total and Pro/Total (r = -0.60) as well as between Asn/Total and Thr/Total (r = -0.60) **(Figure 3.1d, Table S3.2c).** In general, only few FAA relative composition had multiple

significant moderate to strong correlations. These amino acids included Asn, Pro, Ala, Val, and Tyr, which exhibit high, moderate, and low abundance. Interestingly, Asn/Total showed the most significant negative correlations to all other FAAs. In general, my results demonstrated that both FAA and their relative composition demonstrate extensive natural variation, but while the absolute FAAs are strongly and positively correlate with each other, their relative composition exhibit a different pattern.

3.2 Functional analysis of my genome wide association study candidate gene list reveals four key functional categories

A total of 109 FAAs GWAS were performed using both FarmCPU (X. Liu et al., 2016) and GAPIT MLM models (Lipka et al., 2013). My SNP datasets included both Illumina MaizeSNP50 BeadChip (Cook et al., 2012) and GBS (Elshire et al., 2011) as described in (Lipka et al., 2013). SNPs were filtered using minor allele frequency greater than 0.05 and a total of 422,161 SNPs were used in my analysis. Results from the MLM model were not significant therefore the results presented are from FarmCPU model. This model proved to successfully uncover key genes underlying FAA related traits in *Arabidopsis* (Slaten, Yobi, et al., 2020a). I conducted a comprehensive GWAS by using absolute FAA levels (nmole/mg), relative composition and biochemical metabolic ratios; relative composition of each trait represented as the ratio of each FAA to the total sum of the FAA measures (i.e. Ala/Total) and metabolic ratios that are based on the potential relationship within each amino acid metabolic family (i.e. Val/ (Ile+Leu+Val= the BCAA family pathway)) **(Table S3.3).** A similar use of amino acid derived traits in GWAS was previously used and described in Angelovici et al., (2016) and Deng et al. (2017)**.** For brevity from this point forward, we will use the one letter code to describe the FAA

biochemical ratio traits; for example, the ratio above will be V/ILV. All the abbreviations of the FAA can be found in **(Table S3.3).**

My analysis revealed 78 traits (out of 109) with significant SNP-trait associations at 5% FDR correction (Benjamini & Hochberg, 1995) and a total of 549 unique (non-redundant) SNPs were found **(Table 3.1, Supplemental Data S3.2)**. A visualization of the distribution of the significant SNP-trait associations using a 1 Mb window span showed several potential hotspots on chromosomes 1, 2, 4, 5, 6, 8, and 9 (**Figure S3.1a)**. The partition of the 526 unique SNPs across the ten maize chromosomes is shown in **Figure 3.2a** and the FAA trait categorical summary of the GWAS results by amino acids family is summarized in **Table 3.1**. The highest number of SNP-trait associations was for the Aspartate family related traits (31%) and the lowest was for the Serine related traits (1%) **(Table 3.1)**. The highest number of candidate genes was also from Aspartate family related traits (32%) followed by FAA absolute levels (23%) and the lowest one was from the Serine (1%) **(Table 3.1)**.

I extracted 2779 unique candidate genes from 200kb intervals centered around the significant SNPs identified for all the 78 significant traits (out of 109) used for my GWAS (100kb up and 100kb downstream) **(Supplemental Data S3.2)** as was described in previous studies and to compensate for a low marker coverage (Ching et al., 2002; Flint-Garcia et al., 2003; J. Yan et al., 2009). A previous GWAS study on maize protein bound amino acids by Deng et. al (2017) has also used 100Kb up and downstream from the peak SNP to extract the candidate gene list The gene name, gene frequency, gene start and stop position and their annotations are provided **(Supplemental Data S3.2)**. The top three highly frequent genes identified were GRMZM2G427672 (RNA-binding KH

domain-containing protein), repeated 8 times across the FAA traits, some traits are duplicated with different SNPs (M/I, T/M, A, F, V, A, D, and T); GRMZM2G138097 (hypothetical protein LOC100278893), repeated 7 times (R, E/R, R/E, R/EHPRQ, R, E/R and K); and GRMZM2G178460 (Putative envelope ADPATP carrier protein chloroplastic), repeated 7 times across the FAA traits (E/R, R/E, R/EHPRQ, R, E/R, K and R) **(Supplemental Data S3.2)**. Also, the three most significant genes identified based on the lowest p-values are GRMZM2G164352 (pvalue 2.23E-16) associated with Met; GRMZM2G386095 (pvalue 2.23E-16) also associated with Met and GRMZM2G178182 (pvalue 2.23E-15) associated with Tyr **(Supplemental Data S3.2)**.

I further tested whether any specific biological process or pathway were enriched in the 2779 candidate genes, I performed an enrichment analysis using both AgriGO (Tian et al., 2017) and STRING analyses (Szklarczyk et al., 2019), but no enrichments were found. Hence, to assess the functional categorization of the candidate genes, I used MapMan version3.6 (Lohse et al., 2014). The resulting proportion of genes assigned to the respective functional bins is visualized using pie chart **(Figure 3.2b)**. The results showed that protein, RNA, signaling, and transport had the highest contributions with 11.9%, 11.2%, 5.3%, and 3.71 respectively. **(Figure 3.2b)**. Overall, my results suggest the FAA genetic architecture may be dominated by these four functional categories.

3.3 Individual FAA absolute levels overall decreases during kernel maturation while its relative composition showed mixed trends.

I set up an orthogonal experiment where I collected FAA and proteomic data from 10 different seed filling stages of inbred line B73 with a rationale to compare these results with my previous GWAS results to generate a high confidence candidate gene list

that is involved in seed FAA regulation and homeostasis. I collected 15 developing seeds from maize ear with three biological replicates from 10 time-points. Collection of samples was done every 4 days starting at 10 days after pollination (DAP) through desiccation (46 DAP-dry kernels). I started my sampling at 10 DAP since it is the stage when the seeds begin to accumulate storage compounds, especially the SSPs (Larkins, 2017; Prioul et al., 2008).

Next, I quantified the FAA levels and relative composition in each sample in the same manner described above **(Supplemental Data S3.3)**. Hierarchical clustering was used to cluster the trends of FAA accumulation of both absolute levels and relative composition across the 10 seed developmental stages **(Figure 3.3a-d)**. There are three observed patterns of FAA absolute levels (nmol/mg) **(Figure 3.3a, b)**. Interestingly, all FAA levels (nmol per mg seed) had an initial drop from 10 days to 18 DAP. However, the FAAs in cluster I (Gly, Cys, Met, and Trp) displayed a slight increase from 18 DAP up to 26 DAP and then decreased gradually for the most parts from that point on, while in cluster II (Asn, Arg, Pro, Asp and Gln), the levels plummeted from 10 DAP to 18 DAP and then decreased gradually until maturity. Similarly, Cluster III (Lys, Thr, Ile, Leu, Val, His, Tyr, Glu, Phe, Ala and Ser) levels showed a sharp decline from 10 to 18 DAP but then remain stable until 26 DAP before declining again to the lowest levels at 46 DAP **(Figure 3.3a, b)**. Nevertheless, since seed weights and density changes across development, the FAA relative composition dynamics is more biologically relevant and is more representative of the metabolic alteration in the seed. Indeed, when I analyzed the pattern of accumulation of the FAA relative composition, I detected different clustering and patterns. The trends of FAA relative composition indicated that Gly relative

composition was unique as it increased gradually to peak at 34 DAP before decreasing slowly towards the end of seed desiccation (Cluster I), while the relative composition of Ala, Ile, and Met increased towards the 38 and 42 DAP and decreased later on (Cluster II). Interestingly, the majority of FAA relative composition (i.e., Trp, Asn, Cys, Tyr, Phe, His, Pro, Leu, Arg, and Lys) showed mixed trends until 26 DAP and then showed very sharp increase from 34 DAP to 42 DAP before decreasing slightly at the end of seed desiccation (Cluster III), while Asp and Gln showed highest abundance at 10 DAP followed by a decrease (Cluster IV). The relative composition of Thr, Val, Glu, and Ser showed mixed patterns with first peak at 18 DAP, then a decrease at low levels till 34 DAP, and an increase again till 42 DAP followed by a slight decrease at desiccation (Cluster V) **(Figure 3.3c, d).**

3.4 WGCNA reveals strong associations between specific protein co-expression modules and FAA relative composition.

I analyzed protein expression levels from ten seed filling time points using a shotgun proteomic approach (see Material and Methods). The analysis altogether identified 6361 proteins. Proteins with low spectral count and poor reproducibility were removed (see material and Methods) leaving 2648 good quality proteins **(Supplemental Data S2.4)**. Weighted gene co-expression network analysis (WGCNA) (Langfelder & Horvath, 2008) was performed on the filtered proteins and an undirected and weighted protein co-expression network was constructed using optimum soft threshold **(Figure S2.2)**.

Altogether, the 2648 proteins were assigned to 8 distinct modules: blue, turquoise, brown, green, yellow, black, red and grey **(Figure 3.4a)**. The list of proteins found in

each module and their respective annotations can be found in **(Supplemental Data S2.5).** I found that the turquoise module consists of highest number of proteins (853 proteins) followed by grey (805), blue (356), brown (201), yellow (190) , green (107), red (70) and black (66) **(Figure 3.4a; Supplemental Data S2.5).** Visualization of the expression pattern of eigen proteins of a particular module indicated that the blue and turquoise module proteins generally decreased towards the maturity. However, while in the blue module, I observed a sharp decrease from 10 DAP till 18 DAP followed by a gradual decrease, in the turquoise module, I observed a gradual decrease from 18 DAP **(Figure 3.4b)**. Proteins in brown modules increased gradually towards the maturity and then maintained constant levels, whereas in the green modules, the proteins showed a large increase during late maturation. Proteins in the yellow module peaked in 18 DAP and for the most part reduced thereafter; proteins in black modules increased to peak at 30 DAP followed by a slight reduction close to the maturity; proteins in red modules had their main expression peak at 26 DAP, which then decreased during maturity, while proteins in grey modules had mixed patterns of increase and decrease of expression throughout seed development **(Figure 3.4b)**.

To identify genes that are highly associated with the dynamics of FAA, the Eigengene-based module connectivity or module membership (kME) for a particular protein within a given modules was calculated. The full list of proteins and their kME within a respective module can be found in **Supplemental Data S2.5.**

However, my main focus was to further reveal functional relationships between the seed FAA relative composition with the protein expression pattern across the seed developmental stages. Unlike PBAA relative composition, the correlation of FAA

composition traits with protein modules appears to be vague and difficult to interpret **(Figure 3.4a**); I therefore used both FAA absolute and relative composition traits to correlate with protein co-expressed modules. I used WGCNA to perform a correlation between the absolute and relative FAAs composition traits with the protein expression modules. To create modules-FAA association, module Eigen protein (ME), the first principal component of a given module was calculated. The correlations between MEs of each module and the FAA composition traits are shown in **Figure 3.4a.** The five modules (the blue, turquoise, brown, green, and black) demonstrated significant correlations (r = ~ 0.7-0.94) with multiple FAA absolute and compositional traits. The grey, yellow, and red modules have relatively low correlations with the FAA relative composition traits and so they are excluded from my downstream analysis **(Figure 3.4a)**. Hence, I limited my further analysis to those five modules: the blue, turquoise, brown, green, and black modules.

GO enrichment analysis of the proteins identified from five modules in WGCNA were carried out using AgriGO_V2 (Tian et al., 2017). Enrichment for the blue, turquoise, brown, and black module proteins can be found in **Table S2.4**. I did not find enrichment with the proteins from the green modules. Interestingly, I found enrichments specific to modules. For examples, the blue module was enriched with nucleoside metabolic process; the turquoise module was enriched with translation and protein metabolic process; the brown module was enriched with response to light intensity and several stress; the and black module was enriched with carbohydrate and branched chain amino acid metabolic process **(Table S2.4)**. Hence, altogether I have a list of 1583 candidate proteins from the five WGCNA modules that have high correlation with

relative FAA compositional traits. The 1583 candidate proteins were later converted into their respective gene ID and used as orthogonal gene list to compare with previous GWAS candidate gene list to prioritize the high confidence candidate gene list.

3.5 Intersecting the GWAS and WGCNA candidate gene lists revealed 120 High confidence candidate genes.

My underlying assumption for the integration of my two omics approaches was that the genes that appear in both GWAS and WGCNA analyses are genes that do not randomly overlap, and therefore are key regulatory genes. Hence, I compiled a high confidence candidate gene list by intersecting genes that were highly associated with our traits in both approaches. The overlap between the two analyses resulted into 120 genes (**Figure 3.5)**, referred from here on as high confidence candidate genes (HCCG) **(Table S3.4).**

3.6 STRING analysis of 120 HCCG reveals complex genetic architecture of FAAs.

I constructed and visualized the PPI network associated with 120 HCCG using the Search Tool for the Retrieval of Interacting Genes/Proteins database (STRING V11.0) (Szklarczyk et al., 2019). STRING integrates both known and predicted PPIs, which can be applied to predict functional interactions of proteins.

The PPI network generated by STRING consists of 120 nodes (53 connected and 67 unconnected-not shown in figure) and 124 edges, with an average node degree of 2.07 **(Figure 3.6a)**. Each node represents the HCCG and the edges represent the interaction between them. The number of edges is significantly larger (p-value 1.25e-05) than the expected for random network of the same size. To investigate the strong interaction among the nodes, I performed the clustering of the PPI network using MCL clustering

algorithm in STRING, which results in 9 clusters **(Figure 3.6a)**. The genes and the clusters are summarized in **Table S3.4.** I used MapMan version 3.6 (Lohse et al., 2014) to assess the functional categorization of the 9 clusters **(Figure 3.6a, b and Table S3.4).** MapMan functional categorization shows many major biological processes that includes cluster 1 genes (red) categorized as protein synthesis, RNA transcription, and transport; cluster 2 (sandy brown) as TCA and lipid metabolism; cluster 3 (brown) as protein degradation (proteasome); cluster 4 (green yellow) as CHO metabolism; cluster 5 (green) as protein post translational modification (PPTM); cluster 6 (cyan) as cell wall related proteins; cluster 7 (dark cyan) as sulfur assimilation; cluster 8 (cornflower blue) as amino acid metabolism; and cluster 9 (pink) as cell vesical transport **(Figure 3.6a, b and Table S3.4)** which indicates that the FAAs are complex traits and these interconnected gene and gene clusters regularly coordinated to maintain the FAAs homoeostasis in the maize kernels.

4. Discussion

The FAAs are important for multiple biological processes including seed germination, desiccation, longevity, and nutrition. Several studies have shown that the manipulation of specific FAAs can have a pleiotropic, systemwide effect on overall growth and germination indicating that their metabolism and regulation is intertwined with other key biological and metabolic processes in the seeds (Amir et al., 2018; Gad Galili & Amir, 2013). A fundamental understanding of the genetic and metabolic bases of seed FAA composition in crops in general and in maize specifically is essential for the development of effective strategies for seed amino acid biofortification. In this study, I show that using a multi-omics integration approach that utilizes both the genetic basis of

FAA natural variation and their association with proteome dynamics during seed development and maturation identified several biological process which play a key role in seed FAA regulation: RNA transcription, protein metabolism and transport, amino acids metabolism, TCA cycle and carbohydrate metabolism,.

