UNCOVERING THE GENETIC BASIS OF SEED AMINO ACID COMPOSITION IN ARABIDOPSIS USING A MULTI-OMICS INTEGRATIVE APPROACH

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfilment

of the Requirements for the Degree

Doctor of Philosophy

by

MARIANNE LYNN SLATEN

Dr. Ruthie Angelovici, Dissertation Supervisor

MAY 2021

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

UNCOVERING THE GENETIC BASIS OF SEED AMINO ACID COMPOSITION IN ARABIDOPSIS USING A MULTI-OMICS INTEGRATIVE APPROACH

Presented by Marianne Lynn Slaten, a candidate for the degree of doctor of philosophy, and hereby certify that, in their opinion, it is worthy of acceptance.

____

Dr. Ruthie Angelovici

____

Dr. J. Chris Pires

____

Dr. Elizabeth King

____

Dr. Trupti Joshi

DEDICATION

"Here's to strong women. May we know them. May we be them. May we raise them."

– *Unknown*

ACKNOWLEDGEMENTS

I suppose…at long last…this dissertation signals the end of the road for my PhD. It has at times felt like a very long road, but one that I feel very privileged to have walked.

I acknowledge, however, that I have not walked it alone, rather have a long list of thank you's to extend. For it was along this journey that I have been given opportunities to be among a group of talented, hardworking, passionate people that continue to strive for continued learning and a vision of a better world; I have met some of my closest of confidants; I been honored to learn from researchers that I greatly admire; and I have been reminded that there is no substitute for hard work and perseverance.

First and foremost, I'd like to thank my PhD advisor, Dr. Ruthie Angelovici, who took a chance on me when I reached out one day and inquired about PhD positions. I have learned so much about science, about life, and about who I am, owed largely to the vast opportunities given to me beyond the textbook and the lab bench. **Gracias**

I'd also like to thank my committee members Dr. Chris Pires, Dr. Libby King, and Dr. Trupti Joshi, to whom I greatly look up to for not only their research programs, but also their willingness to share expertise with students. I am so thankful for their irreplaceable, continued advice and guidance throughout the past years. **Merci**.

Of course, a lab is but nothing if it is not filled with great people. I'd like to thank all those who have assisted with my projects, who have added to the vibrant lab culture on a daily basis and have made me a cup of joe during those much needed coffee breaks. Dr. Makenzie Mabry, Dr. Vivek Shrestha, Dr. Abou Yobi, Dr. Sarah Turner-Hissong, Yen On Chan, Sam Holden, Gabi Akrap, Clement Bagaza and so many more. **Xiexie**.

I would be remiss if I did not also acknowledge my beginnings. It all started in the lab of Dr. Mark Campbell at the humble Truman State University. Dr. Campbell was the first to drag me into a lab. In his organized (albeit quite unorganized) chaos, he taught me not only to find a passion for agricultural systems but empowered me with a pipette and a protocol. He taught me to believe in myself and that I could do more than I ever thought I could. The passion for which he wakes and brings to the world every day is one that I hope one day I can share with others**. Shukran**

To my family. Those loving individuals who have put up with me talking about school for oh so very long. Their continued love and unwavering faith of me and my abilities have served as a foundation for my strength in the most trying times. **Danke**

And last, but definitely not least, to Chris. Your love and patience has kept me afloat in the roughest waters. You’ve loved me through the chaos, always accepting me and never doubting my success. All those “Dr Grip Pens” you gave me as reminders of the prized Dr. title that lie at the end of the PhD journey feel even more satisfying to write with now. On to the next adventure.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

UNCOVERING THE GENETIC BASIS OF SEED AMINO ACID COMPOSITION IN ARABIDOPSIS USING A MULTI-OMICS INTEGRATIVE APPROACH

Marianne Slaten

Dr. Ruthie Angelovici, Dissertation Supervisor

ABSTRACT

Seeds are a vital source of protein in the diet of humans and livestock. However, protein composition in the seed is low, comprising about 10% of the total composition in the seed. Additionally, protein quality in the seed is poor due to low concentrations of certain essential amino acids (EAA). Since the body is unable to produce EAA, they must be consumed in the diet and failure to do so has detrimental, potentially irreversible, health implications that can result in death. In developing counties where meat and dairy are lacking, protein-energy malnutrition frequently occurs. In contrast, in developing countries large portions of seeds are used in the diet of livestock which must be supplemented with costly synthetic amino acids. Collectively seed amino acid composition of major crops are not sufficient to meet dietary requirements.

Protein in the seed is comprised of free amino acids (FAA) and protein bound amino acids (PBAA) which have both been the targets of manipulation in order to create a seed with a more balanced amino acid profile. However, upon perturbations to the proteome, mutant seeds have demonstrated a rebalancing phenomenon where even large alterations to the amino acid composition activate a compensation mechanism that returns amino acid levels to a comparable composition to the wild-type. Although a lot is known about amino acid metabolic pathways, what regulates such rebalancing mechanisms is still unknown. However, despite the tight regulation, natural variation does exist in seed FAA

and PBAA across *Arabidopsis* ecotypes with a unique composition specific to each ecotype; this suggests rebalancing has a genetic basis. Thus, the first step in seed biofortification efforts must be to first increase the fundamental understanding of the genetic basis of both FAA and PBAA composition in the seed.

Chapter One of this dissertation gives a more in-depth introduction that elaborates on amino acid composition in the seed, the challenges identified in previous experimentation, and how the content of Chapter Two through Chapter Four builds upon and adds value to the area of seed amino acid research as a whole.

Chapter Two focuses on uncovering the genes and biological processes that underly the regulation of free Glutamine which belongs to the Glutamate Family (Arginine, Proline, Glutamine, and Glutamate). Although Glutamine is not an EAA, it is a major nitrogen-containing amino acid that is transported to the seed; thus it's regulatory control is of particular interest. I harness the natural variation of Glutamine in a 360 *Arabidopsis* diversity panel to uncover key regulatory genes. Later, I validate observations from GWAS using both a quantitative trait locus (QTL) analysis and reverse genetic approaches to identify a unique, seed-specific Glutamine-glucosinolate relationship that alters nitrogen and sulfur homeostasis in the seed in the *Arabidopsis* 360 population. Such finds were substantial as they link primary and secondary metabolism in the seed.

Chapter Three focuses on uncovering the genetic basis underlying PBAA composition in dry *Arabidopsis* seeds while expanding upon the work completed in Chapter Two. 576 high confidence candidate genes (HCCGs) are found through integration of GWAS using PBAA traits and transcriptomic analysis across seed development of two mutants showing active rebalancing. To reveal the underlying biological process, I further subject the HCCGs to a protein-protein interaction (PPI) network that strongly suggests that ribosomal genes and potentially other translational machinery may be in the heart of PBAA composition homeostasis and the proteomic rebalancing response.

Chapter Four addresses the need of a comprehensive tool to efficiently and automatically analyze many biochemical derived-traits in GWAS, while also completing pre and post-GWAS analysis. Here, I present the R tool HAPPI GWAS, describing each step in the pipeline, and giving an example of its implementation.

Lastly, Chapter Five reiterates the contributions of this dissertation to the field of seed amino acid research and provides insight into future direction and research projects. The results from this work are vital steps in understanding the complex regulatory mechanisms underlying amino acid composition in the seed which can be used in manipulating the amino acid pools in future translational crop research.

CHAPTER 1: INTRODUCTION

1.1. AMINO ACID COMPOSITION IN THE SEED

Crop seeds, such as cereals and legumes, play a critical role as a primary food source in the diet of humans and livestock (FAO). Commonly, seeds are consumed as a source of energy in the form of carbohydrates but are also important sources of protein in developing countries and in the livestock industry of developed countries. Unfortunately, most of the crop grains are insufficient to meet the dietary requirements of essential amino acids (EAA) for humans and monogastric livestock. Failure to consume sufficient levels of EAA in the diet can lead to protein-energy malnutrition which severely affects the immune, gastrointestinal, nervous, and cardiovascular systems (Grover & Ee, 2009). Protein malnutrition is often a chronic condition in developing countries where meat is lacking in the diet. Even in the U.S., where protein is more readily available, the unique relationship between undernutrition and overweight persists (WHO). Furthermore, millions of dollars are spent every year on supplementing synthetic amino acids in the diet of livestock (Research, 2017). Thus, there is an imminent need for a seed with a balanced EAA composition to help alleviate malnutrition and provide economic advantages for human and livestock.

In general, the seed contains two pools of amino acids (AAs): free amino acids (FAA) and protein bound amino acids (PBAA). FAA composes approximately 7% of total amino acid in *Arabidopsis* seed and approximately 1-10% of total amino acids in maize seed ((Amir et al., 2018). In the seed, FAA are mainly funneled toward protein synthesis (Amir et al., 2018), but also serve as precursor molecules for primary and secondary metabolites (Galili & Amir, 2013) and play a role in plant signaling (Angelovici et al., 2009; Angelovici et al., 2010; Holding & Messing, 2013). In contrast, PBAA composes approximately 90-99% of the total protein in *Arabidopsis* seed and approximately 95% of the protein in maize seed. Depending on the plant species, approximately 50-70% of the PBAA is stored in the form of seed storage proteins (SSP)(Sherry A. Flint-Garcia et al., 2009). These highly abundant SSP in most crop species have low levels of EAA compared to other seed proteins making them the culprit in the poor nutritional quality.

1.2. TARGETED ENGINEERING OR BREEDING FOR LOW SSPS RESULT IN PROTEOMIC REBALANCING IN MOST PLANT SPECIES

Due to their high abundancy and thus integral role in the overall composition in the seed, many attempts have been carried out in the past two decades to discover natural knock-outs and/or to experimentally knock-out or knock-down the various SSPs. These efforts resulted in the identification of a phenomenon termed proteomic rebalancing in which the elimination of SSPs (which can account for over half of the proteomic sink) activates a compensation mechanism that leads to a similar total amino acid level and composition as the wild-type (WT). The mechanism appears to be conserved across plant species and has been documented in soybean, maize, rice, wheat, barley, and *Arabidopsis* (Bose et al., 2020; Hagan et al., 2003; Herman, 2014; Schmidt et al., 2011; Scossa et al., 2008; Tan-Wilson & Wilson, 2012; Withana-Gamage et al., 2013; Wu & Messing, 2014). For example, in the naturally occurring maize *opaque-2* (*o2*) null mutant (Mertz et al., 1964), deletion of a BZIP family transcription factor and subsequent removal of 60-70% of endosperm protein in the seed (in the form of SSP), led to major proteomic reprogramming and proportional increases in high lysine protein fractions; it was established that such increases were the results of increases in non -SSP that have a larger proportion of lysine compared to the SSP in maize. Nevertheless, the overall protein composition was only slightly altered compared to wild-type maize (Coleman & Larkins, 1999; Herman, 2014; Mertz et al., 1964; Schmidt et al., 2011; Withana-Gamage et al., 2013; Wu & Messing, 2014; Wu et al., 2012). Similarly, in barley, no significant protein abundance differences were observed between a WT and D-hordein null line (Bose et al., 2020). However, both studies showed that the agronomic properties of the seed are often negatively affected making it difficult to grow and propagate (Lambert et al., 1969). Similar rebalancing effects were also observed when quality SSP were overexpressed in the seeds of transgenic canola (Altenbach et al., 1992), tobacco (Altenbach et al., 1989; Shaul & Galili, 1992), rice (Lee et al., 2001), and soybean (Dinkins et al., 2001). In soybean, overexpression of cytosolic *O*-acetylserine sulfhydrylase (OASS) resulted in 58-74% increase in protein-bound cysteine levels compared to the WT (Kim et al., 2012). Overall, the study did successfully increase concentrations of sulfur-containing amino acids in the seed. Similar studies showed less success, where increases in FAA were at the expense of other sulfur-rich proteins (Amir & Tabe, 2006; Hagan et al., 2003; Molvig et al., 1997; Tabe & Droux, 2002). Issues with protein instability have also been observed (Altenbach & Simpson, 1990; De Clercq et al., 1990; Forsyth et al., 2005; Galili & Höfgen, 2002; Hoffman et al., 1988; Tabe & Higgins, 1998; Torrent et al., 1997; Wenefrida et al., 2009). Collectively, the body of literature revolving around PBAA manipulation in the seed emphasizes the challenges in altering PBAA as they appear to be tightly regulated.

Since the PBAA alteration did not yield any major success, attempts at perturbing the content of FAA in the seed have also been made. The pathways of most AA have been elucidated and leveraged to successfully increase the FAA pool. Several of the AAs control and are controlled by other AAs and inhibitory loops that receive feedback regulation by their products (Binder, 2010; Gu et al., 2010; Ingle, 2011; Toubiana et al., 2012). Additionally, the increases in FAA biosynthesis are tempered by an increase in catabolism (Zhu & Galili, 2004). By introducing feedback insensitive enzymes, increases in the biosynthesis of FAA has been observed. To that end, several studies have increased biosynthesis of lysine by knocking out the rate limiting step and inhibiting the feedback loop (Falco et al., 1995; Karchi et al., 1993; Mazur et al., 1999; Zhu & Galili, 2003, 2004). For example, Zhu et al (2003) found that free lysine could be substantially increased by enhancing synthesis and reducing catabolism in *Arabidopsis* seed (Zhu & Galili, 2003). These results show that in general, FAA are much more amendable to alterations in overall composition compared to PBAA. However in some cases, seeds with altered FAA content showed plummeting rates of germination, seedling establishment, and general plant growth (Bright et al., 1983; Ghislain et al., 1994; Heremans & Jacobs, 1997; Zhu & Galili, 2003). In the case of lysine accumulation in the seed, retardation of germination after enhancing synthesis and blocking catabolism was traced back to an impaired tricarboxylic acid (TCA) cycle affecting cellular energy homeostasis (Angelovici et al., 2009; Angelovici et al., 2010; Galili, 2011; Zhu & Galili, 2003). Despite observed increases in FAA, no PBAA changes result, suggesting FAA availability is not a limiting factor in PBAA accumulation. Because the FAA makes up such a small proportion of the total protein in the seed, even successful increases in FAA offer limited changes to the overall proteome. Collectively, the tight regulation of FAA and disconnect between PBAA and the FAA precursors has led to undesirable, and at times limited, alterations in the FAA pool that has resulted in an incomplete understanding of the regulatory mechanisms of FAA composition in the seed.

1.3. THE REGULATORY MECHANISM UNDERLYING PROTEIN REBALANCING REMAIN UNCLEAR

The underlying molecular mechanism underlying the PBAA homeostasis in seeds or the rebalancing phenomenon is unclear. Null mutant screens have been largely unsuccessful in identifying major regulatory components because of resulting lethality or activation of the rebalancing mechanisms yielding no distinct AA phenotype. The phenomenon has however, been well characterized by two main trends: 1) many small elevations of many non-SSPs and 2) pronounced elevation of specific proteins. For example, RNA-silencing techniques of soybean SSP resulted in a remodeling of the proteome due to increases in major proteins such as Kunitz trypsin inhibitor, soybean lectin, and the immunodominant soybean allergen P34 (Schmidt et al., 2011). In addition to large changes in select proteins, global increases in a variety of other non-SSPs were observed (Schmidt et al., 2011). Similar results have been observed in maize where proteomic analysis of six opaque endosperm maize mutants found general increases in non-zein proteins and more specific increases in proteins with relatively higher content of lysine (Morton et al., 2016). However, the mechanism by which these changes are occurring have yet to be elucidated. Some studies demonstrate a disconnect with the rebalanced proteome and the transcriptome where no increases in transcript abundance are observed (Herman, 2014; Schmidt et al., 2011), while other studies observe differences in the transcriptional landscape of mutants undergoing active rebalancing compared to the wild-type (WT) (Hunter et al., 2002). The exact regulatory mechanism and relationship between transcription and proteins remains unclear and needs further investigating. To date, no major regulator has been found that enables PBAA manipulation in a specific genetic line. Lack of this basic understanding impedes any significant alterations of AA composition and major seed biofortification efforts. However, natural variation does exist across natural populations which indicates that rebalancing is genetically determined. This topic is discussed in more detail in the next section.

1.4. HARNESSING NATURAL VARIATION TO UNCOVER THE GENETIC BASIS OF AMINO ACID COMPOSITION

It is evident that a complementary approach to mutant analysis is required to uncover the genetic basis of AA regulation in the seed and gain a greater understanding of rebalancing and reprogramming effects. To that end, variation in maize kernel quality and seed characteristics, including protein content and quality, have been shown to be targets of selection during domestication and crop improvement (Sherry A Flint-Garcia et al., 2009). Such selection was made possible due to natural variation in AA composition found across natural populations (Angelovici et al., 2017; Ruthie Angelovici et al., 2013; Deng et al., 2017; Hou et al., 2005; Vaughn et al., 2014; Withana-Gamage et al., 2013). This also suggests that the regulatory genes controlling these AA traits may have been evolutionarily selected for optimization and maintenance of AA composition. This explains the diversity in PBAA across populations, where PBAA composition is fairly fixed for any one particular genotype. Such natural variation in AA composition is well suited to be leveraged in a metabolic Genome Wide Association Study (GWAS) linking locations in the genome with various AA traits. Unlike mutant analysis, GWAS analyzes entire populations, assessing SNP variation across the genome to find statistical associations (i.e. genetic control) with complex traits of interest.

The majority of GWAS work has been completed 360 *Arabidopsis* diversity panel using FAA measurements in dry seed. Two such prominent examples are (Ruthie Angelovici et al., 2013) and (M. L. Slaten et al., 2020). In (Ruthie Angelovici et al., 2013), GWAS was performed on FAA branched chain amino acids (BCAA) traits using the 360 *Arabidopsis* panel. Branched chain amino acid transferases; *BCAT1* and *BCAT2,* were found to have a strong association with the BCAA traits. Additionally, (M. L. Slaten et al., 2020) analyzed the Glutamate Family FAA in a GWAS using the 360 *Arabidopsis* diversity panel. Results showed an unexpected association with the analyzed FAA traits and the aliphatic glucosinolate biosynthetic pathway. Findings suggested a relationship between secondary metabolites and primary FAA metabolites in the seed previously undocumented.

1.5. CHALLENGES ASSOCIATED WITH GWAS USING METABOLIC TRAITS

GWAS studies have been used to successfully dissect complex metabolic traits in *Arabidopsis* (Wu et al., 2016), maize (Schaefer et al., 2018), soybean (Zhaoming Qi et al., 2018), peanut (Hui Zhang et al., 2019), and rapeseed (Min Yao et al., 2020). However, one of the main challenges associated with any GWAS is the identification of an extensive list of genes that is frequently difficult to parse and subset in a biologically meaningful way. To overcome these challenges, researchers have incorporated GWAS results with other multi-omics data to further validate and increase confidence in a small number of genes. One solution has been through the use of metabolite-transcript correlation (Wu et al., 2016). For example, Wu et al (2016) identified novel metabolite-related genes in *Arabidopsis thaliana* by using an integrative network strategy combining metabolite and transcript data with quantitative genetic mapping approaches. Other work has leveraged the framework of co-expression networks to integrate the output of GWAS and improve confidence in a fewer number of genes (Schaefer et al., 2018; H. Zhang et al., 2019).

Another main challenge of GWAS is of the interpretation of the results. A list of significantly associated SNPs is identified through GWAS, which alone offers limited biological insight; these SNPs can then be associated with genes that fall within region of high linkage disequilibrium (LD) with the significant SNP. Yet still, it is hard to elucidate gene-gene interactions and gene-trait interactions, particularly across many metabolomic traits that are interconnected. Instead, candidate genes identified in GWAS can be used in a network-based analysis to identify key mechanisms and/or pathways (Angelovici et al., 2017; W. Chen et al., 2016; Wu et al., 2016). For example, Wu et al (Wu et al., 2016) compared two orthogonal datasets in *Arabidopsis* (GWAS candidate genes versus a metabolite-transcript correlation network) to identify a list of high confidence regulatory features. Additionally, Schaefer et al (Schaefer et al., 2018) present the tool Camoco and demonstrated the integration of GWAS output with gene expression networks to prioritize genes in maize. Tools such as WGCNA (weighted gene co-expression network analysis) (Langfelder & Horvath, 2008) have also been used to create gene expression networks, that can be used to identify hub genes that are compared to the results of orthogonal dataset (DiLeo et al., 2011; Z. Qi et al., 2018), or used to incorporate metabolite data in gene-metabolite networks (Pei et al., 2017). Collectively, by harnessing the natural variation in AA composition and identifying interactions through a multi-omics approach, AA regulators that have been previously unidentifiable through mutant analysis or GWAS alone may be uncovered.

Therefore, in this dissertation I take a systems biology approach to uncover genetic control of PBAA and FAA composition in the *Arabidopsis* seed. To that end, I perform GWAS using the Glutamate Family FAA in the *Arabidopsis* 360 population and on 284 PBAA biochemical, derived ratio traits using the *Arabidopsis* 1001 population. To facilitate running such a large number of traits, I co-create HAPPI GWAS, an R-based tool to automate pre-GWAS, GWAS, and post-GWAS analyses. Finally, I analyze the transcriptome of two SSP mutants across seven stage of seed filling, in addition to analyzing the proteome and PBAA of the three genotypes in the dry seed. Collectively, this work furthers the goal of unraveling the complex process of AA regulation in the seed by identifying genes and putative mechanisms that can be perturbed in future experimentation and translated into more complex crop species with greater economic impact.

1.6. REFERENCES

Altenbach, Susan B, Chiung-Chi Kuo, Lisa C Staraci, Karen W Pearson, Connie Wainwright, Anca Georgescu, and Jeffrey Townsend. 1992. 'Accumulation of a Brazil nut albumin in seeds of transgenic canola results in enhanced levels of seed protein methionine', *Plant molecular biology*, 18: 235-45.

Altenbach, Susan B, Karen W Pearson, Gabrielle Meeker, Lisa C Staraci, and Samuel SM Sun. 1989. 'Enhancement of the methionine content of seed proteins by the expression of a chimeric gene encoding a methionine-rich protein in transgenic plants', *Plant molecular biology*, 13: 513-22.

Altenbach, Susan B, and Robert B Simpson. 1990. 'Manipulation of methionine-rich protein genes in plant seeds', *Trends in Biotechnology*, 8: 156-60.

Amir, R, and L Tabe. 2006. 'Molecular approaches to improving plant methionine content'.

Amir, Rachel, Gad Galili, and Hagai Cohen. 2018. 'The metabolic roles of free amino acids during seed development', *Plant Science*.

Angelovici, Ruthie, Albert Batushansky, Nicholas Deason, Sabrina Gonzalez-Jorge, Michael A Gore, Aaron Fait, and Dean DellaPenna. 2017. 'Network-guided GWAS improves identification of genes affecting free amino acids', *Plant Physiology*, 173: 872-86.

Angelovici, Ruthie, Aaron Fait, Xiaohong Zhu, Jedrzej Szymanski, Ester Feldmesser, Alisdair R Fernie, and Gad Galili. 2009. 'Deciphering transcriptional and metabolic networks associated with lysine metabolism during Arabidopsis seed development', *Plant Physiology*, 151: 2058-72.

Angelovici, Ruthie, Gad Galili, Alisdair R Fernie, and Aaron Fait. 2010. 'Seed desiccation: a bridge between maturation and germination', *Trends in plant science*, 15: 211-18.

Angelovici, Ruthie, Alexander E Lipka, Nicholas Deason, Sabrina Gonzalez-Jorge, Haining Lin, Jason Cepela, Robin Buell, Michael A Gore, and Dean DellaPenna. 2013. 'Genome-wide analysis of branched-chain amino acid levels in Arabidopsis seeds', *The Plant Cell*, 25: 4827-43.

Binder, Stefan. 2010. 'Branched-chain amino acid metabolism in Arabidopsis thaliana', *The Arabidopsis book/American Society of Plant Biologists*, 8.

Bose, U., J. A. Broadbent, K. Byrne, M. J. Blundell, C. A. Howitt, and M. L. Colgrave. 2020. 'Proteome Analysis of Hordein-Null Barley Lines Reveals Storage Protein Synthesis and Compensation Mechanisms', *J Agric Food Chem*, 68: 5763-75.

Bright, Simon WJ, Joseph SH Kueh, and Sven E Rognes. 1983. 'Lysine transport in two barley mutants with altered uptake of basic amino acids in the root', *Plant Physiology*, 72: 821-24.

Chen, Wei, Wensheng Wang, Meng Peng, Liang Gong, Yanqiang Gao, Jian Wan, Shouchuang Wang, Lei Shi, Bin Zhou, and Zongmei Li. 2016. 'Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals', *Nature communications*, 7: 12767.

Coleman, Craig E, and Brian A Larkins. 1999. 'The prolamins of maize.' in, *seed Proteins* (Springer).

De Clercq, Ann, Martine Vandewiele, Jozef Van Damme, Philippe Guerche, Marc Van Montagu, Joël Vandekerckhove, and Enno Krebbers. 1990. 'Stable Accumulation of Modified 2S Albumin Seed Storage Proteins with Higher Methionine Contents in Transgenic Plants', *Plant Physiology*: 970-79.

Deng, Min, Dongqin Li, Jingyun Luo, Yingjie Xiao, Haijun Liu, Qingchun Pan, Xuehai Zhang, Minliang Jin, Mingchao Zhao, and Jianbing Yan. 2017. 'The genetic architecture of amino acids dissection by association and linkage analysis in maize', *Plant biotechnology journal*, 15: 1250-63.

DiLeo, M. V., G. D. Strahan, M. den Bakker, and O. A. Hoekenga. 2011. 'Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome', *PLoS One*, 6: e26683.

Dinkins, Randy D, MS Srinivasa Reddy, Curtis A Meurer, Bo Yan, Harold Trick, Françoise Thibaud-Nissen, John J Finer, Wayne A Parrott, and Glenn B Collins. 2001.

'Increased sulfur amino acids in soybean plants overexpressing the maize 15 kDa zein protein', *In Vitro Cellular & Developmental Biology-Plant*, 37: 742-47.

Falco, SC, T Guida, M Locke, J Mauvais, C Sanders, RT Ward, and P Webber. 1995. 'Transgenic canola and soybean seeds with increased lysine', *Bio/technology*, 13: 577.

FAO. 'Staple foods: What do people eat? ', Accessed June 11. (http://www.fao.org/3/u8480e/u8480e07.htm.

Flint-Garcia, Sherry A, Anastasia L Bodnar, and M Paul Scott. 2009. 'Wide variability in kernel composition, seed characteristics, and zein profiles among diverse maize inbreds, landraces, and teosinte', *Theoretical and Applied Genetics*, 119: 1129-42.

Forsyth, Jane L, Frederic Beaudoin, Nigel G Halford, Richard B Sessions, Anthony R Clarke, and Peter R Shewry. 2005. 'Design, expression and characterisation of lysine-rich forms of the barley seed protein CI-2', *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1747: 221-27.

Galili, G., and R. Amir. 2013. 'Fortifying plants with the essential amino acids lysine and methionine to improve nutritional quality', *Plant Biotechnol J*, 11: 211-22.

Galili, Gad. 2011. 'The aspartate-family pathway of plants: linking production of essential amino acids with energy and stress regulation', *Plant signaling & behavior*, 6: 192-95.

Galili, Gad, and Rainer Höfgen. 2002. 'Metabolic engineering of amino acids and storage proteins in plants', *Metabolic engineering*, 4: 3-11.

Ghislain, Marc, Valérie Frankard, Dirk Vandenbossche, Benjamin F Matthews, and Michel Jacobs. 1994. 'Molecular analysis of the aspartate kinase-homoserine dehydrogenase gene from Arabidopsis thaliana', *Plant molecular biology*, 24: 835-51.

Grover, Zubin, and Looi C Ee. 2009. 'Protein energy malnutrition', *Pediatric Clinics*, 56: 1055-68.

Gu, Liping, A Daniel Jones, and Robert L Last. 2010. 'Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant', *The Plant Journal*, 61: 579-90.

Hagan, ND, N Upadhyaya, LM Tabe, and TJV Higgins. 2003. 'The redistribution of protein sulfur in transgenic rice expressing a gene for a foreign, sulfur-rich protein', *The Plant Journal*, 34: 1-11.

Heremans, Betty, and Michel Jacobs. 1997. 'A Mutant of Arabidopsis thaliana lpar; L.) Heynh. with Modified Control of Aspartate Kinase by Threonine', *Biochemical genetics*, 35: 139-53.

Herman, Eliot M. 2014. 'Soybean seed proteome rebalancing', *Frontiers in plant science*, 5: 437.

Hoffman, Leslie M, Debra D Donaldson, and Eliot M Herman. 1988. 'A modified storage protein is synthesized, processed, and degraded in the seeds of transgenic plants', *Plant molecular biology*, 11: 717-29.

Holding, David, and Joachim Messing. 2013. 'Evolution, structure, and function of prolamin storage proteins', *Seed genomics*: 138-58.

Hou, Anfu, Kede Liu, Niramol Catawatcharakul, Xurong Tang, Vi Nguyen, Wilfred A Keller, Edward WT Tsang, and Yuhai Cui. 2005. 'Two naturally occurring deletion mutants of 12S seed storage proteins in Arabidopsis thaliana', *Planta*, 222: 512-20.

Hunter, B. G., M. K. Beatty, G. W. Singletary, B. R. Hamaker, B. P. Dilkes, B. A. Larkins, and R. Jung. 2002. 'Maize opaque endosperm mutations create extensive changes in patterns of gene expression', *Plant Cell*, 14: 2591-612.

Ingle, Robert A. 2011. 'Histidine biosynthesis', *The Arabidopsis book/American Society of Plant Biologists*, 9.

Karchi, Hagai, Orit Shaul, and Gad Galili. 1993. 'Seed-specific expression of a bacterial desensitized aspartate kinase increases the production of seed threonine and methionine in transgenic tobacco', *The Plant Journal*, 3: 721-27.

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

Lambert, RJ, DE Alexander, and JW Dudley. 1969. 'Relative Performance of Normal and Modified Protein (Opaque-2) Maize Hybrids 1', *Crop Science*, 9: 242-43.

Langfelder, Peter, and Steve Horvath. 2008. 'WGCNA: an R package for weighted correlation network analysis', *BMC bioinformatics*, 9.

Lee, Soo In, Hyun Uk Kim, Yeon-Hee Lee, Suk-Cheol Suh, Yong Pyo Lim, Hyo-Yeon Lee, and Ho-Il Kim. 2001. 'Constitutive and seed-specific expression of a maize lysine-feedback-insensitive dihydrodipicolinate synthase gene leads to increased free lysine levels in rice seeds', *Molecular Breeding*, 8: 75-84.

Mazur, Barbara, Enno Krebbers, and Scott Tingey. 1999. 'Gene discovery and product development for grain quality traits', *Science*, 285: 372-75.

Mertz, Edwin T, Lynn S Bates, and Oliver E Nelson. 1964. 'Mutant gene that changes protein composition and increases lysine content of maize endosperm', *Science*, 145: 279-80.

Molvig, Lisa, Linda M Tabe, Bjorn O Eggum, Andrew E Moore, Stuart Craig, Donald Spencer, and Thomas JV Higgins. 1997. 'Enhanced methionine levels and increased nutritive value of seeds of transgenic lupins (Lupinus angustifolius L.) expressing a sunflower seed albumin gene', *Proceedings of the National Academy of Sciences*, 94: 8393-98.

Morton, K. J., S. Jia, C. Zhang, and D. R. Holding. 2016. 'Proteomic profiling of maize opaque endosperm mutants reveals selective accumulation of lysine-enriched proteins', *J Exp Bot*, 67: 1381-96.

Pei, G., L. Chen, and W. Zhang. 2017. 'WGCNA Application to Proteomic and Metabolomic Data Analysis', *Methods Enzymol*, 585: 135-58.

Qi, Z., Z. Zhang, Z. Wang, J. Yu, H. Qin, X. Mao, H. Jiang, D. Xin, Z. Yin, R. Zhu, C. Liu, W. Yu, Z. Hu, X. Wu, J. Liu, and Q. Chen. 2018. 'Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development', *Plant Cell Environ*, 41: 2109-27.

Qi, Zhaoming, Zhanguo Zhang, Zhongyu Wang, Jingyao Yu, Hongtao Qin, Xinrui Mao, Hongwei Jiang, Dawei Xin, Zhengong Yin, and Rongsheng Zhu. 2018. 'Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development', *Plant, cell & environment*, 41: 2109-27.

Research, Infinium Global. 2017. 'Amino Acid Market: Global Industry Analysis, Trends, Market Size & Forecast to 2023 ', Accessed June 9. ttps://www.researchandmarkets.com/research/xsjjhs/amino_acid.

Schaefer, Robert J, Jean-Michel Michno, Joseph Jeffers, Owen Hoekenga, Brian Dilkes, Ivan Baxter, and Chad L Myers. 2018. 'Integrating coexpression networks with GWAS to prioritize causal genes in maize', *The Plant Cell*, 30: 2922-42.

Schmidt, Monica A., W. Brad Barbazuk, Michael Sandford, Greg May, Zhihong Song, Wenxu Zhou, Basil J. Nikolau, and Eliot M. Herman. 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*, 156: 330-45.

Scossa, F., D. Laudencia-Chingcuanco, O. D. Anderson, W. H. Vensel, D. Lafiandra, R. D'Ovidio, and S. Masci. 2008. 'Comparative proteomic and transcriptional profiling of a bread wheat cultivar and its derived transgenic line overexpressing a low molecular weight glutenin subunit gene in the endosperm', *Proteomics*, 8: 2948-66.

Shaul, Orit, and Gad Galili. 1992. 'Increased lysine synthesis in tobacco plants that express high levels of bacterial dihydrodipicolinate synthase in their chloroplasts', *The Plant Journal*, 2: 203-09.

Slaten, M. L., A. Yobi, C. Bagaza, Y. O. Chan, V. Shrestha, S. Holden, E. Katz, C. Kanstrup, A. E. Lipka, D. J. Kliebenstein, H. H. Nour-Eldin, and R. Angelovici. 2020. 'mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis', *Plant Physiol*, 183: 483-500.

Tabe, Linda, and TJV Higgins. 1998. 'Engineering plant protein composition for improved nutrition', *Trends in plant science*, 3: 282-86.

Tabe, Linda M, and Michel Droux. 2002. 'Limits to sulfur accumulation in transgenic lupin seeds expressing a foreign sulfur-rich protein', *Plant Physiology*, 128: 1137-48.

Tan-Wilson, A. L., and K. A. Wilson. 2012. 'Mobilization of seed protein reserves', *Physiol Plant*, 145: 140-53.

Torrent, Margarita, Iñaki Alvarez, M Isabel Geli, Ionara Dalcol, and Dolors Ludevid. 1997. 'Lysine-rich modified γ-zeins accumulate in protein bodies of transiently transformed maize endosperms', *Plant molecular biology*, 34: 139-49.

Vaughn, Justin N, Randall L Nelson, Qijian Song, Perry B Cregan, and Zenglu Li. 2014. 'The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations', *G3: Genes, Genomes, Genetics*, 4: 2283-94.

Wenefrida, Ida, Herry S Utomo, Sterling B Blanche, and Steve D Linscombe. 2009. 'Enhancing essential amino acids and health benefit components in grain crops for improved nutritional values', *Recent patents on DNA & gene sequences*, 3: 219-25.

WHO. 'Global Database on Child Growth and Malnutrition ', Accessed June 9. http://www.who.int/ nutgrowthdb/about/introduction/en/.

Withana-Gamage, Thushan S, Dwayne D Hegedus, Xiao Qiu, Peiqiang Yu, Tim May, Derek Lydiate, and Janitha PD Wanasundara. 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: 901-12.

Wu, Si, Saleh Alseekh, Álvaro Cuadros-Inostroza, Corina M Fusari, Marek Mutwil, Rik Kooke, Joost B Keurentjes, Alisdair R Fernie, Lothar Willmitzer, and Yariv Brotman. 2016. 'Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in Arabidopsis thaliana', *Plos genetics*, 12: e1006363.

Wu, Yongrui, and Joachim Messing. 2014. 'Proteome balancing of the maize seed for higher nutritional value', *Frontiers in plant science*, 5: 240.

Wu, Yongrui, Wenqin Wang, and Joachim Messing. 2012. 'Balancing of sulfur storage in maize seed', *BMC Plant Biology*, 12: 77.

Yao, Min, Mei Guan, Zhenqian Zhang, Qiuping Zhang, Yixin Cui, Hao Chen, Wei Liu, Habib U Jan, Kai P Voss-Fels, and Christian R Werner. 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.