4.1 Genome wide association studies revealed proteins and RNA metabolism as major functional categories in the resulting candidate gene list.

Seed FAAs are complex traits which display natural variation across natural and artificial populations in maize as well as in other crops (Angelovici et al., 2016; Slaten, Yobi, et al., 2020b). My study shows that both absolute levels (nmol/mg) and relative composition (FAA/TFAA) express significant natural variation **(Figure 3.1a, b).** A correlation analysis of both traits across the association panel showed that, despite a strong and significant positive correlation between the FAA absolute levels, only a handful of relative FAA composition were significant and showed both positive and negative correlations **(Figure 3.1c, d).** The relative composition of Ile and Val as well as Ile and Leu demonstrate a very strong positive correlation **(Figure 3.1d).** The amino acids Ile, Leu, and Val are biosynthetically and chemically related as they all have branched hydrocarbon side chains (Binder, 2010) and fall under BCAA family. BCAAs are essential amino acids and hence essential nutrients for humans and animals. It is reported that the genetic disruption of BCAA catabolism has little effect on leaf amino acid profiles, but severely influences amino acid metabolism in seeds (Angelovici et al., 2013; Gu et al., 2010). Interestingly, the relative composition of Asn and Pro has demonstrated negative correlation with most of the other FAAs **(Figure 3.1d).** It might be because these amino acids are the most abundant in the dry seeds as compared to the

other amino acids, however the biological meaning for this negative correlation in the dry seeds needs to be further investigated. Asn is one of the most important AAs related to nitrogen metabolism as it has N:C ratio of 2:4, which makes it an efficient molecule for the storage and transport of nitrogen in living cells (Lea, Sodek, Parry, Shewry, & Halford, 2007) and can represent 50% of the total free amino acids in the developing cotyledon (Hernandez-Sebastia et al., 2005), hence it might be one of the reasons for its high abundance in the dry seeds. Pro and Asn contribute the most to the FAA pool in dry seeds as Asn is used as a nitrogen storage that is used later during germination, while Pro is the osmoprotectant, mainly used for seed desiccation and maturation (Amir et al., 2018).

I included in my GWAS all the relevant metabolic ratios of FAA based on their biochemical affiliation and considered the candidate genes as part of a comprehensive architecture of the FAA genetic architecture. A similar approach proved very effective in other metabolic studies in both maize and *Arabidopsis* (Angelovici et al., 2013; Deng et al., 2017; Slaten, Yobi, et al., 2020b). My GWAS and candidate gene extraction approach yielded relatively a large number of unique candidate genes for all the FAA relative traits that is likely due to the relatively permissive statistical correction. However, since I aimed to intersect the GWAS candidate gene list with an orthogonal candidate gene list, the latter did not pose a concern. Nevertheless, functional analysis indicated that both "proteins" as well as "RNA" were the largest identified categories **(Figure 3.2b)**, eluding that protein metabolism as well as gene expression may be in the heart of the genetic basis of seed FAA too. Although the functional category for regulating the FAA and

PBAA seems to be consistent, it will be interesting to further investigate whether there are same or distinct genes underlying those categories for FAA and PBAA.

4.2 Multi-omics integration revealed one hundred and twenty high confidence candidate genes

One of the disadvantages of performing large omics analysis is that often we end up with a long list of candidate genes that is difficult to prioritize in a significant or meaningful way, which makes it very hard to uncover biological processes that underlie our traits of interest and identify key regulatory genes for further analysis. To address this challenge, I used two orthogonal approaches and intersected them to pinpoint the genetic and metabolic bases of FAA composition in seed. Previous studies have shown that intersecting GWAS with network analysis from a time development or orthogonal experiment is very efficient (Qi et al., 2018; Schaefer et al., 2018; S. Wu et al., 2016). My underlying assumptions is that the genes that are associated in both analyses are key regulatory genes determining FAA composition in seeds. Overall, I have identified 120 HCCG.

Using a protein-protein interaction analysis to uncover the functional relationship among these genes revealed several biological processes that were previously shown to be involved in seed FAA composition. The previously characterized ones were amino acid metabolism, sulfur assimilation, CHO metabolism, and cell vesical transport **(Figure 3.6a).** For instance, I found Glutamate decarboxylase (GAD) gene cluster (GRMZM2G017110 and GRMZM2G098875) as one of the major amino acid metabolism gene in one of my HCCG clusters **(Figure 3.6a, b - Cluster8 and Table S3.4).** The GWAS results showed that GRMZM2G017110 was significantly associated

with BCAA traits (Val, Leu) and with Ala while GRMZM2G098875 was significantly associated aspartate related traits (D/IMNTDK, N/D, N/IMNTDK) **(Supplemental Data S3.2)**. GAD carries out the alpha-decarboxylation of glutamate to yield the nonprotein amino acid gamma-aminobutyrate (GABA) and carbon dioxide (Singh, 1998). The level of GABA has been reported to increase during the maturation stage of seed development and was suggested to play an important role in seeds (Fait et al., 2006). Studies have also shown that a rapid formation of GABA in developing seeds will degrades to succinate on the onset of germination (Tuin & Shelp, 1994). Therefore, accumulation of GABA in the mature dry seeds are suggested to supply available energy to seeds for later germination (Fait et al., 2011). Hence, GAD is important for the biosynthesis of GABA and the latter plays an important role in regulating the carbon nitrogen balance, amino acid biosynthesis as well as storage reserve accumulation during seed development and germination by modulating the flux of carbon and energy through the TCA cycle (Fait, Fromm, Walter, Galili, & Fernie, 2008). Another important gene identified from this analysis is the cysteine synthase (GRMZM2G005887), also known as O-acetyl-L-serine (OAS) or o-acetyl serine sulfhydrylase (OASS) **(Figure 3.6a, b-Cluster 7 and Table S3.4).** The gene was found to be associated with FAA absolute traits, Isoleucine in my GWAS result **(Supplemental Data S3.2),** which is interesting as previous studies have highlighted the potential regulation of this genes with sulfur related amino acids such as Cys and Met, but not Ile (Kim et al., 2012). Therefore, the connection of OAS with Ile needs to be further investigated**.** Several studies in soybean seeds have shown the promising nature of this gene in improving the amino acid composition in seeds. OAS is reported to be a very important metabolic gene that regulates the sulfur and nitrogen

regulation of soybean seed storage protein composition (Kim et al., 1999). Kim et. al (2012) demonstrated that the overexpression of cytosolic OASS resulted in a 58-74% increase in protein-bound cysteine levels compared with non-transformed wild type soybean seeds, and a 22-32% increase in free cysteine levels, which was important to improve the sulfur related essential amino acids in soybean (Kim et al., 2012). I found another gene, sulfite reductase (GRMZM2G090338), within the same cluster **(Cluster 7-Figure 3.6a, b),** which is also related to sulfur assimilation (Xia, Wang, & Xu, 2018) indicating their association in the similar biological processes.

Interestingly, like my PBAA, my multi-omics approach analysis also supported the core protein metabolism and translation machinery role in seed FAA composition. My PPI analysis revealed that 120 HCCG included a tightly connected group dominated by genes related to protein metabolism with 7 ribosomal proteins and 3 translational initiation factor eiF2a, eIF2b and EIF3L **(Table S3.4).** Interestingly, the protein expression of all the proteins in this cluster (falls under turquoise module excluding the eIF3L that falls under WGCNA Brown module – **Supplemental Data S2.4, Figure 3.4b**) displayed reduction during maturation and desiccation, which indicated that some of the FAA relative composition traits such as Asp and Gln are positively associated with this reduction, while Arg, Asn, Leu and Lys are negatively associated with this reduction **(Figure 3.4a,b)** indicating some sort of switch between the protein and FAA metabolites. However, this attenuation of proteins and their correlation with FAA metabolites is not as clear as compared with PBAA correlation with these protein modules and needs further investigation. This might be due to the multi-faceted role of FAAs making it more complex to understand and interpret, while PBAA functions mainly as storage protein

reservoir. Hence, it might be useful to further breakdown FAA traits into its several amino acids biochemical families to better visualize and interpret this complex analysis.

My PPI analysis and integrative approach indicated that one of the key factors in shaping the seed FAA is specific translation attenuation, most probably driven by alteration in specific components of the translational machinery namely the ribosomal proteins. Although the ribosomal proteins are found to be associated with both PBAA and FAA, it will be interesting to further investigate whether the same or distinct ribosomal proteins regulate the seed FAA and PBAA composition. Previous studies have shown the increase in FAA, especially Lys, is associated with expression of genes encoding ribosomal proteins, as well as those encoding translation initiation and elongation factors, all of which are associated with protein synthesis (Angelovici et al., 2009). In recent years, increasing evidence have shed light on the adaptive nature of RPs as well as their potential selective regulation of translation. Multiple RP mutants revealed the functional role they have in developmental and stress related processes (Ma & Dooner, 2004; Tzafrir et al., 2004; H. Yan et al., 2016; J. Zhang et al., 2016). Several studies also revealed that many RPs are transcriptionally regulated during stress such in sucrose feeding (Gamm et al., 2014), cold, heat, and UV-B stress (Sáez-Vásquez & Delseny, 2019; Sormani et al., 2011) as reviewed in (Martinez-Seidel et al., 2020).

In addition, support for the key role that the translational regulation play in seed FAA composition is the identification of several eIF as my key HCCG. I found eIF2α, eIF2β and eIF3 as my HCCG. Previous studies has reported that elongation factor 1a (EF-1α) concentration is highly correlated with the lysine content in maize endosperm (Habben et al., 1995; M. Jia et al., 2013). Habben et al (1995) reported that the mRNA

levels of EF-1α was significantly increased in the endosperm on *W64Ao2* plants compared with its normal wild type counterpart (Habben et al., 1995). The author mentioned that EF-1a is a lysine rich protein that consists of around 10% Lys and binds to the aminoacyl-tRNAs to the ribosome (Habben et al., 1995). Hence, it will be interesting to investigate the correlation of eIF2α and other eIF HCCG in high lysine maize lines. I also found one of the eIF3 as one of my key HCCG. eIF3 is reported to consist of 12 subunits and is the most complex and largest initiation factor, participating in nearly all major steps of translation initiation (Browning & Bailey-Serres, 2015; Merchante et al., 2017). Studies have shown that alteration in eIFs can have drastic effect on translation, either global repression or upregulation (Merchante et al., 2017). Various enhanced expressed proteins that include several EIFs such as eIF2 and GAPDH have been identified as high-lysine containing proteins that add a substantial contribution to the overall lysine elevation in high lysine maize (M. Jia et al., 2013). In addition to this evidence on translation regulation, a recent study on the genomic prediction of FAAs using prior biological processes in *Arabidopsis* dry seeds revealed that protein metabolism, amino acid metabolic pathway and specialized metabolism contribute to the genetic architecture of FAAs (Turner-Hissong et al., 2020).

5. Conclusion

This study demonstrates the effectiveness of undertaking integrative multi-omics approach to uncover key biological processes that are involved in seed FAA composition in maize. Using this approach, I have identified and highlighted some previously well characterized amino acid metabolic genes, several TCA cycle, and CHO metabolism related genes. Finally, I uncovered the complex dynamic and heterogeneity of the

translational machinery, especially RPs in shaping the FAA composition in seeds, which will lead to new and exciting avenues for seed amino acid biofortification. The biological processes for regulating both the PBAA and FAA are similar with genes related to the functional category of protein metabolism and translation as well as RNA are in the heart of PBAA and FAA regulation, however investigating further on whether the same or distinct genes underlying those functional category will shed light in the inter-regulation of seed PBAA and FAA.

6. References

Amir R, Galili G, Cohen H (2018) The metabolic roles of free amino acids during seed development. Plant Science

Angelovici R, Batushansky A, Deason N, Gonzalez-Jorge S, Gore MA, Fait A, DellaPenna D (2016) Network-guided GWAS improves identification of genes affecting free amino acids. Plant physiology:pp. 01287.02016

Angelovici R, Fait A, Fernie AR, Galili G (2011) A seed high-lysine trait is negatively associated with the TCA cycle and slows down Arabidopsis seed germination. New Phytologist 189 (1):148-159

Angelovici R, Fait A, Zhu X, Szymanski J, Feldmesser E, Fernie AR, Galili G (2009) Deciphering transcriptional and metabolic networks associated with lysine metabolism during Arabidopsis seed development. Plant Physiology 151 (4):2058-2072

Angelovici R, Lipka AE, Deason N, Gonzalez-Jorge S, Lin H, Cepela J, Buell R, Gore MA, DellaPenna D (2013) Genome-wide analysis of branched-chain amino acid levels in Arabidopsis seeds. The Plant cell 25 (12):4827-4843

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological):289-300

Binder S (2010) Branched-chain amino acid metabolism in Arabidopsis thaliana. The Arabidopsis Book/American Society of Plant Biologists 8

Browning KS, Bailey-Serres J (2015) Mechanism of cytoplasmic mRNA translation. The Arabidopsis book/American Society of Plant Biologists 13

Ching A, Caldwell KS, Jung M, Dolan M, Smith OSH, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC genetics 3 (1):19

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia SA (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant physiology 158 (2):824-834

Deng M, Li D, Luo J, Xiao Y, Liu H, Pan Q, Zhang X, Jin M, Zhao M, Yan J (2017) The genetic architecture of amino acids dissection by association and linkage analysis in maize. Plant biotechnology journal 15 (10):1250-1263

Dizigan MA, Kelly RA, Voyles DA, Luethy MH, Malvar TM, Malloy KP (2007) High lysine maize compositions and event LY038 maize plants. Google Patents,

Dong J, Horvath S (2007) Understanding network concepts in modules. BMC systems biology 1 (1):24

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6 (5):e19379

Fait A, Angelovici R, Less H, Ohad I, Urbanczyk-Wochniak E, Fernie AR, Galili G (2006) Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. Plant physiology 142 (3):839-854

Fait A, Fromm H, Walter D, Galili G, Fernie AR (2008) Highway or byway: the metabolic role of the GABA shunt in plants. Trends in plant science 13 (1):14-19

Fait A, Nesi AN, Angelovici R, Lehmann M, Pham PA, Song L, Haslam RP, Napier JA, Galili G, Fernie AR (2011) Targeted enhancement of glutamate-to-γ-aminobutyrate conversion in Arabidopsis seeds affects carbon-nitrogen balance and storage reserves in a development-dependent manner. Plant physiology 157 (3):1026-1042

Flint-Garcia SA, Thornsberry JM, Buckler IV ES (2003) Structure of linkage disequilibrium in plants. Annual review of plant biology 54 (1):357-374

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. The Plant Journal 44 (6):1054-1064

Galili G, Amir R (2013) Fortifying plants with the essential amino acids lysine and methionine to improve nutritional quality. Plant biotechnology journal 11 (2):211-222

Galili G, Avin-Wittenberg T, Angelovici R, Fernie AR (2014) The role of photosynthesis and amino acid metabolism in the energy status during seed development. Frontiers in plant science 5:447

Galili G, Hofgen R (2002) Metabolic engineering of amino acids and storage proteins in plants. Metab Eng 4 (1):3-11. doi:10.1006/mben.2001.0203

Gamm M, Peviani A, Honsel A, Snel B, Smeekens S, Hanson J (2014) Increased sucrose levels mediate selective mRNA translation in Arabidopsis. BMC plant biology 14 (1):306