Zhang, H., M. L. Wang, R. Schaefer, P. Dang, T. Jiang, and C. Chen. 2019a. 'GWAS and Coexpression Network Reveal Ionomic Variation in Cultivated Peanut', *J Agric Food Chem*, 67: 12026-36.

Zhang, Hui, Ming Li Wang, Robert Schaefer, Phat Dang, Tao Jiang, and Charles Chen. 2019b. 'GWAS and coexpression network reveal ionomic variation in cultivated peanut', *Journal of Agricultural and Food Chemistry*, 67: 12026-36.

Zhu, Xiaohong, and Gad Galili. 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: 845-53.

CHAPTER 2: mGWAS UNCOVERS GLN-GLUCOSINOLATE SEED-SPECIFIC INTERACTIONS AND ITS ROLE IN METABOLIC HOMEOSTASIS

Marianne L Slaten[1], Abou Yobi[1], Clement Bagaza[1], Yen On Chan[1], Vivek Shrestha[1], Samuel Holden[1], Ella Katz[2], Christa Kanstrup[3], Alexander E Lipka[4], Daniel J Kliebenstein[2], Hussam Hassan Nour-Eldin[3], and Ruthie Angelovici[1]

1 Division of Biological Sciences, Interdisciplinary Plant Group, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA

2 Department of Plant Sciences, University of California Davis, Davis, California 95616, USA

3 DynaMo Center, Copenhagen Plant Science Centre, Department of Plant and Environmental Sciences, University of Copenhagen, 1871 Frederiksberg, C, Denmark

4 Department of Crop Sciences, University of Illinois, Urbana Illinois, 61801 , USA

Slaten, M. L., Yobi, A., Bagaza, C., Chan, Y. O., Shrestha, V., Holden, S., ... & Nour-Eldin, H. H. (2020). mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis. Plant Physiology, 183(2), 483-500.

2.1 ABSTRACT

Gln is a key player in plant metabolism. It is one of the major free amino acids that is transported into the developing seed and is central for nitrogen metabolism. However, Gln natural variation and its regulation and interaction with other metabolic processes in seeds remain poorly understood. To investigate the latter, I performed a metabolic genome-wide association study (mGWAS) of Gln-related traits measured from the dry seeds of the *Arabidopsis* (*Arabidopsis thaliana*) diversity panel using all potential ratios between Gln and the other members of the Glu family as traits. This semi-combinatorial approach yielded multiple candidate genes that, upon further analysis, revealed an unexpected association between the aliphatic glucosinolates (GLS) and the Gln-related traits. This finding was confirmed by an independent quantitative trait loci mapping and statistical analysis of the relationships between the Gln-related traits and the presence of specific GLS in seeds. Moreover, an analysis of *Arabidopsis* mutants lacking GLS showed an extensive seed-specific impact on Gln levels and composition that manifested early in seed development. The elimination of GLS in seeds was associated with a large effect on seed nitrogen and sulfur homeostasis, which conceivably led to the Gln response. This finding indicates that both Gln and GLS play key roles in shaping the seed metabolic homeostasis. It also implies that select secondary metabolites might have key functions in primary seed metabolism. Finally, this study shows that an mGWAS performed on dry seeds can uncover key metabolic interactions that occur early in seed development.

2.2 INTRODUCTION

Gln is a free amino acid (FAA) that belongs to the Glu family, which also includes Glu, gamma-aminobutyric acid (GABA), Pro, and Arg (Majumdar et al., 2016; Okumoto et al., 2016; Skokut et al., 1978). This amino acid family plays a key role in plant cell core metabolism by providing an entry point for inorganic nitrogen. Briefly, ammonium derived from nitrate or absorbed directly from the soil can be assimilated into Gln via the Gln synthase (GS)/Gln oxoglutarate aminotransferase (GOGAT) cycle (Lea & Miflin, 1974). GS/GOGAT is the primary nitrogen assimilation pathway in plants (Lea, 1998) and is involved in the remobilization of nitrogenous compounds and the assimilation of large amounts of ammonium generated by photorespiration in C3 plants (Foyer et al., 2009).

Gln plays an important role in seed metabolism; as one of the main nitrogen carriers, it is transported via the xylem and phloem to sink tissues, including developing seeds (Besnard et al., 2016; Zhang et al., 2010; Zhang et al., 2015). A study of maturing *Brassica napus* seeds showed that embryos import nitrogen in the form of amino acids (mainly Gln and Ala) to synthesize other amino acids via transamination/deamination reactions and then incorporation into seed storage proteins (SSP) (Schwender et al., 2014). Consistently, studies in *Arabidopsis* (*Arabidopsis thaliana*) have shown that Gln levels are highly elevated before the onset of SSP synthesis (Baud et al., 2002; Aaron Fait et al., 2006) and then drop substantially during seed maturation (Aaron Fait et al., 2006). Even though the majority of seed Gln comes from transport, several Gln synthase isozymes are expressed during seed development in the micropillar, chalaza embryo, and seed coat, which suggests that Gln is also actively synthesized in seeds (Winter et al., 2007). The content of Gln in dry seeds, therefore, may be the result of a balance between its incorporation into SSP, active synthesis, and degradation. However, its composition may also reflect the environmental conditions encountered by the maternal plant. High levels of Gln have been reported in *Arabidopsis* plants facing sulfur deprivation (Nikiforova et al., 2006) and in tobacco plants grown under high nitrogen conditions (Geiger et al., 1999), whereas low levels of Gln have been reported in *Arabidopsis* seedlings grown under nitrate-deficit conditions (Scheible et al., 2004). Interestingly, extensive variation in free Gln content in dry *Arabidopsis* seeds has been reported across the various accessions belonging to the *Arabidopsis* diversity panel (Angelovici et al., 2017) but the genetic architecture regulating this trait remains poorly understood. Knowledge regarding the genes that underlie Gln levels, composition, and seed partitioning would shed light on its potential seed specific functions, its interaction with other biological processes, and its role in downstream metabolism.

In recent years, genome-wide association studies (GWAS) as well as quantitative trait loci (QTL) mapping experiments have facilitated the identification of many loci for both primary and secondary metabolites (Angelovici et al., 2017; Ruthie Angelovici et al., 2013; Eva KF Chan et al., 2011; Chen et al., 2014; Gonzalez-Jorge et al., 2013; Riedelsheimer et al., 2012; Verslues et al., 2014; A. M. Wentzell et al., 2007). In-depth analyses of these QTLs have facilitated the further discovery of key structural and regulatory genes that underlie the natural variation of metabolic traits and the identification of various cellular processes involved in metabolic homeostasis. Although GWAS and QTL mapping have been conducted on FAAs in both vegetative and seed tissues across several species, no major QTLs have been identified for Gln (Chen et al., 2014; Riedelsheimer et al., 2012; Wen et al., 2014). The lack of any identifiable loci implies that

Gln either has a complex genetic architecture or that these studies possibly utilized “underpowered” association panels, or both.

The use of metabolic ratios as traits in GWAS has been useful for dealing with several such calcitrant metabolites. The approach, which relies on biochemical pathways and/or represent relationships uncovered by a metabolic network correlation analysis, has yielded several significant associations even when the absolute levels of metabolites have not (Angelovici et al., 2017; Ruthie Angelovici et al., 2013; Gonzalez-Jorge et al., 2013; Alexander E Lipka et al., 2013; A. M. Wentzell et al., 2007). It has been postulated that metabolic ratios are less complex (since they only represent the metabolite partitioning within biochemical pathways) and, therefore, are more tractable in association mapping studies (Angelovici et al., 2017). Still, even this approach has failed to identify QTLs for Gln in dry seeds (Angelovici et al., 2017). A different approach is clearly needed to uncover the genetic architecture of Gln. Notably, the metabolic ratios used in previous studies do not represent all the potential ratios of Gln-related traits since they were based principally on a priori pathway information, which is often incomplete.

In theory, performing a metabolic genome-wide association study (mGWAS) on all possible Gln-related metabolic ratios would potentially resolve its genetic architecture. In practice, however, such an endeavor would be challenging given the enormous number of metabolic ratios that could be derived from the relationships between Gln and all 20 proteogenic amino acids. Therefore, as a point of departure from previous studies, I derived all possible metabolic ratios of Gln only to its proteogenic amino acid family members, thus theoretically representing all potential biologically relevant partitioning/relationship of Gln within the Glu family (Figure 1). By combining this approach with a Fixed and

Random model Circulating Probability Unification (FarmCPU), which uses fixed and random effect models for powerful and efficient GWAS studies (Liu et al., 2016), I uncovered many significant QTLs for various Gln-derived traits in dry seeds. More importantly, analysis of the candidate genes revealed a surprising enrichment for genes residing in the glucosinolate (GLS) biosynthesis pathway, suggesting a potential interplay between two metabolic pathways that are not known to be directly linked (Figure. 1). I validated this association by using an independent QTL mapping approach as well as by characterizing Gln and other FAAs in mutant plants that have a disrupted GLS composition and loading to the seeds. Data support an association between GLS natural diversity and Gln levels and composition in seeds, and also reveal that GLS loading to the seeds has a profound effect on seed nitrogen and sulfur homeostasis as well as Gln levels and composition. Results strongly suggest that an interaction between Gln and GLS plays a key role in seed metabolic homeostasis.

2.3 RESULTS

2.3.1 The Four Glu Family Members Vary in Abundance, Relative Composition, and Broad-Sense Heritability Across the Arabidopsis Diversity Panel

In a previous study, the Angelovici lab quantified and described the natural variation of 18 out of the 20 proteogenic FAAs (excluding Cys and Asn) measured from dry seeds of three biological repeats of a 313-accession Arabidopsis diversity panel (Angelovici et al., 2017; R. Angelovici et al., 2013). In the current study, I used that data to assess the natural variation among only the proteogenic FAAs in the Glu family (i.e. Glu, Pro, Gln, and Arg).

Analysis showed that the four Glu family members vary in abundance, relative composition, and broad-sense heritability (Supplemental Table S1A). Glu was the most abundant amino acid with a relative composition mean of 0.35, whereas Gln was the least abundant with a relative composition mean of 0.015. Relative composition is defined as the ratio of an individual amino acid to the sum of the 18 measured amino acids (e.g. Gln/Total, Glu/Total). Arg and Pro had a relative composition means (Arg/Total; Pro/ Total) of 0.04 and 0.017, respectively. Gln demonstrated moderate heritability (0.52) along with Pro and Glu (0.48 and 0.63, respectively), whereas Arg had the highest heritability (0.74). Interestingly, Gln had the largest relative SD, whereas Glu had the smallest despite its high abundance (61% and 23% relative standard deviation, respectively).

To evaluate the relationship between Gln and the other Glu family members, I performed a correlation based network analysis among the four FAAs and visualized the results using Cytoscape version 3.6.1 (Supplemental Figure S1). All correlations (r) were significant at a α 0.001 and ranged from 0.12 to 0.54. Gln was moderately correlated with Arg and Glu and weakly correlated with Pro, which had a significant but weak correlation with all Glu family members.

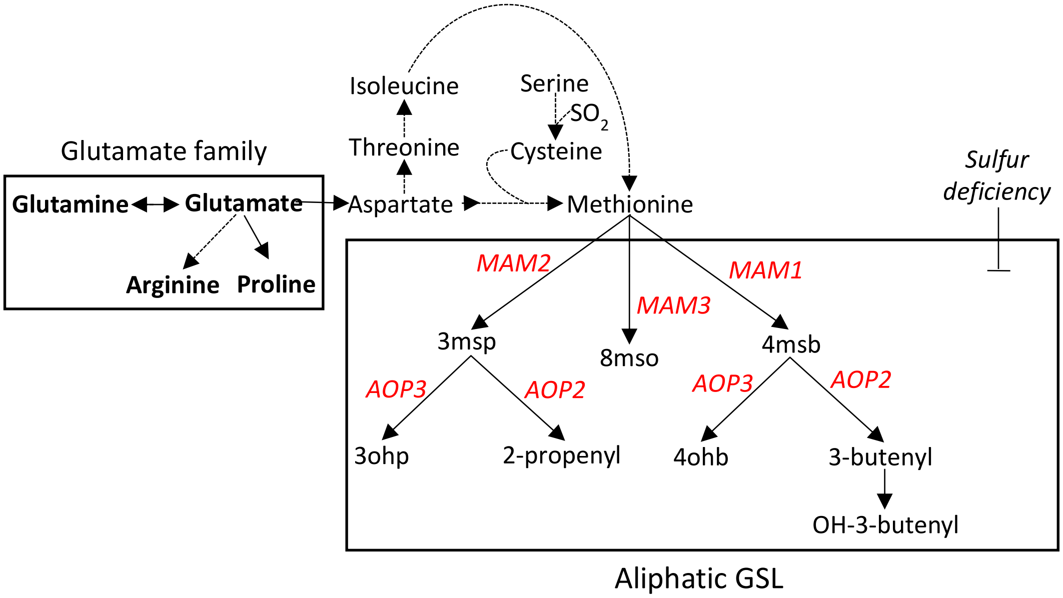


Figure 2. 1 . Simplified metabolic pathways of the Glu amino acid family and the aliphatic GLS. Genes that are within the genomic region of GS-ELONG or GS-AOP loci are represented in red and include MAM1, MAM2, and MAM3 as well as AOP2 and AOP3. These gene products are responsible for most of the GLS natural diversity. MAM1 is responsible for the production of C4 GLS 4-methylsulfinybutyl GLS (4msb); MAM2 (in the absence of MAM1) is required for the production of C3 GLS 3-methylsulfinypropyl GLS (3msp); MAM3 is responsible for the production of C8 GLS 8-methylsulfinyloctyl (8mso). C3 GLS 3msp can be further converted to 3-hydroxypropyl GLS (3ohp) by AOP3 or 2-propenyl by AOP2. C4 GLS 4msb can be converted to 4-hydrozybutyl GLS (4ohb) by AOP3 or 3-butenyl and then OH-3-butenyl by AOP2.

2.3.2 *mGWAS Identified Significant Single Nucleotide Polymorphism-Trait Associations for Six Gln-Related Traits*

In a previous study, no significant associations were identified when seed Gln traits or any Gln-related traits derived from a priori knowledge of the Glu metabolic pathway or correlation-based network analysis were used for the mGWAS (Angelovici et al., 2017). Therefore, I took a slightly different approach in this study by using all possible Gln metabolic ratios that could be derived from Gln relationships with the other members of the Glu family. The various relationships were represented by calculating all the possible ratios in which Gln is the numerator and is divided by a sum of every combination of the four Glu family members, including Gln itself [i.e. Gln/(Gln|Arg|Pro|Gu) | = (and /or)]. This is considered a semi-combinatorial approach since it relies on both a priori knowledge of the Glu family as well as all the possible combinations of the Glu family FAAs in the denominator. The traits and their corresponding means, ranges, and broad-sense heritability scores are listed in Supplemental Table S1B. For simplicity, I used a one letter code in trait representations. The sum of the FAA in the denominator of each trait is represented by a string of one letter codes. For example, Q/EP is Gln divided by the sum of Glu and Pro. This approach yielded 16 Gln-related traits: 14 ratio-based traits (Supplemental Table S1B), one free Gln absolute level, and the Gln relative composition (Gln/Total; Supplemental Table S1A). Of all these 16 traits, Q/QP had the highest heritability (0.53) and Q/RP had the lowest (0.35). In general, the derived traits had low to moderate heritability.

I used the FarmCPU package in R (version 1.02; Liu et al., 2016) to perform an mGWAS on the 16 Gln-related traits. Because FarmCPU may be prone to a type I error, I chose to use the more conservative Bonferroni multiple testing correction procedure instead of the Benjamini and Hochberg (1995) false discovery rate controlling procedure. I also considered single nucleotide polymorphism (SNP)-trait associations significant only at an a α 0.01 Bonferroni correction level. At this significance threshold, I identified 21 SNP–trait associations for six traits: Q/P, Q/R, Q/QP, Q/RP, Q/RQ, and Q/RQP (Figure 2; Supplemental Dataset S1); only 16 SNPs were identified from the 21 signals. None of the six traits included Glu in their denominator but did include either Arg or Pro, or both. The heritability of these six traits ranged from low to moderate (0.35–0.53; Supplemental Table S1B). No significant associations were observed on chromosome 1. One was observed on chromosome 2, and three on chromosome 3. The majority of significant SNPs were identified on either chromosome 4 or 5 (Figure 2; Supplemental Dataset S1). The five SNPs with the lowest P values were located on chromosomes 4 or 5 (Table 1); three of these SNPs fell within a gene, whereas the remaining two were located in a transposable element and an intragenic region. The three genes are annotated as encoding Brassinosteroid suppressor 1 (BSU1), a MATE efflux family protein, and methylthioalkylmalate synthase 1 (MAM1).

2.3.3 *Genes Within Haploblocks Spanning Significant SNPs are Enriched for Glucosinolate Biosynthetic Process*

I compiled a candidate gene list based first on genes that contain a significant SNP. I then expanded the list to include those genes that are in strong linkage disequilibrium (LD; defined as regions with nonrandom associations calculated using a 95% confidence bounds on D prime) with the significant SNPs identified by mGWAS, since significant SNPs identified by GWAS may tag causal variants in neighboring genes that are in LD (Atwell et al., 2010). To that end, I identified haploblocks that spanned the 16 SNPs using Haploview version 4.2 (See "Materials and Methods") (Barrett et al., 2004) and considered all spanned genes as candidates. If a haploblock was not identified for a given SNP and did not fall within a gene, then the gene directly upstream or downstream was recorded. Overall, 27 unique genes were found. The entire list of genes associated with all 16 SNPs is summarized in Supplemental Table S2A.

Next, agriGO was used (http://bioinfo.cau.edu.cn/ agriGO/) to perform a Gene Ontology (GO) enrichment analysis of the 27 genes. All genes identified across the six traits were analyzed since, collectively, they represent the potential genetic architecture of the Gln partition within the Glu family and its relationships to the other members. The analysis revealed a significant enrichment for the following terms: secondary metabolic process, carbohydrate metabolic process, sulfur metabolic process, S-glycoside biosynthetic process, and glucosinolates biosynthetic process (Supplemental Table S2B).

All the significant enrichment terms resulted from three genes: MAM1 (AT5G23010), AOP1 (AT4G03070), and AOP3 (AT4G03050), all of which are annotated as involved in the biosynthesis of aliphatic GLS. Notably, one of the top five significant SNPs fell within MAM1 (Q/P; Table 1). AOP1 was associated with traits Q/RQ and Q/RQP, and AOP3 was associated with trait Q/RQ (Figure 2; Supplemental Dataset S1). Although these genes are located in three different haploblocks, AOP1 and AOP3 are in very close proximity within the genome; the end of AOP3 and the beginning of AOP1 are 11,831 bp apart (Figure 3). The three genes are located in two well-characterized QTLs, GS-ELONG and GS-AOP (Figs. 3 and 4). The GS-ELONG locus controls variation in the side chain length of aliphatic GLS and is characterized by three genes: MAM1, MAM2, and MAM3 (previously MAM-L) (Kroymann et al., 2003; Kroymann et al., 2001). GS-AOP is the collective name of two tightly linked loci, GS-ALK and GS-OHP, and controls GLS side chain modifications (Kliebenstein et al., 2001). The GS-AOP locus represents the branching point in the biosynthesis of aliphatic GLS that includes two 2-oxoglutarate dependent dioxygenases: AOP2, localized in the GS-ALK locus, and AOP3, localized in the GS-OHP locus. The presence/absence of genes in the GS-AOP and GS-ELONG loci account for much of the natural variation in aliphatic GLS profiles in *Arabidopsis* (Figure 1). Thus, despite having significant SNPs directly associated with MAM1, AOP1, and AOP3, because of the high degree of LD in these regions, MAM2, MAM3, and AOP2 are also putative genes of interest.

Next to determine whether the three significant SNPs (i.e. S127050, S127076, and S175365) identified in the two GLS-related QTLs tagged the additional GLS genes in the GS-ELONG and GS-AOP regions. To that end, pairwise LD analysis was performed between the three identified SNPs and the SNPs 65 kb to either side of the first and last MAM or AOP genes in the GS-ELONG and GS-AOP regions (i.e. flanking the regions), respectively (Supplemental Figures S2 and S3). SNP S127076, which resides within the BSU1 gene but is located within the haploblock containing AOP1, is in high LD with AOP1 (S127071 and S127075; r^2 5 0.934 and 0.934) as well as with the SNPs residing in both AOP2 (S127058; r^2 5 0.918) and AOP3 (S127048, S127050, and S127050; r^2 5 0.902, 0.918, and 0.918, respectively). The high LD with neighboring SNPs suggests that this SNP may tag a causal variation in one or both of these AOP genes (Supplemental Figure

S2A). Similarly, SNP S127050, which resides in the same haploblock as AOP3, is in perfect LD with a SNP from AOP2 (S127058; r^2=1) and in high LD with SNPs in AOP1 (S127071, S127075, and S127076; r^2=0.983, 0.983, and 0.918, respectively), which suggests that this SNP may tag the additional AOP genes in the region (Supplemental Figure S2B). Finally, SNP S175365, which resides in the same haploblock as MAM1, is in strong to moderate LD with SNPs associated with MAM2 (S175355; r^2=0.908) and MAM3 (S175394; r^2=0.649; Supplemental Figure S3).

Overall, six genes involved in aliphatic GLS biosynthesis were identified as having moderate (>0.5) to strong (>0.8) LD with three of significant SNPs in the region. It is likely that either one or an allelic combination of all six genes contributes to the natural variation of free Gln and its related traits in dry seeds.

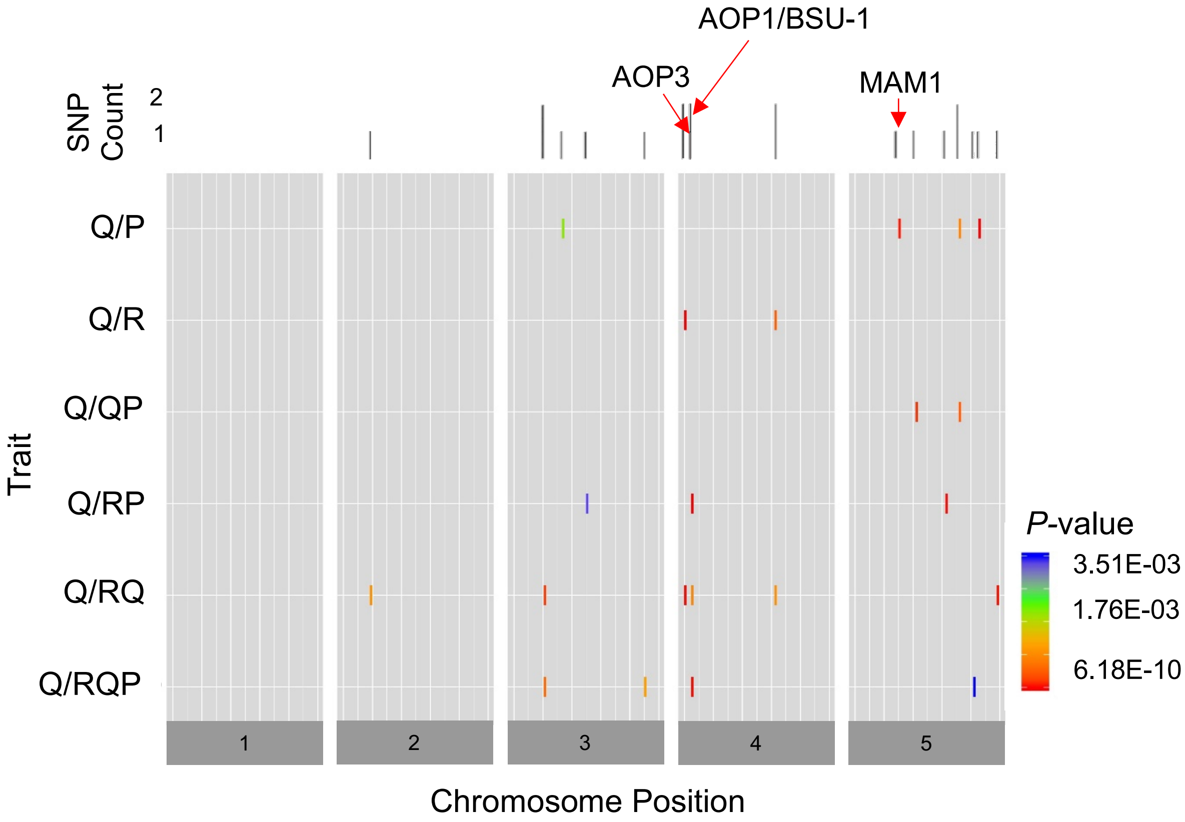


Figure 2. 2. The chromosomal distribution of the 21 significant SNP-trait associations identified by the mGWAS. A total of 21 significant SNPs-trait associations were identified at an α = 0.01 Bonferroni for six traits. SNPs are represented by short lines that are color-coded based on their P-value. The histogram on top of the heatmap illustrates the number of occurrences of each SNP identified across all traits. Arrows indicate SNPs located within a gene or haploblock of interest. An asterisk (*) designates traits that have a significant association with a SNP in a GLS gene or in high LD with GLS genes. Q, glutamine; R, arginine; P, proline.

Table 2. 1. Top five SNP-trait associations with the lowest p-values. Traits, SNP chromosomal position, p-value, relevant haploblock coordinates, the genes that contain significant SNPs or are within a haploblock containing a significant SNP, and their annotation are summarized for each SNP. An asterisk (*) designates a gene containing the SNP. If a SNP is intergenic and falls within a haploblock with additional genes, the additional genes are listed. If the SNP is intergenic and does not fall within a haploblock, the genes directly upstream (L) and downstream (R) are recorded.

SNP	Trait	Chr	Position	*p-value*	Haploblock range	Genes within haploblock	Functional annotation
S204486	Q/P	5	21544586	6.18E-15	None	AT5G53135*	transposable element
						AT5G53140	protein phosphatase 2C family protein
S127076	Q/RP	4	1360042	6.17E-14	1356197-1364333	AT4G03063	Pseudogene of AT4G03070; *AOP1* (2-oxoglutarate-dependent dioxygenase 1.1)
	Q/RQP					AT4G03070	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein (*AOP1*)
						AT4G03080*	*BRI1* suppressor 1 (*BSU1*)-like 1
S124151	Q/R	4	152626	5.32E-13	152604-152640	AT4G00350*	MATE efflux family protein
	Q/RQ						
S190878	Q/RP	5	15830557	5.22E-11	15829858-15830557	AT5G39530 (L)	hypothetical protein (DUF1997)
						AT5G39532 (R)	pseudogene of Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein (computational description)
S175365	Q/P	5	7704845	3.79E-10	7703584-7705070	AT5G23010*	methylthioalkylmalate synthase 1 (*MAM1*)

2.3.4 *QTL Analysis of the Bayreuth-0 and Shahdara Mapping Population Supports the GWAS Finding*

The finding of an association between Gln and GLS in dry seeds was surprising. Glucosinolates are not synthesized in seeds but rather are transported to the seed from the maternal plant (Magrath & Mithen, 1993). Therefore, to independently confirm results from the mGWAS and to further support the association between Gln and the two GLS-related QTLs, biparental QTL mapping using the Bayreuth-0 (Bay) and Shahdara (Sha) recombinant inbred population (Loudet et al., 2002) was performed. Previous work has shown that Bay and Sha segregate at the GS-ELONG and GS-AOP loci and have an epistatic relationship (Kliebenstein et al., 2007; Kliebenstein et al., 2001; Kroymann et al., 2003; Textor et al., 2004; A. M. Wentzell et al., 2007). I hypothesized that if these GLS-related QTLs are indeed responsible for the natural variation of Gln in dry seeds, then the Bay x Sha mapping population should recapitulate the QTL for the Gln-related traits.

To test this hypothesis, the FAA quantifications from 158 recombinant inbred lines of the Bay 3 Sha population were used, as described previously (Angelovici et al., 2017; R. Angelovici et al., 2013), and performed a QTL analysis of 16 Gln-related traits using Multiple QTL Mapping (MQM) in the R/qtl2 package in R (Arends et al., 2010). This approach yielded a total of 25 QTLs for eight traits (for the full list see Supplemental Dataset S2). Six traits had significant log of the odds (LOD) maxima on chromosome 5 at marker MSAT5.14 (position 7498509 bp): Q/RQ, Q/RQP, Q/R, Q/RP, Q/QP, and Q/P. The supporting interval overlapped with the GS-ELONG locus (Table 2). Both the highest percent of total phenotypic variation and the highest LOD were observed for Q/QP and

Q/P. These two traits also had a LOD maxima on chromosome 4 at marker MSAT4.43 with supporting intervals spanning the GS-AOP locus.

Interaction between the two QTLs has been observed previously in GLS traits (Kliebenstein et al., 2007; Kliebenstein et al., 2001). Therefore, I tested whether interactions between the two loci existed for Gln-related traits. Visual inspection of the interaction plots between markers MSAT4.43 and MSAT5.14 clearly indicated interaction between these markers that seem to heavily influence the Q/QP and Q/P trait means (Supplemental Figure S4).

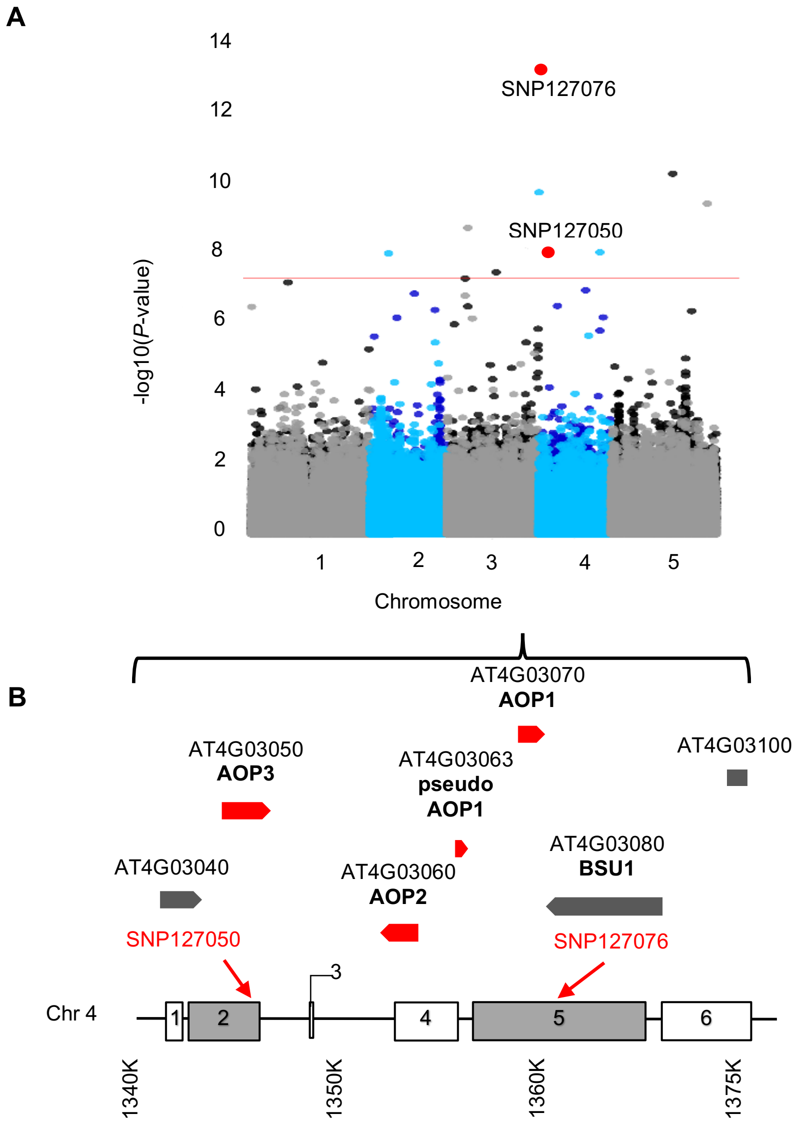


Figure 2. 3. mGWAS of traits Q/RP and Q/RQ. These two representative traits have significant associations with SNPs in LD with genes AOP1 and AOP3, respectively. **A,** Scatterplots for Q/RP (dark blue and black) and Q/RQ (light blue and grey) show significant associations among multiple SNPs, including S127076 and S127050, respectively (red), that reside in haploblocks that contain GLS biosynthesis genes. Negative log10-transformed p-values are plotted against the genomic physical position. The red line represents the 1% Bonferroni cutoff. All SNPs above the red line are significantly associated with other SNPs at that level. **B**, A graphical representation of genes and haploblocks within the genomic region (Chr4: 1340000–1375000 bp) spanning SNPs S127076 and S127050 (red arrows). This region also spans the GS-AOP QTL. S127050 falls within haploblock 2, which spans AOP3. S127076 falls within haploblock 5, which spans BSU1 and AOP1. Genes are represented by wide grey arrow unless they are associated with GLS biosynthesis and marked in red. Haploblocks in the region are represented by numbered boxes and shaded grey if they contain a SNP of interest.

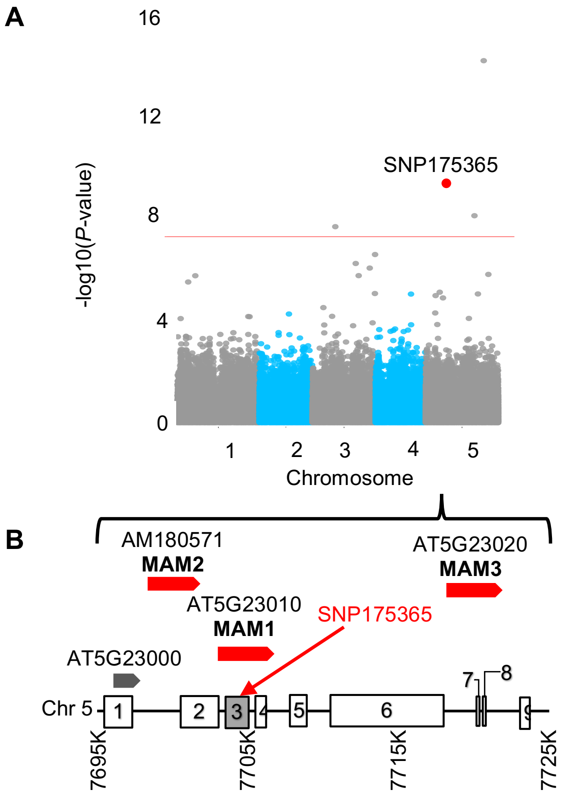


Figure 2. 4. mGWAS of Q/P. **A**, A scatterplot for Q/P shows the significant associations among several SNPs including S175365 (marked in red). Negative log10-transformed p-values are plotted against the genomic physical position. The red line represents the 1% Bonferroni cutoff. All SNPs above the red line are significantly associated with SNPs at that level. **B**, A graphical representation of genes and haploblocks within the genomic region spanning SNP S175365 (red arrow) and GLS genes in close proximity (Chr4:7695000–7725000 bp). This region spans the GS-ELONG QTL. Genes are represented by wide grey arrows unless they are associated with GLS biosynthesis and marked in red. Haploblocks in the region are represented by numbered boxes and shaded grey if they contain a SNP of interest.

2.3.5 *The Presence or Absence of Specific GLS has a Significant Effect on the Levels of the Gln-Related Traits in Dry Seeds*

To further validate the association between GLS natural variation and the Gln-related traits, 133 accessions from the Arabidopsis diversity panel were grown and measured both FAA and GLS levels in the dry seeds (Supplemental Dataset S3). Next, I tested whether the presence or absence of one of the four GLS, which result from the different allelic combinations at the GS-ELONG and GS-AOP loci (Figure 1), was associated with high or low levels of the traits of interest (i.e. the 16 Gln-related traits analyzed in mGWAS). The four GLS analyzed for presence/absence were 3ohp (requiring the presence of MAM2 and AOP3), 2-propenyl (requiring the presence of MAM2 and AOP2), 4ohb (requiring the presence of MAM1 and AOP3), and 3butenyl/OH-3-butenyl (requiring the presence of MAM1 and AOP2). To evaluate this association, t-tests were performed on the levels of the Gln-related traits measured from accessions that either had a specific GLS chemotype (i.e. 3ohp or 4ohb) or completely lacked it (see "Materials and Methods" for more details regarding the statistical analysis). Results showed that Gln absolute levels were significantly less in the presence of 2-propenyl (Supplemental Table S3). However, the presence/absence of both 3ohp and 4ohb had the most significant effect on the Gln traits. The presence of 3ohp had a negative effect on most of the Gln-related ratios and had a positive effect on the absolute levels of Arg, Glu, and Pro. By contrast, the presence of 4ohb had the opposite effect on most of the Gln-related traits in addition to the absolute levels of Glu and Pro (Supplemental Table S3). Taken collectively, these results both confirm that GLS diversity supports the association between these two pathways.