Gu L, Jones AD, Last RL (2010) Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant. The Plant Journal 61 (4):579-590

Hernandez-Sebastia C, Marsolais F, Saravitz C, Israel D, Dewey RE, Huber SC (2005) Free amino acid profiles suggest a possible role for asparagine in the control of storage-product accumulation in developing seeds of low- and high-protein soybean lines. Journal of experimental botany 56 (417):1951-1963. doi:10.1093/jxb/eri191

Holland JB, Nyquist WE, Cervantes-Martínez CT (2003) Estimating and interpreting heritability for plant breeding: an update. Plant breeding reviews 22

Hunter BG, Beatty MK, Singletary GW, Hamaker BR, Dilkes BP, Larkins BA, Jung R (2002) Maize opaque endosperm mutations create extensive changes in patterns of gene expression. The Plant cell 14 (10):2591-2612

Hurkman WJ, Tanaka CK (1986) Solubilization of plant membrane proteins for analysis by two-dimensional gel electrophoresis. Plant Physiol 81 (3):802-806

Jia M, Wu H, Clay KL, Jung R, Larkins BA, Gibbon BC (2013) Identification and characterization of lysine-rich proteins and starch biosynthesis genes in the opaque2mutant by transcriptional and proteomic analysis. BMC plant biology 13 (1):60

Kim H, Hirai MY, Hayashi H, Chino M, Naito S, Fujiwara T (1999) Role of O-acetyl-L-serine in the coordinated regulation of the expression of a soybean seed storage-protein gene by sulfur and nitrogen nutrition. Planta 209 (3):282-289

Kim W-S, Chronis D, Juergens M, Schroeder AC, Hyun SW, Jez JM, Krishnan HB (2012) Transgenic soybean plants overexpressing O-acetylserine sulfhydrylase accumulate enhanced levels of cysteine and Bowman–Birk protease inhibitor in seeds. Planta 235 (1):13-23

Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC systems biology 1 (1):54

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics 9 (1):559

Langfelder P, Zhang B, Horvath S (2007) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24 (5):719-720

Larkins BA (2017) Maize Kernel Development. CABI,

Lawrence CJ, Dong Q, Polacco ML, Seigfried TE, Brendel V (2004) MaizeGDB, the community database for maize genetics and genomics. Nucleic acids research 32 (suppl_1):D393-D397

Lea PJ, Sodek L, Parry MA, Shewry PR, Halford NG (2007) Asparagine in plants. Annals of Applied Biology 150 (1):1-26

Lipka AE, Gore MA, Magallanes-Lundback M, Mesberg A, Lin H, Tiede T, Chen C, Buell CR, Buckler ES, Rocheford T (2013) Genome-wide association study and pathway level analysis of tocochromanol levels in maize grain. G3: Genes, Genomes, Genetics:g3. 113.006148

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28 (18):2397-2399

Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS genetics 12 (2):e1005767

Lohse M, Nagel A, Herter T, May P, Schroda M, Zrenner R, Tohge T, Fernie AR, Stitt M, Usadel B (2014) M ercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant, cell & environment 37 (5):1250-1258

Ma Z, Dooner HK (2004) A mutation in the nuclear-encoded plastid ribosomal protein S9 leads to early embryo lethality in maize. The Plant Journal 37 (1):92-103

Martinez-Seidel F, Beine-Golovchuk O, Hsieh Y-C, Kopka J (2020) Systematic review of plant ribosome heterogeneity and specialization. Frontiers in Plant Science 11:948

Merchante C, Stepanova AN, Alonso JM (2017) Translation regulation in plants: an interesting past, an exciting present and a promising future. The Plant Journal 90 (4):628-653

Muehlbauer G, Gengenbach B, Somers D, Donovan C (1994) Genetic and amino-acid analysis of two maize threonine-overproducing, lysine-insensitive aspartate kinase mutants. Theoretical and applied genetics 89 (6):767-774

Prioul JL, Méchin V, Lessard P, Thévenot C, Grimmer M, Chateau-Joubert S, Coates S, Hartings H, Kloiber-Maitz M, Murigneux A (2008) A joint transcriptomic, proteomic and metabolic analysis of maize endosperm development and starch filling. Plant biotechnology journal 6 (9):855-869

Qi Z, Zhang Z, Wang Z, Yu J, Qin H, Mao X, Jiang H, Xin D, Yin Z, Zhu R (2018) Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development. Plant, cell & environment 41 (9):2109-2127

Sáez-Vásquez J, Delseny M (2019) Ribosome biogenesis in plants: from functional 45S ribosomal DNA organization to ribosome assembly factors. The Plant cell 31 (9):1945-1967

Schaefer RJ, Michno J-M, Jeffers J, Hoekenga O, Dilkes B, Baxter I, Myers CL (2018) Integrating coexpression networks with GWAS to prioritize causal genes in maize. The Plant cell 30 (12):2922-2942

Singh BK (1998) Plant amino acids: biochemistry and biotechnology. CRC Press,

Slaten ML, Chan YO, Shrestha V, Lipka AE, Angelovici R (2020a) HAPPI GWAS: Holistic Analysis with Pre and Post Integration GWAS. Bioinformatics. doi:10.1093/bioinformatics/btaa589

Slaten ML, Yobi A, Bagaza C, Chan YO, Shrestha V, Holden S, Katz E, Kanstrup C, Lipka AE, Kliebenstein DJ (2020b) mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis. Plant Physiology

Sormani R, Masclaux-Daubresse C, Daniele-Vedele F, Chardon F (2011) Transcriptional regulation of ribosome components are determined by stress according to cellular compartments in Arabidopsis thaliana. PloS one 6 (12):e28070

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research 47 (D1):D607-D613

Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z (2017) agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic acids research 45 (W1):W122-W129

Tuin LG, Shelp BJ (1994) In situ [14C] glutamate metabolism by developing soybean cotyledons I. Metabolic routes. Journal of plant physiology 143 (1):1-7

Turner-Hissong SD, Bird KA, Lipka AE, King EG, Beissinger TM, Angelovici R (2020) Genomic prediction informed by biological processes expands our understanding of the genetic architecture underlying free amino acid traits in dry Arabidopsis seeds. G3: Genes, Genomes, Genetics

Tzafrir I, Pena-Muralla R, Dickerman A, Berg M, Rogers R, Hutchens S, Sweeney TC, McElver J, Aux G, Patton D (2004) Identification of genes required for embryo development in Arabidopsis. Plant physiology 135 (3):1206-1220

Wu S, Alseekh S, Cuadros-Inostroza Á, Fusari CM, Mutwil M, Kooke R, Keurentjes JB, Fernie AR, Willmitzer L, Brotman Y (2016) Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in Arabidopsis thaliana. PLoS genetics 12 (10):e1006363

Xia Z, Wang M, Xu Z (2018) The maize sulfite reductase is involved in cold and oxidative stress responses. Frontiers in plant science 9:1680

Yan H, Chen D, Wang Y, Sun Y, Zhao J, Sun M, Peng X (2016) Ribosomal protein L18aB is required for both male gametophyte function and embryo development in Arabidopsis. Scientific reports 6:31195

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PloS one 4 (12):e8451

Yobi A, Bagaza C, Batushansky A, Shrestha V, Emery ML, Holden S, Turner-Hissong S, Miller ND, Mawhinney TP, Angelovici R (2020) The complex response of free and bound amino acids to water stress during the seed setting stage in Arabidopsis. The Plant Journal 102 (4):838-855

Zhang J, Yuan H, Yang Y, Fish T, Lyi SM, Thannhauser TW, Zhang L, Li L (2016) Plastid ribosomal protein S5 is involved in photosynthesis, plant development, and cold stress tolerance in Arabidopsis. Journal of experimental botany 67 (9):2731-2744

7. Figures

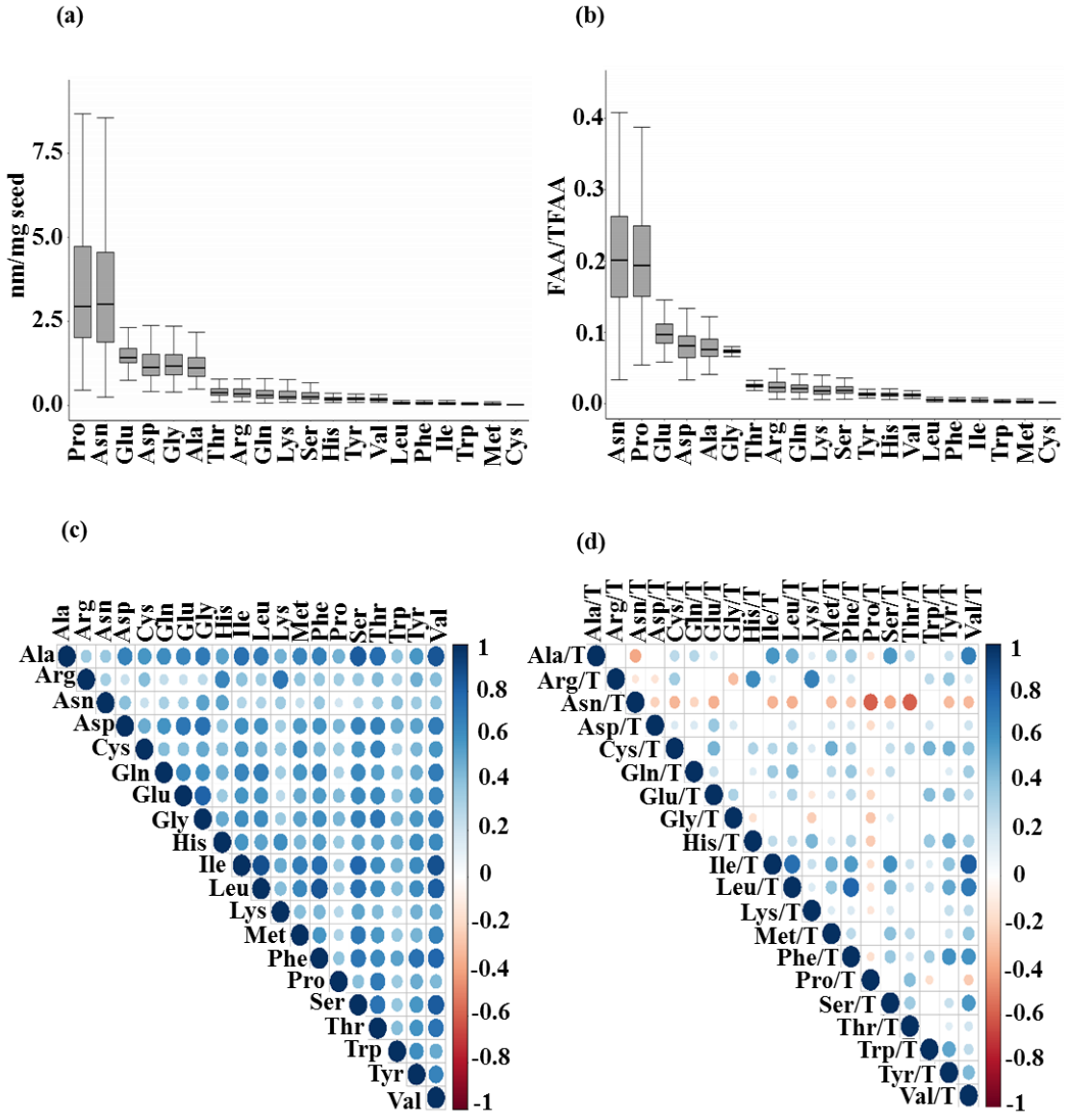


Figure 3.1 **The natural variation and relationships of FAA traits measured from the diversity panel.** Boxplot showing the FAA (a) absolute levels and (b) relative compositional distribution in the 279 taxa from Goodman-Buckler maize association panel. (c) Pairwise Pearson correlation analysis was done between the absolute FAA levels and (d) relative compositions using back transformed BLUPs of 279 taxa from Goodman-Buckler maize association panel. The correlation matrix was visualized in R v.3.4.3 (R Core Team). Each dot represents a significant correlation coefficient at qFDR values <0.05. Blue dots indicate positive correlation while red indicates negative correlations.

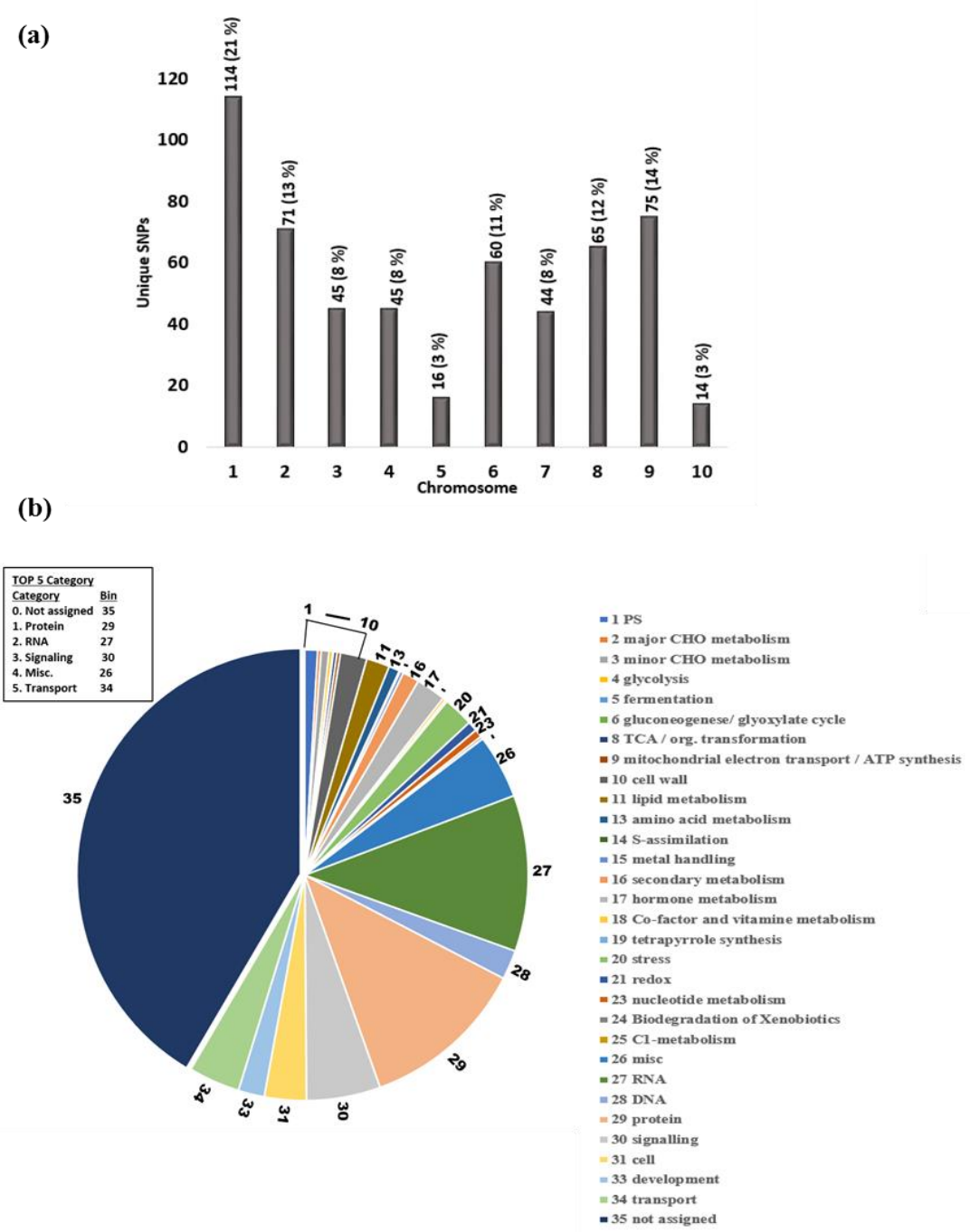


Figure 3.2 **The genomic distribution of the significant unique SNPs found in GWAS and the functional categorization of the extracted candidate. (a)** The partition of the significant unique SNPs across the 10 chromosomes in maize. (**b**) Pie chart representing the functional categorization of the 2779 GWAS candidate genes using MapMan version 3.6. The number in the parenthesis represent the proportion of genes that falls into a functional category. The top 4 categories are highlighted and include: protein, RNA, signaling, and misc.

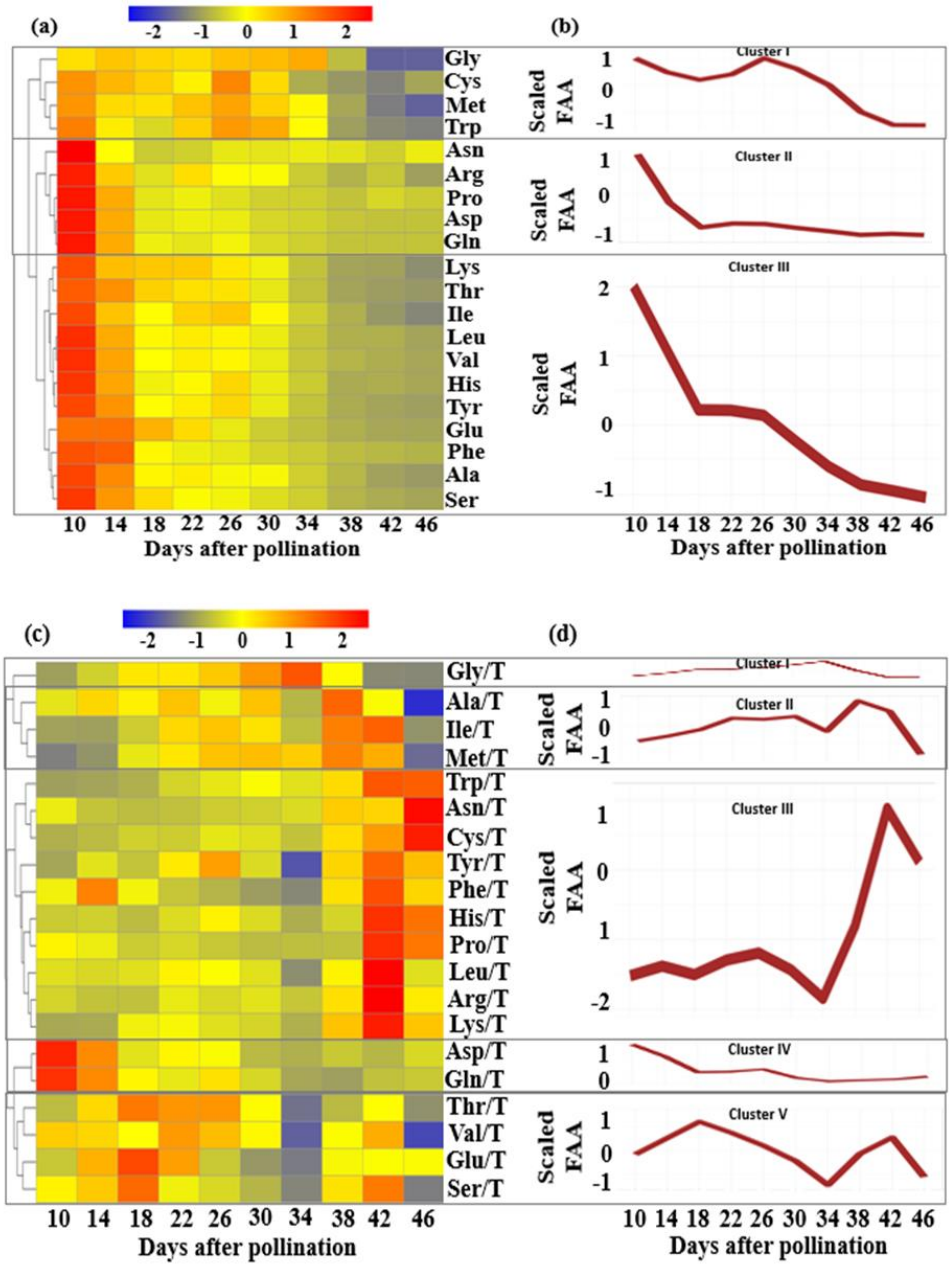


Figure 3.3: **Seed FAA composition dynamics during development and maturation.** Heatmap **(a)** and hierarchical clustering trends (**b**) of the FAA absolute levels across ten seed filling time points of maize inbred B73. Heatmap **(c)** and hierarchical clustering trends (**d**) of the FAAs relative composition across ten seed filling time points of maize inbred B73. Average values of three biological replicates from each time point were scaled and used to create heatmap (*n* =3). Heatmap blue color indicates low values for FAA accumulation while the red color indicates high accumulation. The red line indicates **the average** expression pattern of individual FAAs accumulation within a cluster (*n* = X) and was created using geom_line "mean" function in ggplot2 package in R v.3.4.3.

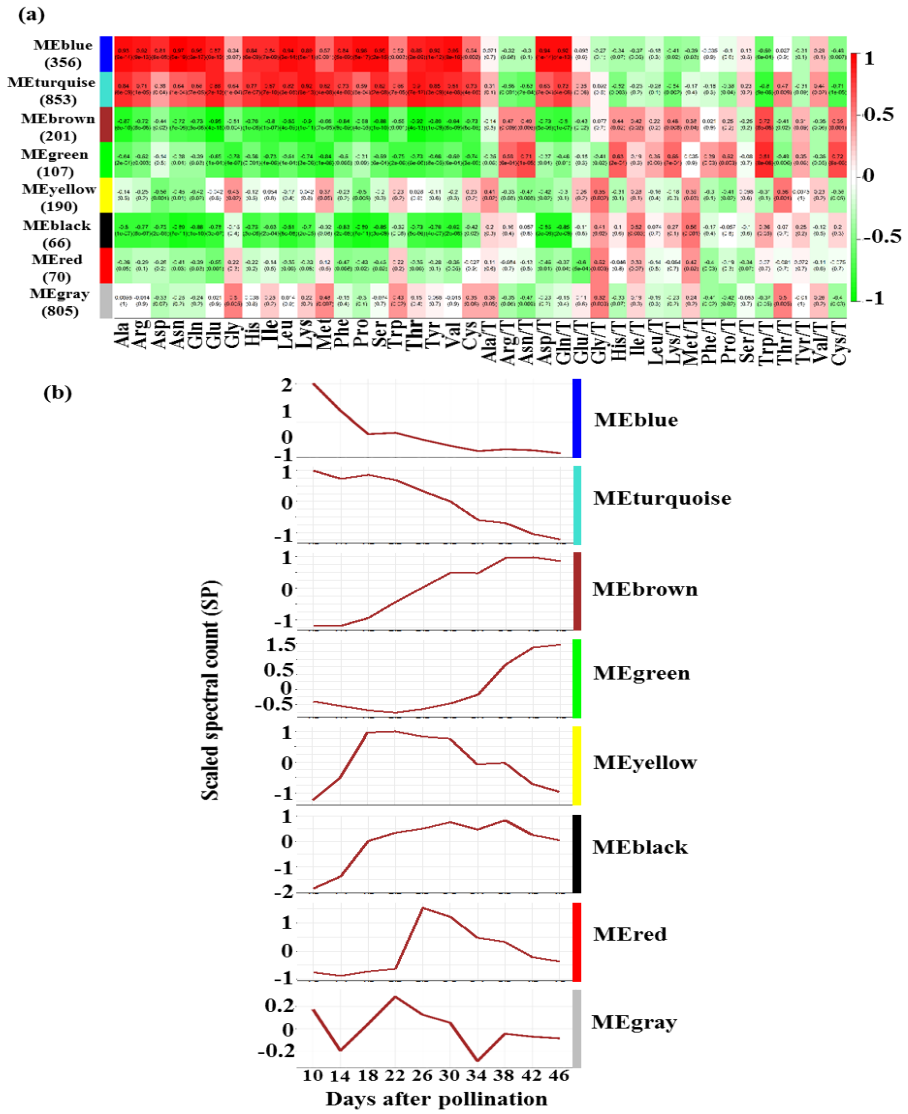


Figure 3.4: **The relationship among the protein co-expression modules and the FAA composition dynamics during seed maturation. (a)** Module-trait relationships (MTRs) from the WGCNA analysis. Modules names are displayed as rows on the left in the y-axis (i.e. MEblue denotes modules eigen protein for blue module) and the FAA absolute and relative composition traits (e.g. Ala/T, which is the ratio of Ala/Sum total of all FAAs levels) are displayed in columns as x-axis. The total number of proteins in the respective module is described in the parentheses along the y-axis. The correlation coefficients between modules Eigen protein (ME)-FAA traits are shown in the top of each row whereas the corresponding p-values are displayed at the bottom of each row within parentheses. The MTRs are colored based on their correlation; red is a strong positive correlation while green is a strong negative correlation. **(b)** The expression trend of Eigen protein found in the corresponding modules across the 10 different seed developing time points in days after pollination (DAP). The x-axis is the 10 DAPs while the y-axis is the expression of module eigen protein using the scaled spectral count data.

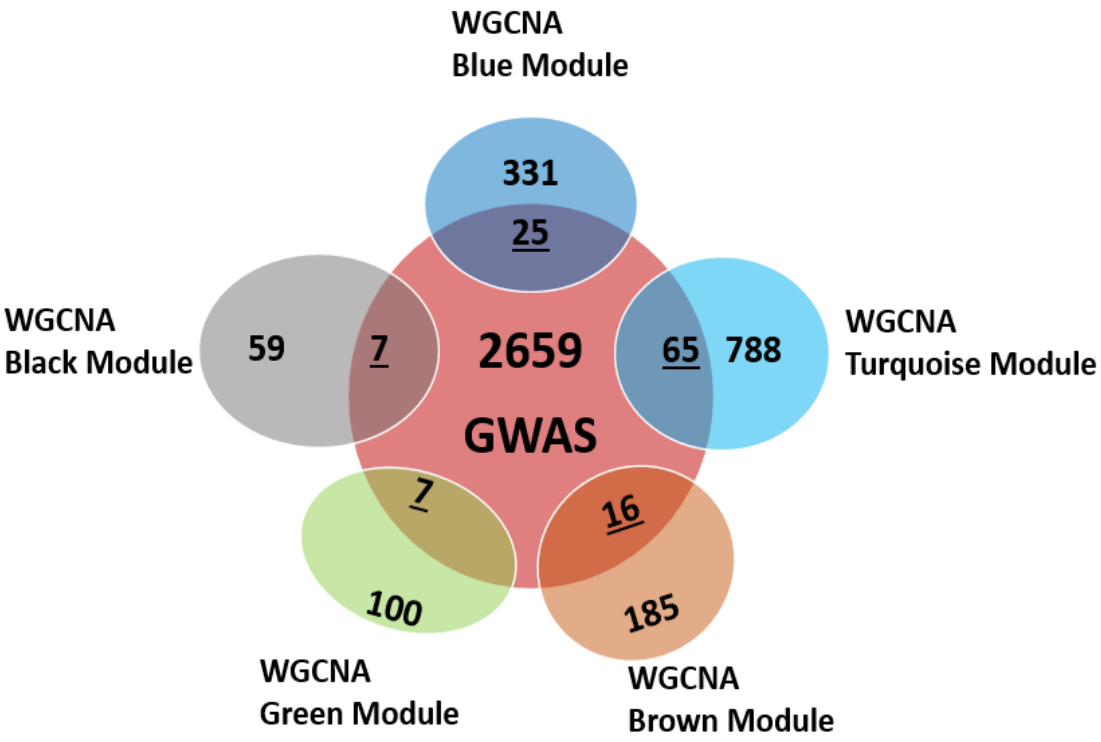
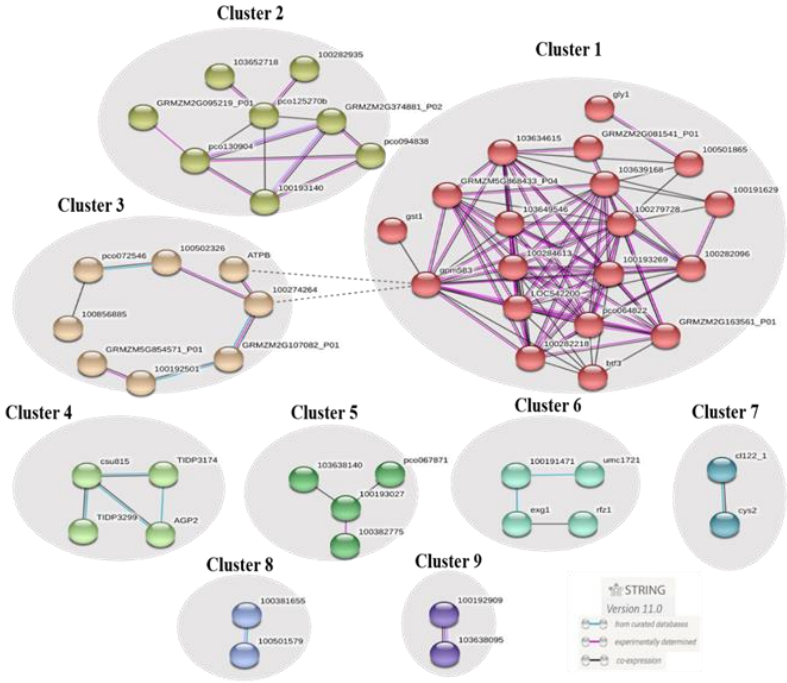


Figure 3.5: **Comparison between the candidate genes lists of WGCNA and GWAS.** Venn diagram depicting the 120 genes that overlap with both the GWAS candidate gene list and the relevant proteomic co-expression modules (turquoise, brown, green, black and blue modules). I refer to the overlapping genes as high confidence candidate genes (HCCG).

(a)



(b)

STRING			MAPMAN CATEGORY
Cluster number	Cluster color	Gene count	Bin name
1	Red	19	RNA transcription, Protein synthesis and transport
2	Sandy Brown	8	TCA cycle, lipid, and amino acid metabolism
3	Brown	8	Protein Degradation
4	Green Yellow	4	CHO metabolism
5	Green	4	Cell Vesical transport and Protein post-translaiton modification
6	Cyan	4	Cell wall proteins
7	Dark Cyan	2	Sulphur assimilation
8	Cornflower Blue	2	Amino acid metabolism
9	Blue	2	Cell Vesical transport

Figure 3.6: **Protein-Protein interaction (PPI) of 120 HCCG. (**a) Protein-protein interaction of 120 HCCG created using STRING (v11.0). Proteins are indicated by nodes labeled with the encoding protein symbol from STRING while interaction between nodes were indicated by edges. Smooth line edges indicate intra-cluster interaction whereas dotted edges indicate inter-cluster interactions. Cluster analysis using MCL algorithm resulted in 9 distinct clusters. (b) Table representation of cluster numbers, cluster color, gene count within each cluster using STRING, and the bin name is the functional category of the clusters using MapMan version 3.6.0.