2.3.6 *FAA Characterization of Mutants in GLS Genes Present in the GS-ELONG and GS-AOP Showed Only Small Effects on Gln-Related Traits in the Col-0 Background*

A transgenic approach was performed to further confirm the association between aliphatic GLS and Gln content in dry *Arabidopsis* seeds. Null and overexpression (OX) mutants were obtained of the six relevant genes located in the GS-ELONG or GS-AOP locus and involved in aliphatic GLS biosynthesis. All plants were grown to maturity, and their dry seeds harvested and analyzed for FAA content and composition. Also obtained and quantified was the dry seed FAA content of a bsu1 null mutant, which lacks the BSU1 genes that contain the significant SNP (i.e. S127076) identified for traits Q/RP and Q/RQP (Figure 4; Table 1). The T-DNA insertion lines were ordered from the SALK and WISC T-DNA collections and included insertions in the AT4G03070 (*aop1*), AT4G03050 (*aop3*), AT5G23020 (*mam3*), and AT4G03080 (*bsu1*) genes. The T-DNA insertion locations are summarized in Supplemental Figure S5. Null homozygous mutants were isolated and confirmed by the absence of the full transcript in a tissue of high expression (Supplemental Figures S5 and S6). Based on the eFP browser expression data (Schmid et al., 2005; Winter et al., 2007), AOP1 expression was evaluated in imbibed seeds, AOP3 was evaluated in young siliques, MAM1 and MAM3 were evaluated in seedlings, and BSU1 was evaluated in leaves. The reverse transcription PCR primers used are listed in Supplemental Table S4. Interestingly, all genes, excluding AOP2, showed some transcript expression during seed development, despite a lack of GLS synthesis at the seed level. MAM2 does not exist in the Columbia-0 (Col-0) ecotype and does not have any publicly available expression profiles.

In addition to null mutants, mutants with altered GLS composition in the Col-0 background were also obtained. These mutants included *gsm*1, which accumulates C3 GLS and has large reductions in 4-methyl sulfinylbutyl and 6-methylsulfinyl glucosinolates (Haughn et al., 1991; Kroymann et al., 2001). Because the Col-0 accession does not contain MAM2 and has a truncated nonfunctional AOP2 protein (Jensen et al., 2015; Kroymann et al., 2001; A. M. Wentzell et al., 2007), analysis also included a previously characterized *AOP2* overexpression mutant in the Col-0 background that accumulates alkene GLS (Burow et al., 2015; Rohr et al., 2006). Collectively, these mutants represent some of the potential GLS composition alterations that can occur in the Col-0 background. The ability of any single gene mutant to capture the diversity of GLS is limited because it arises from a complex allelic combination (Kliebenstein et al., 2001).

Dry seed FAA was quantified for each of these single gene mutants and the fold change (FC) was assessed, as compared with its respective wild-type control (Col-0 or Col-3), for 16 Gln-related traits (Supplemental Dataset S4A). Gln absolute levels in the *aop1*, *aop3*, and *AOP2-OX* mutants did not change significantly. An elevated amount of Arg in the *aop3* mutant led to reductions in two Gln-related traits, Q/R and Q/RQ (0.54 and 0.75 FC, respectively; Figure 5; Supplemental Table S5A; Supplemental Dataset S4B). In addition, Glu and Pro were reduced in the *AOP2-OX* mutant but did not lead to any significant changes in the Gln-related ratios (Figure 5B; Supplemental Table S5B). The *bsu1* mutant had significantly high levels of Arg and Glu (a 1.62 and 1.43 FC, respectively), but the levels of Gln and related ratios were unchanged (Figure 5; Supplemental Table S5B). The FAA quantifications of the AOP-related mutants showed that, in addition to minor alterations in the Glu family FAAs, few other FAAs changed significantly (Figure

5A; Supplemental Table S5B). The analysis of the MAM-related mutants showed that levels of Gln, Glu, and Pro were slightly elevated (a 1.39, 1.19, and 1.35 FC, respectively) in the *gsm1* mutant, which led to slight increases in nine traits Gln related ratios (Figure 5B; Supplemental Table S5). In sum, the single gene mutants showed only a small effect of the altered GLS composition on the Gln-related traits.

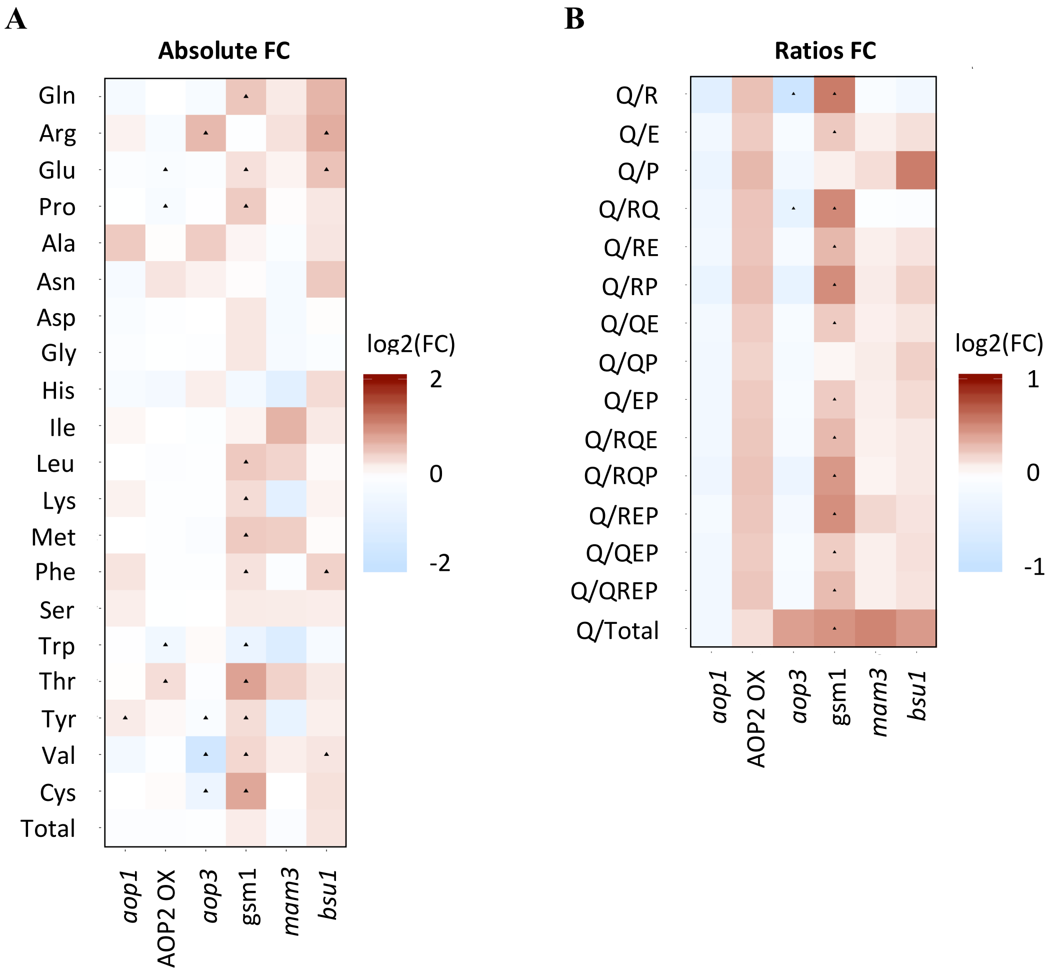


Figure 2. 5. Heatmap of the fold changes of FAA and Gln-related traits average in GLS null and OX mutants and the Col-0 ecotype. Measurements were obtained from dry seeds and used to calculate the log2 Fold Change (FC) with respect to the Col-0 ecotype for each FAA (**A**) and each Gln-related trait (**B**). Blue indicates that the FAA or Gln-related trait decreased relative to Col-0, whereas red indicates that the FAA or Gln-related trait increased relative to Col-0. A t-test was performed to determine the significance of the changes between each mutant and Col-0 (n = 3–4). Asterisks (*) represent significant difference between the mutant and the control at a α = 0.05 level. Q, glutamine; E, glutamate; R, arginine; P, proline.

Table 2. 2. QTL analysis of the (Gln)-related traits from the Bay x Sha mapping population. Only QTL that span the GS-ELONG and/or GS-AOP region are shown. Genotypic and phenotypic data from 158 recombinant inbred lines were analyzed using the R/qtl2 package to identify significant QTLs. An asterisk (*) indicates traits with significant SNP-trait associations in GWAS.

Trait	Chr	Marker name	Marker position (cM)	Marker position (Mbp)	Suporting interval (cM)	Supporting interval (Mbp)	Log of the odds	Phenotypic variance %	# GSL genes in supporting interval	
									Gene name	Gene location
*Q/RQ	5	MSAT5.14	26.6	7.499	17.7-45.4	4.67-13.96	3.49	9.67	AT5G23010 (*MAM1*)	7703041-7706973
									MAM2	
									AT5G23020 (*MAM3*)	7718118-7721866
									AT5G26000	9079452-9082384
*Q/RQP	5	MSAT5.14	26.6	7.499	17.7-41.8	4.67-12.11	7.82	20.38	AT5G23010 (*MAM1*)	7703041-7706973
									MAM2	
									AT5G23020 (*MAM3*)	7718118-7721866
									AT5G26000	9079452-9082384
Q/R	5	MSAT5.14	26.6	7.499	17.7-45.4	4.67-13.96	2.80	7.84	AT5G23010 (*MAM1*)	7703041-7706973
									MAM2	
									AT5G23020 (*MAM3*)	7718118-7721866
									AT5G26000	9079452-9082384
*Q/RP	5	MSAT5.14	26.6	7.499	17.7-41.8	4.67-12.11	5.97	15.97	AT5G23010 (*MAM1*)	7703041-7706973
									MAM2	
									AT5G23020 (*MAM3*)	7718118-7721866
									AT5G26000	9079452-9082384

Table 2.2 COTINUED

Q/QP	4	MSAT4.43	10.7	2.576	2-24.2	0.41-7.55	3.97	10.93	AT4G03050 (*AOP3*)	1343845-1346436
									AT4G03060 (*AOP2*)	1351688-1354096
									AT4G03063 (*AOP1*)	1355667-1356018
									AT4G03070 (*AOP1*)	1358267-1359698
									AT4G12030	7210807-7213376
	5	MSAT5.14	26.6	7.499	17.7- 41.8	4.67-12.11	15.79	36.89	AT5G23010 (*MAM1*)	7703041-7706973
									MAM2	
									AT5G23020 (*MAM3*)	7718118-7721866
									AT5G26000	9079452-9082384
*Q/P	4	MSAT4.43	10.7	2.576	2-15.8	0.41-5.63	5.88	15.75	AT4G03050 (*AOP3*)	1343845-1346436
									AT4G03060 (*AOP2*)	1351688-1354096
									AT4G03063 (*AOP1*)	1355667-1356018
									AT4G03070 (*AOP1*)	1358267-1359698
	5	MSAT5.14	26.6	7.499	17.7- 41.8	4.67-12.11	12.10	29.72	AT5G23010 (*MAM1*)	7703041-7706973
									MAM2	
									AT5G23020 (*MAM3*)	7718118-7721866
									AT5G26000	9079452-9082384

2.3.7 *Elimination of* **Aliphatic** *GLS Triggers a Strong Seed-Specific Increase in Free Gln*

To further characterize the association between aliphatic GLS and the Gln-related traits, absolute levels of each FAA in the dry seeds of three null mutants (*myb28/29*, *myb34/51*, and *grt1/2*) with altered GLS compositions and the Col-0 ecotype was also quantified. The log2 of the average FC, defined as the ratios between individual amino acid levels in the mutants and their levels in their respective controls, were calculated and used to create heat maps of the FAAs (Figure 6; Supplemental Dataset S4). The *myb28/29* double knockout mutant is a null mutant of two transcription factors that regulate the aliphatic GLS in Arabidopsis: *MYB28* (AT5G61420) and *MYB29* (AT5G07690). This double knockout eliminates all aliphatic GLS from the entire plant, including the seed (Sonderby et al., 2007). A double knockout of *GTR1* (AT3G47960) and *GTR2* (AT5G62680), resulting in the *gtr1/2* mutant, abolishes the transport of all GLS to the seeds (H. H. Nour-Eldin et al., 2012). Finally, a double knockout of the two transcription factors *MYB51* (AT1G18570) and *MYB34* (AT5G60890), resulting in the *myb34/51* mutant, eliminates the indole GLS from the entire plant (H. Frerigmann & T. Gigolashvili, 2014).

The FAA analysis revealed that Gln levels were significantly higher in the *myb28/29* and *gtr1/2* mutants, but not in the *myb34/51* mutant, as compared to Col-0 (Figure 6; Supplemental Table S5A; Supplemental Dataset S4A). In fact, Gln showed the most pronounced FC among all FAAs measured: a 97 FC in the *myb28/29* mutant and a 598 FC in the *gtr1/2* mutant (Figure 6; Supplemental Table S5A, B). In addition to Gln, three other Glu family members increased significantly in the *myb28/29* and *gtr1/2* mutants: a 35.1 and 64.5 FC for Arg, a 3.3 and 4.7 FC for Glu, and a 1.3 and 4 FC for Pro, respectively (Supplemental Table S5A, B). Alterations in these Glu family FAAs led to significant FC increases in all Gln-related ratios, ranging from a 1.5–1.9 FC in Q/RQ and a 76.3 and 150.7 FC in Q/P in the *myb28/29* and *gtr1/2* mutants, respectively (Figure 6B; Supplemental Table S5A). In the *myb28/29* and *gtr1/2* mutants, increases in Asn (10.40 and 9.87 FC, respectively) and His (8.78 and 47.28 FC, respectively) were also observed. Glu and Asp also showed a consistent elevation (~3–5 FC) in both mutants (Figure 6A; Supplemental Table S5B). The total sum of the FAAs (TFAA) measured also increased significantly in both *myb28/29* and *gtr1/2*, by 4.73 and 12.58, respectively (Supplemental Table S5B).

Since TFAA changed in both mutants, I also calculated the percent of each FAA to the sum of the TFAA measured (%FAA/TFAA, w/w) in all genotypes including Col-0 (Supplemental Dataset S4C; Supplemental Table S5C). In both mutants, the largest increase was in the relative composition of Gln, which increased from ~1% in Col-0 to 22.82% in the *myb28/29* mutant and to 53.10% in the *gtr1/2* mutant (Figure 6C; Supplemental Table S5C). Arg and His were the only other FAAs that consistently increased in both the *myb28/29* and *gtr1/2* mutants: from ~1% of the total FAA in Col-0 to 8.82% and 6.10%, respectively, for Arg and to 2.44% and 4.95%, respectively, for His. The relative compositions of the remaining FAAs were consistently lower in both mutants (excluding Asn, which showed opposite trends in the two mutants) (Figure 6C; Supplemental Table S5C). The largest decreases were in the two most abundant FAAs in the Col-0 seeds, Glu and Gly, which had relative abundances of 28.81% and 18.77% in Col-0, 19.94% and 10.65% in *myb28/29*, and 6.66% and 2.83% in *gtr1/2*, respectively (Figure 6C; Supplemental Table S5C).

Next, I tested whether a reduction in GLS (rather than its complete elimination) would result in significant alterations in Gln levels. The dry seed FAA levels from the *myb28* and *myb29* single mutants were quantified, which have approximately half the seed GLS as the Col-0 ecotype (Francisco et al., 2016). The *myb28* mutant had significant FCs only in Pro levels (a 1.23 FC increase) (Supplemental Table S5A, B). The *myb29* mutant, by contrast, showed minor but significant increases in both Gln absolute levels (1.55 FC) and relative composition (Gln/Total 1.26 FC) as well as FCs (1.7–1.47) in several Gln-related traits (i.e., Q/REP, Q/E, Q/P, Q/RE, Q/QE, Q/QP, Q/EP Q/RQE, Q/QEP, Q/RQEP) in the *myb29* mutant (Figure 6B; Supplemental Table S5A). Nevertheless, levels of Asp, Gly, Leu, and Phe were also elevated significantly in this mutant, with FCs of 1.23–1.42 (Figure 6A; Supplemental Table S5B). Collectively, this genetic analysis indicated to us that Gln levels were extensively altered in response to a complete absence of aliphatic GLS either in the plant or specifically in the seed.

To evaluate if the response was seed specific, I analyzed the FAA content in the rosette leaves and stems of the *myb28/29* and *gtr1/2* double mutants and the respective Col-0 control. Tissues were collected approximately 20 days after bolting in order to capture the metabolic steady state of the FAA in these tissues during seed setting and filling. Neither mutant had significant fold changes in Gln levels in either its leaves or stems (Supplemental Dataset S5; Supplemental Table S6). In contrast to the seeds, no elevation in TFAA (as explained above) in either mutant was found. The results support the genetic evidence that the elevated Gln levels in the mutant seeds are occurring at the seed level rather than resulting from specific increases in the maternal tissue.

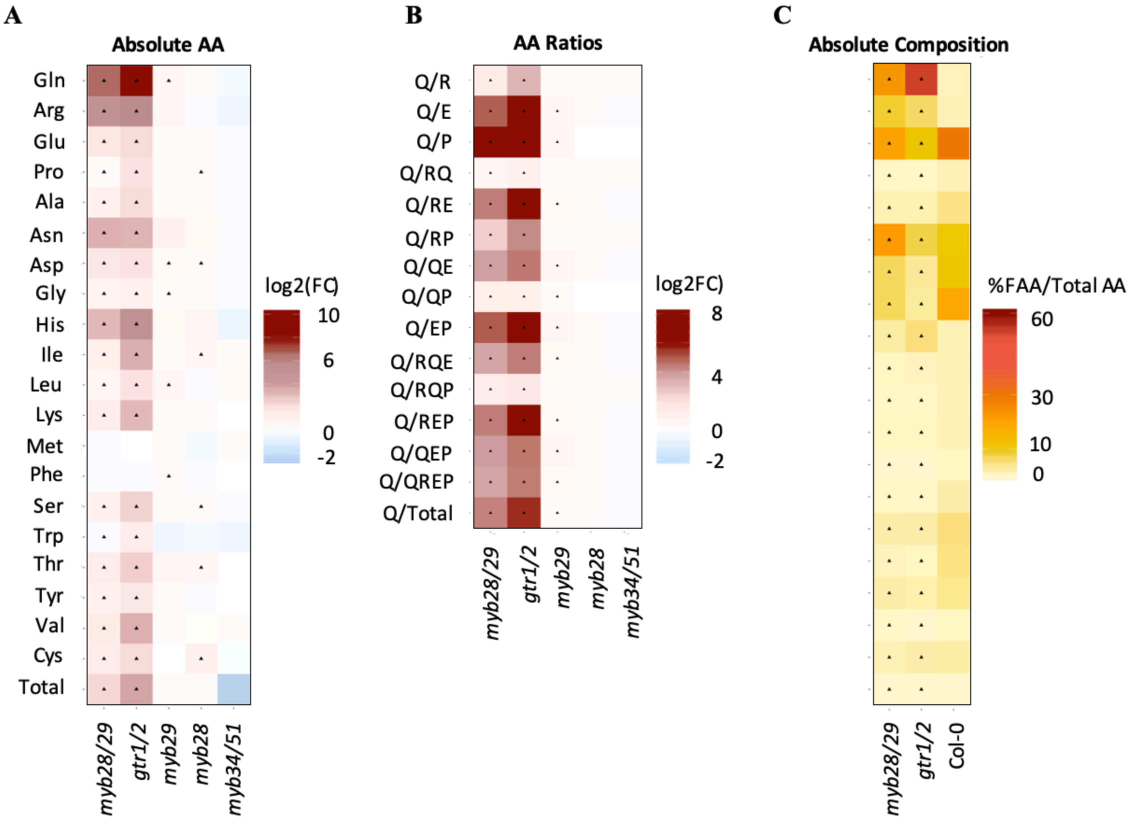


Figure 2. 6. Heatmap of the fold changes of FAA and Gln-related traits average in the GLS single and double knockout mutants and the Col-0 ecotype. **A and B,** Measurements were obtained from dry seeds and used to calculate the average log2 fold change (FC) with respect to the Col-0 ecotype for each FAA (**A**) and each Gln-related trait (**B**). Blue indicates that the FAA or Gln-related trait decreased relative to Col-0, whereas red indicates that the FAA or Gln-related trait increased relative to Col-0. **C**, A heatmap of the FAA composition of each amino acid (% FAA/Total AA) calculated per genotype using the FAA measurements from seeds. Red indicates that the amino acid represents a higher percentage of the total composition relative to Col-0, whereas yellow indicates that the amino acid represents a lower percentage of the total composition relative to Col-0. Asterisks (*) indicate a significant difference between the mutant and the Col-0 as deduced by t-test at an α = 0.05 level (n = 3–4). Q, glutamine; E, glutamate; R, arginine; P, proline.

2.3.8 *Gln Levels Are Elevated During Early Seed Maturation in Both the myb28/29 and the gtr1/2 Mutants.*

During seed maturation, FAAs (especially Gln) are incorporated into the SSPs especially during seed filling/maturation (A. Fait et al., 2006). Hence, it was assessed whether Gln levels are elevated during the early stages of seed development. To do this, I isolated developing seeds at 12, 14, 16, and 18 days after flowering (DAF) and at the dry seed stage from the *myb28/29* and *gtr1/2* mutants and the Col-0 ecotype and analyzed the FC in FAA levels across these time points (Supplemental Dataset S6). Analysis indicated that, as compared to the Col-0 control, the seeds from both mutants had substantial increases in Gln as early as 12 DAF (Figure 7; Supplemental Table S7). At 12 DAF, there was a 24 FC increase of Gln in the *myb28/29* mutant and a 37 FC increase in the *gtr1/2* mutant (Supplemental Table S7). Gln levels were higher across all the developmental time points in both mutants. Although Gln levels in all genotypes showed an overall reduction trend, the FC observed in the mutants continued to increase as the seed progressed to desiccation. (Figure 7A, B; Supplemental Table S7). Gln absolute levels at all time points exceeded the levels of any other amino acid (Supplemental Dataset S6).

Since the TFAA changed in both mutants, the changes in FAA relative composition as described above were also evaluated. The relative composition of Gln dropped from 9.5% (12 DAF) to ~1.11% (dry seed) in the Col-0 and dropped from ~54.1% (12 DAF) to 22.82% (dry seed) in the *myb28/29* mutant (Supplemental Table S7B). Surprisingly, the Gln content in the *gtr1/2* mutant remained between 54.53% and 61.40% throughout the entire seed maturation process, despite a drop in Gln absolute levels (Figure 7C; Supplemental Table S7B). Hence, Gln is only a minor amino acid in Col-0 but the most abundant one in the mutants. By contrast, Glu is most abundant in the seeds, and its levels increased from 21.3% (12 DAF) to 28.8% (dry seed) in the Col-0, remained constant at ~20% in the *myb28/29* mutant throughout development, and decreased from 13.9% (12 DAF) to 10.6% (dry seed) in the *gtr1/2* mutant (Supplemental Table S7B). Very pronounced changes were also recorded in the composition of Gly, which had a lower relative composition as compared to the Col-0 throughout seed development (Figure 7C; Supplemental Table S7). Notably, at all seed developmental stages, the FC never exceeded 2 for Gly or 6 for Glu (Supplemental Table S7A).

Collectively, these results show that compositional alteration to FAAs in the glucosinolate mutants occurs very early in seed maturation and persists in the dry seeds.

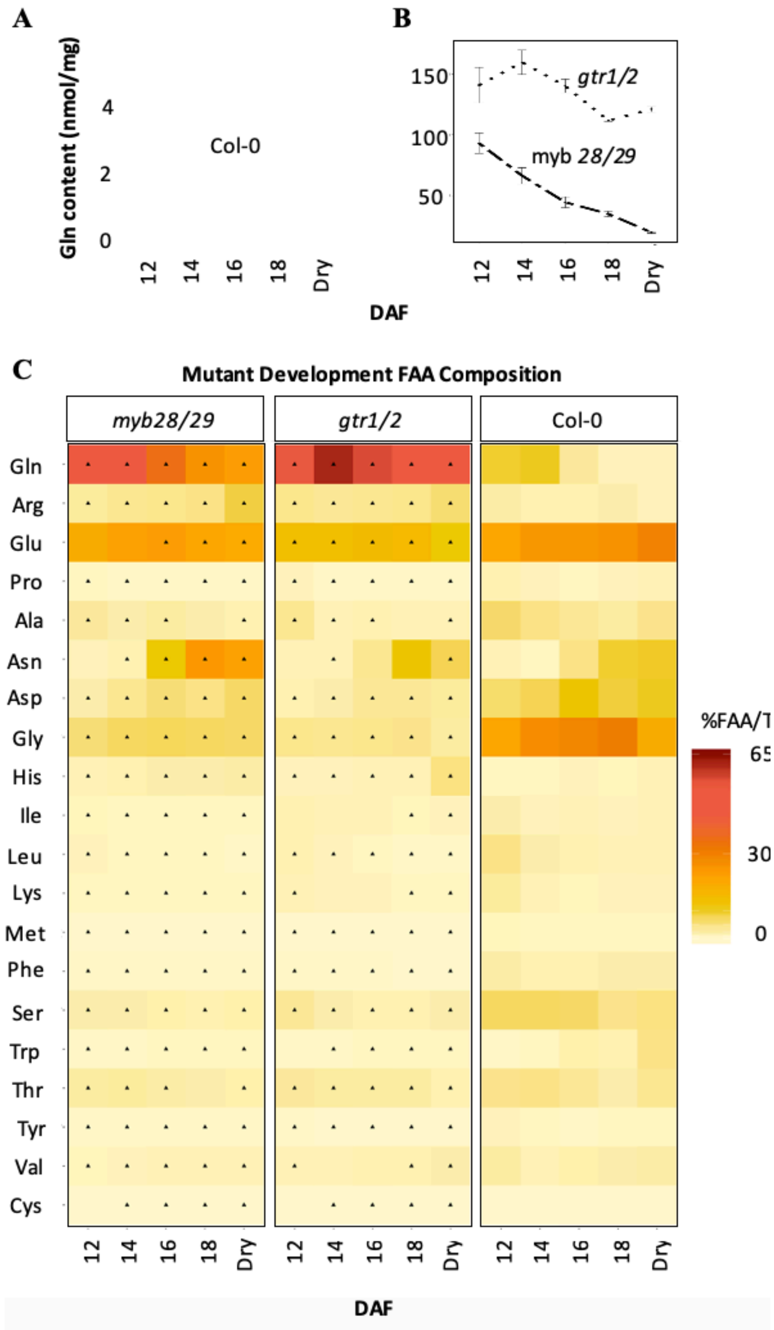


Figure 2. 7. Gln levels and the relative composition of free amino acids across maturation in the GLS double knockout mutants (myb28/29, grt1/2) and the Col-0 ecotype. Seeds were harvested at 12, 14, 16, and 18 days after flowering (DAF) and at full maturation (dry seed) and FAA levels and composition were analyzed across these developmental time points. **A**, Gln levels (nmol/mg) across seed maturation in Col-0. **B**, Gln levels in the double mutants. **C**, Composition analysis of the 20 FAAs across the developmental time-points for the three genotypes. Relative composition is presented as % of each FAA to TFAA in the seed. TFAA represents a sum of the 20 FAA measurements. Red indicates that the amino acid represents a higher percentage of the total composition relative to Col-0, whereas yellow indicates that the amino acid represents a lower percentage of the total composition relative to Col-0. Asterisks (*) indicate a significant difference between the mutant and Col-0 as deduced by t-test at an α = 0.05 level (n = 4, except for 18 DAF which is n = 2 for gtr1/2).

2.3.9 *Both Sulfur and Nitrogen Significantly Changed in Seeds that Lacked GLS*

GLS are high in nitrogen and sulfur compounds. A lack of GLS in seeds may cause a change in their homeostasis, which is known to have a substantial impact on Gln levels (Nikiforova et al., 2006; Nikiforova et al., 2005). To test this possibility, nitrogen, carbon, and sulfur in the *myb28/29* and *gtr1/2* mutants and in the Col-0 control were measured (Table 3).

As compared to Col-0, nitrogen was higher in both mutants (by 8% and 15%, respectively), sulfur was significantly lower (by 79% and 90%, respectively), and carbon was unaltered (Table 3). Finally, it was assessed whether the elevated levels of Gln and other FAAs reflected any changes in the levels or composition of proteins. To do this, the protein-bound amino acids (PBAA) in the dry seeds of the two mutants and in Col-0 was analyzed. The analysis revealed no significant or consistent alterations in PBAA levels (Supplemental Dataset S7; Supplemental Table S8).

Table 2. 3. Nitrogen, carbon, and sulfur absolute levels measured from the dry seeds of Arabidopsis myb28/29 and gtr1/2 double mutants and the Col-0 ecotype (n =5). The table lists the averages of the absolute levels, the percentage of each absolute level in the mutants relative to Col-0, the percentage of increase or decrease, and the significance of the difference after using a Duncan's Multiple Range Test, with different lower-case letters indicating significant differences at the α = 0.05 level.

Line	**Av. Absolute levels (μg/mg)**	**% of control**	**Difference**	***Duncan***
Col-0	34.15 ± 0.32	100	0	b
myb28/29	36.94 ± 0.48	108.17	8.17	a
gtr1/2	39.18 ± 0.19	114.73	14.73	a
Col-0	525.14 ± 4.16	100	0	a
myb28/29	519.1 ± 5.18	98.85	-1.15	a
gtr1/2	508.76 ± 2.94	96.88	-3.12	a
Col-0	5.37 ± 0.03	100	0	a
myb28/29	1.15 ± 0.01	21.43	-78.57	b
gtr1/2	0.54 ± 0.02	9.96	-90.04	c

2.4 DISCUSSION

Genome-wide association studies have successfully uncovered many genes involved in the natural variation and regulation of various metabolic traits, including FAAs in seeds (R. Angelovici et al., 2013; E. K. Chan et al., 2011; Diepenbrock et al., 2017; Alexander E. Lipka et al., 2013; Magrath, 1994; Parkin et al., 1994). Yet, none of these studies have identified any significant SNP associations with free Gln in dry seeds. The intractability of this trait would suggest that Gln has a highly complex genetic architecture. When faced with such complex metabolic traits, some researchers have enlisted metabolic ratios based on *a priori* knowledge or unbiased network analysis, an approach that has yielded additional QTLs that could not be retrieved using direct measurements of the absolute traits (Angelovici et al., 2017; R. Angelovici et al., 2013; Diepenbrock et al., 2017). Unfortunately for free Gln in seeds, neither absolute measurements nor specific metabolic ratios have resulted in significant associations.

In this study, I used a semi-combinatorial approach to formulate metabolic ratios as traits in a mGWAS. Unlike previous studies, this approach yielded several novel SNP-trait associations. Interestingly, I identified unique SNP-trait associations across the different Gln-related traits, suggesting a slightly different genetic architecture for each metabolic ratio (Figure 2; Supplemental Dataset 1). Since all the traits represent the Gln partition or a relationship to the other Glu family members, all the SNPs were treated as contributing to one genetic architecture of Gln metabolism. This collective analysis enabled us to compile a comprehensive candidate gene list that, upon further analysis, revealed a strong association between Gln and an unexpected metabolic pathway, the GLS biosynthesis. This approach could help elucidate the genetic basis of other complex metabolites and further reveal unexpected metabolic pathway associations.

2.4.1 *Unexpected Association Between the Gln-Related Traits and the Aliphatic GLS Natural Diversity is Supported by Multiple Independent Lines of Evidence*

The semi-combinatorial mGWAS analysis revealed that the natural variation of the Gln-related traits measured from dry seeds is strongly associated with natural variation of aliphatic GLS. Not only did I identify an enrichment of GLS biosynthesis genes in the collective candidate gene list, but also identified two aliphatic GLS biosynthetic genes in the top significant SNP-trait associations analysis (Table 1; Supplemental Table 2B). This association is surprising because GLS biosynthesis has three main steps (chain elongation of either methionine, branched chain, or aromatic amino acids; core structure formation; secondary modifications) (Kliebenstein et al., 2001), none of which involve Gln. In general, GLS are nitrogen- and sulfur-containing compounds that likely evolved from cyanogen glucosides but are largely limited to the Brassicales (Halkier & Gershenzon, 2006). Their breakdown products display a variety of biological activities explaining their defensive roles (Johnson et al., 2009). Although GLS accumulate to very high levels in seeds, they are synthesized in the vegetative tissue and transported from the maternal plant to the seed (Magrath & Mithen, 1993). Nevertheless, this study provides multiple lines of evidence confirming an association between the natural variation of Gln-related traits and the natural diversity of aliphatic GLS. Firstly, it is important to note that the three significant SNPs associated with aliphatic GLS fell within two well characterized QTLs: the *GS-ELONG* and the *GS-AOP* (Magrath, 1994). Previous studies have shown that the presence and absence of five genes within these QTLs account for much of the diversity in the aliphatic GLS profile in *Arabidopsis*. These genes are *MAM1–3*, *AOP2*, and *AOP3* (Halkier & Gershenzon, 2006). Pairwise LD analysis of the three significant SNPs identified in these two regions revealed that these SNPs are likely tagging all five genes within these two key QTLs (Supplemental Figure S2 and Supplemental Figure S3). Secondly, an independent QTL mapping of the Gln-related traits measured from the Bay/Sha mapping population (which segregates for these two key QTLs, (A. M. Wentzell et al., 2007) also identified significant associations of both *GS-ELONG* and *GS-AOP* loci with several Gln-related traits (Table 2; Supplemental Dataset 2). Lastly, the presence/absence of various chemotypes, arising from different allelic combinations of the *MAM* and *AOP* genes (Figure 1), resulted in significantly different levels in the Gln-related traits (Supplemental Table S3). GLS 3ohp and 4ohb, in particular, showed strong associations with the Gln-related traits and are among the most abundant class of GLS in seeds (Petersen et al., 2002; Velasco et al., 2008). In addition, the aliphatic GLS are the most abundant GLS in *Arabidopsis* seeds (Kliebenstein et al., 2001). Interestingly, their precise function in this tissue is unclear. Taken together, results show that, although unexpected, the pathway level association revealed by the mGWAS approach is strongly supported by multiple independent approaches.

2.4.2 *The Nature of the Association Between the Gln-Related Traits and the GLS Natural Diversity is Complex and Seed Specific*

The precise nature of the association between GLS and the Gln-related traits is unclear. The data indicate that the association is not simple. Analysis of known single gene mutants of the genes related to GLS in the *GS-ELONG* and *GS-AOP* regions in the Col-0 background (which lacks the expression of *AOP2* and *MAM2*) (Kroymann et al., 2001) showed relatively small changes in the Gln-related traits (Figure 5; Supplemental Table 5). This finding is perhaps not surprising since GLS diversity relies on the presence of a complex epistatic interaction network of different GLS QTLs (Burow et al., 2010), and the ability of a single gene elimination in a set genotypic background to capture all the potential allelic combinations is very limited. In addition, a reduction of about half of the aliphatic GLS through single mutations in either the *myb28* or *myb29* mutants (Francisco et al., 2016) did not result in any large effects on the Gln-related traits (Figure 6; Supplemental Table 5). However, the elimination of all GLS transported to the seeds in the *gtr1/2* double mutant or removal of the aliphatic GLS in the *myb28/29* from the entire plant had a profound effect on the composition of all FAAs and most prominently on Gln (Figure 6; Supplemental Table 5). These findings emphasize that the association between Gln and GLS relies on a complete elimination of specific GLS in the seed. This observation is further supported by statistical analysis of the association between levels of the Gln-related traits and the presence/absence of specific GLS in a natural population (Supplemental Table S3). More importantly, lack of FAA alteration in the stem and leaf measured from the double mutant clearly showed that the association between GLS and Gln is seed specific and is not the cause of a pleotropic effect that could arise from a lack of GLS in the mother plant or a direct interaction of the MYB genes with any Gln-related pathway genes (Supplemental Table S6). In line with the observation, a study of the perturbation of aliphatic GLS biosynthesis in *Arabidopsis* showed mild alteration in leaf FAA, including free Gln; in fact, the study found that Gln levels in leaves slightly decreased (Chen et al., 2012). Interestingly, FAA analysis performed during early seed maturation further indicated that the response of Gln to the lack of GLS, especially aliphatic, occurs early

(Figure 7; Supplemental Table 7). Overall, this early seed-specific interaction strongly suggests that both GLS and Gln have key functions in seed metabolic homeostasis that are not manifested in the vegetative tissues. Moreover, it also demonstrates that an mGWAS of FAA in dry seeds can reveal associations of biological processes taking place in early development.

2.4.3 ***The Association between Gln and GLS Is Likely Indirect and Induced by Alterations in the Seed Metabolic Homeostasis***

The molecular mechanism that underlies the interaction between GLS and Gln in the seeds is not clear. The Gln response appears to depend on the presence/absence of aliphatic GLS that is manifested in a specific tissue and is not dosage dependent. This suggests that the interaction is likely indirect and is potentially mediated through alteration of signaling/sensing pathways or other aspects of cell metabolism. Consistently, previous studies in *Arabidopsis* leaves have shown that perturbation of the aliphatic GLS alter several proteins and metabolites involved in various physiological processes, including photosynthesis, oxidative stress, hormone metabolism, and specific amino acids (Chen et al., 2012). It also has been shown in *Arabidopsis* specific that indole GLS activation products can interact with the conserved TIR auxin receptor to alter auxin sensitivity (Katz et al., 2015). Furthermore, exogenous application of a specific aliphatic GLS (3ohp) causes an alteration in root meristem growth in an array of plant lineages, even those that have never been reported to produce GLS (Malinovsky et al., 2017). These authors have established that this response is due to the interaction between GLS and the TOR pathway, which is a key primary metabolic sensor that controls growth and development and is conserved back to the last common eukaryotic ancestor (Henriques et al., 2014). These findings highlight the potential interactions of aliphatic GLS with primary metabolism and a conserved sensing mechanism. Consistent with these observations, data show that the presence of specific GLS compounds has a significant effect on the levels of the Gln-related ratios: 3ohp had a negative effect on most of the Gln-related ratios, whereas 4ohb had the opposite effect (Supplementary Table S3). These two GLS may possibly interact with distinct conserved metabolic regulatory pathways that affect Gln metabolism.

Data also indicate that the strong seed-specific association between the Gln-related traits and GLS in the seeds lacking aliphatic GLS (i.e., *myb28/29* and *gtr1/2*) may be induced due to substantial alteration in the overall cell metabolic homeostasis. Analysis of the carbon, nitrogen, and sulfur contents of the two double mutants lacking aliphatic GLS in seeds support this hypothesis. The results show that carbon remains relatively stable whereas both the nitrogen and sulfur homeostasis is severely altered: total sulfur is dramatically decreased and nitrogen is increased (Table 3). GLS are compounds rich in both nitrogen and sulfur, which are present in high levels in seeds. It was previously suggested that GLS may function as a sulfur storage, due to the large induction of the GLS breakdown pathway during broccoli (*Brassica oleracea* var. *italic*) seed germination (Gao et al., 2014). Gln is also known to increase upon both high nitrogen availability and sulfur deficiency (Nikiforova et al., 2006; Nikiforova et al., 2005). A study of sulfur starvation in *Arabidopsis* seedlings showed that plants convert the accumulated excess nitrogen into nitrogenous compounds, including Gln (reviewed in (Nikiforova et al., 2006)). Hence, it is possible that the lack of stored sulfur in the form of GLS in seeds may lead to sulfur deficiency, in turn leading to an elevation in FAAs, especially Gln. It is worth mentioning that no coherent pattern of alteration of the PBAA composition was observed in the *myb28/29* and the *gtr1/2* mutants as compared to the Col-0 ecotype, indicating that the elevation in Gln is not due to a lack of incorporation of Gln into SSP (Supplemental Table 8). The latter finding further supports the conclusions that sulfur reduction is due mainly to GLS reduction and that the interaction between the pathways is mediated through signaling/sensing cascades that are induced in response to the alterations to seed metabolic homeostasis.

2.5 CONCLUSIONS

This study demonstrated that free glutamine in *Arabidopsis* seeds is strongly affected by glucosinolate diversity and presence in this organ. This finding clearly highlights that the presence of specific secondary metabolites can profoundly affect primary metabolism in seeds and that selected specialized metabolites may play a larger role in the metabolic homeostasis of this tissue than originally believed. Evolutionary theory predicts that the diversity and composition of plant defense compounds, such as the glucosinolates, in the different plant tissues reflect past selection pressures imposed on plants by their environment (Jones & Firn, 1991), pressures that are believed to be key driving forces of compound diversity and composition (Benderoth et al., 2006). This study supports this claim and further suggests that the GLS effect on core metabolism may have played a role in shaping its diversity and composition; further studies are needed to reveal the extent of this phenomenon and its implication for seed fitness. This study also aligns with previous work that has shown that although defense mechanisms, such as GLS, although evolutionarily more recent and often species- and taxa-specific, have established connections with conserved regulatory/signaling pathways involved in core metabolism and other essential cellular processes. The latter was suggested to be evolutionarily advantageous in helping plants coordinate both defense metabolism and growth (Malinovsky et al., 2017). Finally, this study demonstrates that performing a semi-combinatorial ratio based mGWAS using metabolites measured in dry seeds can capture events occurring early in seed development. This finding has practical implications for future metabolic analyses since it is easier to perform an mGWAS on dry seeds than on developing seeds.

2.6 MATERIALS AND METHODS

Plant growth and seed collection

All *Arabidopsis* (*Arabidopsis thaliana*) genotypes were grown at 22°C/24°C (day/night) under long-day conditions (16 h of light/8 h of dark). Growth of the *Arabidopsis* diversity panel (Horton et al., 2012; Nordborg et al., 2005; Platt et al., 2010) was as described (R. Angelovici et al., 2013).

Seed and tissue collection

Developing siliques were marked to track their developmental stage. Siliques were harvested at 12, 14, 16, or 18 days after flowering (DAF) as well as from dry seeds, flash frozen in liquid nitrogen upon collection, and stored at -80°C. Siliques were lyophilized, and the seeds were isolated and ground for the metabolic analysis.

Sample leaf and stem tissues were collected from the same plants at approximately 20 days after bolting. Only green tissue was collected. Tissues were flash frozen in liquid nitrogen upon collection and stored at -80°C. Tissues were lyophilized and ground for the metabolic analysis.

Isolation of T-DNA insertion mutants and genotypic characterization

The mutant lines SAIL_181_F06 (*aop1*), SALK_001655C (*aop3*), SALK_004536C (*mam3*), and WiscDsLoxHs043_06G (*bsu1*) were obtained from the Arabidopsis Biological Resource Center (https://abrc.osu.edu). The SALK and WiscDsLoxHs043_06G insertions are in the Col-0 background, and the SAIL_181_F06 mutant is in the Col-3 background. Homozygous mutant lines were validated by genomic PCR using gene-specific primers in combination with the T-DNA left border primer. Primers spanning the full-length transcript were used to confirm lack of transcripts for respective genes. The list of primers can be found in Supplemental Table S4.

The *AOP2* overexpression line (Burow et al., 2015), the *myb28* and *myb29* single mutants, the *myb28/29* and *myb34/51* knockout mutants (Henning Frerigmann & Tamara Gigolashvili, 2014; Sonderby et al., 2010), and the *GSM1* mutant (Haughn et al., 1991) were provided by Dr. Dan Kliebenstein with the University of California, Davis. The GLS transporter mutant *gtr1/2* (Hussam Hassan Nour-Eldin et al., 2012) was provided by Dr. Hussam Hassan Nour-Eldin with Copenhagen University.

Transcript analysis

Total RNA extracted from dry and developing seeds was isolated using a hot borate method (Birtić & Kranner, 2006) and purified using Direct-zol RNA Miniprep Plus filter columns (Zymo Research). Total RNA from leaves was extracted using the Direct-zol RNA

Miniprep Plus Kit (Zymo Research). First-strand cDNA was synthesized from 1 μg of purified, total RNA using the iScript cDNA Synthesis Kit (Bio-rad). RT-PCR was used to determine transcript levels using the Dream Taq Green Master Mix (Thermo Fisher Scientific) with *ACTIN2* (AT3G18780) as a control. (For primers, see Supplemental Table S4.)

Data analysis and mGWAS

mGWAS was performed using the genotypic data from the 250K single nucleotide polymorphism (SNP) chip that was performed on the *Arabidopsis* diversity panel (Horton et al., 2012). SNPs with a minor allele frequency (MAF) <0.05 were removed, leaving 214,052 SNPs for association mapping. The 14 derived ratios and glutamine absolute value and relative composition were used as trait inputs. The association tests were conducted in R 3.3.2 (Team, 2014) using the R package Fixed and Random Model Circulating Probability Unification (FarmCPU) (Liu et al., 2016). The Bonferroni correction was used to control the experiment wise type I error rate at $\alpha = 0.01$

Statistical analysis of the Gln-related trait levels in accessions harboring specific GLS

A subset of 133 accessions from the 313-accession population were grown in single replicates and analyzed for FAA and GLS contents. Accessions were grouped based on the presence or absence of four GLS chemotypes: 3ohp, 2-propenyl, 4ohb, and 3-butenyl/OH-3-butenyl. A *t*-test was used to test for the presence/absence of the group based on levels of Gln-related traits. Because sample sizes of some groups were small, a 1,000 permutations procedure was conducted to determine statistical significance at $\alpha = 0.05$ (Churchill & Doerge, 1994).

GO enrichment

For the candidate genes associated with the SNPs, a GO enrichment analysis was performed using agriGO (Du et al., 2010) with the following parameters: a hypergeometric test with a $\alpha = 0.05$ *FDR* correction ($n = 3$), an agriGO suggested background, *Arabidopsis* as the select organism, and a complete GO ontology.

Correlation analysis

A Spearman's rank correlation was used to calculate an *r* correlation matrix between all raw absolute levels of FAA in the glutamate family after one round of outlier removal was performed in R. A *p-value* was calculated to reflect the significance of each correlation. Results were filtered using a $r^2 = 0.1$ threshold and a p-value = 0.001. Results were visualized with Cystoscope (Shannon et al., 2003) using the method previously described in (Batushansky et al., 2016).

Haplotype analysis and pairwise LD analysis

Since the average LD in *Arabidopsis* is 10 kb (S. Kim et al., 2007), the haploblock analysis was performed on a 10-kb window, where the 5 kb to the left and right of each significant SNP were used as inputs. The haploblock analysis was completed using Haploview version 4.2 (Barrett et al., 2004). Pairwise LD values (r^2) were calculated between the significant SNP of interest and neighboring (upstream and downstream) SNPs in a +/-5 kb window.

All SNPs were filtered at a 5% MAF. Default Gabriel block parameters were used, resulting in blocks that contained at least 95% of the SNPs in strong LD. Any genes contained, or partially contained, in the haploblock with a significant SNP were saved as putative genes of interest. If no haploblock was identified for a respective SNP, then the genes immediately upstream and downstream of the SNP were saved.

QTL analysis

A QTL analysis was conducted on the 158 RILs from the Bay x Sha population (Loudet et al., 2002) in R 3.3.2 (Team, 2014) using the R/qtl2 package (Broman et al., 2019). Data were previously quantified and described in (R. Angelovici et al., 2013). Publicly available genotype markers spanned the five chromosomes for a total of 69 markers. The *mqmaugment* function in R was used to calculate genotype probabilities using a step value and an assumed genotyping error rate of 0.001. Missing values were replaced with the most probable values using the *fill.geno* function; unsupervised cofactor selection was completed through backward elimination. Genome-wide LOD significance thresholds were determined using 1,000 permutations for each trait. For each QTL, confidence intervals were determined by a 1.5 LOD drop from peak marker. Percent variance explained (PVE) was calculated using the following formula: PVE = 1 - 10^(-(2/n)*LOD), where n is the sample size. Epistatic interactions were explored using the *effectplot* function.

Metabolic analyses

Amino acid analysis: Amino acids were analyzed from four biological replicates (n = 3–4) for seed tissues and from five biological replicates (n = 5) for leaf and stem tissues. The PBAAs were extracted from ~3 mg of dry seed by performing acid hydrolysis (200 µl 6N HCl at 110°C for 24 h) followed by the FAA extraction method described in (A. Yobi & R. Angelovici, 2018). FAAs were extracted with 1 mM of perfluoroheptanoic acid (PFHA) from ~6 mg tissue, as described previously (Yobi et al., 2019). The analyses were performed using an ultra-performance liquid chromatography-tandem mass spectrometer (UPLC-MS/MS) instrument (Waters Corporation, Milford, MA), as detailed previously (A. Yobi & R. Angelovici, 2018; Yobi et al., 2019).

GLS analysis: GLS identification and quantification were completed using high-performance liquid chromatography with diode-array detection (HPLC/DAD), as previously described previously (Kliebenstein et al., 2001).

Nitrogen and carbon analyses: Nitrogen and carbon levels were determined using an ECS 4010 CHNSO analyzer (Costech Analytical Technologies, Inc.), following the instructions in the manufacturer's manual. Briefly, ~2 mg tissue from five biological replicates (n = *5*) were placed in tin capsules (Costech Analytical Technologies, Inc.) and analyzed along with 2 mg of 2,5-bis-2-(5-tert-butylbenzoxazolyl)thiophene (BBOT; Thermo Fisher Scientific) as an internal standard. Helium was used as a carrier gas, and separation was performed on a GC column maintained at 110°C. Detection was based on a TCD detector, and quantification was carried out by plotting against external standards for both molecules.

Sulfur analysis: Sulfur measurements were performed using the procedure described in (Ziegler et al., 2013b). Briefly, ~50 mg of seed from six biological replicates (n = *6*) were digested with a concentrated HNO3 with an internal standard. Seeds were soaked at room temperature overnight and then incubated at 105°C for 2 h. After cooling, the samples were diluted and analyzed with an Elan 6000 DRC (dynamic reaction cell)-e mass spectrometer (PerkinElmer SCIEX) connected to a Perfluoroalkoxy (PFA) microflow nebulizer (Elemental Scientific) and Apex HF desolvator (Elemental Scientific) as described in (Ziegler et al., 2013b).

Accession numbers

At4g03070 (*AOP1*), At4g03060 (*AOP2*), At4g03050 (*AOP3*), At5g23010 (*MAM1*), At5g23020 (*MAM3*), At4g03063 (*BSU1*), At5g61420 (*MYB28*), At5g07690 (*MYB29*), At3g47960 (*GTR1*), At5g62680 (*GTR2*), AT1G18570 (*MYB51*), AT5G60890 (*MYB34*).

2.7 SUPPLEMENTAL INFORMATION

Supplemental Figures S1-S7, Supplemental Tables S1-S8 and Supplemental Datasets S1-7 are available with the original *Plant Physiology* publication downloadable by accessing the publication https://doi.org/10.1104/pp.20.00039 or downloading a zip file directly from http://www.plantphysiol.org/content/suppl/2020/04/21/pp.20.00039.DC1

2.8 ACKNOWLEDGMENTS

The authors wish to acknowledge Melody Kroll for assistance with editing the manuscript and Ivan Baxter and Greg Ziegler for sulfur analysis.

2.9 REFERENCES

Altenbach, Susan B, Chiung-Chi Kuo, Lisa C Staraci, Karen W Pearson, Connie Wainwright, Anca Georgescu, and Jeffrey Townsend. 1992. 'Accumulation of a Brazil nut albumin in seeds of transgenic canola results in enhanced levels of seed protein methionine', *Plant molecular biology*, 18: 235-45.

Altenbach, Susan B, Karen W Pearson, Gabrielle Meeker, Lisa C Staraci, and Samuel SM Sun. 1989. 'Enhancement of the methionine content of seed proteins by the expression of a chimeric gene encoding a methionine-rich protein in transgenic plants', *Plant molecular biology*, 13: 513-22.

Altenbach, Susan B, and Robert B Simpson. 1990. 'Manipulation of methionine-rich protein genes in plant seeds', *Trends in Biotechnology*, 8: 156-60.

Amir, R, and L Tabe. 2006. 'Molecular approaches to improving plant methionine content'.

Amir, Rachel, Gad Galili, and Hagai Cohen. 2018. 'The metabolic roles of free amino acids during seed development', *Plant Science*.

Angelovici, R., A. E. Lipka, N. Deason, S. Gonzalez-Jorge, H. Lin, J. Cepela, R. Buell, M. A. Gore, and D. Dellapenna. 2013. 'Genome-wide analysis of branched-chain amino acid levels in Arabidopsis seeds', *Plant Cell*, 25: 4827-43.

Angelovici, Ruthie, Albert Batushansky, Nicholas Deason, Sabrina Gonzalez-Jorge, Michael A Gore, Aaron Fait, and Dean DellaPenna. 2017. 'Network-guided GWAS improves identification of genes affecting free amino acids', *Plant Physiology*, 173: 872-86.

Angelovici, Ruthie, Aaron Fait, Xiaohong Zhu, Jedrzej Szymanski, Ester Feldmesser, Alisdair R Fernie, and Gad Galili. 2009. 'Deciphering transcriptional and metabolic networks associated with lysine metabolism during Arabidopsis seed development', *Plant Physiology*, 151: 2058-72.

Angelovici, Ruthie, Gad Galili, Alisdair R Fernie, and Aaron Fait. 2010. 'Seed desiccation: a bridge between maturation and germination', *Trends in plant science*, 15: 211-18.

Angelovici, Ruthie, Alexander E Lipka, Nicholas Deason, Sabrina Gonzalez-Jorge, Haining Lin, Jason Cepela, Robin Buell, Michael A Gore, and Dean DellaPenna. 2013. 'Genome-wide analysis of branched-chain amino acid levels in Arabidopsis seeds', *The Plant Cell*, 25: 4827-43.

Arends, D., P. Prins, R. C. Jansen, and K. W. Broman. 2010. 'R/qtl: high-throughput multiple QTL mapping', *Bioinformatics*, 26: 2990-2.

Barrett, Jeffrey C, B Fry, JDMJ Maller, and Mark J Daly. 2004. 'Haploview: analysis and visualization of LD and haplotype maps', *Bioinformatics*, 21: 263-65.

Batushansky, A., D. Toubiana, and A. Fait. 2016. 'Correlation-based network generation, visualization, and analysis as a powerful tool in biological studies: a case study in cancer cell metabolism', *Biomed Res Int*, 2016: 1-9.

Baud, Sébastien, Jean-Pierre Boutin, Martine Miquel, Loïc Lepiniec, and Christine Rochat. 2002. 'An integrated overview of seed development in Arabidopsis thaliana ecotype WS', *Plant Physiology and Biochemistry*, 40: 151-60.

Benderoth, M., S. Textor, A. J. Windsor, T. Mitchell-Olds, J. Gershenzon, and J. Kroymann. 2006. 'Positive selection driving diversification in plant secondary metabolism', *Proceedings of the National Academy of Sciences of the United States of America*, 103: 9118-23.

Besnard, J., R. Pratelli, C. Zhao, U. Sonawala, E. Collakova, G. Pilot, and S. Okumoto. 2016. 'UMAMIT14 is an amino acid exporter involved in phloem unloading in Arabidopsis roots', *Journal of Experimental Botany*, 67: 6385-97.

Binder, Stefan. 2010. 'Branched-chain amino acid metabolism in Arabidopsis thaliana', *The Arabidopsis book/American Society of Plant Biologists*, 8.

Birtić, Simona, and Ilse Kranner. 2006. 'Isolation of high-quality RNA from polyphenol-, polysaccharide-and lipid-rich seeds', *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques*, 17: 144-48.

Bose, U., J. A. Broadbent, K. Byrne, M. J. Blundell, C. A. Howitt, and M. L. Colgrave. 2020. 'Proteome Analysis of Hordein-Null Barley Lines Reveals Storage Protein Synthesis and Compensation Mechanisms', *J Agric Food Chem*, 68: 5763-75.

Bright, Simon WJ, Joseph SH Kueh, and Sven E Rognes. 1983. 'Lysine transport in two barley mutants with altered uptake of basic amino acids in the root', *Plant Physiology*, 72: 821-24.

Broman, Karl W, Daniel M Gatti, Petr Simecek, Nicholas A Furlotte, Pjotr Prins, Śaunak Sen, Brian S Yandell, and Gary A Churchill. 2019. 'R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations', *Genetics*, 211: 495-502.

Burow, M., B. A. Halkier, and D. J. Kliebenstein. 2010. 'Regulatory networks of glucosinolates shape Arabidopsis thaliana fitness', *Curr Opin Plant Biol*, 13: 348-53.

Burow, Meike, Susanna Atwell, Marta Francisco, Rachel E Kerwin, Barbara A Halkier, and Daniel J Kliebenstein. 2015. 'The glucosinolate biosynthetic gene AOP2 mediates feed-back regulation of jasmonic acid signaling in Arabidopsis', *Molecular plant*, 8: 1201-12.

Chan, E. K., H. C. Rowe, J. A. Corwin, B. Joseph, and D. J. Kliebenstein. 2011a. 'Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana', *PLoS Biol*, 9.

Chan, Eva KF, Heather C Rowe, Jason A Corwin, Bindu Joseph, and Daniel J Kliebenstein. 2011b. 'Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana', *PLoS biology*, 9: e1001125.

Chen, W., Y. Gao, W. Xie, L. Gong, K. Lu, W. Wang, Y. Li, X. Liu, H. Zhang, H. Dong, W. Zhang, L. Zhang, S. Yu, G. Wang, X. Lian, and J. Luo. 2014. 'Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism', *Nature Genetics*, 46: 714-21.

Chen, Wei, Wensheng Wang, Meng Peng, Liang Gong, Yanqiang Gao, Jian Wan, Shouchuang Wang, Lei Shi, Bin Zhou, and Zongmei Li. 2016. 'Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals', *Nature communications*, 7: 12767.

Chen, Y. Z., Q. Y. Pang, Y. He, N. Zhu, I. Branstrom, X. F. Yan, and S. Chen. 2012. 'Proteomics and metabolomics of Arabidopsis responses to perturbation of glucosinolate biosynthesis', *Mol Plant*, 5: 1138-50.

Churchill, G. A., and R. W. Doerge. 1994. 'Empirical threshold values for quantitative trait mapping', *Genetics*, 138: 963-71.

Coleman, Craig E, and Brian A Larkins. 1999. 'The prolamins of maize.' in, *seed Proteins* (Springer).

De Clercq, Ann, Martine Vandewiele, Jozef Van Damme, Philippe Guerche, Marc Van Montagu, Joël Vandekerckhove, and Enno Krebbers. 1990. 'Stable Accumulation of Modified 2S Albumin Seed Storage Proteins with Higher Methionine Contents in Transgenic Plants', *Plant Physiology*: 970-79.

Deng, Min, Dongqin Li, Jingyun Luo, Yingjie Xiao, Haijun Liu, Qingchun Pan, Xuehai Zhang, Minliang Jin, Mingchao Zhao, and Jianbing Yan. 2017. 'The genetic architecture of amino acids dissection by association and linkage analysis in maize', *Plant biotechnology journal*, 15: 1250-63.

Diepenbrock, C. H., C. B. Kandianis, A. E. Lipka, M. Magallanes-Lundback, B. Vaillancourt, E. Gongora-Castillo, J. G. Wallace, J. Cepela, A. Mesberg, P. J. Bradbury, D. C. Ilut, M. Mateos-Hernandez, J. Hamilton, B. F. Owens, T. Tiede, E. S. Buckler, T. Rocheford, C. R. Buell, M. A. Gore, and D. DellaPenna. 2017. 'Novel Loci Underlie Natural Variation in Vitamin E Levels in Maize Grain', *Plant Cell*, 29: 2374-92.

DiLeo, M. V., G. D. Strahan, M. den Bakker, and O. A. Hoekenga. 2011. 'Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome', *PLoS One*, 6: e26683.

Dinkins, Randy D, MS Srinivasa Reddy, Curtis A Meurer, Bo Yan, Harold Trick, Françoise Thibaud-Nissen, John J Finer, Wayne A Parrott, and Glenn B Collins. 2001. 'Increased sulfur amino acids in soybean plants overexpressing the maize 15 kDa zein protein', *In Vitro Cellular & Developmental Biology-Plant*, 37: 742-47.

Du, Z., X. Zhou, Y. Ling, Z. Zhang, and Z. Su. 2010. 'agriGO: a GO analysis toolkit for the agricultural community', *Nucleic Acids Res*, 38: W64-70.

Fait, A., R. Angelovici, H. Less, I. Ohad, E. Urbanczyk-Wochniak, A. R. Fernie, and G. Galili. 2006a. 'Arabidopsis seed development and germination is associated with temporally distinct metabolic switches', *Plant Physiol*, 142: 839-54.

Fait, Aaron, Ruthie Angelovici, Hadar Less, Itzhak Ohad, Ewa Urbanczyk-Wochniak, Alisdair R Fernie, and Gad Galili. 2006b. 'Arabidopsis seed development and germination is associated with temporally distinct metabolic switches', *Plant physiology*, 142: 839-54.

Falco, SC, T Guida, M Locke, J Mauvais, C Sanders, RT Ward, and P Webber. 1995. 'Transgenic canola and soybean seeds with increased lysine', *Bio/technology*, 13: 577.

FAO. 'Staple foods: What do people eat? ', Accessed June 11. (http://www.fao.org/3/u8480e/u8480e07.htm.

Flint-Garcia, Sherry A, Anastasia L Bodnar, and M Paul Scott. 2009. 'Wide variability in kernel composition, seed characteristics, and zein profiles among diverse maize inbreds, landraces, and teosinte', *Theoretical and Applied Genetics*, 119: 1129-42.

Forsyth, Jane L, Frederic Beaudoin, Nigel G Halford, Richard B Sessions, Anthony R Clarke, and Peter R Shewry. 2005. 'Design, expression and characterisation of lysine-rich forms of the barley seed protein CI-2', *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1747: 221-27.

Foyer, C. H., A. J. Bloom, G. Queval, and G. Noctor. 2009. 'Photorespiratory metabolism: genes, mutants, energetics, and redox signaling', *Annu Rev Plant Biol*, 60: 455-84.

Francisco, M., B. Joseph, H. Caligagan, B. Li, J. A. Corwin, C. Lin, R. Kerwin, M. Burow, and D. J. Kliebenstein. 2016. 'The Defense Metabolite, Allyl Glucosinolate, Modulates Arabidopsis thaliana Biomass Dependent upon the Endogenous Glucosinolate Pathway', *Frontiers in Plant Science*, 7: 774.

Frerigmann, H., and T. Gigolashvili. 2014a. 'MYB34, MYB51, and MYB122 distinctly regulate indolic glucosinolate biosynthesis in Arabidopsis thaliana', *Mol Plant*, 7: 814-28.

Frerigmann, Henning, and Tamara Gigolashvili. 2014b. 'MYB34, MYB51, and MYB122 distinctly regulate indolic glucosinolate biosynthesis in Arabidopsis thaliana', *Molecular plant*, 7: 814-28.

Galili, G., and R. Amir. 2013. 'Fortifying plants with the essential amino acids lysine and methionine to improve nutritional quality', *Plant Biotechnol J*, 11: 211-22.

Galili, Gad. 2011. 'The aspartate-family pathway of plants: linking production of essential amino acids with energy and stress regulation', *Plant signaling & behavior*, 6: 192-95.

Galili, Gad, and Rainer Höfgen. 2002. 'Metabolic engineering of amino acids and storage proteins in plants', *Metabolic engineering*, 4: 3-11.

Gao, J., X. Yu, F. Ma, and J. Li. 2014. 'RNA-seq analysis of transcriptome and glucosinolate metabolism in seeds and sprouts of broccoli (Brassica oleracea var. italic)', *Plos One*, 9: e88804.

Ghislain, Marc, Valérie Frankard, Dirk Vandenbossche, Benjamin F Matthews, and Michel Jacobs. 1994. 'Molecular analysis of the aspartate kinase-homoserine dehydrogenase gene from Arabidopsis thaliana', *Plant molecular biology*, 24: 835-51.

Gonzalez-Jorge, S., S. H. Ha, M. Magallanes-Lundback, L. U. Gilliland, A. Zhou, A. E. Lipka, Y. N. Nguyen, R. Angelovici, H. Lin, J. Cepela, H. Little, C. R. Buell, M. A. Gore, and D. Dellapenna. 2013. 'Carotenoid cleavage dioxygenase4 is a negative regulator of β-carotene content in Arabidopsis seeds', *Plant Cell*, 25: 4812-26.

Grover, Zubin, and Looi C Ee. 2009. 'Protein energy malnutrition', *Pediatric Clinics*, 56: 1055-68.

Gu, Liping, A Daniel Jones, and Robert L Last. 2010. 'Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant', *The Plant Journal*, 61: 579-90.

Hagan, ND, N Upadhyaya, LM Tabe, and TJV Higgins. 2003. 'The redistribution of protein sulfur in transgenic rice expressing a gene for a foreign, sulfur-rich protein', *The Plant Journal*, 34: 1-11.

Halkier, B. A., and J. Gershenzon. 2006. 'Biology and biochemistry of glucosinolates', *Annu Rev Plant Biol*, 57: 303-33.

Haughn, George W, Laurence Davin, Michael Giblin, and Edward W Underhill. 1991. 'Biochemical genetics of plant secondary metabolites in Arabidopsis thaliana: the glucosinolates', *Plant Physiology*, 97: 217-26.

Henriques, R., L. Bogre, B. Horvath, and Z. Magyar. 2014. 'Balancing act: matching growth with environment by the TOR signalling pathway', *Journal of Experimental Botany*, 65: 2691-701.

Heremans, Betty, and Michel Jacobs. 1997. 'A Mutant of Arabidopsis thaliana lpar; L.) Heynh. with Modified Control of Aspartate Kinase by Threonine', *Biochemical genetics*, 35: 139-53.

Herman, Eliot M. 2014. 'Soybean seed proteome rebalancing', *Frontiers in plant science*, 5: 437.

Hoffman, Leslie M, Debra D Donaldson, and Eliot M Herman. 1988. 'A modified storage protein is synthesized, processed, and degraded in the seeds of transgenic plants', *Plant molecular biology*, 11: 717-29.

Holding, David, and Joachim Messing. 2013. 'Evolution, structure, and function of prolamin storage proteins', *Seed genomics*: 138-58.

Horton, Matthew W, Angela M Hancock, Yu S Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N Wayan Muliyati, Alexander Platt, F Gianluca Sperone, and Bjarni J Vilhjálmsson. 2012. 'Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel', *Nature genetics*, 44: 212.

Hou, Anfu, Kede Liu, Niramol Catawatcharakul, Xurong Tang, Vi Nguyen, Wilfred A Keller, Edward WT Tsang, and Yuhai Cui. 2005. 'Two naturally occurring deletion mutants of 12S seed storage proteins in Arabidopsis thaliana', *Planta*, 222: 512-20.

Hunter, B. G., M. K. Beatty, G. W. Singletary, B. R. Hamaker, B. P. Dilkes, B. A. Larkins, and R. Jung. 2002. 'Maize opaque endosperm mutations create extensive changes in patterns of gene expression', *Plant Cell*, 14: 2591-612.

Ingle, Robert A. 2011. 'Histidine biosynthesis', *The Arabidopsis book/American Society of Plant Biologists*, 9.

Jensen, L. M., D. J. Kliebenstein, and M. Burow. 2015. 'Investigation of the multifunctional gene AOP3 expands the regulatory network fine-tuning glucosinolate production in Arabidopsis', *Front Plant Sci*, 6: 762.

Johnson, S.D., M.E. Griffiths, C.I. Peter, and M.J. Lawes. 2009. 'Pollinators, “mustard oil” volatiles, and fruit production in flowers of the dioecious tree Drypetes natalensis (Putranjivaceae)', *American Journal of Botany*, 96: 2080–86.

Jones, C. G., and R. D. Firn. 1991. 'On the Evolution of Plant Secondary Chemical Diversity', *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 333: 273-80.

Karchi, Hagai, Orit Shaul, and Gad Galili. 1993. 'Seed-specific expression of a bacterial desensitized aspartate kinase increases the production of seed threonine and methionine in transgenic tobacco', *The Plant Journal*, 3: 721-27.

Katz, E., S. Nisani, B. S. Yadav, M. G. Woldemariam, B. Shai, U. Obolski, M. Ehrlich, E. Shani, G. Jander, and D. A. Chamovitz. 2015. 'The glucosinolate breakdown product indole-3-carbinol acts as an auxin antagonist in roots of Arabidopsis thaliana', *Plant J*, 82: 547-55.

Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark, S. Ossowski, J. R. Ecker, D. Weigel, and M. Nordborg. 2007. 'Recombination and linkage disequilibrium in Arabidopsis thaliana', *Nat Genet*, 39.

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

Kliebenstein, D. J., J. C. D'Auria, A. S. Behere, J. H. Kim, K. L. Gunderson, J. N. Breen, G. Lee, J. Gershenzon, R. L. Last, and G. Jander. 2007. 'Characterization of seed-specific benzoyloxyglucosinolate mutations in Arabidopsis thaliana', *Plant J*, 51: 1062-76.

Kliebenstein, D. J., J. Kroymann, P. Brown, A. Figuth, D. Pedersen, J. Gershenzon, and T. Mitchell-Olds. 2001. 'Genetic control of natural variation in Arabidopsis glucosinolate accumulation', *Plant Physiol*, 126: 811-25.

Kroymann, J., S. Donnerhacke, D. Schnabelrauch, and T. Mitchell-Olds. 2003. 'Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus', *Proc Natl Acad Sci U S A*, 100 Suppl 2: 14587-92.

Kroymann, J., S. Textor, J. G. Tokuhisa, K. L. Falk, S. Bartram, J. Gershenzon, and T. Mitchell-Olds. 2001. 'A gene controlling variation in Arabidopsis glucosinolate composition is part of the methionine chain elongation pathway', *Plant Physiol*, 127: 1077-88.

Lambert, RJ, DE Alexander, and JW Dudley. 1969. 'Relative Performance of Normal and Modified Protein (Opaque-2) Maize Hybrids 1', *Crop Science*, 9: 242-43.

Langfelder, Peter, and Steve Horvath. 2008. 'WGCNA: an R package for weighted correlation network analysis', *BMC bioinformatics*, 9.

Lea, P. J., and B. J. Miflin. 1974. 'Alternative route for nitrogen assimilation in higher plants', *Nature*, 251: 614-6.

Lea, Peter John. 1998. "The enzymes of glutamine. Glutamate, asparagine, and aspartate metabolism." In, 63-124. *Plant amino acids*.

Lee, Soo In, Hyun Uk Kim, Yeon-Hee Lee, Suk-Cheol Suh, Yong Pyo Lim, Hyo-Yeon Lee, and Ho-Il Kim. 2001. 'Constitutive and seed-specific expression of a maize lysine-feedback-insensitive dihydrodipicolinate synthase gene leads to increased free lysine levels in rice seeds', *Molecular Breeding*, 8: 75-84.

Lipka, Alexander E, Michael A Gore, Maria Magallanes-Lundback, Alex Mesberg, Haining Lin, Tyler Tiede, Charles Chen, C Robin Buell, Edward S Buckler, and Torbert Rocheford. 2013. 'Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain', *G3: Genes, Genomes, Genetics*, 3: 1287-99.

Lipka, Alexander E., Michael A. Gore, Maria Magallanes-Lundback, Alex Mesberg, Haining Lin, Tyler Tiede, Charles Chen, C. Robin Buell, Edward S. Buckler, Torbert Rocheford, and Dean Dellapenna. 2013. 'Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain', *G3*, 3: 1287-99.

Liu, Xiaolei, Meng Huang, Bin Fan, Edward S Buckler, and Zhiwu Zhang. 2016. 'Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies', *Plos genetics*, 12: e1005767.

Loudet, O, S Chaillou, C Camilleri, D Bouchez, and F Daniel-Vedele. 2002. 'Bay-0× Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis', *Theoretical and Applied Genetics*, 104: 1173-84.

Magrath, R., Bano, F., Morgner, M. . 1994. 'Genetics of aliphatic glucosinolates I. Side chain elongation in Brassica napus and Arabidopsis thaliana', *Heredity* 72: 290-99.