8. Supplementary Figures

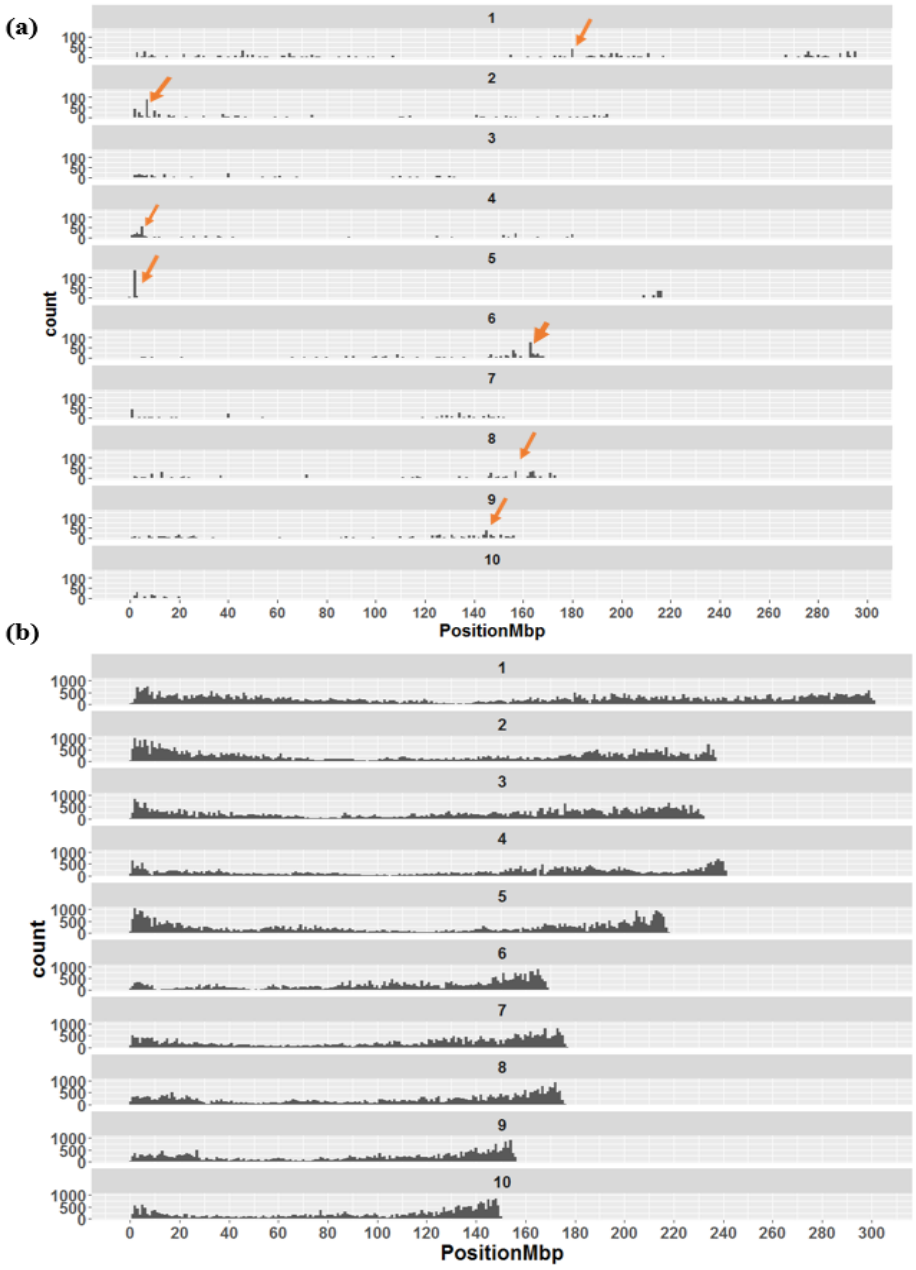


Figure S3.1 **Genomic distribution of significant and all SNPs across the genome (a)** Genome wide distribution of all the significant SNP found in our GWAS. **(b)** Genome wide SNP distribution of all the SNPs that were used for our analysis. The x-axis represents the position in Mbp across the genome while the y-axis represents the count in numbers of SNPs within the window size of 1 Mb. Several potential hotspot of SNP associations are detected and marked by the orange arrows in chromosome 1, 2, 4, 5, 6, 8, and 9.

9. Tables

Table 3.1 **A summary of 109 FAA GWAS results.** The data is summarized by FAA trait category i.e. FAA absolute levels, relative composition, aspartate family related ratios, BCAA family related ratios, glutamate family related ratios, and shikimate family related ratio traits. The table presents both the absolute numbers (n) and the percentage per family for the following trait parameters: total number of traits analyzed for GWAS studies, number of significant traits in GWAS at 5% FDR, number of unique (non-redundant) SNPs, average SNPs per trait per family, and number of unique candidate genes from 200kb window of peak SNP.

FAA Traits Category	**Total number of traits analyzed**		**No. of sig. traits**		**No. of unique SNPs**		**Average SNP per trait**	**No. of unique Candidate genes**	
	n	%	n	%	n	%		n	%
Absolute levels	20	18	19	24	147	27	8	664	23
Relative composition	20	18	12	15	76	14	6	397	14
Aspartate family related ratios	23	21	20	26	170	31	9	889	32
BCAA family related ratios	14	13	8	10	43	8	5	266	10
Glutamate family related ratios	17	16	10	13	62	11	6	277	10
Serine family related ratios	3	3	3	4	6	1	2	19	1
Shikimate family related ratios	12	11	6	8	45	8	8	267	10
Total sum of the FAA	109	100	78	100	549	100	44	2779	100

10. Supplementary Tables

Table S3.1 **Statistical and heritability summary of the FAA absolute and relative composition traits.** The mean, standard deviation (SD), relative standard deviation (RSD), range and broad sense heritability (BSH) of 20 FAA absolute levels measured and calculated from the dry seeds of Goodman-Buckler maize association panel and the 20 calculated FAA absolute and relative composition traits are described in (a) and (b), respectively. Statistics (mean, SD, RSD and range) were calculated using the dry seed FAA measurement from back- transformed BLUPs of the 279 taxa of Goodman-Buckler maize association panel while BSH was calculated using replicated data (raw data after outlier removal) from two years and two replicas.

(a)		Backtransformed BLUPs					
Trait	Trait Symbol	Mean	SD	Relative SD %	Min	Max	Broad Sense Heritability
Pro	P	3.47	1.92	55.44	0.45	9.23	0.93
Asn	N	3.38	1.84	54.40	0.25	8.99	0.81
Glu	E	1.50	0.33	21.74	0.75	2.69	0.78
Asp	D	1.29	0.61	47.54	0.41	4.22	0.92
Gly	G	1.23	0.44	35.29	0.40	2.44	0.77
Ala	A	1.21	0.49	40.66	0.49	3.31	0.89
Thr	T	0.41	0.15	35.93	0.11	0.93	0.42
Arg	R	0.38	0.18	45.56	0.11	1.53	0.77
Gln	Q	0.34	0.18	51.44	0.07	1.27	0.92
Lys	K	0.34	0.23	67.19	0.09	1.48	0.94
Ser	S	0.31	0.18	56.95	0.07	1.29	0.94
His	H	0.21	0.08	40.38	0.09	0.49	0.86
Tyr	Y	0.21	0.06	31.03	0.09	0.67	0.91
Val	V	0.19	0.06	33.64	0.09	0.44	0.90
Leu	L	0.08	0.03	39.56	0.04	0.22	0.92
Phe	F	0.08	0.03	39.62	0.03	0.20	0.92
Ile	I	0.07	0.03	47.50	0.03	0.24	0.92
Trp	W	0.05	0.02	29.68	0.02	0.13	0.93
Met	M	0.05	0.03	60.53	0.01	0.20	0.48
Cys	C	0.02	0.00	13.67	0.02	0.04	0.63

(b)		Backtransformed BLUPs					
Trait	Trait Symbol	Mean	SD	Relative SD %	Min	Max	Broad Sense Heritability
Asn/T	N/T	0.21	0.08	37.38	0.03	0.43	0.77
Pro/T	P/T	0.20	0.07	36.60	0.05	0.45	0.89
Glu/T	E/T	0.10	0.02	19.06	0.06	0.15	0.67
Asp/T	D/T	0.08	0.02	26.57	0.03	0.16	0.82
Ala/T	A/T	0.08	0.02	22.41	0.04	0.13	0.78
Gly/T	G/T	0.07	0.00	4.17	0.06	0.08	0.04
Thr/T	T/T	0.03	0.00	13.13	0.01	0.04	0.11
Arg/T	R/T	0.02	0.01	39.62	0.01	0.06	0.73
Gln/T	Q/T	0.02	0.01	34.71	0.01	0.05	0.83
Lys/T	K/T	0.02	0.01	55.77	0.01	0.08	0.96
Ser/T	S/T	0.02	0.01	37.85	0.01	0.07	0.92
Tyr/T	Y/T	0.01	0.00	20.56	0.01	0.02	0.89
His/T	H/T	0.01	0.00	26.08	0.01	0.03	0.78
Val/T	V/T	0.01	0.00	19.25	0.01	0.02	0.83
Leu/T	L/T	0.01	0.00	27.03	0.00	0.01	0.88
Phe/T	F/T	0.01	0.00	25.72	0.00	0.01	0.86
Ile/T	I/T	0.00	0.00	30.84	0.00	0.01	0.88
Trp/T	W/T	0.00	0.00	29.37	0.00	0.01	0.93
Met/T	M/T	0.00	0.00	50.17	0.00	0.01	0.94
Cys/T	C/T	0.00	0.00	22.74	0.00	0.00	0.58

Table S3.2 **Correlation among the absolute FAA levels and their relative composition.** (a) Pairwise Pearson correlation coefficients between the 20 absolute FAA levels using back transformed BLUPs of 279 taxa from Goodman-Buckler maize association panel. (b) Significance of correlation analysis of absolute FAA using corr.test function in R and the corresponding adjusted p-value using FDR correction at 5% level of significance. (c) Pairwise Pearson correlation coefficient between 20 relative (compositional) FAA using back transformed BLUPs of 279 taxa from Goodman-Buckler maize association panel. (d) Significance of relative FAA correlation analysis using corr.test and the corresponding adjusted p-value using FDR correction at 5% level of significance. T stands for Total (E.g.: Ala/T=Ala/Total).

(a)	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	1																			
Arg	0.36	1.00																		
Asn	0.38	0.33	1.00																	
Asp	0.68	0.23	0.41	1.00																
Cys	0.59	0.42	0.22	0.49	1.00															
Gln	0.63	0.24	0.32	0.59	0.42	1.00														
Glu	0.66	0.23	0.34	0.73	0.42	0.63	1.00													
Gly	0.70	0.27	0.54	0.75	0.50	0.61	0.81	1.00												
His	0.54	0.66	0.52	0.39	0.43	0.48	0.36	0.50	1.00											
Ile	0.75	0.41	0.29	0.60	0.56	0.66	0.58	0.61	0.58	1.00										
Leu	0.70	0.35	0.30	0.60	0.48	0.69	0.59	0.62	0.53	0.88	1.00									
Lys	0.48	0.72	0.36	0.34	0.39	0.35	0.26	0.36	0.62	0.50	0.43	1.00								
Met	0.66	0.39	0.23	0.56	0.63	0.59	0.51	0.54	0.45	0.70	0.64	0.42	1.00							
Phe	0.68	0.38	0.37	0.60	0.45	0.66	0.56	0.61	0.55	0.76	0.85	0.46	0.58	1.00						
Pro	0.48	0.34	0.18	0.45	0.37	0.32	0.43	0.46	0.37	0.35	0.36	0.27	0.30	0.39	1.00					
Ser	0.84	0.37	0.27	0.66	0.57	0.64	0.64	0.67	0.52	0.79	0.74	0.53	0.71	0.73	0.42	1.00				
Thr	0.77	0.44	0.34	0.68	0.67	0.55	0.64	0.72	0.52	0.63	0.63	0.43	0.58	0.63	0.71	0.73	1.00			
Trp	0.41	0.33	0.41	0.50	0.33	0.41	0.41	0.47	0.47	0.47	0.43	0.31	0.38	0.54	0.36	0.38	0.43	1.00		
Tyr	0.59	0.48	0.35	0.54	0.45	0.49	0.56	0.58	0.62	0.63	0.63	0.48	0.46	0.75	0.50	0.59	0.59	0.62	1.00	
Val	0.85	0.43	0.40	0.69	0.58	0.70	0.66	0.70	0.63	0.88	0.83	0.51	0.67	0.80	0.44	0.83	0.73	0.51	0.67	1.00

(b)	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0																			
Arg	1.58E-08	0																		
Asn	1.53E-10	1.21E-09	0																	
Asp	2.28E-37	4.16E-05	2.48E-13	0																
Cys	8.33E-27	7.27E-13	0.0001	1.29E-18	0															
Gln	2.24E-31	6.51E-06	3.32E-08	1.71E-27	1.46E-13	0														
Glu	5.54E-35	3.27E-05	3.75E-09	5.66E-47	7.10E-14	1.24E-32	0													
Gly	5.00E-41	2.00E-07	5.57E-23	2.23E-51	9.10E-20	1.87E-29	5.04E-66	0												
His	3.68E-21	2.00E-37	1.57E-21	1.45E-12	5.44E-14	8.86E-18	1.50E-10	2.00E-20	0											
Ile	6.44E-50	3.27E-13	5.01E-07	4.21E-29	7.94E-24	1.17E-35	2.80E-26	9.73E-30	1.16E-27	0										
Leu	7.92E-41	2.44E-08	2.61E-07	2.09E-29	1.77E-17	4.57E-40	1.15E-27	1.12E-30	2.92E-21	1.46E-88	0									
Lys	2.84E-16	7.10E-49	1.05E-10	1.32E-09	4.71E-12	1.83E-10	2.33E-06	3.74E-11	1.40E-32	5.21E-20	2.58E-13	0								
Met	3.44E-35	1.06E-11	5.31E-05	5.69E-24	3.64E-32	1.15E-27	3.98E-20	1.82E-22	8.49E-16	1.37E-41	5.23E-33	4.29E-14	0							
Phe	1.17E-37	1.20E-09	1.73E-10	5.80E-28	2.05E-15	3.34E-35	2.40E-24	2.46E-29	1.58E-22	9.93E-53	9.67E-79	3.68E-15	4.46E-26	0						
Pro	8.55E-17	1.01E-11	0.00047	4.57E-16	6.28E-11	1.14E-08	6.01E-15	2.84E-17	9.84E-12	5.94E-10	6.61E-10	8.06E-08	6.91E-08	1.74E-11	0					
Ser	1.10E-73	6.47E-11	8.27E-06	2.29E-35	2.24E-25	3.55E-33	1.58E-32	4.35E-37	2.22E-20	6.81E-60	2.41E-49	1.74E-21	3.48E-44	5.18E-45	2.44E-13	0				
Thr	1.44E-53	1.00E-14	2.61E-09	1.91E-37	1.54E-36	4.78E-23	8.02E-33	2.04E-45	1.43E-20	3.78E-31	2.37E-30	6.81E-15	2.96E-26	2.49E-31	2.45E-43	1.53E-46	0			
Trp	3.86E-12	6.20E-11	1.40E-13	1.34E-18	8.88E-09	2.31E-13	4.03E-13	1.95E-17	3.56E-18	8.86E-17	1.46E-13	7.49E-10	1.12E-11	2.82E-21	1.19E-11	2.57E-11	1.19E-14	0		
Tyr	2.45E-26	6.78E-17	3.45E-09	1.20E-21	6.42E-15	7.67E-20	2.97E-24	1.42E-25	5.00E-31	1.16E-33	1.27E-32	5.45E-18	1.63E-16	7.07E-50	8.00E-19	6.00E-28	5.38E-27	9.74E-31	0	
Val	1.03E-76	1.50E-14	1.56E-11	2.72E-39	8.35E-26	3.91E-42	8.20E-35	7.92E-41	2.59E-32	1.47E-89	1.17E-70	3.98E-20	1.63E-37	7.05E-61	9.04E-15	9.95E-73	2.04E-47	7.27E-20	5.29E-39	0