Magrath, R., and R. Mithen. 1993. 'Maternal Effects on the Expression of Individual Aliphatic Glucosinolates in Seeds and Seedlings of Brassica-Napus', *Plant Breeding*, 111: 249-52.

Majumdar, R, B. Barchi, S.A. Turlapati, M. Gagne, R. Minocha, S. Long, and S.C. Minocha. 2016. 'Glutamate, Ornithine, Arginine, Proline, and Polyamine Metabolic Interactions: The Pathway Is Regulated at the Post-Transcriptional Level', *Frontiers in Plant Science*, 7: 78.

Malinovsky, F. G., M. F. Thomsen, S. J. Nintemann, L. M. Jagd, B. Bourgine, M. Burow, and D. J. Kliebenstein. 2017. 'An evolutionarily young defense metabolite influences the root growth of plants via the ancient TOR signaling pathway', *Elife*, 6.

Mazur, Barbara, Enno Krebbers, and Scott Tingey. 1999. 'Gene discovery and product development for grain quality traits', *Science*, 285: 372-75.

Mertz, Edwin T, Lynn S Bates, and Oliver E Nelson. 1964. 'Mutant gene that changes protein composition and increases lysine content of maize endosperm', *Science*, 145: 279-80.

Molvig, Lisa, Linda M Tabe, Bjorn O Eggum, Andrew E Moore, Stuart Craig, Donald Spencer, and Thomas JV Higgins. 1997. 'Enhanced methionine levels and increased nutritive value of seeds of transgenic lupins (Lupinus angustifolius L.) expressing a sunflower seed albumin gene', *Proceedings of the National Academy of Sciences*, 94: 8393-98.

Morton, K. J., S. Jia, C. Zhang, and D. R. Holding. 2016. 'Proteomic profiling of maize opaque endosperm mutants reveals selective accumulation of lysine-enriched proteins', *J Exp Bot*, 67: 1381-96.

Nikiforova, V. J., M. Bielecka, B. Gakiere, S. Krueger, J. Rinder, S. Kempa, R. Morcuende, W. R. Scheible, H. Hesse, and R. Hoefgen. 2006. 'Effect of sulfur availability on the integrity of amino acid biosynthesis in plants', *Amino Acids*, 30: 173-83.

Nikiforova, V. J., J. Kopka, V. Tolstikov, O. Fiehn, L. Hopkins, M. J. Hawkesford, H. Hesse, and R. Hoefgen. 2005. 'Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of Arabidopsis plants', *Plant Physiol*, 138: 304-18.

Nordborg, Magnus, Tina T Hu, Yoko Ishino, Jinal Jhaveri, Christopher Toomajian, Honggang Zheng, Erica Bakker, Peter Calabrese, Jean Gladstone, and Rana Goyal. 2005. 'The pattern of polymorphism in Arabidopsis thaliana', *PLoS biology*, 3: e196.

Nour-Eldin, H. H., T. G. Andersen, M. Burow, S. R. Madsen, M. E. Jorgensen, C. E. Olsen, I. Dreyer, R. Hedrich, D. Geiger, and B. A. Halkier. 2012. 'NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds', *Nature*, 488: 531-34.

Nour-Eldin, Hussam Hassan, Tonni Grube Andersen, Meike Burow, Svend Roesen Madsen, Morten Egevang Jørgensen, Carl Erik Olsen, Ingo Dreyer, Rainer Hedrich, Dietmar Geiger, and Barbara Ann Halkier. 2012. 'NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds', *nature*, 488: 531.

Okumoto, S., D. Funck, M. Trovato, and G. Forlani. 2016. 'Editorial: Amino Acids of the Glutamate Family: Functions beyond Primary Metabolism', *Frontiers in Plant Science*, 7: 318.

Parkin, I., R. Magrath, D. Keith, A. Sharpe, R. Mithen, and D. Lydiate. 1994. 'Genetics of Aliphatic Glucosinolates .2. Hydroxylation of Alkenyl Glucosinolates in Brassica-Napus', *Heredity*, 72: 594-98.

Pei, G., L. Chen, and W. Zhang. 2017. 'WGCNA Application to Proteomic and Metabolomic Data Analysis', *Methods Enzymol*, 585: 135-58.

Petersen, B. L., S. Chen, C. H. Hansen, C. E. Olsen, and B. A. Halkier. 2002. 'Composition and content of glucosinolates in developing Arabidopsis thaliana', *Planta*, 214: 562-71.

Platt, Alexander, Matthew Horton, Yu S Huang, Yan Li, Alison E Anastasio, Ni Wayan Mulyati, Jon Ågren, Oliver Bossdorf, Diane Byers, and Kathleen Donohue. 2010. 'The scale of population structure in Arabidopsis thaliana', *PLoS genetics*, 6: e1000843.

Qi, Z., Z. Zhang, Z. Wang, J. Yu, H. Qin, X. Mao, H. Jiang, D. Xin, Z. Yin, R. Zhu, C. Liu, W. Yu, Z. Hu, X. Wu, J. Liu, and Q. Chen. 2018. 'Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development', *Plant Cell Environ*, 41: 2109-27.

Qi, Zhaoming, Zhanguo Zhang, Zhongyu Wang, Jingyao Yu, Hongtao Qin, Xinrui Mao, Hongwei Jiang, Dawei Xin, Zhengong Yin, and Rongsheng Zhu. 2018. 'Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development', *Plant, cell & environment*, 41: 2109-27.

Research, Infinium Global. 2017. 'Amino Acid Market: Global Industry Analysis, Trends, Market Size & Forecast to 2023 ', Accessed June 9. ttps://www.researchandmarkets.com/research/xsjjhs/amino_acid.

Riedelsheimer, Christian, Jan Lisec, Angelika Czedik-Eysenberg, Ronan Sulpice, Anna Flis, Christoph Grieder, Thomas Altmann, Mark Stitt, Lothar Willmitzer, and Albrecht E Melchinger. 2012. 'Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize', *Proceedings of the National Academy of Sciences*, 109: 8872-77.

Rohr, F., C. Ulrichs, T. Mucha-Pelzer, and I. Mewis. 2006. 'Variability of aliphatic glucosinolates in Arabidopsis and their influence on insect resistance', *Commun Agric Appl Biol Sci*, 71: 507-15.

Schaefer, Robert J, Jean-Michel Michno, Joseph Jeffers, Owen Hoekenga, Brian Dilkes, Ivan Baxter, and Chad L Myers. 2018. 'Integrating coexpression networks with GWAS to prioritize causal genes in maize', *Plant Cell*, 30: 2922-42.

Scheible, W. R., R. Morcuende, T. Czechowski, C. Fritz, D. Osuna, N. Palacios-Rojas, D. Schindelasch, O. Thimm, M. K. Udvardi, and M. Stitt. 2004. 'Genome-wide reprogramming of primary and secondary metabolism, protein synthesis, cellular growth processes, and the regulatory infrastructure of Arabidopsis in response to nitrogen', *Plant Physiol*, 136: 2483-99.

Schmidt, Monica A, W Brad Barbazuk, Michael Sandford, Greg May, Zhihong Song, Wenxu Zhou, Basil J Nikolau, and Eliot M Herman. 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*, 156: 330-45.

Schwender, Jörg, Christina König, Matthias Klapperstück, Nicolas Heinzel, Eberhard Munz, Inga Hebbelmann, Jordan O Hay, Peter Denolf, Stefanie De Bodt, and Henning Redestig. 2014. 'Transcript abundance on its own cannot be used to infer fluxes in central metabolism', *Frontiers in plant science*, 5: 668.

Scossa, F., D. Laudencia-Chingcuanco, O. D. Anderson, W. H. Vensel, D. Lafiandra, R. D'Ovidio, and S. Masci. 2008. 'Comparative proteomic and transcriptional profiling of a bread wheat cultivar and its derived transgenic line overexpressing a low molecular weight glutenin subunit gene in the endosperm', *Proteomics*, 8: 2948-66.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Res*, 13: 2498-504.

Shaul, Orit, and Gad Galili. 1992. 'Increased lysine synthesis in tobacco plants that express high levels of bacterial dihydrodipicolinate synthase in their chloroplasts', *The Plant Journal*, 2: 203-09.

Skokut, T.A., C.P. Wolk, J. Thomas, J.C. Meeks, and P.W. Shaffer. 1978. 'Initial organic products of assimilation of [N]ammonium and [N]nitrate by tobacco cells cultured on different sources of nitrogen', *Plant Physiology*, 62: 299-304.

Slaten, M. L., A. Yobi, C. Bagaza, Y. O. Chan, V. Shrestha, S. Holden, E. Katz, C. Kanstrup, A. E. Lipka, D. J. Kliebenstein, H. H. Nour-Eldin, and R. Angelovici.

2020. 'mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis', *Plant Physiol*, 183: 483-500.

Sonderby, I. E., M. Burow, H. C. Rowe, D. J. Kliebenstein, and B. A. Halkier. 2010. 'A Complex Interplay of Three R2R3 MYB Transcription Factors Determines the Profile of Aliphatic Glucosinolates in Arabidopsis1[C][W][OA]', *Plant Physiology*, 153: 348-63.

Sonderby, I. E., B. G. Hansen, N. Bjarnholt, C. Ticconi, B. A. Halkier, and D. J. Kliebenstein. 2007. 'A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates', *Plos One*, 2: e1322.

Tabe, Linda, and TJV Higgins. 1998. 'Engineering plant protein composition for improved nutrition', *Trends in plant science*, 3: 282-86.

Tabe, Linda M, and Michel Droux. 2002. 'Limits to sulfur accumulation in transgenic lupin seeds expressing a foreign sulfur-rich protein', *Plant Physiology*, 128: 1137-48.

Tan-Wilson, A. L., and K. A. Wilson. 2012. 'Mobilization of seed protein reserves', *Physiol Plant*, 145: 140-53.

Team, RC. 2014. 'A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria2014', *URL:(https://www. R-project. org)*.

Textor, S., S. Bartram, J. Kroymann, K. L. Falk, A. Hick, J. A. Pickett, and J. Gershenzon. 2004. 'Biosynthesis of methionine-derived glucosinolates in Arabidopsis thaliana: recombinant expression and characterization of methylthioalkylmalate synthase, the condensing enzyme of the chain-elongation cycle', *Planta*, 218: 1026-35.

Torrent, Margarita, Iñaki Alvarez, M Isabel Geli, Ionara Dalcol, and Dolors Ludevid. 1997. 'Lysine-rich modified γ-zeins accumulate in protein bodies of transiently transformed maize endosperms', *Plant molecular biology*, 34: 139-49.

Vaughn, Justin N, Randall L Nelson, Qijian Song, Perry B Cregan, and Zenglu Li. 2014. 'The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations', *G3: Genes, Genomes, Genetics*, 4: 2283-94.

Velasco, P., P. Soengas, M. Vilar, M. E. Cartea, and M. del Rio. 2008. 'Comparison of glucosinolate profiles in leaf and seed tissues of different Brassica napus crops', *Journal of the American Society for Horticultural Science*, 133: 551-58.

Verslues, Paul E, Jesse R Lasky, Thomas E Juenger, Tzu-Wen Liu, and M Nagaraj Kumar. 2014. 'Genome-wide association mapping combined with reverse genetics identifies new effectors of low water potential-induced proline accumulation in Arabidopsis', *Plant physiology*, 164: 144-59.

Wen, W., D. Li, X. Li, Y. Gao, W. Li, H. Li, J. Liu, H. Liu, W. Chen, J. Luo, and J. Yan. 2014. 'Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights', *Nat Commun*, 5: 3438.

Wenefrida, Ida, Herry S Utomo, Sterling B Blanche, and Steve D Linscombe. 2009. 'Enhancing essential amino acids and health benefit components in grain crops for improved nutritional values', *Recent patents on DNA & gene sequences*, 3: 219-25.

Wentzell, A. M., H. C. Rowe, B. G. Hansen, C. Ticconi, B. A. Halkier, and D. J. Kliebenstein. 2007. 'Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways', *PLoS Genet*, 3: 1687-701.

WHO. 'Global Database on Child Growth and Malnutrition ', Accessed June 9. http://www.who.int/ nutgrowthdb/about/introduction/en/.

Winter, D., B. Vinegar, H. Nahal, R. Ammar, G. V. Wilson, and N. J. Provart. 2007. 'An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets', *PLoS One*, 2: e718.

Withana-Gamage, Thushan S, Dwayne D Hegedus, Xiao Qiu, Peiqiang Yu, Tim May, Derek Lydiate, and Janitha PD Wanasundara. 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: 901-12.

Wu, Si, Saleh Alseekh, Álvaro Cuadros-Inostroza, Corina M Fusari, Marek Mutwil, Rik Kooke, Joost B Keurentjes, Alisdair R Fernie, Lothar Willmitzer, and Yariv Brotman. 2016. 'Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in Arabidopsis thaliana', *Plos genetics*, 12: e1006363.

Wu, Yongrui, and Joachim Messing. 2014. 'Proteome balancing of the maize seed for higher nutritional value', *Frontiers in plant science*, 5: 240.

Wu, Yongrui, Wenqin Wang, and Joachim Messing. 2012. 'Balancing of sulfur storage in maize seed', *BMC Plant Biology*, 12: 77.

Yao, Min, Mei Guan, Zhenqian Zhang, Qiuping Zhang, Yixin Cui, Hao Chen, Wei Liu, Habib U Jan, Kai P Voss-Fels, and Christian R Werner. 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., and R. Angelovici. 2018. 'A High-throughput absolute-level quantification of protein-bound amino acids in seeds', *Curr Protoc Plant Biol*: e20084.

Yobi, Abou, Albert Batushansky, Melvin J. Oliver, and Ruthie Angelovici. 2019. 'Adaptive responses of amino acid metabolism to the combination of desiccation and low nitrogen availability in Sporobolus stapfianus', *Planta*, 249: 1535-49.

Zhang, H., M. L. Wang, R. Schaefer, P. Dang, T. Jiang, and C. Chen. 2019a. 'GWAS and Coexpression Network Reveal Ionomic Variation in Cultivated Peanut', *J Agric Food Chem*, 67: 12026-36.

Zhang, Hui, Ming Li Wang, Robert Schaefer, Phat Dang, Tao Jiang, and Charles Chen. 2019b. 'GWAS and coexpression network reveal ionomic variation in cultivated peanut', *Journal of Agricultural and Food Chemistry*, 67: 12026-36.

Zhang, L., Q. Tan, R. Lee, A. Trethewy, Y. H. Lee, and M. Tegeder. 2010. 'Altered xylem-phloem transfer of amino acids affects metabolism and leads to increased seed yield and oil content in Arabidopsis', *Plant Cell*, 22: 3603-20.

Zhang, Yuanyuan, Baohua Li, Dongxin Huai, Yongming Zhou, and Daniel J Kliebenstein. 2015. 'The conserved transcription factors, MYB115 and MYB118, control expression of the newly evolved benzoyloxy glucosinolate pathway in Arabidopsis thaliana', *Frontiers in Plant Science*, 6: 343.

Zhu, Xiaohong, and Gad Galili. 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: 845-53.

Ziegler, G., A. Terauchi, A. Becker, P. Armstrong, K. Hudson, and I. Baxter. 2013. 'Ionomic Screening of Field-Grown Soybean Identifies Mutants with Altered Seed Elemental Composition', *Plant Genome*, 6.

CHAPTER 3: INTEGRATIVE NETWORK-BASED APPROACH REVEALS THE GENETIC ARCHITECTURE AND PUTATIVE REGULATORY MECHANSIMS CONTROLLING SEED PROTEIN BOUND AMINO ACID COMPOSITION IN ARABIDOPSIS

3.1 ABSTRACT

Seeds are an important source of protein around the world for human and livestock. Despite their prevalent consumption as a main food source, seeds are deficient in several essential amino acids, making them a target for biofortification efforts. Previous studies focusing on alteration of the seed proteome have identified a rebalancing phenomenon that poses a great challenge in altering the protein bound amino acid (PBAA) composition in the seed. In this study, I used a systems biology approach to shed understanding on the genetic regulation of PBAA in the *Arabidopsis* seed. To that end, I harnessed the genetic and phenotypic diversities of the *Arabidopsis* 1001 population to complete GWAS on 274 derived biochemical traits. In parallel, I analyzed the metabolome, proteome, and transcriptome of two seed storage protein (SSP) mutants showing active rebalancing. I compared the orthogonal datasets and analyzed overlapping genes for functional enrichment. I found translation to be the most constantly enriched biological process in many of the analyses. In addition, I found overrepresentation of cyclin genes suggesting that the cell cycle may play a role in rebalancing. Overall, these findings strongly suggest that translational machinery and the cell cycle may be a target for future efforts in

manipulating the seed proteome, although further experimentation and analysis are required to validate these results.

3.2 INTRODUCTION

Protein bound amino acids are vital to the seed. They serve as storage reserves for nitrogen, carbon and sulfur and are utilized during seed germination and beginning stages of seedling development (Angelovici et al., 2011; Aaron Fait et al., 2006; Rajjou et al., 2012; Tan-Wilson & Wilson, 2012). These processes are paramount to the overall fitness of the plant and are most likely the reason for their tight regulation. In the seed, PBAA are stored mainly in the form of seed storage proteins (SSP) encompassing ~95 % of the total seed proteome. The remaining ~5% is made up of free amino acids (FAA), which serve as precursors for protein synthesis and other important plant processes. Previous work analyzing the seed proteome has shown that the elimination of SSP results in a rebalancing of the seed proteome, resulting in an overall seed PBAA composition comparable to the wild-type (WT) (Herman, 2014; Wu & Messing, 2014). This has been observed in several plant species including soybean (Schmidt et al., 2011), maize (Morton et al., 2016), rice (Hagan et al., 2003), wheat (Altenbach et al., 2014), and *Arabidopsis* (Withana-Gamage et al., 2013), suggesting an evolutionary conserved process is at play. Mutant experimentation aimed at increasing and decreasing main seed proteins of several plant species has demonstrated the challenges of overcoming this rebalancing phenomenon. For example in maize, reduction of the main SSP, zeins, was followed by increases in non-SSP and other FAA, while the PBAA composition remained mostly unchanged (Jia et al., 2013; Kodrzycki et al., 1989). It was later found that the Opaque2 protein, a BZIP transcription factor, is responsible for the accumulation of non-zein proteins by altering zein transcription at the zein promotor (Schmidt et al., 1990). Reports in *Arabidopsis* seed showed increases in 12 FAAs when a leucine catabolic gene was knocked-out, but once again the PBAA composition remained unchanged (Gu et al., 2010). In soybean, RNA-silencing techniques were used to knock-down two main SSPs, eliminating over two-thirds of the protein content (Herman, 2014). The mutations resulted in accumulation of compensating proteins up to 11X more than the WT; however a parallel increase in transcript abundance was not observed. Despite overall alterations to the proteome, the total PBAA content was unchanged (Herman, 2014). Feeding studies have also been performed in hopes of seeing alterations in PBAA composition. In *Arabidopsis*, it was shown that feeding higher endogenous methionine did not necessarily lead to increases in SSP (Cohen et al., 2016). Additionally, transcript abundance did not necessarily correlate with observed protein abundance; significant elevation in the three genes encoding the 2S-albumins were observed, however no significant changes were observed in the proteome (Cohen et al., 2016). Nonetheless, in other studies in maize, analysis of gene expression patterns of different opaque mutants with a common phenotype, are shown to be quite distinct in their transcript abundance and are pleiotropic (Hunter et al., 2002). Such results beg to question if the rebalancing phenomenon is impacted by transcriptional and translational controls. Collectively, these studies reiterate the complexity of the rebalancing mechanism and need to further define regulatory mechanisms and processes by comparing different biological levels using a systems biology approach.

Despite this tight regulation, natural variation in PBAA content does exist across populations (Angelovici et al., 2017; Ruthie Angelovici et al., 2013; M. L. Slaten et al., 2020). Interestingly, perturbation of the proteome of specific genetic lines shows a return to a specific protein composition (Withana-Gamage et al., 2013, Schmidt et al., 1990), which suggests that the trait may be genetically set for each genotype. Previous reports have also suggested that protein content is determined by genotype (Herman, 2014). This suggests that key regulatory genes that act as metabolic sensors may be contributing to the rebalancing of PBAA and that these same genes have also undergone natural selection leading to the genotypic and phenotypic diversities observed across populations.

By understanding the genes that maintain PBAA in the seed, the regulation of such genes could be altered and new composition could be achieved. In this study, I took a systems biology approach by creating comprehensive, orthogonal datasets using the 1001 *Arabidopsis* diversity panel and two SSP mutants showing active rebalancing. Datasets are then compared to identify high confidence genes that were shared across orthogonal datasets.

Previous GWAS on seed amino acid composition show success in identifying candidate genes of interest that could be later validated (Ruthie Angelovici et al., 2013; M. L. Slaten et al., 2020). For example in *Arabidopsis*, analysis of seed metabolic traits revealed a novel catabolic role of BCAT2 in branched chain amino acid metabolism (Ruthie Angelovici et al., 2013). In the first part of this experiment, I harness the phenotypic and genotypic diversities of 869 ecotypes in the 1001 *Arabidopsis* diversity panel by running 274 metabolic traits and over 1 million high quality SNP calls to identify many candidate genes. However, reaching conclusions in the long list of genes resulting from GWAS can be difficult. Additionally, GWAS on the dry seed may fail to identify interesting candidate genes at specific developmental stages. Therefore, in the second part of this experiment, I harness differences occurring in two SSP mutants compared to the

WT Columbia-0 (Col-0). The mutants chosen for this study have been previously described in (Withana-Gamage et al., 2013) and include a triple knockout line of 12S SSP (CRU-), which eliminates all of the three main storage proteins in *Arabidopsis* (cruciferins) (Nguyen et al., 2015) and an RNAi knockdown line devoid of 2S napins (napin-RNAi), the second most abundant proteins in *Arabidopsis* seeds after the cruciferins (Wanasundara, 2011). Seeds of the CRU- triple-knockout show phenotypic differences in the size and distribution of protein storage vacuoles, but contain similar protein levels to the WT control; in contrast, overall protein content of the napin-RNAi lines are slightly different than the WT (Withana-Gamage et al., 2013). In short, the dry seed PBAA phenotype does not reflect the huge perturbations to the proteome as a result of such mutations. This suggests that both mutants undergo proteomic rebalancing, however, do so by activating different genetic and proteomic alterations. To address such questions, I analyzed the metabolome and proteome of the two SSP mutants and WT in the dry seed. Additionally, I analyzed the transcriptome in the three genotypes in seven timepoints across seed development.

Lastly, in order to further define a list of high confidence genes, comparison of GWAS orthogonal datasets with the SSP transcriptomic and proteomic datasets identify overlapping genes. Integrative analyses comparing two or more orthogonal datasets have previously yielded informative results (Schaefer et al., 2018; Wu et al., 2016; M. Yao et al., 2020; H. Zhang et al., 2019). Past research revealed key regulatory elements identified in *Arabidopsis* by comparing genome-scale quantitative genetic mapping with metabolite-transcript correlation networks (Wu et al., 2016). More specifically, in (Wu et al., 2016), 94 primarily metabolic traits revealed associations with SNPs in 314 *Arabidopsis* accessions. Orthogonal datasets originating from significantly changing transcripts and metabolic data across 23 timepoints and eight stress conditions yielded condition-specific networks. Highly robust correlations were identified and compared with GWAS candidate genes, in addition to candidate genes from a previously published QTL dataset to identify novel and previously reported regulatory genes for both primary and secondary metabolism. Results revealed previously identified, as well as novel regulatory genes. Similarly, by comparing candidate genes from each analysis in this study, I filter large gene lists to uncover a subset of high confidence candidate genes. Through functional analysis on such a subset of high confidence genes, a better understanding of mechanistic controls of PBAA regulation in the seed is achieved.

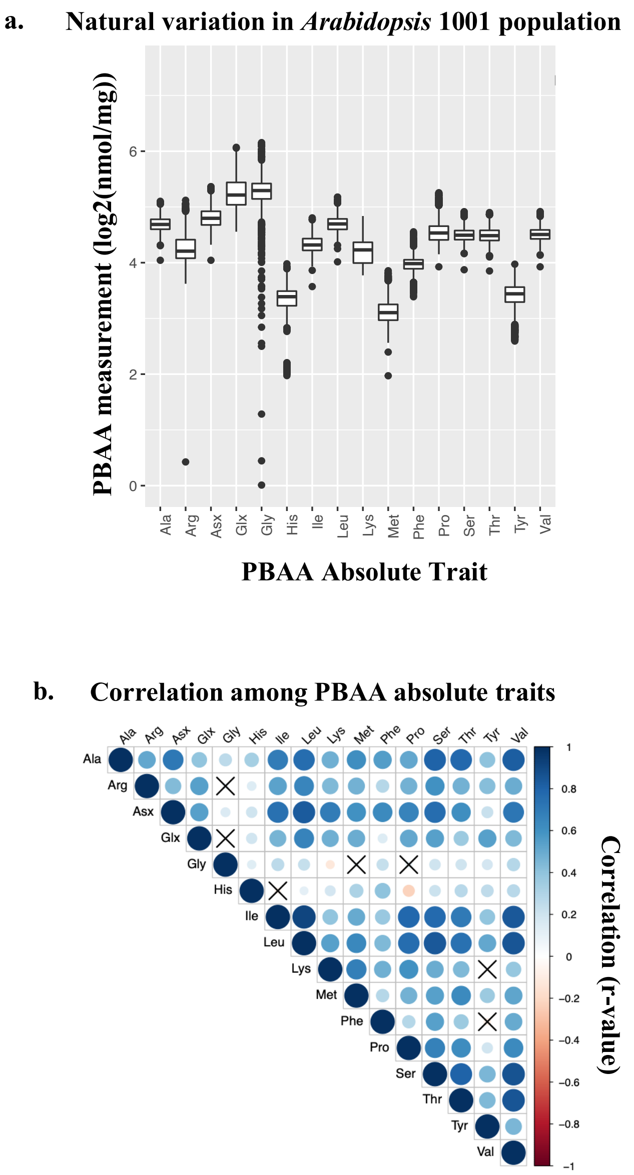
3.3 RESULTS

3.3.1 *Seed PBAA Absolute Levels, Correlation, and Heritability*

The PBAA content was measured in ecotypes found in the 1001 *Arabidopsis* diversity panel (Alonso-Blanco et al., 2016). Due to delays in flowering only 869 of the 1135 ecotypes could be harvested and phenotyped. The ecotypes were grown out in two independent replicates. Seeds were harvested at maturity and analyzed for PBAA content (See “Materials and Methods”). Due to limitations in the extraction protocol, the 16 PBAAs measured represent a total of 18 PBAAs. During extraction, Asn and Gln hydrolyze to Asp and Glu and thus cannot be differentiated; they are denoted as Asx and Glx, respectively (See (Abou Yobi & Ruthie Angelovici, 2018) for additional information).

Analysis of the PBAA content across the population showed considerable natural variation across the 16 PBAA absolute levels (Figure 3.1a). Additionally, for each measured PBAA level, variation existed across ecotypes. Overall, the average content for each PBAA can be grouped into one of three groups: high, medium, and low (ranging from 23.05-30.83, 53.12-96.29, and 109.51-197.34 nmol/mg, respectively) (Supplemental Table S3.1).

A pairwise Spearman’s rank correlation analysis was performed to evaluate the relationship among the individual PBAA absolute levels (Figure 3.1b). All the significant pairwise correlations were positive except for Pro-His. Interestingly, Gly and His had very low correlations with most of the PBAA.



a. **Natural variation in *Arabidopsis* 1001 population**

b. **Correlation among PBAA absolute traits**

Figure 3. 1. a. Boxplot showing natural variation in PBAA absolute traits across the 1001 *Arabidopsis* population. **b.** Correlation of the PBAA absolute traits from the 1001 *Arabidopsis* population. Pairwise Spearman correlation analysis was done between all absolute PBAA traits. Blue dots represent positive correlation and red dots represent negative correlations. The size of the dot indicates the strength of correlation, with larger dots having stronger correlations. FDR significance threshold α = 0.05 was used. 'X' designates samples that do not have a significant correlation.

3.3.2 *mGWAS of the PBAA Identified 1652 Significant SNP-Trait Associations*

In order to fully explore the diversity in seed PBAA composition, absolute levels were used to create a total of 274 traits derived from within family relationships of PBAA. Derived ratio traits were calculated where each absolute PBAA served as a numerator in the ratio and all possible combinations of the PBAA in that family were used as a denominator. Absolute and compositional traits (PBAA/Total PBAA) were also included. These ratios include:

1) absolute levels of PBAA (in nmol mg $^{-1}$ dry seeds),

2) relative composition (AA/Total),

3) biochemical (ratios based on metabolic pathways). (For a full list of traits see Supplemental Table S3.2) (For more on trait calculation, see "Materials and Methods"). Previous studies have demonstrated that using such derivative traits is more effective in identifying causative genes than the use of absolute measurements alone (Angelovici et al., 2017; Ruthie Angelovici et al., 2013; M. L. Slaten et al., 2020). Broad-sense heritability calculations of all traits were estimated and ranged from low to moderate. A full list of heritability estimates can be found in (Supplemental Table S3.3).

GWAS was performed on 274 traits using HAPPI GWAS (Marianne L. Slaten et al., 2020) as described in (M. L. Slaten et al., 2020). Publicly available GBS data for the 1001 *Arabidopsis* population was used (Alonso-Blanco et al., 2016). Association studies were performed using the FarmCPU (Liu et al., 2016) and the MLM models (Lipka et al., 2012). FarmCPU (Fixed And Random Model Circulating Probability Unification) is a recently developed model selection algorithm that demonstrates improved statistical power and computation efficiency by iterating between a Fixed Effect Model (FEM) and a

Random Effect Model (REM) (Liu et al., 2016). Multiple associated markers (i.e. QTNs) are used as covariates in a mixed linear model (MLM). A conservative Bonferroni cutoff threshold was used (0.05/#SNPS) to filter significant SNP-trait associations. The MLM model fits population structure (fixed effect) and kinship (random effect) (Lipka et al., 2012). An FDR of 0.05 is used as the significance threshold. A complete of Best Linear Unbiased Estimators (BLUEs) used as input in GWAS can be found in Supplemental Dataset S3.1.

Association studies using FarmCPU yielded a total of 1652 significant SNP-trait associations across all ratio traits (987 unique significant SNPs in 210 traits) (Table 3.1; Supplemental Table S3.4a). Significant SNPs identified via GWAS were then expanded to include all genes that were in the same haploblock containing the SNP. Haploblock is defined here as a non-random association quantified using a 95% confidence bounds on D prime (i.e. a region of high linkage disequilibrium (LD) with the SNP). Genes within the haploblock spanning the significant SNP are of interest because they may tag casual neighboring genes in high LD (Atwell et al., 2010). After haploblock analysis, the 1652 SNP-trait associations were found to be in high LD with 1521 unique candidate genes. One SNP was of particular interest because it was associated with the highest number of traits and was also among the most significant associations. Among all identified significant SNPs, SNP S3_8965104 was the SNP with the lowest *p-value* (x_vs_mtx trait; *p-value* 2.42e-40). Interestingly, SNP S3_8965104 was identified as the SNP with the lowest *p-value* in four out of the six PBAA families. Additionally, SNP S3_8965104 was the most frequently identified SNP across all traits. In the Aspartate Family alone, the SNP was identified in 32 traits; however, interestingly, it was not identified in any Shikimate Family traits (Table 3.1; Supplemental Table S3.4a). Further analysis showed that SNP S3_8965104 is in a haploblock containing four genes: AT3G24550 (proline extensin-like receptor kinase 1), AT3G24560 (Adenine nucleotide alpha hydrolases-like superfamily protein), AT3G24570 (Peroxisomal membrane 22 kDa (Mpv17/PMP22) family protein), and AT3G04755 (Natural antisense transcript overlaps with AT3G24570).

Table 3. 1. Summary of FarmCPU GWAS results. Results presented per family including number of significant SNPs and genes across all traits in each PBAA family, number of SNPs duplicated across traits, most occurring SNP, SNP with lowest *p-value*, and number of SNPs unique across all traits. Bonferroni test-correction was implemented to identify significant SNP-trait associations (α = 0.05; adjusted *p-value* = 4.729e-08). Abbreviations: PyrFam = Pyruvate Family, GluFam = Glutamate Family, AspFam = Aspartate Family, SerFam = Serine Family, BCAA = Branched Chain Amino Acid Family, ShikFam = Shikimate Family.

Family	**# of traits**	**# traits w/ sig SNPs**	**Total # sig SNPs**	**Total # genes**	**# unique SNPs**	**# unique genes**	**# duplicated SNPs**	**# duplicated genes**	**SNP w/ lowest p-value**	**SNP most occurring**
AspFam	155	118	960	1874	539	837	421	1037	2.42e-40 (S3_8965104)	32 (S3_8965104)
BCAAFam	25	22	227	423	156	258	71	165	1.83e-40 (S3_8965104)	18 (S3_8965104)
GluFam	64	51	290	529	213	350	77	179	8.66e-27 (S3_8965104)	5 (S1_24861714) 5 (S3_8965104) 5 (S4_6831718)
PyrFam	24	21	209	428	148	265	61	163	1.83e-40 (S3_8965104)	17 (S3_8965104)
SerFam	8	4	39	78	39	76	0	2	2.83e-16 (S4_14295678)	No duplicated SNPs
Shikfam	8	3	21	39	20	39	1	0	6.29e-13 (S2_4789695)	2 (S4_5982200)

To determine if there was any enrichment for biological function, a Gene Ontology (GO) enrichment analysis was completed using the 1521 unique candidate genes using agriGO (Tian et al., 2017). The rationale behind analyzing all traits together initially is due to the fact that all 16 PBAA traits are highly connected, as observed by identification of duplicated significant SNPs across families. Thus collectively, the traits make-up the seed proteome and share genetic architecture. Interestingly, no significant enrichment was found. Functional analysis using the candidate genes per family also yielded no significant enrichments. To further characterize the genes, the candidate genes were mapped to MapMan categories (see "Materials and Methods") (Figure 3.2). The top three most abundant MapMan categories included protein (21% of genes), RNA (14% of genes) and signaling (11% of genes) (Figure 3.2).

All GWAS analyses, biological enrichment, and functional characterization were completed in parallel using the MLM model in HAPPI GWAS (Supplemental Document S1). Although not discussed here, output from the MLM model for the traits with significant SNP associations was compared with the candidate genes from the FarmCPU model in detail in Supplemental Document S1 and Supplemental Table S3.4b.

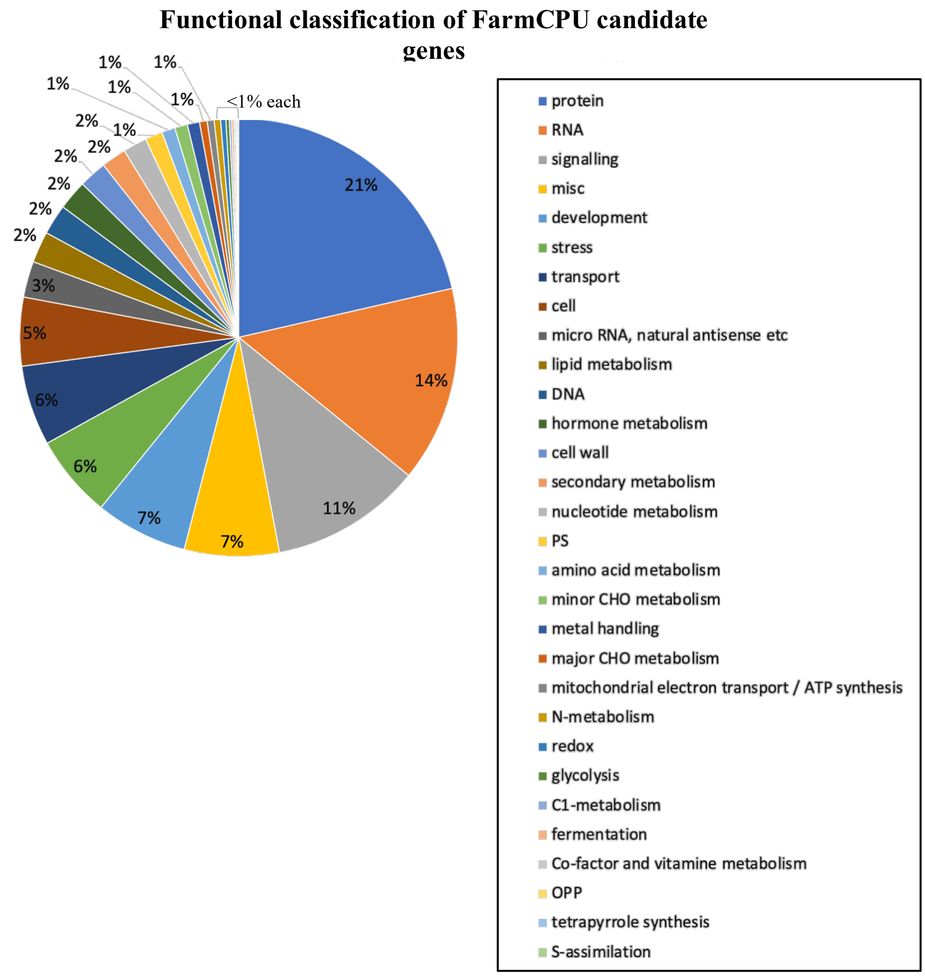


Figure 3. 2. Pie chart representing the functional categorization of the 1521 GWAS candidate genes classified using MapMan functional categories. Genes classified as 'NA' or 'not assigned' were removed prior to visualization.