(c)	Ala/T	Arg/T	Asn/T	Asp/T	Cys/T	Gln/T	Glu/T	Gly/T	His/T	Ile/T	Leu/T	Lys/T	Met/T	Phe/T	Pro/T	Ser/T	Thr/T	Trp/T	Tyr/T	Val/T
Ala/T	1.00																			
Arg/T	0.00	1.00																		
Asn/T	-0.39	-0.14	1.00																	
Asp/T	0.11	-0.14	-0.22	1.00																
Cys/T	0.27	0.35	-0.34	0.16	1.00															
Gln/T	0.28	-0.06	-0.22	0.16	0.09	1.00														
Glu/T	0.18	0.04	-0.35	0.37	0.47	0.23	1.00													
Gly/T	0.04	-0.30	-0.07	0.17	-0.06	0.12	0.31	1.00												
His/T	0.08	0.63	-0.01	-0.11	0.30	0.14	0.07	-0.17	1.00											
Ile/T	0.58	0.13	-0.34	0.14	0.29	0.35	0.15	0.03	0.26	1.00										
Leu/T	0.47	0.12	-0.35	0.19	0.33	0.43	0.31	0.03	0.25	0.75	1.00									
Lys/T	0.13	0.67	-0.09	-0.11	0.15	-0.01	-0.12	-0.24	0.45	0.22	0.15	1.00								
Met/T	0.33	0.17	-0.32	0.16	0.48	0.27	0.16	0.04	0.16	0.48	0.38	0.14	1.00							
Phe/T	0.38	0.17	-0.27	0.17	0.32	0.37	0.27	0.02	0.31	0.57	0.79	0.16	0.27	1.00						
Pro/T	-0.14	-0.04	-0.60	-0.19	-0.10	-0.16	-0.20	-0.27	-0.26	-0.16	-0.16	-0.14	-0.07	-0.17	1.00					
Ser/T	0.58	0.02	-0.39	0.16	0.23	0.24	0.07	0.14	0.09	0.61	0.47	0.16	0.41	0.37	-0.09	1.00				
Thr/T	0.27	0.10	-0.60	-0.02	0.31	0.03	0.02	0.02	-0.06	0.22	0.19	0.04	0.27	0.19	0.42	0.35	1.00			
Trp/T	0.03	0.26	-0.09	0.19	0.47	0.02	0.43	-0.08	0.38	0.13	0.23	0.08	0.12	0.34	-0.18	-0.02	-0.04	1.00		
Tyr/T	0.22	0.39	-0.31	0.15	0.49	0.14	0.41	0.02	0.51	0.41	0.51	0.24	0.19	0.61	-0.12	0.23	0.14	0.53	1.00	
Val/T	0.69	0.17	-0.32	0.20	0.40	0.34	0.26	-0.01	0.36	0.82	0.71	0.26	0.40	0.59	-0.26	0.57	0.20	0.26	0.45	1.00

(d)	Ala/T	Arg/T	Asn/T	Asp/T	Cys/T	Gln/T	Glu/T	Gly/T	His/T	Ile/T	Leu/T	Lys/T	Met/T	Phe/T	Pro/T	Ser/T	Thr/T	Trp/T	Tyr/T	Val/T
Ala/T	0.00																			
Arg/T	0.98	0.00																		
Asn/T	0.00	0.03	0.00																	
Asp/T	0.10	0.02	0.00	0.00																
Cys/T	0.00	0.00	0.00	0.01	0.00															
Gln/T	0.00	0.39	0.00	0.02	0.18	0.00														
Glu/T	0.01	0.60	0.00	0.00	0.00	0.00	0.00													
Gly/T	0.54	0.00	0.34	0.01	0.32	0.07	0.00	0.00												
His/T	0.38	0.00	0.80	0.05	0.00	0.03	0.36	0.01	0.00											
Ile/T	0.00	0.03	0.00	0.07	0.00	0.00	0.05	0.76	0.00	0.00										
Leu/T	0.00	0.06	0.00	0.01	0.00	0.00	0.00	0.75	0.00	0.00	0.00									
Lys/T	0.04	0.00	0.18	0.07	0.02	0.89	0.03	0.00	0.00	0.00	0.02	0.00								
Met/T	0.00	0.01	0.00	0.01	0.00	0.00	0.02	0.60	0.02	0.00	0.00	0.02	0.00							
Phe/T	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.01	0.00	0.00						
Pro/T	0.03	0.55	0.00	0.00	0.15	0.01	0.00	0.00	0.00	0.02	0.01	0.03	0.30	0.01	0.00					
Ser/T	0.00	0.78	0.00	0.02	0.00	0.00	0.34	0.04	0.23	0.00	0.00	0.01	0.00	0.00	0.20	0.00				
Thr/T	0.00	0.13	0.00	0.75	0.00	0.59	0.76	0.75	0.38	0.00	0.01	0.56	0.00	0.00	0.00	0.00	0.00			
Trp/T	0.69	0.00	0.17	0.00	0.00	0.75	0.00	0.19	0.00	0.03	0.00	0.19	0.06	0.00	0.01	0.77	0.53	0.00		
Tyr/T	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.03	0.00	0.00	
Val/T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table S3.3 **List of 109 seed FAA traits used for GWAS**. These traits included from the quantifications of 20 absolute levels, their relative composition and known biochemical interactions (based on their affiliation with their respective amino acid families: Aspartate, Glutamate, BCAA, Shikimate and Serine).

Abs & Relative composition to total AAs	Biochemistry based metabolic ratios, grouped by AA families' affiliation		Free Amino acids one letter code	
AA-Abs & AA/T T= Sum of 20 AA	Asp Family =Ile, Met, Asn,Thr, Asp Lys (IMNTDK)	BCAA Family= Ile, Val, Leu (IVL) Pyr Family=Leu, Ala, Val (LAV)	Ala	A
Ala	A/E	A/LAV	Arg	R
Ala/T	D/IMNTDK	E/V	Asn	N
Arg	D/K	I/IVL	Asp	D
Arg/T	E/I	I/L	Cys	C
Asn	E/K	I/LAV	Gln	Q
Asn/T	I/IMNTDK	I/V	Glu	E
Asp	IMNTDK	IVL	Gly	G
Asp/T	K/IMNTDK	L/IVL	His	H
Cys	M/I	L/LAV	Ile	I
Cys/T	M/IMNTDK	L/V	Leu	L
Gln	M/K	LAV	Lys	K
Gln/T	N/D	V/A	Met	M
Glu	N/E	V/IVL	Phe	F
Glu/T	N/IMNTDK	V/LAV	Pro	P
Gly	N/K		Ser	S
Gly/T	N/M		Thr	T
His	N/Q	Shikimate Fam =Trp, Phe,Tyr (WFY)	Trp	W
His/T	N/T	D/W	Tyr	Y
Ile	T/G	E/W	Val	V
Ile/T	T/I	F/E		
Leu	T/IMNTDK	F/WFY		
Leu/T	T/K	Q/W		
Lys	T/M	S/W		
Lys/T		W/F		
Met	Glu Family= Glu, His, Pro, Arg, Gln (EHPRQ)	W/WFY		
Met/T	D/EHPRQ	WFY		
Phe	E/EHPRQ	Y/F		
Phe/T	E/P	Y/W		
Pro	E/R	Y/WFY		
Pro/T	EHPRQ			
Ser	H/D	Ser Family=Ser, Gly, Cys		
Ser/T	H/E	C/SGC		
Thr	H/EHPRQ	G/SGC		
Thr/T	H/M	S/SGC		
Trp	H/Q			
Trp/T	H/W			
Tyr	P/EHPRQ			
Tyr/T	Q/E			
Val	Q/EHPRQ			
Val/T	R/E			
	R/EHPRQ			
	R/P			

Table S3.4 **The 120 HCCG STRING and functional analysis.** String analysis of the 120 HCCG including cluster number, cluster color, Protein ID, STRING Protein name and protein description. MCL clustering method was used in STRING V11.0 resulting into 9 distinct cluster in the PPI network derived from 120 HCCG. "Unconnected" label in the cluster number indicates those proteins that are not connected in the given PPI analysis. MapMan functional categorization of the 120 HCCG and their respective bin, bin code, and bin name are also presented.

STRING PPI Cluster analysis						MapMan Annotations		
cluster number	**cluster color**	**gene count**	**Protein**	**STRING Protein name**	**Protein description**	**Bin**	**BinCode**	**BinName**
1	Red	19	GRMZM2G006178_P01	103639168	ABC transporter E family member 2	27	27.2	RNA.transcription
1	Red	19	GRMZM2G024354_P01	100282218	Ribosomal protein L15 ; Belongs to the eukaryotic ribosomal protein eL15 family	29	29.2.1.2.2.15	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L15
1	Red	19	GRMZM2G030016_P02	100193269	40S ribosomal protein S7 ; Belongs to the eukaryotic ribosomal protein eS7 family	29	29.2.1.2.1.7	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S7
1	Red	19	GRMZM2G038032_P01	gpm583	Guanine nucleotide-binding protein beta subunit-like protein	30	30.5	signalling.G-proteins
1	Red	19	GRMZM2G041881_P01	btf3	Nascent polypeptide-associated complex subunit beta	27	27.3.67	RNA.regulation of transcription.putative transcription regulator
1	Red	19	GRMZM2G056462_P01	100282096	Eukaryotic translation initiation factor 2 beta subunit	29	29.2.3	protein.synthesis.initiation
1	Red	19	GRMZM2G058138_P01	100279728	Eukaryotic translation initiation factor 2 subunit alpha	29	29.2.3	protein.synthesis.initiation
1	Red	19	GRMZM2G067303_P02	LOC542200	40S ribosomal protein S20 ; Belongs to the universal ribosomal protein uS10 family	29	29.2.1.2.1.20	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S20

1	Red	19	GRMZM2G078143_P01	gly1	Serine hydroxymethyltransferase; Interconversion of serine and glycine; Belongs to the SHMT family	1	1.2.5	PS.photorespiration.serine hydroxymethyltransferase
1	Red	19	GRMZM2G081541_P01	GRMZM2G081541_P01	ABC transporter F family member 1	34	34.16	transport.ABC transporters and multidrug resistance systems
1	Red	19	GRMZM2G109121_P02	103634615	annotation not available	27	27.2	RNA.transcription
1	Red	19	GRMZM2G116273_P01	gst1	Glutathione S-transferase 1; Conjugation of reduced glutathione to a wide number of exogenous and endogenous hydrophobic electrophiles. Involved in the detoxification of certain herbicides; Belongs to the GST superfamily. Phi family	26	26.9	misc.glutathione S transferases
1	Red	19	GRMZM2G131473_P02	100191629	Methionine aminopeptidase 2; Cotranslationally removes the N-terminal methionine from nascent proteins. The N-terminal methionine is often cleaved when the second residue in the primary sequence is small and uncharged (Met-Ala-, Cys, Gly, Pro, Ser, Thr, or Val)	29	29.5.7	protein.degradation.metalloprotease
1	Red	19	GRMZM2G133121_P01	100501865	Phosphoribosylaminoimidazole carboxylase family protein / AIR carboxylase family protein	23	23.1.2.6	nucleotide metabolism.synthesis.purine.AIR carboxylase

1	Red	19	GRMZM2G153181_P01	103649546	Eukaryotic translation initiation factor 3 subunit L; Component of the eukaryotic translation initiation factor 3 (eIF-3) complex, which is involved in protein synthesis of a specialized repertoire of mRNAs and, together with other initiation factors, stimulates binding of mRNA and methionyl-tRNAi to the 40S ribosome. The eIF-3 complex specifically targets and initiates translation of a subset of mRNAs involved in cell proliferation	35	35.2	not assigned.unknown
1	Red	19	GRMZM2G163561_P01	GRMZM2G163561_P01	40S ribosomal protein S23; Uncharacterized protein ; Belongs to the universal ribosomal protein uS12 family	29	29.2.1.2.1.23	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S23
1	Red	19	GRMZM2G427014_P01	100284613	40S ribosomal protein S16 ; Belongs to the universal ribosomal protein uS9 family	29	29.2.1.2.1.16	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S16
1	Red	19	GRMZM5G868433_P04	GRMZM5G868433_P04	60S ribosomal protein L7-2; Uncharacterized protein	29	29.2.1.2.2.7	protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L7
1	Red	19	GRMZM5G899149_P01	pco064822	40S ribosomal protein S26; Uncharacterized protein ; Belongs to the eukaryotic ribosomal protein eS26 family	29	29.2.1.2.1.26	protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S26
2	Sandy Brown	8	GRMZM2G107082_P01	GRMZM2G107082_P01	ATP-citrate synthase alpha chain protein 3	8	8.2.11	TCA / org. transformation.other organic acid transformaitons.atp-citrate lyase

2	Sandy Brown	8	GRMZM2G113408_P01	ATPB	ATP synthase subunit beta, mitochondrial; Mitochondrial membrane ATP synthase (F(1)F(0) ATP synthase or Complex V) produces ATP from ADP in the presence of a proton gradient across the membrane which is generated by electron transport complexes of the respiratory chain. F-type ATPases consist of two structural domains, F(1) - containing the extramembraneous catalytic core, and F(0) - containing the membrane proton channel, linked together by a central stalk and a peripheral stalk. During catalysis, ATP synthesis in the catalytic domain of F(1) is coupled via a rotary mechanism of the c [...]	NA	NA	NA
2	Sandy Brown	8	GRMZM2G131539_P01	100502326	Enolase	4	4.1.13	glycolysis.cytosolic branch.enolase
2	Sandy Brown	8	GRMZM2G154595_P01	100274264	Malate dehydrogenase	8	8.1.9	TCA / org. transformation.TCA. malate DH
2	Sandy Brown	8	GRMZM2G166646_P01	100856885	Putative mitochondrial-processing peptidase subunit alpha-2 chloroplastic/mitochondrial	29	29.3.2	protein.targeting.mitochondria
2	Sandy Brown	8	GRMZM5G833389_P02	pco072546	2,3-bisphosphoglycerate-independent phosphoglycerate mutase; Catalyzes the interconversion of 2-phosphoglycerate and 3-phosphoglycerate	4	4.1.12	glycolysis.cytosolic branch.phosphoglycerate mutase
2	Sandy Brown	8	GRMZM5G848768_P02	100192501	3-ketoacyl-CoA thiolase 2 peroxisomal; Belongs to the thiolase family	11	11.9.4.5	lipid metabolism.lipid degradation.beta-oxidation.acyl-CoA thioesterase

2	Sandy Brown	8	GRMZM5G854571_P01	GRMZM5G854571_P01	annotation not available	13	13.2.6.3	amino acid metabolism.degradation.aromatic aa.tryptophan
3	Brown	8	AC199782.5_FGP001	pco094838	ATPase ASNA1 homolog; ATPase required for the post-translational delivery of tail-anchored (TA) proteins to the endoplasmic reticulum. Recognizes and selectively binds the transmembrane domain of TA proteins in the cytosol. This complex then targets to the endoplasmic reticulum by membrane-bound receptors, where the tail-anchored protein is released for insertion. This process is regulated by ATP binding and hydrolysis. ATP binding drives the homodimer towards the closed dimer state, facilitating recognition of newly synthesized TA membrane proteins. ATP hydrolysis is required for ins [...]	34	34.18.1	transport.unspecified anions.arsenite-transporting ATPase
3	Brown	8	GRMZM2G005080_P01	100193140	Proteasome subunit alpha type; The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH; Belongs to the peptidase T1A family	29	29.5.11.20	protein.degradation.ubiquitin.proteasom
3	Brown	8	GRMZM2G095219_P01	GRMZM2G095219_P01	Pleckstrin homology (PH) domain-containing protein	35	35.2	not assigned.unknown

3	Brown	8	GRMZM2G102596_P01	pco130904	Proteasome subunit beta type; The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH; Belongs to the peptidase T1B family	29	29.5.11.20	protein.degradation.ubiquitin.proteasom
3	Brown	8	GRMZM2G114220_P01	pco125270b	Ubiquitin fusion degradation protein 1	29	29.5.11	protein.degradation.ubiquitin
3	Brown	8	GRMZM2G126453_P01	103652718	AAA-type ATPase family protein	29	29.5.9	protein.degradation.AAA type
3	Brown	8	GRMZM2G159538_P01	100282935	Plant UBX domain-containing protein 10; Fas-associated factor 1-like protein; Uncharacterized protein	29	29.5	protein.degradation
3	Brown	8	GRMZM2G374881_P02	GRMZM2G374881_P02	Proteasome subunit beta type; The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH; Belongs to the peptidase T1B family	29	29.5.11.20	protein.degradation.ubiquitin.proteasom
4	Green Yellow	4	GRMZM2G027955_P01	AGP2	Glucose-1-phosphate adenylyltransferase large subunit 2, chloroplastic/amyloplastic; This protein plays a role in synthesis of starch. It catalyzes the synthesis of the activated glycosyl donor, ADP- glucose from Glc-1-P and ATP	2	2.1.2.1	major CHO metabolism.synthesis.starch.AGPase
4	Green Yellow	4	GRMZM2G032003_P01	csu815	UDP-glucose pyrophosphorylase2	4	4.3.1	glycolysis.unclear/dually targeted.UGPase

4	Green Yellow	4	GRMZM2G099860_P01	TIDP3174	Trehalose-6-phosphate synthase2	3	3.2.3	minor CHO metabolism.trehalose. potential TPS/TPP
4	Green Yellow	4	GRMZM2G328500_P01	TIDP3299	UDP-glucose 6-dehydrogenase	10	10.1.4	cell wall.precursor synthesis.UGD
5	Green	4	GRMZM2G030144_P01	100193027	AP-1 complex subunit gamma-1	31	31.4	cell.vesicle transport
5	Green	4	GRMZM2G122135_P02	103638140	Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A beta isoform	29	29.4	protein.postranslational modification
5	Green	4	GRMZM2G164352_P03	pco067871	Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A beta isoform	29	29.4	protein.postranslational modification
5	Green	4	GRMZM5G836182_P01	100382775	ADP-ribosylation factor A1F; Belongs to the small GTPase superfamily. Arf family	29	29.3.4.99	protein.targeting.secretory pathway.unspecified
6	Cyan	4	GRMZM2G017186_P11	umc1721	Exhydrolase II	10	10.6.1	cell wall.degradation.cellulases and beta -1,4-glucanases
6	Cyan	4	GRMZM2G084812_P01	rfz1	Fasciclin-like arabinogalactan protein 8	10	10.5.1.1	cell wall.cell wall proteins.AGPs.AGP
6	Cyan	4	GRMZM2G111324_P01	100191471	Glucan endo-13-beta-glucosidase 3; Belongs to the glycosyl hydrolase 17 family	26	26.4.1	misc.beta 1,3 glucan hydrolases.glucan endo-1,3-beta-glucosidase
6	Cyan	4	GRMZM2G147687_P04	exg1	Exoglucanase1	10	10.6.1	cell wall.degradation.cellulases and beta -1,4-glucanases
7	Dark Cyan	2	GRMZM2G005887_P03	cys2	Cysteine synthase	13	13.1.5.3.1	amino acid metabolism.synthesis.serine-glycine-cysteine group.cysteine.OASTL

7	Dark Cyan	2	GRMZM2G090338_P01	cl122_1	Sulfite reductase [ferredoxin], chloroplastic; Essential protein with sulfite reductase activity required in assimilatory sulfate reduction pathway during both primary and secondary metabolism and thus involved in development and growth	14	14.3	S-assimilation.sulfite redox
8	Cornflower Blue	2	GRMZM2G017110_P01	100501579	Glutamate decarboxylase; Uncharacterized protein ; Belongs to the group II decarboxylase family	13	13.1.1.1.1	amino acid metabolism.synthesis. central amino acid metabolism.GABA.G lutamate decarboxylase
8	Cornflower Blue	2	GRMZM2G098875_P02	100381655	Glutamate decarboxylase; Belongs to the group II decarboxylase family	13	13.1.1.1.1	amino acid metabolism.synthesis. central amino acid metabolism.GABA.G lutamate decarboxylase
9	Blue	2	AC155622.2_FGP004	100192909	Coatomer subunit alpha-1	31	31.4	cell.vesicle transport
9	Blue	2	GRMZM2G000645_P01	103638095	Coatomer subunit beta'; The coatomer is a cytosolic protein complex that binds to dilysine motifs and reversibly associates with Golgi non-clathrin-coated vesicles, which further mediate biosynthetic protein transport from the ER, via the Golgi up to the trans Golgi network. Coatomer complex is required for budding from Golgi membranes, and is essential for the retrograde Golgi-to-ER transport of dilysine-tagged proteins	31	31.4	cell.vesicle transport
Unconnected	NA	NA	GRMZM2G004138_P01	pco117754	Caffeoyl-CoA O-methyltransferase 1	16	16.2.1.6	secondary metabolism.phenylpr opanoids.lignin

								biosynthesis.CCoAOMT
Unconnected	NA	NA	GRMZM2G007791_P01	ago2a	Argonaute2a; Belongs to the argonaute family	27	27.3.36	RNA.regulation of transcription.Argonaute
Unconnected	NA	NA	GRMZM2G008410_P01	542422	Protein BTR1; Transcribed sequence 1087 protein; Uncharacterized protein	27	27.4	RNA.RNA binding
Unconnected	NA	NA	GRMZM2G014788_P01	100384763	Carbohydrate-binding-like fold	35	35.2	not assigned.unknown
Unconnected	NA	NA	GRMZM2G018950_P01	100273121	annotation not available	29	29.5.5	protein.degradation.serine protease
Unconnected	NA	NA	GRMZM2G019404_P01	mha2	Plasma membrane ATPase	34	34.1	transport.p- and v-ATPases
Unconnected	NA	NA	GRMZM2G020523_P01	100272496	Peroxidase 2 ; Belongs to the peroxidase family. Classical plant (class III) peroxidase subfamily	26	26.12	misc.peroxidases
Unconnected	NA	NA	GRMZM2G020661_P01	100284756	Ras protein Rab11B; Ras-related protein Rab11B; Uncharacterized protein	30	30.5	signalling.G-proteins
Unconnected	NA	NA	GRMZM2G025215_P01	cl31807_1	Putative DUF1421 domain family protein	27	27.3.30	RNA.regulation of transcription.Trihelix, Triple-Helix transcription factor family
Unconnected	NA	NA	GRMZM2G029543_P01	pco095766	Malonyl CoA-acyl carrier protein transacylase	11	11.1.2	lipid metabolism.FA synthesis and FA elongation.Acetyl CoA Transacylase
Unconnected	NA	NA	GRMZM2G031859_P01	cl19897_1	Exosome complex exonuclease RRP44 homolog A; Belongs to the RNR ribonuclease family	31	31.2	cell.division
Unconnected	NA	NA	GRMZM2G032766_P01	100279462	Calcium-dependent lipid-binding (CaLB domain) family protein	20	20.2.2	stress.abiotic.cold
Unconnected	NA	NA	GRMZM2G033555_P01	GRMZM2G033555_P01	Dihydroflavonol-4-reductase	16	16.8.3	secondary metabolism.flavonoids.dihydroflavonols

Unconnected	NA	NA	GRMZM2G033641_P01	umc1462	Patellin-1; Putative patellin family protein; Uncharacterized protein	34	34.99	transport.misc
Unconnected	NA	NA	GRMZM2G034069_P01	100192966	NAD(P)-binding Rossmann-fold superfamily protein	16	16.8.3	secondary metabolism.flavonoids.dihydroflavonols
Unconnected	NA	NA	GRMZM2G039886_P01	GRMZM2G039886_P01	HSP40/DnaJ peptide-binding protein; DnaJ subfamily B member 5; Uncharacterized protein	20	20.2.1	stress.abiotic.heat
Unconnected	NA	NA	GRMZM2G043198_P01	pdh2	Pyruvate dehydrogenase E1 component subunit beta; The pyruvate dehydrogenase complex catalyzes the overall conversion of pyruvate to acetyl-CoA and CO2	8	8.1.1.1	TCA / org. transformation.TCA. pyruvate DH.E1
Unconnected	NA	NA	GRMZM2G048324_P01	nrx1	Nucleoredoxin1	35	35.2	not assigned.unknown
Unconnected	NA	NA	GRMZM2G051050_P01	GRMZM2G051050_P01	Ypt/Rab-GAP domain of gyp1p superfamily protein	30	30.5	signalling.G-proteins
Unconnected	NA	NA	GRMZM2G051219_P01	bf1	Blue fluorescent1	13	13.1.6.5.2	amino acid metabolism.synthesis.aromatic aa.tryptophan.anthranilate phosphoribosyltransferase
Unconnected	NA	NA	GRMZM2G054300_P04	103639279	APx1-Cytosolic Ascorbate Peroxidase ; Belongs to the peroxidase family	21	21.2.1	redox.ascorbate and glutathione.ascorbate
Unconnected	NA	NA	GRMZM2G059299_P01	100382042	G-type lectin S-receptor-like serine/threonine-protein kinase SD2-5; Putative D-mannose binding lectin domain related protein; Uncharacterized protein	26	26.16	misc.myrosinases-lectin-jacalin

Unconnected	NA	NA	GRMZM2G060702_P02	ADF3	Actin-depolymerizing factor 3; Actin-depolymerizing protein. Severs actin filaments (F-actin) and binds to actin monomers	31	31.1	cell.organisation
Unconnected	NA	NA	GRMZM2G061662_P01	100193236	NEDD8-activating enzyme E1 regulatory subunit; Regulatory subunit of the dimeric E1 enzyme. E1 activates RUB1/NEDD8 by first adenylating its C-terminal glycine residue with ATP, thereafter linking this residue to the side chain of the catalytic cysteine, yielding a RUB1-ECR1 thioester and free AMP. E1 finally transfers RUB1 to the catalytic cysteine of RCE1	29	29.5.11.2	protein.degradation.ubiquitin.E1
Unconnected	NA	NA	GRMZM2G061900_P01	100272423	Ras-related protein ARA-3; Uncharacterized protein	30	30.5	signalling.G-proteins
Unconnected	NA	NA	GRMZM2G067225_P01	aos1	Allene oxide synthase1; Putative cytochrome P450 superfamily protein; Uncharacterized protein	17	17.7.1.3	hormone metabolism.jasmonate.synthesis-degradation.allene oxidase synthase
Unconnected	NA	NA	GRMZM2G072240_P02	100383183	Carboxypeptidase; Belongs to the peptidase S10 family	29	29.5.5	protein.degradation.serine protease
Unconnected	NA	NA	GRMZM2G075655_P01	GRMZM2G075655_P01	Myosin heavy chain-related	31	31.1	cell.organisation
Unconnected	NA	NA	GRMZM2G080542_P01	gpm703	Protein arginine N-methyltransferase PRMT10; Belongs to the class I-like SAM-binding methyltransferase superfamily. Protein arginine N-methyltransferase family	26	26.6	misc.O-methyl transferases
Unconnected	NA	NA	GRMZM2G081037_P01	100273215	Metal-dependent protein hydrolase	35	35.2	not assigned.unknown