3.3.3 *Characterization of PBAA Rebalancing in Dry Seed Using Metabolomics and Proteomics in Two SSP Mutants Show Active Rebalancing*

Multi-omics phenotypes of the WT and two mutants dry seeds undergoing protein rebalancing (Withana-Gamage et al., 2013) were first characterized. PBAA in each mutant was measured and compared to the WT using the same protocol previously mentioned (Abou Yobi & Ruthie Angelovici, 2018). In the WT, the three most abundant PBAA are Gly (16%), Glx (15%), and Ala (8%) and the three least abundant are Met (1%), Tyr (2%) and His (2%) (Figure 3.3a). To best visualize changes in PBAA in the dry seed, % PBAA/TPBAA (total Protein Bound Amino Acid) (i.e. PBAA/total PBAA*100) for each PBAA was visualized in a heatmap across development in each genotype (Figure 3.3b; Supplemental Table S3.5). As expected, the two mutant genotypes showed only slight variation for each PBAA compared to the WT. The most notable difference was observed in the CRU- mutant where Lys and His were statistically different than the WT (*p-value*=0.05) (Figure 3.3b; Supplemental Table S3.5). No statistically significant differences were identified for the absolute PBAA traits in the napin-RNAi mutant compared to the WT. Such minimal changes in dry seed PBAA support previous findings of active rebalancing phenomenon in the mutants and the tight regulation of PBAA composition.

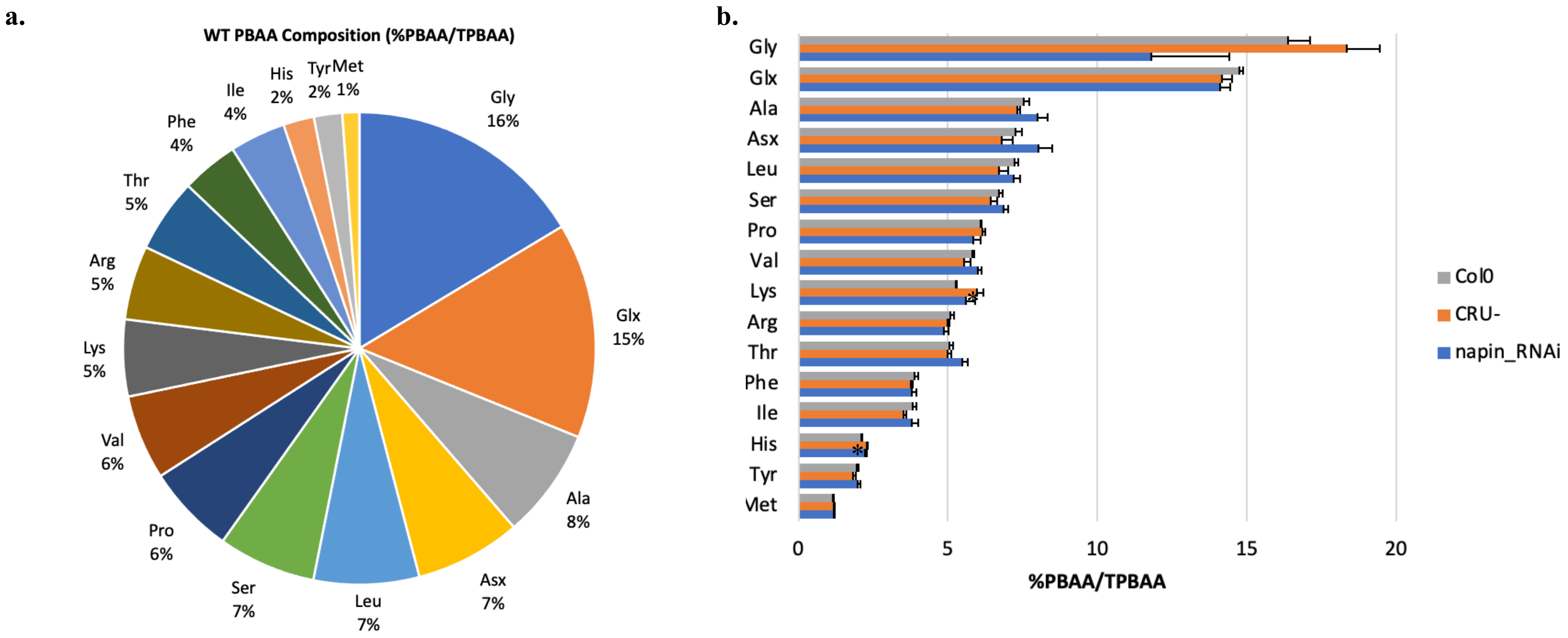


Figure 3. 3. Characterization of protein bound amino acid in the CRU- and napin-RNAi SSP mutants and WT**. a.** Pie chart of PBAA relative levels versus total PBAA (%PBAA/TPBAA) in dry seed of the WT. **b.** Bar chart of WT, CRU-, and napin-RNAi %PBAA/TPBAA. Error bars represent standard errors (n= 6). Statistical significance determined by a t-test at α = 0.05 is indicated by an asterisk (*). Abbreviations: PBAA = protein bound amino acids, TPBAA = total protein bound amino acids. See Supplementary Table 3.2 for trait abbreviations.

Next, the proteome composition of dry seeds in response to the two mutants was evaluated and compared to the WT using a mass spectrometry-based identification of proteins (See “Materials and Methods” for more information). In order to ensure statistical rigor, experiments were carried out in triplicate (one sample from each grow-out). After removal of low spectral counts and poor reproducibility, 1342 and 1292 proteins were identified in the CRU- and napin-RNAi mutants, respectively. Significant differences in each proteome was determined by comparing each mutant to the WT in a t-test at 10% FDR correction. Comparison of proteins yielding an FDR less than the cutoff, were considered differentially expressed proteins (DEP). For the CRU- mutant, a total of 929 DEPs (85% of the identified proteins) were identified as significantly altered. Of these proteins, 882 DEPs were increased and 47 DEPs were decreased (Figure 3.4). In contrast, a total of 20 DEP (11% of the identified proteins) were identified in the napin-RNAi mutant. A total of 15 DEP were increased and 5 DEPs were decreased (Figure 3.4). For a complete list of proteins, t-test results, and fold-change for each mutant, see Supplemental Table S3.6.

Next, DEPs from each mutant were binned based on FC and subject to functional enrichment analysis using agriGO (Tian et al., 2017). In the CRU- mutant, DEPs with FC ranging from 0 to -2 showed a top functional GO enrichment for nuclear outer membrane-endoplasmic reticulum membrane network (GO:0042175; FDR 0.00045) (Supplemental Table S3.7a); CRU- DEPs with a FC ranging from 4 to 6 showed a top functional GO enrichment for intracellular protein transport (GO:0006886; FDR 0.012) (Supplemental Table S3.7b). The napin-RNAi DEPs showed no statistically significant GO enrichment.

Collectively, these results demonstrate that differences are occurring in the two mutants. Despite perturbations to the napin genes, the dry seed proteome of the napin-RNAi remained fairly constant, whereas the CRU- mutant showed large perturbations that suggests protein rebalancing occurred. Nonetheless, regardless of proteomic alteration occurring mainly in the CRU-, PBAA composition was maintained with little change in both the CRU- and napin-RNAi mutants.

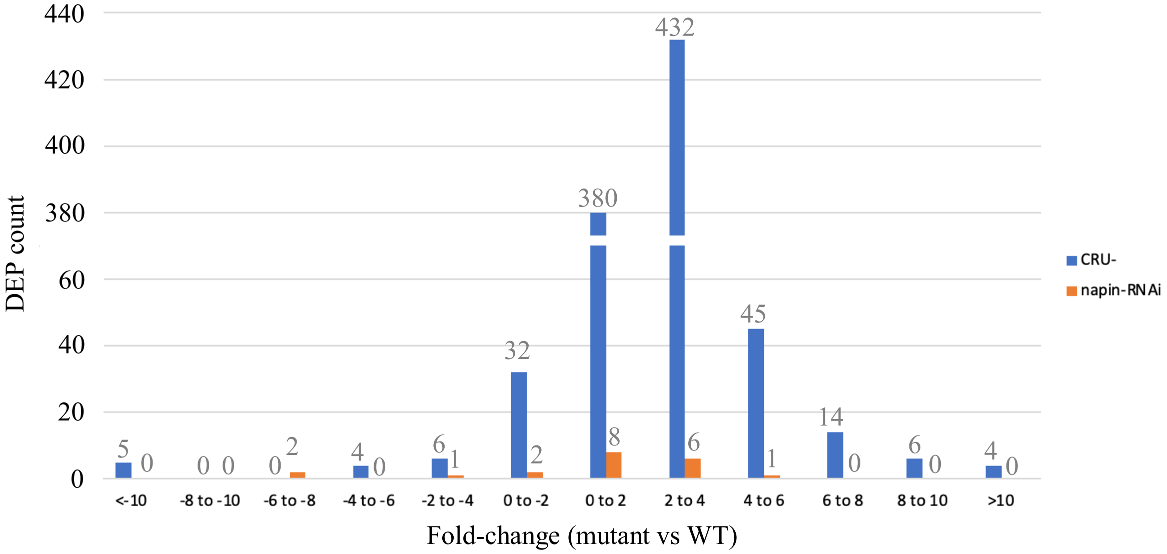


Figure 3. 4. Characterization of fold-change for differentially expressed proteins (DEPs) in the CRU- and napin-RNAi mutants. Fold-change calculated by comparing mutant expression versus wild-type (WT); DEPs were categorized into 12 bins based on magnitude of fold-change.

3.3.4 *Transcriptomics Analysis From the SSP Mutants Across Development Suggests a Role of Transcriptional Regulation in Seed Rebalancing*

To further understand the genetic basis underlying the proteomic re-balancing and generate an orthogonal candidate genes list to compare with the GWAS results, transcriptional changes in the two SSP mutants across development were measured. To achieve this goal, the total RNA of the two SSP mutants and one WT genotype across seed development was extracted and used for a library construction and sequenced using the NovaSeq platform.

After trimming and cleaning, an average of 51,774152 paired end reads of 100 bp length were obtained across all samples (see "Materials and Methods" for additional information on read processing and transcriptome assembly). Reads were aligned to the TAIR 10 *Arabidopsis* reference genome and transcripts were quantified at the gene level using STAR v2.7.6a (Dobin et al., 2013). Multi-mapped reads were discarded. The number of identified transcripts varied at each developmental timepoint but were consistent across genotypes (Supplemental Table S3.8). In total, 22,094 distinct transcripts were identified across samples and developmental timepoints. (For additional information on methodology, see "Materials and Methods"). See Supplemental Dataset S3.4 for a counts per million of all identified genes across all samples.

Changes in the transcriptome of the mutants were compared to the WT at each timepoint to evaluate alterations in regulatory control throughout seed development using edgeR (MD et al., 2010). Genes that are significantly different than the WT (i.e. differentially expressed genes (DEGs)) were characterized individually and across development. (For additional information on methodology, see "Materials and Methods"). A full list of DEGs can be found in Supplemental Table S3.9.

In the CRU- mutant, a total of 8473 DEGs were identified across development, corresponding to 6259 unique genes (Figure 3.5a). At the beginning of development (12-16 DAF) and in the dry seed (22 DAF), most DEGs were increased, but at 18 and 20 DAF most DEGs were decreased (Figure 3.5a). The majority of DEGs were also identified at 18 and 20 DAF timepoints, 4591 (54% of all DEGs identified) and 2547 (30% of all DEGs identified) DEGs, respectively. Collectively, the two timepoints accounted for 84% of all DEGs across development. GO enrichment analysis using agriGO (Tian et al., 2017) revealed GO enrichment terms among all up-regulated DEGs across development for 'translation' (GO:0006412; FDR: 1.10e-63) and a variety of different ribosome and rRNA associated enrichments (Supplemental Table S3.10f). GO enrichment analyses at 16 and 18 DAF up-regulated DEGs showed similar enrichments of 'translation' and 'rRNA processing' (Supplemental Table S3.10c, d). Functional characterization of all the down-regulated CRU- DEGs across development showed an enrichment for 'photosynthesis' (GO:0015979; FDR 5.60e-15); enrichments for a variety of different biological processes were observed at 18 and 20 DAF (Supplemental Table S3.10h-j). A full list of GO enrichment terms for all 6259 genes and DEG at individual timepoints can be found in Supplemental Table S3.10.

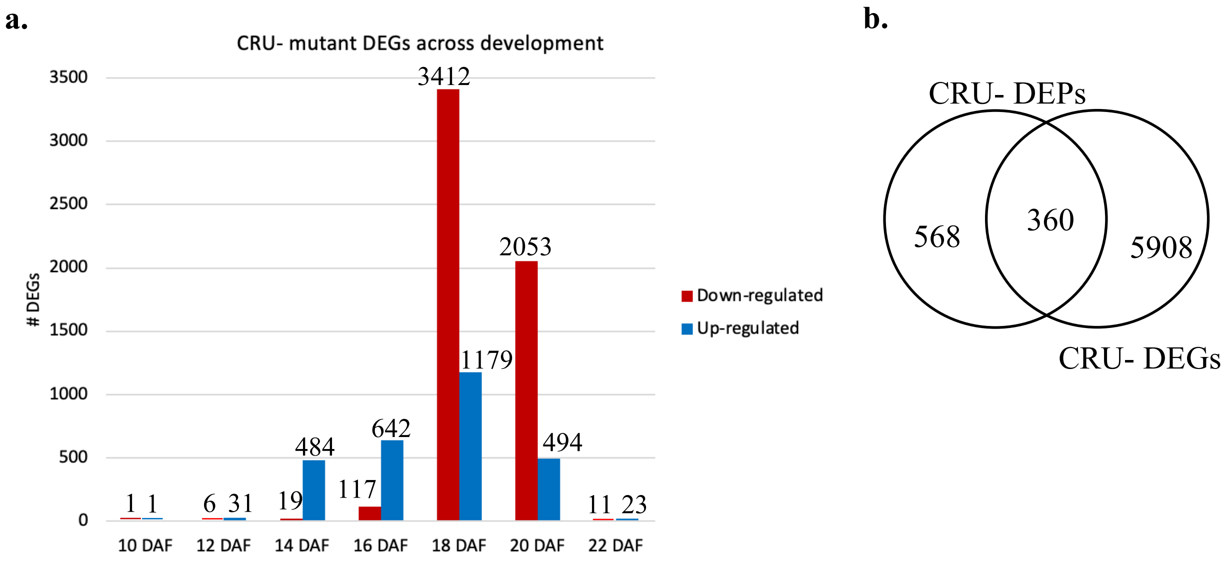


Figure 3. 5. a. Distribution of up-regulated and down-regulated differentially expressed genes (DEGs) from the CRU- mutant at each developmental timepoint. **b.** Overlap between CRU- DEGs and differentially expressed proteins (DEPs).

In the napin-RNAi mutant, a total of 4609 DEGs were identified across development, corresponding to 4290 unique genes (Figure 3.6a). Similar to the CRU- mutant, most of the DEGs were identified at 18 (934 genes) and 20 (3551 genes) DAF; a total of 77% of all DEGs identified in the napin-RNAi mutant occurred at 20 DAF, which is almost 1.5 times more than that identified in the CRU- mutant at 20 DAF. Similar to CRU-, the majority of DEGs were decreased at these timepoints (Figure 3.6a). GO enrichment analysis using agriGO (Tian et al., 2017) for all decreased DEGs across development revealed significant top biological enrichment term for 'fatty acid biosynthetic process' (GO:0006633; FDR: 1.900e-16) (Supplemental Table S3.10p). The same top enrichment term was also identified among decreasing DEGs at the 20 DAF timepoint (FDR: 2.50e-14) (Supplemental Table S3.10o). Decreasing DEGs at 18 DAF showed a top enrichment for photosynthesis (GO:0015979; FDR 2.58e-08) (Supplemental Table S3.10n). Similar analyses using the increased napin-RNAi DEGs at each developmental timepoint showed significant GO enrichment terms at 10 DAF (GO:0048316; seed development; FDR 0.0095) and 20 DAF (GO:0009657; plastid organization; 1.40e-08) (Supplemental Table S3.10k-m). A full list of GO enrichment terms for all 4290 DEGs and DEGs at individual timepoints can be found in Supplemental Table S3.10.

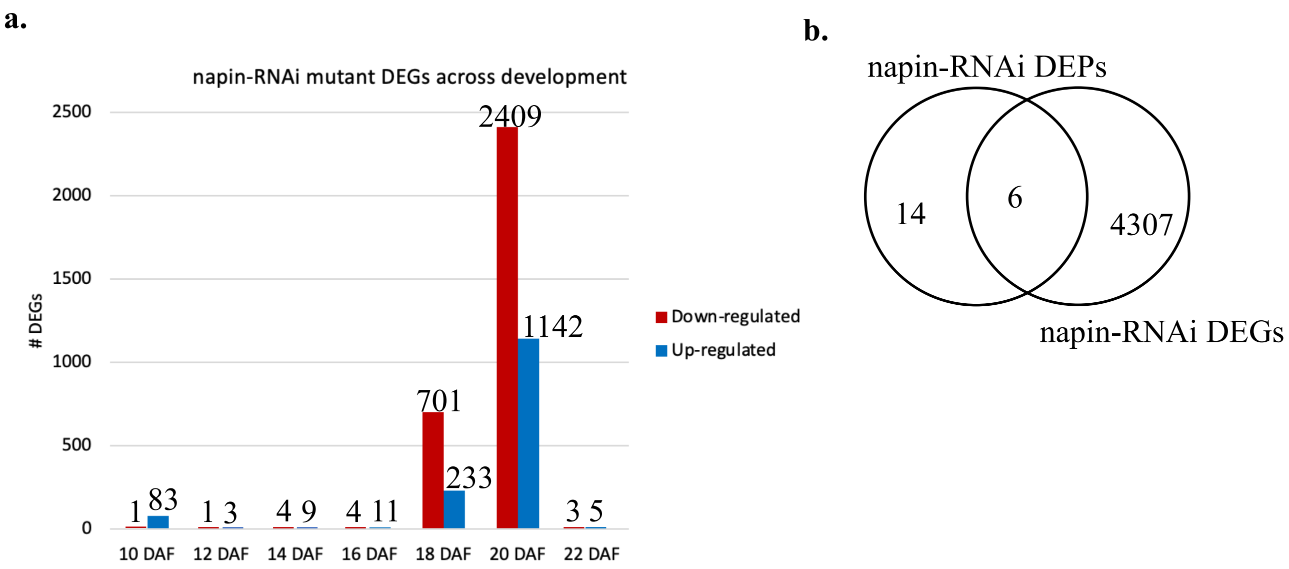


Figure 3. 6. **a.** Distribution of up-regulated and down-regulated differentially expressed genes (DEGs) from the napin-RNAi mutant at each developmental timepoint. **b.** Overlap between napin-RNAi DEGs and differentially expressed proteins (DEPs).

To determine the extent of shared gene expression, DEGs from the two mutants were compared. In both mutants, a very distinct trend in the number of DEGs across development is apparent; comparatively few DEGs were identified from 10-16 DAF, followed by a huge increase in 18-20 DAF, and then by a sharp decrease in DEGs in 22 DAF (Figure 3.5a; Figure 3.6a). The majority of the DEGs identified at 18 and 20 DAF decreased in abundance compared to WT; however, large numbers of up-regulated DEGs were also observed at these timepoints. More specifically, a large number of up-regulated DEGs were observed in the CRU- mutant as early as 14 DAF, going all the way until 20 DAF and were observed in the napin-RNAi mutant at 10 DAF and again at 18 and 20 DAF (Figure 3.65a; Figure 3.6a). Such large number of upregulated DEGs at 18-20 DAF are surprising as 18-20 DAF are considered stages of seed desiccation, where seeds start to turn yellow before completely drying out at 22 DAF. Additionally, this high number of DEGs in both mutants suggests the mutants are responding to the SSP alternation at the transcriptomic level; such changes are observed despite major alterations in the dry seed proteome only being observed in the CRU- mutant (Figure 3.4; Supplemental Table S3.6). The number of DEGs do however vary between mutants. Interestingly, 503 and 759 DEGs were identified in the CRU- mutant at 14 and 16 DAF, respectively, whereas in the napin-RNAi mutant only 13 and 15 DEGs are identified at these time points, respectively (Supplemental Table S3.9). Large transcriptomic responses appear delayed in the napin-RNAi mutant where a large number of DEGs are not identified until 18 DAF, while in the CRU- mutant a substantial number of DEGs are identified starting at 14 DAF.

In order to further define and compare DEGs across development, similarities and differences of mutant DEGs at each developmental timepoint were compared between the two mutants. Comparing all DEGs across development in the CRU- mutant versus the napin-RNA mutant showed an overlap of 2498 genes (Figure 3.7; Supplemental Table S3.11). An enrichment analysis using agriGO (Du et al., 2010) showed the top statistically enriched biological terms for the 2498 shared genes to be fatty acid biosynthetic process (GO:0006633 ; FDR 1.60e-12) (Supplemental Table S3.12a).

Surprisingly, analysis of the unique DEGs for the CRU- (3770 DEGs) and napin-RNAi (1815 DEGs) showed top statistical GO enrichments for peptide biosynthetic process (GO:0043043; FDR 3.72e-21) and transmembrane transporter activity (GO:0022857; FDR 0.0089), respectively (Supplemental Table S3.12b-c). A complete list of shared DEGs between the two mutants and their GO enrichments can be found in Supplemental Table S3.11 and Supplemental Table S3.12.

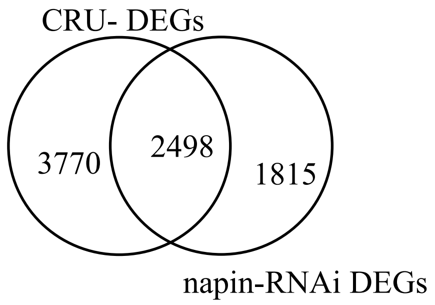


Figure 3. 7. Venn diagram showing overlap of unique DEGs from CRU- and napin-RNAi mutants across all of development. (FDR 0.05; fold-change 1.5)

Lastly to understand the potential interaction between the proteome and the transcriptome, I also analyzed overlap of DEGs across all developmental timepoints with DEPs from the dry seed for each mutant. Genes common to DEGs and DEPs datasets represent genes that deviate from the WT in gene expression and protein content, thus linking the two biological levels. A total of 360 genes were identified as both DEGs and DEPs in the CRU- mutant (Supplemental Table S3.13). Surprisingly, despite most of the DEGs being identified at 18 and 20 DAF, only 5% (217 genes) and 3% (83 genes) of these DEGs were also found as DEPs, respectively. In comparison, a much larger percentage of genes in the other timepoints were identified as DEGs and DEPs: 12 DAF: 24% (9 genes identified as DEGs and DEPs); 14 DAF: 16% (80 genes identified as DEGs and DEPs); 16 DAF: 19% (142 genes identified as DEGs and DEPs); 22 DAF: 12% (4 genes identified as DEGs and DEPs).

A GO enrichment analysis of all shared DEGs/DEPs across development showed a significant functional enrichment for response to cadmium ion (GO:0046686; 1.70e-39) arising from 68 DEGs/DEPs (Supplemental Table S3.14). In comparison, only 6 unique overlapping DEGs/DEPs were identified in the napin-RNAi mutant (Supplemental Table S3.13). For the napin-RNAi mutant, the percent of total DEGs also found as DEPs in each developmental timepoints was low, with 15% (2 genes identified as DEGs and DEPs) being identified at 14 DAF, but these low numbers are mainly due to the low number of DEPs. Although further analysis of these datasets is warranted and additional developmental proteomics data would make for more robust comparisons, in this study, I mainly focus on the intersection of the DEGs from the two SSP mutants to the GWAS to filter down high priority candidate genes.

3.3.5 *576 High Confidence Candidate Genes (HCCGs) Were Identified by Overlapping FarmCPU GWAS Candidate Genes with DEGs Across Mutant Development*

Candidate genes arising from GWAS and transcriptomics datasets were compared for shared gene membership. The rationale behind such comparison is that since both the GWAS analyses and the SSP mutant transcriptomic analyses sought to identify underlying genetic control of PBAA composition in the seed, comparison of the two resulting gene lists should result in a nonrandom list of high confidence genes; such genes will be referred to from here on out as high confidence candidate genes (HCCGs).

DEGs from the CRU- and napin-RNAi mutants at each developmental timepoint were compared with the 1521 candidate genes from FarmCPU GWAS. For the CRU- mutant, a total of 335 genes were shared with GWAS candidate genes at five time points (Figure 3.8a; Supplemental Table S3.15). Most of the HCCGs originated from 18 DAF (244 genes), but HCCGs were also identified in 12 DAF (1 gene), 14 DAF (21 genes), 16 DAF (36 genes), and 20 DAF (134 genes). No HCCGs were identified from 10 DAF and 22 DAF. To determine if there was an enrichment for any biological function, agriGO (Tian et al., 2017) was used to complete a GO enrichment analysis using the HCCGs. Down-regulated HCCGs showed biological enrichment only at 18 DAF (GO:0044699; single-organism process; FDR 0.012) (Supplemental Table S3.16d). Up-regulated HCCGs showed biological enrichment at 16 DAF (GO:0043043; peptide biosynthetic process; FDR 0.0024) (Supplemental Table S3.16b). A full list of all significantly enriched GO terms for shared increased and decreased CRU- DEGs with GWAS candidate genes, refer to Supplemental Table S3.16.

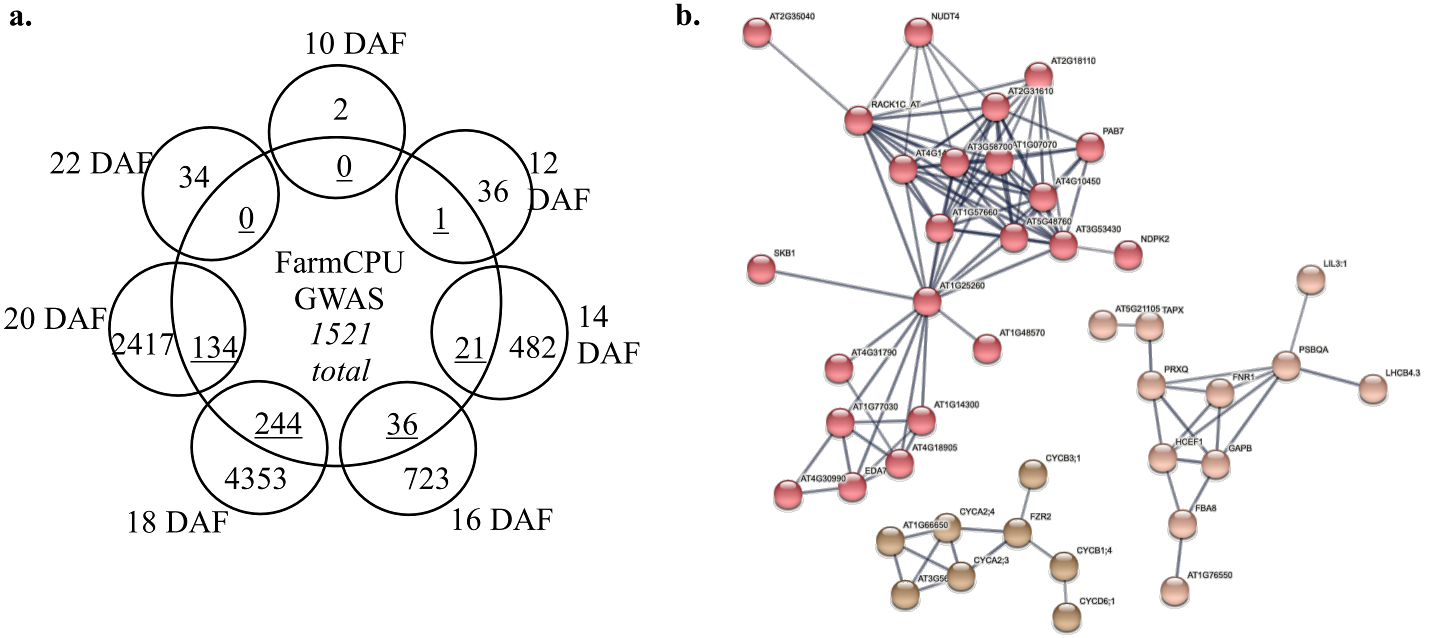


Figure 3. 8. Analysis of overlapping DEGs from the CRU- mutant gene expression analysis and FarmCPU candidate genes. Shared genes are referred to as HCCGs. **a.** Venn diagram showing overlap of unique CRU- DEGs and GWAS FarmCPU candidate genes at each timepoint in development. (FDR 0.05; fold-change 1.5). **b.** PPI of the 335 HCCGs. Proteins are indicated by nodes and interaction between nodes are represented by edges. Smooth edges indicate intra-cluster interaction. Cluster analysis using MCL algorithm (1.5) was used. Only connections of high confidence (interaction score of >0.7) are visualized.

A similar comparison analysis was completed for the napin-RNAi DEGs. Overall, fewer HCCGs were identified for the napin-RNAi mutant compared to the CRU- mutant. A total of 241 HCCGs were identified. Unlike the CRU- mutant, most of HCCGs originated from 20 DAF (204 genes); overlap was also observed in 10 DAF (6 genes) and 18 DAF (45 genes) (Figure 3.9a; Supplemental Table S3.15). Interestingly, only decreased HCCGs from 18 DAF showed a significant biological GO enrichment for single-organism process (GO:0044699; FDR 0.0076). (Supplemental Table S3.16h). GO enrichment analysis of all 241 shared HCCGs showed a statistical enrichment for fatty acid biosynthesis (49 genes) (GO:0006633; FDR 5.20e-07) (Supplemental Table S3.16d). A full list of all significantly enriched GO terms for shared up and down-regulated napin-RNAi DEGs with GWAS candidate genes, refer to Supplemental Table S3.16.

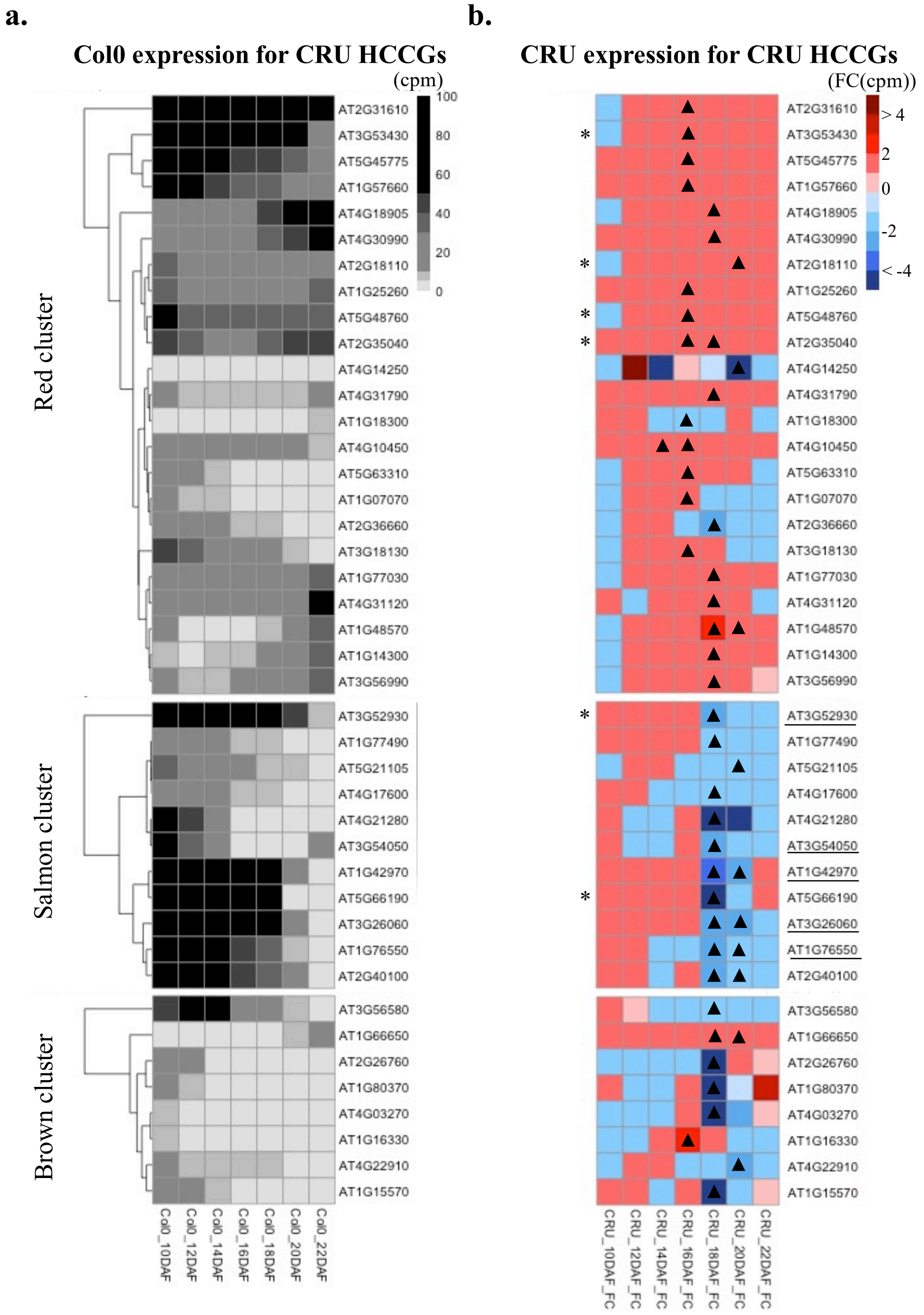


Figure 3. 9. Expression of 42 HCCGs from the CRU- PPI red, salmon and brown clusters. **a.** Col0 expression, represented as cpm, across development for the 42 HCCGs. Within each PPI cluster, genes are clustered based on gene expression, **b.** CRU- expression, represented as FC(cpm) (Col0/CRU-), across development for the 42 HCCGs. Underlined genes represent HCCGs shared between the CRU- and napin-RNAi PPI networks. Asterix (*) designate genes that are also found as CRU- DEPs. Triangles represent genes that are found as CRU- DEGs. Abbreviations: FC = fold-change; cpm = counts per million; HCCGs = high confidence candidate genes.

3.3.6 *STRING Protein-Protein Interaction Network of HCCG Show Distinct, High Confidence Modules*

To understand potential functional interaction by identifying clusters of highly interconnected nodes in each set of HCCGs, high confidence PPI networks were constructed and visualized for the CRU- (335 HCCGs) and napin-RNAi (241 HCCGs) mutants, respectively, using the Search Tool for the Retrieval of Interacting Genes/Proteins database STRING (V11.0) (Szklarczyk et al., 2019) (Figure 3.8b and Figure 3.9b). To decrease false positive interactions, networks were filtered so that only high confidence interaction scores remained (>0.7) (interaction score refers the probability an interaction between nodes describes an actual functional linkage between two proteins); additionally, all modules were filtered to contain at least three nodes. The resulting PPI network for each respective mutant was analyzed individually.