Unconnected	NA	NA	GRMZM2G083810_P01	hsp18f	17.5 kDa class II heat shock protein	20	20.2.1	stress.abiotic.heat
Unconnected	NA	NA	GRMZM2G093405_P05	cl7341_1a	Uncharacterized protein	35	35.2	not assigned.unknown
Unconnected	NA	NA	GRMZM2G100288_P01	100274007	Receptor-like protein kinase FERONIA; Belongs to the protein kinase superfamily	30	30.2.16	signalling.receptor kinases.Catharanthus roseus-like RLK1
Unconnected	NA	NA	GRMZM2G103771_P01	gpm462	Stress-inducible membrane pore protein	29	29.3.2	protein.targeting.mitochondria
Unconnected	NA	NA	GRMZM2G107665_P01	GRMZM2G107665_P01	Aminomethyltransferase	35	35.2	not assigned.unknown
Unconnected	NA	NA	GRMZM2G107984_P02	100501766	DEAD-box ATP-dependent RNA helicase 53	27	27.1.2	RNA.processing.RNA helicase
Unconnected	NA	NA	GRMZM2G111143_P01	pco087970b	Glycosyl hydrolase superfamily protein; Belongs to the glycosyl hydrolase 17 family	26	26.4.1	misc.beta 1,3 glucan hydrolases.glucan endo-1,3-beta-glucosidase
Unconnected	NA	NA	GRMZM2G116087_P01	cmu1	Chorismate mutase; Uncharacterized protein	13	13.1.6.2.1	amino acid metabolism.synthesis.aromatic aa.phenylalanine and tyrosine.chorismate mutase
Unconnected	NA	NA	GRMZM2G116258_P01	100193482	Putative glutamate-1-semialdehyde 2,1-aminomutase family protein; Uncharacterized protein ; Belongs to the class-III pyridoxal-phosphate-dependent aminotransferase family	19	19.3	tetrapyrrole synthesis.GSA
Unconnected	NA	NA	GRMZM2G119146_P01	GRMZM2G119146_P01	Uncharacterized protein	27	27.4	RNA.RNA binding
Unconnected	NA	NA	GRMZM2G124047_P01	IDP706	Putative serine peptidase S28 family protein	29	29.5	protein.degradation
Unconnected	NA	NA	GRMZM2G124365_P01	GRMZM2G124365_P01	Chorismate mutase; Uncharacterized protein	13	13.1.6.2.1	amino acid metabolism.synthesis.aromatic aa.phenylalanine and

								tyrosine.chorismate mutase
Unconnected	NA	NA	GRMZM2G126002_P01	pco064579	Putative oxidoreductase, aldo/keto reductase family protein	17	17.2.3	hormone metabolism.auxin.induced-regulated-responsive-activated
Unconnected	NA	NA	GRMZM2G133407_P04	pco095801	Soluble epoxide hydrolase	26	26.1	misc.misc2
Unconnected	NA	NA	GRMZM2G141704_P01	GRMZM2G141704_P01	ECA1 (ER-TYPE CA2+-ATPASE 1)	34	34.21	transport.calcium
Unconnected	NA	NA	GRMZM2G151734_P01	pco101658	Haloacid dehalogenase-like hydrolase (HAD) superfamily protein	33	33.99	development.unspecified
Unconnected	NA	NA	GRMZM2G153823_P01	IDP1950	Putative 14-3-3 protein ; Belongs to the 14-3-3 family	30	30.7	signalling.14-3-3 proteins
Unconnected	NA	NA	GRMZM2G154687_P01	GRMZM2G154687_P01	Phospholipase A1-IIgamma; Triacylglycerol lipase; Uncharacterized protein	11	11.9.2.1	lipid metabolism.lipid degradation.lipases.triacylglycerol lipase
Unconnected	NA	NA	GRMZM2G161905_P01	gst25	Glutathione S-transferase GST 25	26	26.9	misc.glutathione S transferases
Unconnected	NA	NA	GRMZM2G169384_P01	100282986	Multiple organellar RNA editing factor 8 chloroplastic/mitochondrial	33	33.99	development.unspecified
Unconnected	NA	NA	GRMZM2G173341_P01	100285211	Putative UDP-arabinopyranose mutase 5	10	10.5.5	cell wall.cell wall proteins.RGP
Unconnected	NA	NA	GRMZM2G176595_P01	pco112411b	Beta-expansin 1a; Beta-expansin 6; Uncharacterized protein ; Belongs to the expansin family	10	10.7	cell wall.modification
Unconnected	NA	NA	GRMZM2G177928_P01	GRMZM2G177928_P01	Strictosidine synthase; Uncharacterized protein	16	16.4.1	secondary metabolism.N misc.alkaloid-like
Unconnected	NA	NA	GRMZM2G179797_P03	cl891_1a	Threonine synthase; Uncharacterized protein	13	13.1.3.2.1	amino acid metabolism.synthesis.aspartate family.threonine.threonine synthase

Unconnected	NA	NA	GRMZM2G180335_P01	100279900	Dynamin-related protein 3A; Belongs to the TRAFAC class dynamin-like GTPase superfamily. Dynamin/Fzo/YdjA family	26	26.17	misc.dynamin
Unconnected	NA	NA	GRMZM2G181018_P01	100281165	Arabinogalactan protein	35	35.2	not assigned.unknown
Unconnected	NA	NA	GRMZM2G307992_P03	IDP643	Uncharacterized protein	35	35.2	not assigned.unknown
Unconnected	NA	NA	GRMZM2G316362_P02	cl27608_1	Acyl-desaturase; Uncharacterized protein	11	11.1.15	lipid metabolism.FA synthesis and FA elongation.ACP desaturase
Unconnected	NA	NA	GRMZM2G341729_P01	GRMZM2G341729_P01	USP family protein	20	20.2.2	stress.abiotic.cold
Unconnected	NA	NA	GRMZM2G428096_P01	GRMZM2G428096_P01	annotation not available	34	34.21	transport.calcium
Unconnected	NA	NA	GRMZM2G458549_P01	pco060326	Galactokinase	3	3.8.1	minor CHO metabolism.galactose.galactokinases
Unconnected	NA	NA	GRMZM2G589579_P01	100381552	Argonaute4a; Putative argonaute family protein; Uncharacterized protein ; Belongs to the argonaute family	27	27.3.36	RNA.regulation of transcription.Argonaute
Unconnected	NA	NA	GRMZM2G701221_P01	AY104923	ER6 protein	20	20.2.99	stress.abiotic.unspecified
Unconnected	NA	NA	GRMZM5G800842_P01	100382860	Ubiquitin-activating enzyme E1 2	29	29.5.11.2	protein.degradation.ubiquitin.E1
Unconnected	NA	NA	GRMZM5G825524_P01	GRMZM5G825524_P01	Vacuolar protein sorting-associated protein 35; Plays a role in vesicular protein sorting	29	29.3.4.3	protein.targeting.secretory pathway.vacuole
Unconnected	NA	NA	GRMZM5G833625_P01	GRMZM5G833625_P01	Uncharacterized protein	35	35.2	not assigned.unknown
Unconnected	NA	NA	GRMZM5G835704_P01	100282630	VIP1 protein	35	35.2	not assigned.unknown

CHAPTER 4: CONCLUSION AND FUTURE WORKS

Summary/Conclusion

Seeds are an important source of both calories and proteins. However, seeds of major staple crops are deficient in one or more of the essential amino acids that we do not synthesize in our bodies and must obtain from diet. Prolonged deficiency of these EAA can lead to severe malnutrition to both humans and livestock. One of the sustainable approaches to tackle this issue is through seed amino acid biofortification. Hence, to do so, we need to better understand the genetic architecture and regulation of absolute seed amino acid levels and relative composition (both PBAAs and FAAs). Maize is the most productive crop in terms of yield per acreage around the world and is used as both food and fiber. The major objective of this research was to identify the key genes that shape both PBAAs and FAAs in maize kernels by harnessing natural variation as well as analysis of proteomic dynamics of maize kernel developmental series.

In this study, I first performed GWAS studies on 76 PBAAs related traits measured and calculated from dry kernels of the 282 maize association panel grown in multiple years and extracted the first candidate gene list. I then compared the candidate genes list in the first analysis to a second one generated by an association analysis between the proteomic expression data and the PBAA composition. This association demonstrated the novel finding on the translational switch during seed maturation which might be contributing to the PBAA dynamics in seed. Integration of these two approaches yielded 80 high confidence candidate genes (HCCG). Functional analysis of these 80 PBAA HCCG revealed several previously studied and well characterized genes such as storage proteins, cell vesicle transport, sulfur assimilation and amino acid, lipid, and

CHO metabolic pathway genes. Using the integrative omics approach, I ranked these 80 PBAA HCCG genes and found that 27kD y zeins and 50kD y zeins, two well characterized genes that have been postulated to play roles in initiation and filling of seed storage proteins, were the most highly ranked genes. The latter supported the validity of my approach and reinforced my novel finding that strongly indicated that the translational machinery especially, ribosomal protein heterogeneity may be in the heart of the mechanism shaping the PBAA composition.

Using the same pipeline as PBAA, I performed GWAS on 109 FAAs related traits measured and calculated from dry kernels of the 282 maize association panel grown in multiple years with multiple replications and extracted the first candidate gene list. As stated before, I compared that list to a second candidate gene list generated from an association analysis between the proteomic expression data and both FAA absolute levels and composition quantified from a B73 maturation developmental series. When comparing the candidate gene lists from the two approaches, I found 120 HCCG for FAA. Functional analysis of the 120 FAA HCCG revealed previously characterized genes and biological processes such as RNA transcription, TCA cycle related genes, lipid, CHO and amino acid metabolism and cell vesicle transport. Interestingly, my result in FAA analysis again highlighted the role of translational machinery especially RPs gene cluster indicating its roles in shaping the FAA regulation and homeostasis in seeds.

The FAA study indicated that FAA are more complex than PBAA which may be due to the multi-faceted roles of FAA in central metabolism. Therefore, it might prove more beneficial to analyze each amino acid family separately. This will help to reduce the data complexity and better interpret the results.

Interestingly, this study revealed a novel interconnected cluster dominated by translational machinery genes, especially RP, for both FAA and PBAA, supporting the key role of translation dynamics in shaping both seed PBAA and FAA composition. My findings suggest that novel seed biofortification strategies, which target translational machinery dynamics should be considered and further explored. My study also suggests that some biological processes such as protein metabolism and translation shape both PBAA and FAA regulations. However, further analysis on the similarity and difference between the two genetic architectures and their relationships is still required.

Future work

In this study, I have identified key processes that are involved in shaping amino acid metabolism as well as multiple high confidence candidate genes. Nevertheless, additional analysis is needed to evaluate their role in shaping amino acid composition in seeds. Below, I suggest several follow up studies that could facilitate this goal.

Functional analysis of selected candidate genes using reverse and forward genetics approaches

One of the traditional approaches and the most common to test gene functionality is by either knocking out or overexpressing genes of interest. Since my study has shown the strong implications of the ribosomal proteins, I suggest that overexpression of selected HCCG RP that are down regulated during seed maturation under either constitutive or seed specific promoters. These specific alterations can help uncover their role in the proteome and amino acid composition in the dry seed. A complementary approach that could also shed some light on the ribosomal protein heterogeneity role in

seeds can include reverse genetic approach that include knock out of selected ribosomal proteins.

Similar approaches can be applied to several EIFs candidate genes that were revealed in my PBAA (eIF3) and FAA (eIF2α, eIF2β and eIF3). Previous studies has reported that the elongation factor 1a (EF-1α) mRNA concentration is highly correlated with the lysine content in maize endosperm of *W64Ao2* plants compared with its normal wild type counterpart (Habben, Moro, Hunter, Hamaker, & Larkins, 1995; Jia et al., 2013), while there are no studies for other eIFs relationships with amino acids as far as I know.

Analysis of translational switch during seed maturation using Ribo-seq approaches

My study suggests that a translational switch may occur during seed maturation and it could be one of the factors determining PBAA composition in seeds. I hypothesize that this switch may be part of a global translational reprogramming or a more specific translational switch. Analyzing the translational event during seed maturation can be done using methods such Ribo-Seq that helps determine the translated transcripts and their translational efficacy.

Integrating an additional orthogonal dataset extracted from SSP mutants

During my study here, in addition to B73 inbred line, I have also generated FAA and PBAA metabolic data from 10 different seed filling stages of the *opaque2* mutant in the B73 background (*o2/B73)*, as well as their proteomics data using shotgun sequencing. I suggest that we can integrate as well as compare the metabolic and proteomic

expressions of *o2/B73* mutant with the wild B73 to further narrow the candidate gene lists that regulate amino acid composition in maize kernels.

Preforming multivariate GWAS on both FAA and PBAA

In this study, I performed univariate GWAS for both PBAA and FAA. However, my correlation analysis of the absolute levels of both PBAA and FAA shows that these traits strongly correlates with each other (PBAA among themselves, FAA among themselves). Multivariate approach was shown to be more powerful to tackle dependent traits (Pitchers et al., 2017; Zhu & Zhang, 2009).

Uncovering the relationship between genetic relationship between PBAA and FAA

Although, FAA contributes only ~1-10% of total amino acid pools, it is one of the most important precursors for the protein bound amino acids. However, not much is known on the inter-regulation of these two AA pools, especially in seeds. It would be interesting to know if there are any common master regulators that can influence both PBAA and FAA in seeds. A preliminary study can be done by overlapping the HCCG from PBAA and FAA and identify the overlapped genes between the two and later use cloning to understand the gene/s specific role in regulating both free and bound amino acid.

The research project in my dissertation and its future works holds a tremendous potential for the development of nutritious and balanced maize kernels and hence could be used to mitigate the malnutrition in both human and livestock around the world. There has been a wealth of metabolic data generated for 282 maize association panel and both metabolic and proteomic data generated for 10 different seed filling stages of wild B73

and o2/B73 mutant, which will be made available to public and could be used by students, faculties, industries, and plant breeders. I believe that the findings from my research would open new avenues towards developing and breeding quality maize protein.

References

Habben, J. E., Moro, G. L., Hunter, B. G., Hamaker, B. R., & Larkins, B. A. (1995). Elongation factor 1 alpha concentration is highly correlated with the lysine content of maize endosperm. *Proceedings of the National Academy of Sciences*, *92*(19), 8640-8644.

Jia, M., Wu, H., Clay, K. L., Jung, R., Larkins, B. A., & Gibbon, B. C. (2013). Identification and characterization of lysine-rich proteins and starch biosynthesis genes in the opaque2mutant by transcriptional and proteomic analysis. *BMC plant biology*, *13*(1), 60.

Merchante, C., Stepanova, A. N., & Alonso, J. M. (2017). Translation regulation in plants: an interesting past, an exciting present and a promising future. *The Plant Journal*, *90*(4), 628-653.

Pitchers, W. R., Nye, J., Márquez, E. J., Kowalski, A., Dworkin, I., & Houle, D. (2017). The power of a multivariate approach to genome-wide association studies: an example with Drosophila melanogaster wing shape. *bioRxiv*, 108308.

Turner-Hissong, S. D., Bird, K. A., Lipka, A. E., King, E. G., Beissinger, T. M., & Angelovici, R. (2020). Genomic prediction informed by biological processes expands our understanding of the genetic architecture underlying free amino acid traits in dry Arabidopsis seeds. *G3: Genes, Genomes, Genetics*.

Zhu, W., & Zhang, H. (2009). Why do we test multiple traits in genetic association studies? *Journal of the Korean Statistical Society, 38*(1), 1-10.

Appendix

Supplementary dataset for Chapter 2 and Chapter 3 can be found at *https://missouri.box.com/s/7igm6gjbsf0tbbgn8bin8dz63af9alc4*

VITA

Vivek Shrestha was born on August 12th, 1990 in Inaruwa, Sunsari, Nepal to the parents of late. Kul Bahadur Shrestha and late. Uma Shrestha. He received his Bachelor of Science in Agriculture with a major in Plant Pathology from Tribhuvan University, Nepal in 2011. After completion of his undergraduate, he volunteered in Nepal Academy of Science and Technology (NAST) as a research assistant to work on PCR based diagnosis of Citrus greening/ Huanglongbing disease in Nepal. Later, he joined the International Maize and Wheat Improvement Center (CIMMYT) as a technical officer where he got involved in a food security project in the rural areas of Nepal and also got familiar with appropriate mechanization, seed marketing and value chain analysis. It was during his work at CIMMYT where he worked with farmers in remote villages of Western and Far-Western Nepal that he developed a deep interest in the field of agriculture to help farmers to improve their livelihood. He realized the importance of improved seed and thus developed an interest in the field of Plant Genetics and Breeding that landed him in a Maize Genetics lab at South Dakota State University (SDSU) for his master's degree. At SDSU, he worked on the projects related to search for modifiers of the maize gametophyte factor (Ga1-s) and quantitative trait polymorphisms emerging from doubled-haploid maize lines. After completion of his MS degree in 2016, he joined the Division of Biological Sciences at the University of Missouri in Dr. Ruthie Angelovici's lab where he worked in uncovering the genetic architecture and metabolic basis of seed amino acid composition in maize kernels using multi-omics integration approach. He will receive his Ph.D. in Biological Sciences in December 2020.