The PPI network of CRU- HCCGs consisted of 334 nodes (77 were connected at least one other protein, and 257 were unconnected) and 130 edges. The MCL clustering algorithm in STRING resulted in three high confidence modules. The red cluster contained 23 genes and was the largest cluster; the salmon cluster has 11 genes; and the brown cluster had eight genes, representing the smallest cluster (Figure 3.8b; Supplemental Table S3.17). To assess any functional characterization, a GO enrichment analysis was performed for each cluster. Strikingly, the top significant GO enrichment for the red cluster was organonitrogen compound metabolic process (GO:1901564; FDR 1.00e-05), but enrichment for 'translation' (GO:0006412; FDR: 3.00e-05) was also identified (Supplemental Table S3.18a). The top biological GO enrichment for the salmon cluster was photosynthesis (GO:0015979; FDR 8.00e-11) (Supplemental Table S3.18a). Since there were less than 10 genes in the brown module, a GO enrichment could not be done in agriGO, however, visual inspection of the proteins showed that five proteins were annotated as cyclin genes (Supplemental Table S3.18a).

Next, to see how the CRU- HCCGs in the red, salmon, and brown clusters behaved in the CRU- mutant and the WT genotypes, gene expression for the HCCGs were visualized in a heatmap where WT expression patterns were plotted in counts per million (cpm) (Figure 3.9a) and CRU- expression was plotted in FC (CRU-/WT) (Figure 3.9b). In the WT as to be expected, most of the HCGGS started with higher expression early in development and gradually decreased as the seed senesced (Figure 3.9a). FC in the CRU- mutant for the 42 HCCGs ranged from 4.89 FC (AT4G14250 annotated as Putative plant UBX domain-containing protein 14; Structural constituent of ribosome) (red cluster). to -43.17 FC (AT2G26760 annotated as Cyclin B1;4 (CYCB1;4); functioning as cyclin-dependent protein kinase regulator activity) (brown cluster). A full list of HCCGs and their description for the red, salmon, and brown modules can be found in Supplemental Table S3.17. Closer analysis of expression dynamics in the WT and CRU- mutant for the ribosomal genes in the red cluster showed expression levels to deviate substantially in the CRU- mutant from about 12-20 DAF, especially at 12-16 DAF where most of the genes increase drastically in expression compared to the WT (Supplemental Figure S1a-l). In comparison, analysis of cpm data for the cyclin genes in the brown cluster showed smaller deviations from the WT (Supplemental Figure S1m-t).

Lastly, to further increase confidence in a smaller number of genes, overlapping HCCGs in the red, salmon, and brown modules with the CRU- DEPs were identified. Six genes were identified and are as follows: AT3G53430 (Ribosomal protein L11 family protein), AT5G48760 (Ribosomal protein L13 family protein), AT2G35040 (AICARFT/IMPCHase bienzyme family protein), AT5G66190 (Ferredoxin--NADP reductase, leaf isozyme 1, chloroplastic), AT3G52930 (Fructose-bisphosphate aldolase 8, cytosolic), and AT2G18110 (Translation elongation factor EF1B/ribosomal protein S6 family protein) (Figure 3.9b).

A second PPI network was also created using the napin-RNAi HCCG and consisted of 241 nodes (36 were connected at least one other protein, and 205 were unconnected) and 32 edges. The MCL clustering algorithm in STRING resulted in three high confidence modules. The red cluster contained 12 genes and was the largest cluster; the brown and dark golden clusters had 3 genes each (Figure 3.10b; Supplemental Table S3.17). GO enrichment analysis of each cluster showed the top significant GO enrichment as generation of precursor metabolites and energy in the red cluster (GO:0006091; FDR 3.10e-09) (Supplemental Table S3.18b). Since there was less than 10 genes in the brown and dark golden cluster, a GO enrichment could not be done in agriGO (Supplemental Table S3.18b).

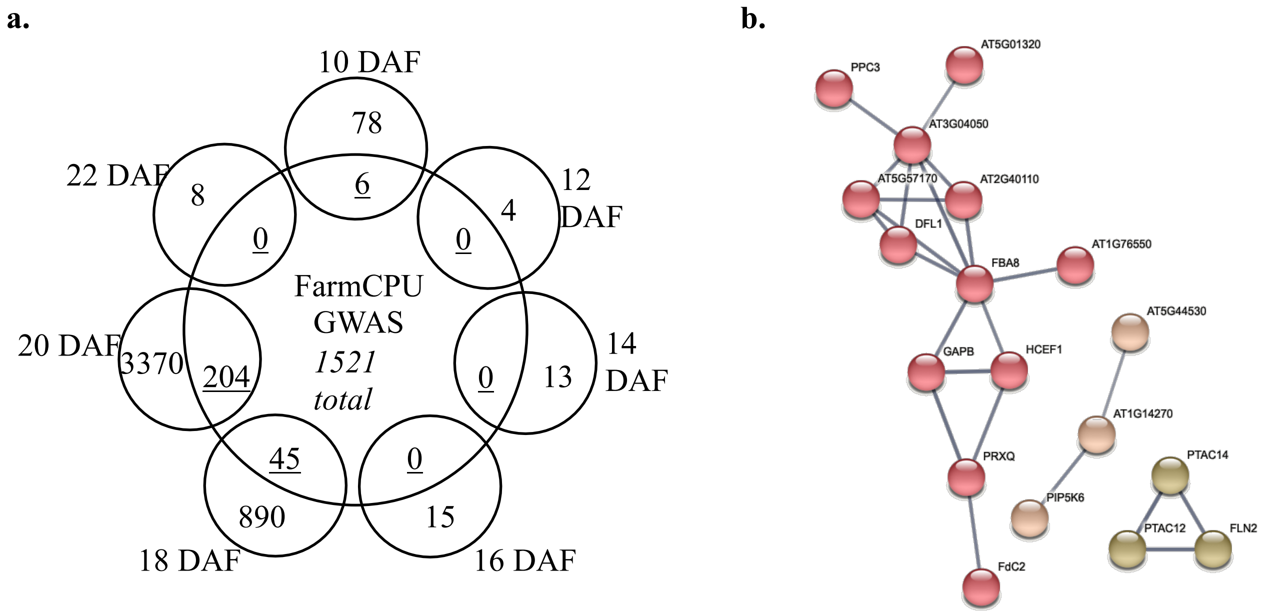


Figure 3. 10. Analysis of overlapping DEGs from the napin-RNAi mutant gene expression analysis and FarmCPU candidate genes. Shared genes are referred to as HCCGs. **a.** Venn diagram showing overlap of unique napin-RNAi DEGs and GWAS FarmCPU candidate genes at each timepoint in development. (FDR 0.05; FC 1.5), **b.** PPI of the 241 HCCGs. Proteins are indicated by nodes and interaction between nodes are represented by edges. Smooth edges indicate intra-cluster interaction. Cluster analysis using MCL algorithm (1.5) was used. Only connections of high confidence (interaction score of >0.7) are visualized.

Next, visualization of HCCG from the three modules were analyzed for their expression patterns in the WT and napin-RNAi mutant where WT expression patterns were plotted in cpm (Figure 3.11a) and CRU- expression was plotted in FC (Figure 3.11b). Similar to the CRU- HCCGs, some HCGGS started with higher expression early in development and gradually decreased in expression as the seed senesced (Figure 3.11a). However, unlike the CRU- HCGGs, approximately one-third of the HCCGs showed fairly constitutive expression across development. FC (WT/CRU-) in the napin-RNAi mutant for the 18 HCCGs ranged from 2.69 FC (AT5G54510; red cluster) to -9.25 FC (AT5G44530; brown cluster). Both genes were found at 20 DAF. AT5G54510 is annotated as an Indole-3-acetic acid-amido synthetase GH3.6. AT5G44530 is annotated as Subtilisin-like protease SBT2.3. A full list of HCCGs and their description for the red, salmon, and brown clusters can be found in (Supplemental Table S3.17). No HCCGs in the red, brown, or dark golden clusters overlapped with napin-RNAi DEPs identified in the proteomics analysis. Closer analysis of the cpm levels in the napin-RNAi mutant and the WT showed quite similar expression trends across development with relatively small deviations occurring primarily at 18 and 20 DAF (Supplemental Figure S1u-y).

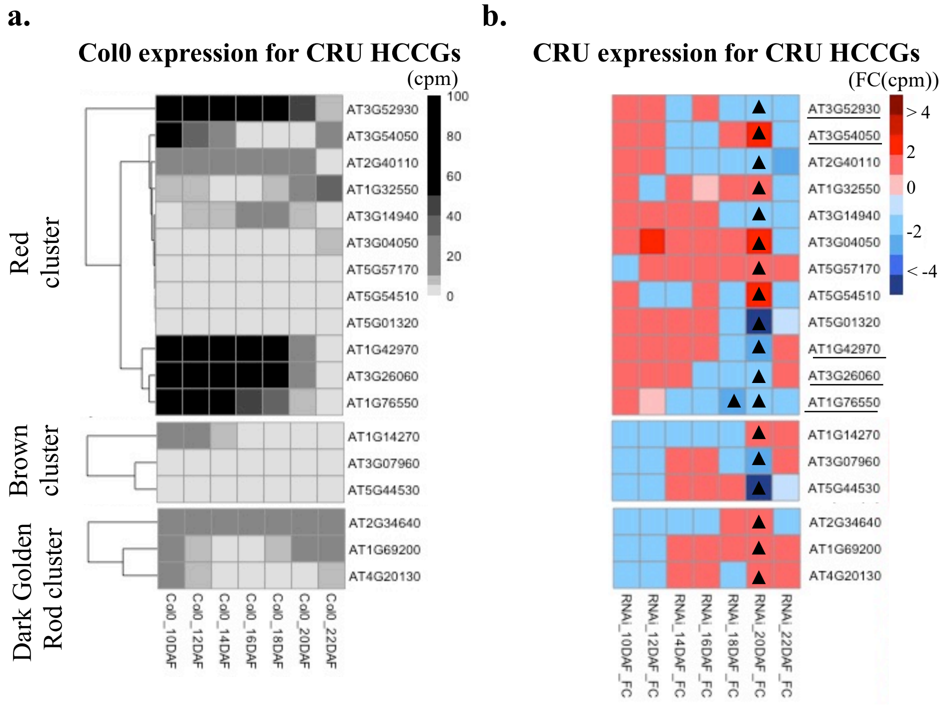


Figure 3. 11. Expression of 18 HCCGs from the napin-RNAi PPI red, brown and dark golden rod clusters. **a.** Col0 expression, represented as cpm, across development for the 18 HCCGs. Within each PPI cluster, genes are clustered based on gene expression, **b.** CRU- expression, represented as FC(cpm) (Col0/CRU-), across development for the 18 HCCGs. Underlined genes represent HCCGs shared between the CRU- and napin-RNAi PPI networks. No shared HCCGs with napin-RNAi DEPs. Triangles represent genes that are found as napin-RNAi DEGs. Abbreviations: FC = fold-change; cpm = counts per million; HCCGs = high confidence candidate genes.

3.4 DISCUSSION

The first step in making meaningful contributions to seed amino acid biofortification is to have a deeper understanding of the regulatory mechanisms underlying PBAA composition in the seed. In this study I shed light on key, putative mechanisms by integrating orthogonal datasets. I am able to do this by capturing both the natural variation in PBAA composition that likely resulted from natural selection across evolutionary history, in addition to exploiting omics data from two SSP actively showing PBAA rebalancing. The datasets created in this study were greatly thought-out and implemented. Presented analyses are simply the vital first steps in adequately harnessing all the information such comprehensive datasets offer. Results shed light on putative mechanistic controls, translational machinery and the cell cycle, as avenues for future inquiry.

3.4.1 *GWAS Identified Many Significant SNP-trait Associations, However Could Not Alone Highlight Key Regulatory Genes or Mechanisms*

Despite tight regulation of PBAA composition in the seed, PBAA composition has been previously documented in natural and artificial populations of maize and other crops (Deng et al., 2017; La et al., 2019). In this study, I demonstrate significant natural variation in dry seed PBAA composition displayed across the 1001 *Arabidopsis* diversity panel (Alonso-Blanco et al., 2016). Correlation among PBAA absolute levels show substantial significant correlation that is almost exclusively positive (Figure 3.1b). It has been proposed that these high correlations are a result of tightly regulated amino acid metabolism that is at least partially controlled by incorporation into SSP (Amir et al., 2018). In this study, I performed GWAS on absolute and compositional PBAA traits, as well as the PBAA metabolic ratio related traits, that represent biochemical relationships within the amino acid biochemical families. Similar approaches have been successfully used in past studies (Ruthie Angelovici et al., 2013; Deng et al., 2017; M. L. Slaten et al., 2020). For example in (M. L. Slaten et al., 2020), derived ratio traits using the Glutamine Family FAA in dry *Arabidopsis* seed were used and collectively analyzed to identify novel SNP-trait associations when absolute values alone could not. Significant SNP-trait associations have also been identified in other work when absolute levels of metabolites failed to do so (Alexander E Lipka et al., 2013; M. L. Slaten et al., 2020; Adam M Wentzell et al., 2007), demonstrating the benefit of using more than just absolute levels. GWAS analysis using the biochemical traits in this work yielded a large number of SNP-trait associations and relevant candidate genes (Supplemental Table S3). Such a large number of candidate genes may reflect the complex nature of PBAA composition that is highly connected with the overall metabolome in the seed. Also, it has been suggested that FarmCPU may be more permissive than other GWAS models (unpublished collaborator's discussions) and results in a higher number of SNP-trait associations (and type-I error). To help correct for this, the Bonferroni adjustment was used. Nonetheless, since the candidate genes are compared with orthogonal datasets in downstream analysis, a resulting large list of candidate genes is less concerning.

A total of 1652 significant SNP-trait associations were identified across all 274 analyzed PBAA traits, with 987 SNPs being unique across all traits (Table 3.1). Identification of duplicate SNPs is to be expected as only 16 absolute traits are used to create a total of 274 derived ratio traits; additionally, absolute traits share biochemical families and thus similar biological pathways. A total of 1521 candidate genes resulted nevertheless in no significantly enriched biological process. Functional characterization of these genes showed the top three categories were of protein (21%), RNA (14%), and signaling (11%) (Figure 3.2). These results suggest that alterations in proteins, particularly at the translational level, may be underlying PBAA regulation. Interestingly, previous reports using PBAA derived traits in maize also showed no statistically significant GO enrichment among GWAS candidate genes, but identified the same top functional characterization of protein, RNA, and signaling (Shrestha, 2020). Such results suggest that further filtering of the GWAS candidate genes in a biologically meaningful way is required to further pinpoint regulatory biological processes.

3.4.2 *CRU- and napin-RNAi Metabolome Reflect Rebalancing, but Show Distinct Dry Seed Proteomes*

The rebalancing mechanism was characterized only in the dry seed of the CRU- and napin-RNAi mutants. Overall, the mutants were characterized by small changes in the metabolome as expected; however, the CRU- had large changes in the proteome while the napin-RNAi had only a few proteomic changes (Figure 3.4; Supplemental Table S3.6). The proteomic changes in the respective mutants were consistent with the metabolic results which showed only His and Lys PBAA compositions to be statistically different in the CRU- mutant compared to the WT and no statistical differences in the napin-RNAi mutant (Figure 3.3b). Increased Lys concentration following SSP perturbations have been previously documented in the *opaque-2* maize mutant (Hunter et al., 2002). Proteomic analysis of six maize mutants found that the increased Lys was the result of both general increases in non-zein proteins and a more specific increase of proteins with relatively higher content of Lys (Morton et al., 2016). In this study, the proteomic data is consistent with the latter hypothesis showing general increases in many proteins, as well as, more specific and pronounced elevation in others (Figure 3.4; Supplemental Table S3.6).

It is still unclear why the CRU- had such an extensive effect on the proteome while the napin-RNAi did not. The lack of proteomic changes observed in the napin-RNAi mutant suggest that rebalancing is different to some degree between the two mutants. It is possible that the CRU- proteomic perturbation is simply much more severe, as the CRU proteins comprise approximately 60% of the total seed amino acids while the napin proteins comprise approximately 20% of the total protein amino acids, and this is reflected in the dry seed proteome. Thus, the rebalancing response may differ not only based on genotype (CRU- versus RNAi-napin), but also on the extent of perturbation (60% versus 20%). The fate of each mutant proteome may be decided early in development and thus the large number of downregulated DEGs could simply be degradation of unused RNA. Collectively, the metabolome and proteome suggest that the two type of proteomic alterations investigated in the CRU- and napin-RNAi mutants elicited two different responses in the seed. Further transcriptomic analysis should shed more light on the differences and similarities of these responses.

3.4.3 *Developmental Transcriptome Analysis of CRU- and napin-RNAi Mutants Show Large Responses in Late Seed Development*

Previous studies have showed conflicting reports of transcriptomic changes in response to SSP perturbations (Schmidt et al., 2011; Tzin & Galili, 2010). For example in several opaque mutants in maize, large changes in proteins were coupled with large changes in gene expression (Hunter et al., 2002). It was concluded that multiple mechanisms led to the maize phenotype, each with differing effects in the overall gene expression landscape

(Hunter et al., 2002). However in soybean, RNAi silencing of storage proteins resulted in increases in compensatory proteins with no upregulation of transcript abundance (Herman, 2014). In the current study, transcriptomic responses were identified in both mutants across development.

A total of 6259 and 4290 DEGs were identified in the CRU- and napin-RNAi mutant, respectively (Figure 3.5a and Figure 3.6a). GO enrichment analysis of all the up-regulated CRU- DEGs across development showed a biological enrichment for 'translation'. 'Translation' was also highly significant among up-regulated DEGs at 16 DAF with an FDR of 1.20e-120 (Supplemental Table S3.10g). Interestingly, 'translation' was not found to be significantly enriched in any napin-RNAi DEG lists, further differentiating the effects of the two SSP mutations and suggesting that the mechanisms of rebalancing are not identical in all mutants. It is possible that the CRU- specific up-regulation of genes (leading to the 'translation' enrichment) is manifesting later in development and contributing to the larger proteomic perturbations identified only in the CRU- dry seed. Previous work comparing candidate genes from a mutant analysis with GWAS candidate genes in maize yielded a similar 'translational' enrichment (Shrestha, 2020); it was suggested that ribosomal heterogeneity may be a key player in shaping PBAA composition in the seed.

The large number of up-regulated DEGs identified at 18 and 20 DAF was identified in both mutants (Figure 3.5a; Figure 3.6a). These results were surprising because this stage of development typically marks the beginning of senescence where the seeds turns yellow and transcriptions decreases (Baud et al., 2002; Boyes et al., 2001). In *Arabidopsis*, previous experimentation show that the majority of SSP accumulate from 10 to 17 DAF

(Baud et al., 2002). Therefore, it is possible that the large number of DEGs observed in this experiment at 18 and 20 DAF are a compensatory response to the lack of SSP, further supported by the significant GO enrichments for a variety of different biosynthetic pathways (Supplemental Table S3.10b,c). These results further support the idea that different processes and pathways are responding and being activated at this point in development.

In this experiment, there is a large disconnect in DEGs occurring across development in the transcriptome with DEPs observed in the dry seed proteome in both the CRU- and napin-RNAi mutants. Only 360 (41% of all DEPs; 6% of all DEGs across development) and 6 genes (13% of all DEPs; less than 0.1% of all DEGs across development) in the CRU- and napin-RNAi mutants, respectively, were identified as both DEGs across all of seed development and DEPs in the dry seed (Figure 3.5c; Figure 3.6c). Previous reports in *Arabidopsis* seed comparing the proteome versus the transcriptome have yielded varied results ranging from mild correlation to large discordance between transcripts and proteins (Griffin et al., 2002; Hajduch et al., 2010; Le Roch et al., 2004; Mergner et al., 2020). In (Hajduch et al., 2010), analysis of five stages of *Arabidopsis* seed filling showed an overall concordance rate of protein to transcript of 56%. It was suggested that certain intermediary metabolic processes were particularly discordant due to hypothesized post-transcriptional regulation of core metabolism during seed development (Hajduch et al., 2010). Among the shared DEGs/DEPs in CRU- mutant were metabolically specialized genes such as those participating in the flow of carbon from sucrose into amino acids including AT4G38970 (Fructose-bisphosphate aldolase 2), AT3G55440 (ATCTIMC, cytosolic isoform triose phosphate isomerase), AT3G26650 (GAPA, glyceraldehyde 3-phosphate dehydrogenase subunit), and AT5G52920 (PKP-BETA1, PKP1, plastidial pyruvate kinase 1). Although a true statistical correlation test is required to determine significant associations between transcripts and proteins in this study, this crude analysis mimics similar reports in *Arabidopsis* seed filing where these same metabolically specialized genes were identified with more statistical rigor (Hajduch et al., 2010). Possible causes for the disconnect between transcripts changing throughout development and altered proteins in the dry seed could be post-transcriptional control of protein translation through increased transcript degradation and/or inactivation of translation (elBaradi et al., 1986). A better understanding of the concordance/disconcordance of proteins and transcripts in these mutants across development may shed light on which regulatory mechanisms are controlling PBAA composition.

3.4.4 *PPI Network of HCCGs Show Tightly Connected Clusters with Overrepresentation for Translation and the Cell Cycle*

Orthogonal datasets comprised of candidate genes from FarmCPU GWAS and DEGs from the CRU- and napin-RNAi developmental transcriptomic analysis were compared for overlapping genes, termed HCCGs. In this analysis, a total of 335 and 241 HCCGs were identified for the CRU- and napin-RNAi mutant, respectively (Figure 3.8; Figure 3.10). Classification of HCCGs revealed interesting putative regulatory mechanisms.

A PPI network analysis of the HCCGS from the CRU- and napin-RNAi mutants revealed high confidence clusters in each network. Significant enrichment for the three GO categories (biological, molecular, and cellular) of the CRU- red cluster reiterated that alterations are occurring at the translational level with enrichment for 'translation', ‘structural constituent of ribosome’, ‘RNA binding’, and additional terms associated with ribosomal subunits/complexes (Supplemental Table S3.18a). Analysis of the gene expression dynamics of the ribosomal genes across development showed many of the ribosomal genes to be drastically different (mainly up-regulated) in the CRU- mutant from 12-20 DAF compared to the WT (Supplemental Figure S1). For example, AT2G31610 (annotated as Ribosomal protein S3 family protein; function is described as structural constituent of ribosome; involved in response to salt stress, translation, response to abiotic stimulus) showed opposite expression trends from the WT, increasing from 12-16 DAF, and decreasing from 16-18 DAF, to then closely parallel the WT expression at the dry seed stage (Supplemental Figure S1h). These stark changes in gene expression that deviate from the WT may suggest metabolic shifts occurring in the seed. Previous reports have demonstrated ribosomal proteins being overexpressed during key developmental transitions in soybean (Shamimuzzaman & Vodkin, 2014). Taken collectively, these results again reiterate that translational machinery and RNA synthesis may be at the core of PBAA rebalancing.

Also of particular interest in the CRU- PPI network was the brown cluster. The cluster had a total of eight genes associated with the cell cycle (Supplemental Table S3.18a). Interestingly, at 18 DAF, six of the eight genes in the brown cluster are differentially expressed (Figure 3.9). These are interesting results as by 18 DAF limited cell division is occurring. Therefore, there may be little relevance to these results in terms of cell division, but they do suggest that some adjustments of the cell cycle and/or protein degradation may be occurring. Analysis of the gene expression dynamics of the cyclin genes across development showed deviations from the WT to be much less dramatic than those observed in the ribosomal genes in the red cluster (Supplemental Figure S1). Nonetheless, the identification of such a high confidence cluster of genes identified in orthogonal datasets warrant mentioning as an interesting result.

3.5 CONCLUSIONS

This study was successful in furthering the limited knowledge base of regulation of PBAA composition in the *Arabidopsis* seed. This study successfully showed that a systems biology, multi-omics approach can be taken to successfully prioritize genes and shed light on biological processes affecting PBAA composition. Analysis of the gene expression landscape of two SSP and a WT throughout seed development showed that transcriptional responses are indeed occurring, albeit at a surprisingly delayed developmental stage. Comparing GWAS and transcriptome orthogonal datasets results suggest shared regulatory control between protein homeostasis and PBAA rebalancing. Such is a novel finding that should be explored more in the future. Furthermore, PPI interaction networks reveal putative regulatory mechanisms controlling PBAA composition in the seed. Additional work is needed to further define rewiring in mutant networks and more advanced statistical models should be implemented to define the relationship between the proteome and transcriptome across development. This study marks the creation of a comprehensive dataset and the first pass of analysis that yielded interesting results that could be further harnessed to advance biological discovery.

3.6 SUPPLEMENTAL INFORMATION

Supplemental Figure 3.1, Supplemental Tables S3.1-S3.18, Supplemental Datasets S3.1-S3.4, and Supplemental Document 3.1 are available

through Microsoft OneDrive:

https://mailmissouri-my.sharepoint.com/:f:/g/personal/mleww8_umsystem_edu/Egv_LCuhTURPvitCsCUf_GoBhzAQicDpq_W23nPlnCnKIg?e=yLiKb7

3.7 MATERIALS AND METHODS

Plant growth and seed collection

All *Arabidopsis* (*Arabidopsis thaliana*) genotypes were grown in controlled growth chambers at 22°C/24°C (day/night) under long-day conditions (16 h of light/8 h of dark).

The 1001 *Arabidopsis* diversity panel (Alonso-Blanco et al., 2016) was grown as two independent replicates under the before-mentioned conditions. Dry seeds were collected at the end of the desiccation period and stored in a desiccator at room temperature. Due to delays in flowering and inadequate seed-set, the seed of 869 accessions were analyzed for PBAA content.

For the SSP project, the three genotypes (Col-0, CRU-, and RNAi (Withana-Gamage et al., 2013)) were grown in three independent replicates (i.e. grow-outs). Siliques were harvested at 10, 12, 14, 16, 18, 20, 22 days after flowing (DAF). Upon harvest, siliques were flash frozen and stored in the -80°C freezer. Seeds were extracted from siliques over liquid nitrogen under a magnifying glass using tweezers. Harvested seeds remained in the -80°C freezer until protein and RNA extractions. Dry seeds (22 DAF) were harvested and stored in a desiccator at room temperature.

Metabolic and proteomics analyses

For the 1001 population and SSP mutants, dry seeds were collected at the end of the desiccation period and stored in a desiccator at room temperature (see *Plant growth and seed collection* of "Materials and Methods" for more details) and PBAA composition and proteome was analyzed. For the 1001 population, seed from two individual grow-outs were collected and replicates (2) from within each grow-out were combined for analysis per accession. For the SSP mutants, seed from the three genotypes from three individual grow-outs were collected in replicates of three. Replicates from each grow-out were combined per genotype for analysis.

For metabolic analysis, the PBAA natural variation in dry seed AA composition was characterized using a protocol developed in the Angelovici lab in which a microscale, 96 well plate, high throughput extraction protocol is used to extract and quantify PBAA using a high-resolution LC MS/MS-MRM protocol as described in (Ruthie Angelovici et al., 2013; Abou Yobi & Ruthie Angelovici, 2018). In the 1001 population, each accession was measured once per grow-out (for a total of two PBAA replicates per accession). For the SSP mutants, each genotypes was measured in two replicates from each of the three grow-outs (for a total of six replicates per genotype)

For proteomics analysis, proteins were extracted from dry seed as described in (Hurkman & Tanaka, 1986; Abou Yobi & Ruthie Angelovici, 2018). Dry seeds were finely ground and weighed into 3 mg samples. The samples were extracted with Tris-HCl buffered phenol and SDS extraction buffer, followed by trypsin digestion and purification. The peptides were analyzed on a Bruker trapped ion mobility spectrometry time-of-flight tandem mas spectroscopy (timsTOF Pro MS/MS) platform. Resulting data underwent peak selection using the PEAKS DB search engine (version 8.5, Bioinformatics Solutions Inc.). Proteins with spectral counts equal to 0, had less than or equal to three spectral count in both mutants, duplicated proteins and proteins with a coefficient of variance (CV) greater than 30 (even when one outlier replicate was removed) were removed from the analysis. Fold change for each protein was calculated per the mutant using the following equation: [positive FC = (mutant/WT)] and [negative FC = (-WT/mutant)]. Significant differentially expressed proteins were identified using FDR 0.1 threshold. For the SSP mutants, seed from the three genotypes were measured in one replicate from each grow-out (for a total of 3 replicates per genotype).

Proteomics SSP Mutant Fold-change plots

Fold-change (FC) in protein quantifications in the DEPs identified in the CRU- and napin-RNAi mutants, respectively, were visualized in a heatmap. FC was calculated using the following equations: [positive FC = (mutant/WT)] and [negative FC = (-WT/mutant)].

Heritability calculation

Heritability estimates were calculated from HAPPI GWAS (Marianne L. Slaten et al., 2020) output (raw data after outliers removed). A linear mode is fit with Line as a random effect using the nlme v3.1-152 package in R.

PBAA correlation analysis

Pearson pairwise correlation analysis was completed using back-transformed BLUEs for all 16 PBAA traits using corr.p function from the psych R package v2.0.12 with options adjust = “fdr” and alpha=0.05. Subsequent correlation (r) and p-value (p) matrices were used in visualization using the corrplot function from the corrplot v0.84 R package.

mGWAS analysis

mGWAS was performed using the genotypic data from a resequencing effort of the 1001 Genomes Accessions (Alonso-Blanco et al., 2016). SNPs were filtered at a MAF >=0.5 and a SNP call rate of 10%, resulting in 1,057,383 biallelic SNPs remaining (from the 10,707,430 original SNPs). The 16 PBAA absolute traits were used to create a total of 274 biochemical traits. PBAA traits were first grouped into their respective biochemical family (Aspartate Family, BCAA Family, Glutamate Family, Pyruvate Family, Serine Family, and Shikimate Family) and then used to make combinatorial ratios traits, where each PBAA trait would serve as the numerator with every combinatorial combination of all the PBAA in the family as the denominator. In addition to the ratio traits, the absolute value traits, as well as the compositional traits (PBAA/Total) were run. Analyses were performed on the 274 traits using HAPPI GWAS (Marianne L. Slaten et al., 2020). In short, HAPPI GWAS incorporated the following pre-GWAS, GWAS and post-GWAS analyses:

Pre-GWAS: Metabolic data was subject to quality control check and preprocessing, including outlier removal (Kutner et al., 2005) and Box-Cox transformation (Box & Cox, 1964; Fox et al., 2012). Best Linear Unbiased Estimates (BLUEs) were calculated to combine replicates across multiple grow-outs; genotype was used as a fixed effect in a linear model.

GWAS: Association studies were performed using FarmCPU (Liu et al., 2016) and GAPIT (Lipka et al., 2012) R packages. FarmCPU SNP-trait associations were filtered using a Bonferroni correction (0.05/#SNPS) to control the experiment wise type I error rate at α = 0.05 was used. A mixed linear model (MLM) was used in GAPIT, fitting population structure (fixed effect) and kinship (random effect) (Lipka et al., 2012); an FDR correction to control the experiment wise type I error rate at α = 0.05 was used.

Post-GWAS: In order to identify SNPs in strong linkage disequilibrium (LD) with each significant SNP, haploblock analysis was performed using Haploview (Barrett et al., 2004). Previously calculated average LD decay in *Arabidopsis* (10kb) (Sung Kim et al., 2007) was used in Haploview to calculate pair-wise LD with the significant SNP and every neighboring SNP 5kb upstream and 5kb downstream. Here haploblocks are defined by an r^2 threshold of 0.8. Any genes contained, or partially contained, in the haploblock containing the significant SNP are saved as putative genes of interest. If no haploblock is identified for a respective SNP, the gene immediately upstream and downstream of the SNP will be saved. Thus, the pipeline should conclude with a list of putative genes of interest to be further analyzed downstream.

RNA extraction

Total RNA extracted from dry and developing seeds was isolated using a hot borate method (Birtić & Kranner, 2006) and purified using Direct-zol RNA Miniprep Plus filter columns (Zymo Research). Samples were subject to sequencing using the NovaSeq platform, resulting in library sizes of approximately 50 million, 100-bp paired-end reads (per sample). This work was performed with the support of the University of Missouri DNA Core Facility.

Gene expression analysis

RNAseq data was trimmed for quality control before analyses. In short, for reads whose 3' ends overlap with the adapter for a minimum of 3 bases with 90% identity, were trimmed at the 3` end for Illumina adapters, for ambiguous nucleotides (N's) and for poly-G artifacts arising from Illumina two-color chemistry using cutadapt version 1.18 (Martin, 2011). After trimming, reads shorter than 10 bp were discarded (with their associated paired read). Reads were then aligned to the TAIR10 genome assembly (Athaliana_447_TAIR10.fa.gz) using STAR v2.7.6a (Dobin et al., 2013). In brief, STAR genomeGenerate mode, TAIR 10 genome assembly and annotation (Athaliana_447_TAIR10.fa.gz and TAIR10_GFF3_genes.gtf) were used to generate genome indices. Reads were then aligned with STAR using --twopassMode Basic, --outSAMtype BAM SortedByCoordinate, and --quantMode GeneCounts options. Reads were quantified on the gene level. Mitochondrial and chloroplast genes were filtered out as contaminates.

Differential gene expression analysis was completed using the glm function quasi-likelihood F-test (Lun et al., 2016; Lund et al., 2012) in the edgeR package in R (DJ et al., 2012; MD et al., 2010) as published in (Y. Chen et al., 2016). Reads were first filtered for lowly expressed counts using the rowSums functions, where a cutoff of 0.2 for the CPM was chosen to roughly equal 10/L (where L= minimum library size in millions (i.e. 50 million reads)) in at least three libraries (i.e. replicates). To eliminate composition bias between libraries, normalization for composition bias was determined by calculating Trimmed Mean of M values (TMM) using the calcNormFactors function. Dispersion estimates were calculated using the estimateDisp function using the robust=TRUE option. A design matrix was created to that each DAF for each genotype is a group as published in (Y. Chen et al., 2016). To identify DEG in each mutant genotype at timepoint in seed development, contrasts between each respective mutant at each DAF were created (12 total contrasts) (contrast = WT/mutant). Differential gene expression analysis was completed using the glmQLFTest function (Lun et al., 2016; Lund et al., 2012). DEG results were filtered by FDR <= 0.05 and logFC >= 0.58 or <=-0.58 (FC=1.5). Unique transcripts were identified from normalized cpm data greater than or equal to five cpm.

GO enrichment

GO enrichment analyses were performed using agriGO v2 (Tian et al., 2017) with the following parameters: a hypergeometric test with an $\alpha = 0.05$ *FDR* correction, *Arabidopsis* as the select organism, and a complete GO ontology. For enrichment analyses of GWAS candidate genes associated with significant SNPs, an agriGO suggested background list for *Arabidopsis* was used (TAIR10). For enrichment analyses of DEGs from the RNAseq data, a background list incorporating genes from the 21,989 transcripts identified across development was used. For enrichment analyses of DEPs from the proteomics data, a background list incorporating proteins identified in the CRU- (1342 proteins) and napin-RNAi (1292 proteins) mutants, respectively were used.

Gene functional characterization using MapMan classifications

Functional characterization of candidate genes was performed using MapMan version 3.6 (Lohse et al., 2014). All genes were mapped to first order MapMan Bins (functional classifications) using the *Arabidopsis* mapping database *Ath_AFFY_ATH1_TAIR10_Aug2012.txt* obtained from https://mapman.gabipd.org/mapmanstore. A total of 40 functional gene categories were in the mapping database.

STRING

Protein-protein interaction (PPI) networks were constructed from the HCCG for the CRU and RNAi mutants (335 genes and 241 genes, respectively) at high confidence (interaction score >0.7). The Multiple Protein Search Tool was used on the STRING V11.0 database (Szklarczyk et al., 2019). In PPI networks, proteins are indicated by nodes and interaction between nodes are represented by edges. The MCL clustering algorithm (MCL inflation parameter = 1.5) within STRING was used to cluster strong interactions in the network. Singletons with no connection and nodes creating a cluster of two were removed from the network. Network statistics and enrichment analyses for the entire network were completed using the STRING Analysis feature. The STRING Export feature was used to export a tabular text output for the network that could be visualized in Cytoscape.

Visualization of HCCG expression in the SSP mutants

Gene expression for the 335 and 241 HCCGs from the CRU- and napin-RNAi PPI networks, respectively, were visualized in a heatmap. First WT expression in counts per million (cpm) was visualized across the seven developmental timepoints for the HCCGs. Next, gene expression for the HCCGs in each respective mutant was visualized using FC(cpm) calculated using the following equation: [positive FC = (WT/mutant)] and [negative FC = (-mutant/WT)].

3.8 REFERNCES

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., . . . Ding, W. (2016). 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell, 166*(2), 481-491.

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. *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. (2017). Network-guided GWAS improves identification of genes affecting free amino acids. *Plant physiology, 173*(1), 872-886.

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., 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. *The Plant Cell, 25*(12), 4827-4843.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., . . . Hu, T. T. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature, 465*(7298), 627.

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2004). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics, 21*(2), 263-265.

Baud, S., Boutin, J.-P., Miquel, M., Lepiniec, L., & Rochat, C. (2002). An integrated overview of seed development in Arabidopsis thaliana ecotype WS. *Plant Physiology and Biochemistry, 40*(2), 151-160. doi:https://doi.org/10.1016/S0981-9428(01)01350-X

Birtić, S., & Kranner, I. (2006). Isolation of high-quality RNA from polyphenol-, polysaccharide-and lipid-rich seeds. *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques, 17*(3), 144-148.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological), 26*(2), 211-243.

Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R., & Görlach, J. (2001). Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. *Plant Cell, 13*(7), 1499-1510. doi:10.1105/tpc.010011

Chen, Y., Lun, A. T., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res, 5*, 1438. doi:10.12688/f1000research.8987.2

Cohen, H., Pajak, A., Pandurangan, S., Amir, R., & Marsolais, F. (2016). Higher endogenous methionine in transgenic Arabidopsis seeds affects the composition of storage proteins and lipids. *Amino acids, 48*(6), 1413-1422.

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.

DJ, M., Y, C., & GK, S. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. In (Vol. 40, pp. 4288-4297): Nucleic Acids Research.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res, 38*(Web Server issue), W64-70. doi:10.1093/nar/gkq310 elBaradi, T. T., van der Sande, C. A., Mager, W. H., Raué, H. A., & Planta, R. J. (1986). The cellular level of yeast ribosomal protein L25 is controlled principally by rapid degradation of excess protein. *Curr Genet, 10*(10), 733-739. doi:10.1007/BF00405095

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. doi:10.1104/pp.106.086694

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., . . . Graves, S. (2012). Package 'car'. *Vienna: R Foundation for Statistical Computing*.

Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., & Aebersold, R. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol Cell Proteomics, 1*(4), 323-333. doi:10.1074/mcp.m200001-mcp200

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.

Hagan, N., Upadhyaya, N., Tabe, L., & Higgins, T. (2003). The redistribution of protein sulfur in transgenic rice expressing a gene for a foreign, sulfur-rich protein. *The Plant Journal, 34*(1), 1-11.

Hajduch, M., Hearne, L. B., Miernyk, J. A., Casteel, J. E., Joshi, T., Agrawal, G. K., . . . Thelen, J. J. (2010). Systems analysis of seed filling in Arabidopsis: using general linear modeling to assess concordance of transcript and protein expression. *Plant Physiol, 152*(4), 2078-2087. doi:10.1104/pp.109.152413

Herman, E. M. (2014). Soybean seed proteome rebalancing. *Frontiers in plant science, 5*, 437.

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. doi:10.1105/tpc.003905

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 opaque2 mutant by transcriptional and proteomic analysis. *BMC plant biology, 13*(1), 60.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., . . . Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature genetics, 39*(9), 1151.

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. doi:10.1105/tpc.1.1.105

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (Vol. 5): McGraw-Hill Irwin Boston.

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.

Le Roch, K. G., Johnson, J. R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., . . . Winzeler, E. A. (2004). Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res, 14*(11), 2308-2318. doi:10.1101/gr.2523904

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, 3*(8), 1287-1299.

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, 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.

Lun, A. T., Chen, Y., & Smyth, G. K. (2016). It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. In (pp. 391-416). *Statistical Genomics*: Humana Press.

Lund, S. P., Nettleton, D., McCarthy, D. J., & Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical applications in genetics and molecular biology, 11*(5).

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. In (Vol. 17, pp. 10-12): EMBnet.journal.

MD, R., DJ, M., & GK, S. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. In (Vol. 26, pp. 139-140): Bioinformatics.

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., . . . Kuster, B. (2020). Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature, 579*(7799), 409-414. doi:10.1038/s41586-020-2094-2

Morton, K. J., Jia, S., Zhang, C., & Holding, D. R. (2016). Proteomic profiling of maize opaque endosperm mutants reveals selective accumulation of lysine-enriched proteins. *J Exp Bot, 67*(5), 1381-1396. doi:10.1093/jxb/erv532

Nguyen, T.-P., Cueff, G., Hegedus, D. D., Rajjou, L., & Bentsink, L. (2015). A role for seed storage proteins in Arabidopsis seed longevity. *Journal of Experimental Botany, 66*(20), 6399-6413.

Rajjou, L., Duval, M., Gallardo, K., Catusse, J., Bally, J., Job, C., & Job, D. (2012). Seed germination and vigor. *Annu Rev Plant Biol, 63*, 507-533. doi:10.1146/annurev-arplant-042811-105550

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

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., . . . Lohmann, J. U. (2005). A gene expression map of Arabidopsis thaliana development. *Nat Genet, 37*(5), 501-506. doi:10.1038/ng1543

Schmidt, M. A., Barbazuk, W. B., Sandford, M., May, G., 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, 156*(1), 330-345.

Schmidt, R. J., Burr, F. A., Aukerman, M. J., & Burr, B. (1990). Maize regulatory gene opaque-2 encodes a protein with a "leuucind-zipper" motif that binds to zeing DNA. In (Vol. 87): National Academy of Sciences of the United States of America.

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. *Funct Integr Genomics, 14*(4), 683-696. doi:10.1007/s10142-014-0388-x

Shrestha, V. (2020). *UNCOVERING THE GENETIC ARCHITECTURE AND METABOLIC BASIS OF AMINO ACID COMPOSITION IN MAIZE KERNELS USING MULTI-OMICS INTEGRATION.* (PhD). University of Missouri, Columbia,

Slaten, M. L., Chan, Y. O., Shrestha, V., Lipka, A. E., & Angelovici, R. (2020). HAPPI GWAS: Holistic Analysis with Pre- and Post-Integration GWAS. In (Vol. 36, pp. 4655–4657). Bioinformatics.

Slaten, M. L., Yobi, A., Bagaza, C., Chan, Y. O., Shrestha, V., Holden, S., . . . Angelovici, R. (2020). mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis. *Plant Physiol, 183*(2), 483-500. doi:10.1104/pp.20.00039

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.

Tan-Wilson, A. L., & Wilson, K. A. (2012). Mobilization of seed protein reserves. *Physiologia Plantarum, 145*(1), 140-153.

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 Res, 45*(W1), W122-W129. doi:10.1093/nar/gkx382

Tzin, V., & Galili, G. (2010). New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. *Molecular plant, 3*(6), 956-972.

Wanasundara, J. P. (2011). Proteins of Brassicaceae oilseeds and their potential as a plant protein source. *Critical reviews in food science and nutrition, 51*(7), 635-677.

Wentzell, A. M., Rowe, H. C., Hansen, B. G., Ticconi, C., Halkier, B. A., & Kliebenstein, D. J. (2007). Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *Plos genetics, 3*(9), e162.

Wi, J., Na, Y., Yang, E., Lee, J. H., Jeong, W. J., & Choi, D. W. (2020). a homolog of mice MPV17, enhances osmotic stress tolerance. *Physiol Mol Biol Plants, 26*(7), 1341-1348. doi:10.1007/s12298-020-00834-x

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.

Yao, M., Guan, M., Zhang, Z., Zhang, Q., Cui, Y., Chen, H., . . . Qian, L. (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), 320. doi:10.1186/s12864-020-6711-0

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, M. L., Schaefer, R., Dang, P., Jiang, T., & Chen, C. (2019). GWAS and Coexpression Network Reveal Ionomic Variation in Cultivated Peanut. *J Agric Food Chem, 67*(43), 12026-12036. doi:10.1021/acs.jafc.9b04939

CHAPTER 4: HAPPI GWAS: HOLISTIC ANALYSIS WITH PRE AND POST INTEGRATION

Marianne L. Slaten[1†], Yen On Chan[1†], Vivek Shrestha[1], Alexander E. Lipka[2], Ruthie Angelovici[1*]

1 Division of Biological Sciences, Interdisciplinary Plant Group, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA

2 Department of Crop Sciences, University of Illinois, Urbana, IL, 61801, USA.

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, 36(17), 4655–4657.

4.1 ABSTRACT

Advanced publicly available sequencing data from large populations have enabled informative genome-wide association studies (GWAS) that associate SNPs with phenotypic traits of interest. Many publicly available tools able to perform GWAS have been developed in response to increased demand. However, these tools lack a comprehensive pipeline that includes both pre-GWAS analysis such as outlier removal, data transformation, and calculation of Best Linear Unbiased Predictions (BLUPs) or Best Linear Unbiased Estimates (BLUEs). In addition, post-GWAS analysis such as haploblock analysis and candidate gene identification are lacking. Here, I present HAPPI GWAS, an open-source tool able to perform pre-GWAS, GWAS, and post-GWAS analysis in an automated pipeline using the command-line interface. HAPPI GWAS is written in R for any Unix-like operating systems and is available on GitHub [https://github.com/angelovicilab/HAPPI_GWAS.git].

4.2 INTRODUCTION

Recent advances and publicly available sequencing data of large populations coupled with the development of improved statistical methods has enabled informative genome-wide association studies (GWAS). As a result, the genetic architecture of many agronomically important traits have been associated with specific genomic loci. Demand to run GWAS, not only on large datasets, but also on a user-friendly, flexible platform has grown and become an increasingly important demand to fulfill.

The increased demand for analyzing large genotypic datasets has been answered with an increase in publicly available tools and methods. A past effort has focused heavily on ease of usability. GWAS programs such as GAPIT (Lipka et al., 2012) incorporate a variety of statistical models into a single R package, while others such as FarmCPU (Liu et al., 2016) implement novel statistical models. Other programs use graphical user interfaces (GUIs) such as TASSEL (Bradbury et al., 2007) or web-based platforms such as GWAPP (Seren et al., 2012) and easyGWAS (Grimm et al., 2017). However, these tools do not provide all crucial steps: pre-GWAS (outlier removal, transformation, and BLUP/BLUE calculations) and post-GWAS (haploblock analysis and gene extraction and identification), in addition to a user-friendly platform with comprehensive output.

Due to the lack of publicly available tools that can conduct GWAS and pre-GWAS and post-GWAS analysis, users are often required to use different tools for each step in the GWAS analysis which can often be a time-intensive, arduous process that requires self-teaching of many tools and extensively formatting output data from one tool to input into the next. Additionally, these tools often require that the analysis is run through web-based platforms and workflows that restrict users to set parameters and models. Thus, in response to these needs, I have co-developed Holistic Analysis with Pre and Post Integration (HAPPI) GWAS which provides a complete GWAS pipeline including pre-GWAS, GWAS, and post-GWAS analysis in a single tool.

HAPPI GWAS is an R-based tool that runs on the command-line interface of any Unix, Linux, and Mac operating system. HAPPI GWAS source code is loaded into the R environment using the R script command in the terminal while outsourcing analysis to external tools (external tools require ImageMagick and Java). HAPPI GWAS is free and publicly available for download. A summary of the main contributions of HAPPI GWAS include: 1) eliminating the need for multiple tools by providing a comprehensive GWAS pipeline for all phases of a GWAS analysis, 2) allowing high-throughput analysis of multiple traits with easy comparison of GWAS results across all traits and concise, publication-ready figures and tables, and 3) allowing user-defined models and threshold parameters specified at the start of the workflow and automatically implemented throughout the pipeline without additional configuration.

4.3 METHODS

HAPPI GWAS is implemented in four main steps: pre-GWAS, GWAS, post-GWAS and Outputs, Summaries and Visualizations (for HAPPI GWAS workflow refer to Figure 4.1). Each step is customizable by the user through a YAML file. The YAML file instructs HAPPI GWAS of the name and location of data input and output and allows for user-defined parameters at each step of the pipeline. Additional information regarding each step, parameter flexibility and tutorial datasets can be found in the HAPPI GWAS manual (Supplemental Document S1) and on the wiki page linked in the GitHub repository

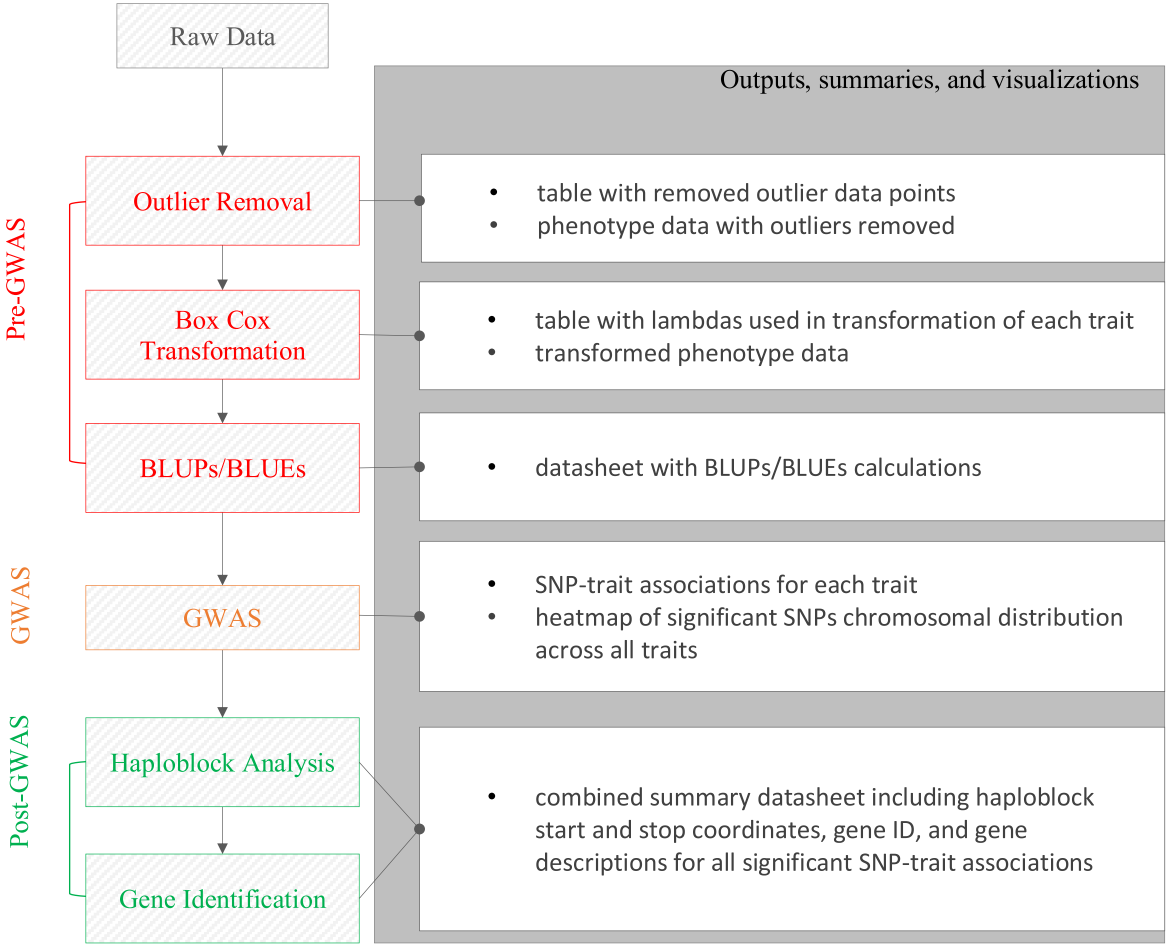

Raw Data
Outputs, summaries, and visualizations
Pre-GWAS
Outlier Removal
table with removed outlier data points
phenotype data with outliers removed
Box Cox Transformation
table with lambdas used in transformation of each trait
transformed phenotype data
BLUPs/BLUEs
datasheet with BLUPs/BLUEs calculations
GWAS
GWAS
SNP-trait associations for each trait
heatmap of significant SNPs chromosomal distribution across all traits
Post-GWAS
Haploblock Analysis
Gene Identification
combined summary datasheet including haploblock start and stop coordinates, gene ID, and gene descriptions for all significant SNP-trait associations

Figure 4. 1. HAPPI GWAS workflow outlining pre-GWAS, GWAS, post-GWAS, and outputs, summaries, and visualizations steps.

4.3.1 Input Data

Phenotypic data: The phenotypic data must be provided by the user in the form of raw data or BLUP/BLUE data. Users are not limited by the number of traits that can be run in HAPPI GWAS. All phenotypes should be saved in a tab-delimited text file (.txt) or comma-separated values file (.csv) with missing data indicated by “NA”. Duplicate values are not allowed in the phenotypic data, so it is necessary to solve duplication problems using mathematical approaches such as *lsmean* or arithmetic mean before feeding the data into the tool.

Genotypic data: The genotypic data must be provided in Hapmap or numeric format. Users can provide private genotypic data in addition to using the formatted genotypic data provided for maize (Flint-Garcia et al., 2005) and *Arabidopsis* (Horton et al., 2012; Nordborg et al., 2005). Genotype files that are too large to import due to memory limitations, can be saved per chromosome.

Step 1: Pre-GWAS

Ensuring raw data meets all assumptions prior to GWAS is vital to reproducible and accurate results but can be difficult to navigate. HAPPI GWAS automatically inputs raw data into pre-GWAS analysis by removing outliers using Studentized deleted residuals (Cook, 1977) and transforming data using the Box-Cox procedure (Box & Cox, 1964). The variance between replicates is evaluated via mixed models to calculate BLUPs and BLUEs. Model flexibility allows users to define specific random and fixed effects in these mixed models. If preprocessing is completed externally, the option to skip the pre-GWAS step is available.

Outlier removal: Population-wide outlier removal is automatically completed with the *generateBLUP* and *generateBLUE* options. Data are fit using a mixed linear model where the influence of data points is determined by using the Studentized deleted residuals (Kutner et al., 2005). That is, a Studentized deleted residual is calculated for every experimental unit. All experimental units where the Studentized residual exceeds critical value of a null *t*-distribution (for testingH_0 : A given experimental unit is an outlier) with *n-3* degrees of freedom (estimating the intercept, the line variance component, and the population variance component) at a Bonferroni-corrected experimental-wise type I error rate of 0.05 are declared to be an outlier and are removed from the data set. Output phenotype files contain data with outliers replaced with "NA"; in addition, a data file with a list of removed outlier points and a file with outlier residuals can be retrieved in the "generateBLUP" directory created in the output folder.

Box-Cox transformation: Box-Cox transformation is used to transform each trait to meet normality assumptions. A lambda is calculated per trait and used to transform each trait. The model used to calculate BLUPs/BLUEs will be used at this step. The *powerTransform* function is part of the *car* package (Fox et al., 2012) and works by using the maximum likelihood-like approach of Box and Cox (Box & Cox, 1964) to select a transformation. Within the function, the options for *Family* should be set to "bcPower" and *Lambdas* set to -2 to 2. The output phenotype file contains transformed phenotypic data. A list of lambdas for each trait in addition to a data file with the transformed data can be retrieved in the "generateBLUP" directory created in the output folder. Box-Cox transformation is automatically run with the *generateBLUP* and *generateBLUE* options.

BLUPs/BLUEs calculations: After outlier removal and transformation, genetic values (either as random or fixed effects) are estimated using the *generateBLUP* and *generateBLUE* options, respectively. A general mixed linear model combines information from all relatives measured to improve estimates (see equation below). In doing so, replicates per accession within a given trait are eliminated, and only one value per accession per trait remains.

$$y = x\beta + Z\mu + e$$

Where

y = vector of observation (phenotypes)

x = matrix of fixed effects

β = vector of fixed effects to be estimated (i.e. year, location, treatment effects)

Z = matrix of random effects

μ = vector of random effects to be estimated (genetic values)

e = vector of residual errors

The user can bypass outlier removal and transformation steps and input externally calculated BLUPs/BLUEs.

BLUPs (Best Linear Unbiased Prediction): A linear mixed model is used to predict random effects (û). In BLUP calculations, the Accession ID/Taxa name will be considered a random effect. All additional variables in the model are random. A file containing the BLUP data can be found in the “generateBLUP” directory created in the output folder.

BLUEs (Best Linear Unbiased Estimates): A linear mixed model is used to estimate fixed effects (β-hat). In BLUE calculations, the Accession ID/Taxa name will be considered a fixed effect. All additional variables are random. A file containing the BLUE data can be found in the "generateBLUP" directory created in the output folder.

Step 2: GWAS

The GWAS step accepts user-defined phenotype data and genotype data. Provided genotype files can be used in combination with user-supplied phenotype data. Multiple traits can be stored in each phenotype dataset and run consecutively. Configuration files are edited by the user to supply values for mandatory defined variables that are called when the program is invoked. All GWAS analysis is performed by calling the GAPIT v3 R package (Wang & Zhang, 2018). The *GAPIT* option is required to run GWAS. The *extractHaplotype* and *searchGenes* (post-GWAS) options are optional.

Models:

1. Generalized Linear Model (GLM): model including only fixed effects. Population structure is defined (Q matrix). Both the marker and population structure are defined as fixed effects in the model. No random effects are found in the model.
2. Mixed Linear Model (MLM): model including both fixed and random effects. Relatedness is conveyed through a kinship matrix (K) as a random effect and population structure (Q matrix) is accounted for as fixed effect using STRUCTURE (Pritchard et al., 2000) or PCA.
3. Multiple Locus Mixed Linear Model (MLMM): model including forward-backward stepwise linear mixed-model to estimate variance components (Segura et al., 2012).

4. Settlement of MLM Under Progressively Exclusive Relationship (SUPER): a model that extracts a small subset of SNPs and uses them in FaST-LMM (Wang et al., 2014).
5. Farm-CPU (FarmCPU): a model using pseudo QTNs is used to iterate between fixed and random effect models (Liu et al., 2016).

Step 3: Post-GWAS

Most GWAS packages output a list of significant SNP-trait associations. However, a list of obscure SNP IDs is often uninformative until associated with genes. In the post-GWAS step, a list of significant SNPs is fed directly into a haploblock analysis in Haploview (Barrett et al., 2004). The Haploblock analysis filters SNPs at a 5% minor allele frequency (MAF) and quantifies the degree of linkage disequilibrium (LD) using D prime in the surrounding genomic region to estimate a haploblock (defined as regions of high LD) described with a start and stop location. Genes contained or partially contained within haploblocks are identified and output with respective gene descriptions in the final summary datasheet (Supplemental Table S1-S3). If no genes overlap the haploblock or the significant SNP does not fall within a haploblock, the gene directly upstream and downstream of the significant SNP is given. HAPPI GWAS allows users to define the window size for LD calculations to increase gene identification in a larger interval or to skip the post-GWAS step entirely in species where limited genomic information is available.

Haploblock analysis: When interpreting significant SNP-trait associations from GWAS, it is beneficial to focus beyond the identified SNP and determine the extent of LD surrounding the SNP. SNPs (and genes) contained within this region of high LD are all of the putative interests. When the *extractHaplotype* option is used, for each significant SNP identified in GWAS, pairwise LD is calculated between the significant SNP and every neighboring SNP in a user-defined window using Haploview (Barrett et al., 2004). Regions of high LD (95% confidence bounds on D prime) (i.e. haploblocks) are identified and automatically used downstream in the *searchGenes* section. SNPs are filtered at a 5% minor allele frequency (MAF) and LD is calculated using D prime. Genes contained or partially contained within haploblocks are identified and output with respective gene descriptions in the final summary datasheet. If no genes overlap the haploblock or the significant SNP does not fall within a haploblock, the gene directly upstream and downstream of the significant SNP is given. If no gene annotation file is available, haploblock analysis and gene identification steps can be skipped entirely. In species with limited genomic information available that prevents accurate LD calculations, the haploblock analysis can be skipped (by removing the *extractHaplotype* option, while the *searchGenes* option is still used) and genes contained in a user-defined window, flanking the significant SNP, can be output

Identify genes: Haploblock information for each SNP is automatically used in the *searchGene* option where genes contained in or overlapping with the calculated haploblock (from the *extractHaplotype* section) are identified. By identifying each gene associated with the GWAS significant SNPs, HAPPI GWAS is able to output a list of genes, rather than SNPs, which is more informative in determining the complex SNP-trait relationships. Files required to run Haploview in each of the Demo datasets are provided.

LD parameters can be altered through the *GAPIT_LD_number* option in the YAML file. At the *Identify Genes* step, a GFF file is also required. *M*aize GFF files are included in the tool packages. *Arabidopsis* GFF files can be downloaded from Cyverse. User-defined GFF files can also be used and designated in the YAML file.

Step 4: Outputs, Summaries, and Visualizations

GWAS results from all traits found in the phenotype file are summarized concisely in tables and figures as part of the automatic summary output. Collective analysis of related traits can be powerful in the detection of pleiotropy. HAPPI GWAS compiles GWAS results creating a combined GWAS results summary that includes significant SNP IDs, gene names, gene descriptions, and haploblock information (Supplemental Table S1). Two additional summary tables are created: a table summarizing the top five SNP-trait associations with the lowest *P*-value across all analyzed traits (Supplemental Table S2) and a table summarizing the most recurring SNP-trait associations across all analyzed traits (Supplemental Table S3). Lastly, a unique HAPPI GWAS visualization representation of the chromosomal distribution of all significant SNP-trait associations (Supplemental Figure S2) for all traits is provided. This figure is unique to HAPPI GWAS and differs from other multi-trait GWAS visualizations such as Zbrowse (Ziegler et al., 2013a) because only significant SNPs are visualized. This novel format allows for easier comparison of genome-wide SNP distributions across the traits as compared to overlapping Manhattan plots. Automatic GAPIT output, as found in the GAPIT3 manual (Wang & Zhang, 2018), is also included in the output. For more information regarding each step, refer to the HAPPI GWAS manual (Supplemental Document S1). These files include:

1. A text file containing outlier residuals produced from Studentized Deleted Residuals analysis (Outliers_residuals.txt).

2. A data file containing phenotype data with outliers removed (Outlier_removed_data.csv).
3. A list of data points identified as outliers through Studentized Deleted Residuals analysis (Outlier_data.csv).
4. A list of lambda values calculated for each trait during Box-Cox transformation (Lambda_values.csv).
5. A data file containing Box-Cox transformed phenotype data (Boxcox_transformed_data.csv).
6. A data file containing calculated BLUPs (if *generateBLUP* flag is used) (BLUP.csv).
7. A data file containing calculate BLUEs (if *generateBLUE* flag is used) (BLUE.csv).

Example HAPPI GWAS output files can be accessed through the original Bioinformatics publication (https://doi.org/10.1093/bioinformatics/btaa589) or through the Angelovici lab GitHub (https://github.com/Angelovici-Lab/HAPPI.GWAS).

4.4 RESULTS

4.4.1 Performance Test

The average computing time per trait is directly related to the total number of individuals in the population and the size of the genotype file. To show the effect of sample size and SNP number on runtime, one trait using filtered genotypic data from the *Arabidopsis* 1001 dataset (Alonso-Blanco et al., 2016) with varying sample size and SNP number with one processing core is analyzed. The filtered data has a total of 1,057,383 SNPs and the phenotypic data is from 901 individuals in replicates of two. To run the entire dataset through HAPPI GWAS takes 6 hours and 39 minutes. As the number of individuals in the population decreases, run time remains relatively constant. Using the full genotypic data with 1,057,383 SNPs but decreasing the number of individuals by half (450 individuals) results in a runtime of around 6 hours and 6 minutes. Conversely, decreasing the number of SNPs to half (i.e., to 528,692 SNPs) while maintaining a population size of 901 results in a shorter runtime of 2 hours and 7 minutes. The tool was tested on the CentOS machine with 500 GB of RAM and 30 TB of disk space.

Example Analysis in Maize

All necessary inputs to run the maize demo data are provided within the cloned repository from GitHub. Additional input demo data for other species can be downloaded from Cyverse (See "Data Availability" section). Files include phenotype data, genotype data, Haploview files, and GFF gene annotation files.

Maize Demo files can be found in the *Demo* folder within the cloned HAPPI GWAS GitHub repository. Hapmap, Haploview, and GFF files are provided and split by chromosome into 10 files, respectively. BLUPs have been externally calculated. All phenotypic data was obtained from (Flint-Garcia et al., 2005). For this tutorial, first navigate into the cloned HAPPI GWAS repository. Please refer to the following commands:

```
cd <absolute path identified before the HAPPI_GWAS repository cloning process>
        #navigate into the user-created HAPPI_GWAS folder
cd HAPPI_GWAS      # navigate into the cloned repository
```

To run the Maize Demo data follow these steps:

Step 1: Edit the Demo_GLM.yaml file:

Edit the "BLUP or BLUE" section. Ensure the path and file name are correct.

In this tutorial, externally calculated BLUPs (i.e. the “Raw Data” section is blank) will be used. The first column in the BLUPs file is the Line ID, and subsequent columns (starting with column 2) are the phenotypic data in the form of BLUPs.

Edit the “GAPIT3” section. Ensure the path at line “GAPIT_genotype_file_path” is correct (i.e. using the correct absolute path). The mdp_genotype_chr[1-10].hmp.txt files will be used. Note how the *GLM* is the selected model as all other model options are ignored by the addition of #. SNP MAF is filtered at 0.05 with a significant FDR threshold of 0.05. A desired window size of 100,000 bp on each side of the significant SNP is defined by editing GAPIT_LD_number: 100000.

Edit the “Haploview” section. Ensure the path at line “Haploview_file_path” is correct. The mdp_genotype_haploview_chr[1-10].txt files will be used.

Edit the “Match Gene Start and Stop” section. Ensure the path at line “GFF_file_path” is correct. The gene_chr[1-10].gff.txt files will be used.

Edit the “Output Directory” section. Ensure the path on line “output” is correct.

Step 2: Run HAPPI GWAS using the following command:

```
Rscript HAPPI_GWAS.R Demo_GLM.yaml -GAPIT -extractHaplotype -searchGenes
```

Step 3: Access output data at the following (See Figure 4.2):

```
cd <user defined output path found in the “Output Directory” section of the YAML file>
```

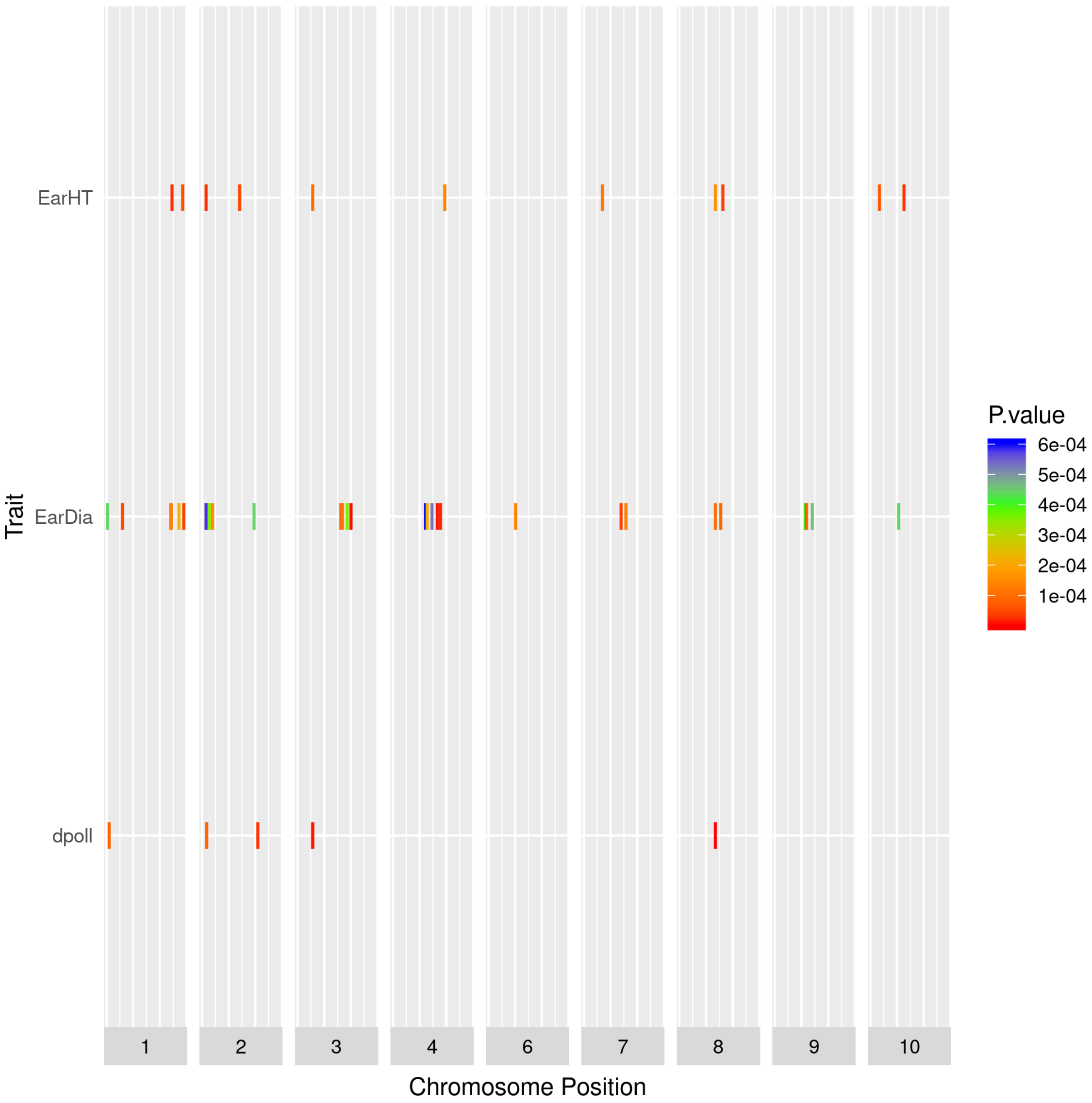


Figure 4. 2. Summary heatmap from HAPPI GWAS demo data illustrating significant SNP distribution across chromosomes and all traits. The x-axis shows only chromosomes containing significant SNPs; the y-axis shows all traits with significant SNP-trait associations. Rectangles represent SNPs that are color-coded based on P-value.

4.4.2 Code and Data Availability

Formatted genotypic data from the Goodman Buckler maize association population (Flint-Garcia et al., 2005) are provided with the HAPPI GWAS download. Formatted genotypic data from the *Arabidopsis* 360 population (Horton et al., 2012; Nordborg et al., 2005) and *Arabidopsis* 1001 population (Alonso-Blanco et al., 2016) is available for download at the Angelovici CyVerse account found here: /iplant/home/Angelovici_lab/HAPPI_GWAS

4.5 CONCLUSIONS

HAPPI GWAS is a holistic tool that integrates pre-GWAS, GWAS, post-GWAS, and outputs, summaries, and visualizations without sacrificing ease of use or flexibility. Incorporation of all four steps leads to a comprehensive pipeline that aims to be computationally approachable, regardless of user background. It improves upon past GWAS tools by increasing the scope of analysis and plasticity of defined parameters.

4.6 SUPPLEMENTAL INFORMATION

Supplemental Table S1-S3 and Supplemental Document S1 (HAPPI GWAS manual) are available with the original *Bioinformatics* publication downloadable by accessing the publication https://doi.org/10.1093/bioinformatics/btaa589 or downloading a zip file directly from

https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/bioinformatics/36/17/10.1093_bioinformatics_btaa589/8/btaa589_Supplemental_data.zip?Expires=1615274141&Signature=zJJIOz

0yV1RobrwKj63ter~V0jLDQ47w4CyzHrCox~Z2Cmom5jlemdcjCY-HROomf2MoKwhrYcUTm2QSUB~Il2-l4LzvJOOwF2zSXXsfApFikDQyIeuaIBxfbrgqgxE2I2bTFjF8bp3MM9ZYA9WmH1~-E1SnGKOUKdPJNLQrSj-BsHp1N~hIuwuG6ur9B6e1Zs~TQMqyxhcCTkQxDfrpuMQkF1OKXiovidgesJMCaxEPbB9c1jbx2zhRoJIoP0-759QXkKMKLbG9u-nxALLHl46cKfkZx3JTwl49PcqGJgxWw9~JfBd~gqN9HEFPQvjrEPNVG3MtwY5NgDbbyR3LIA__&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA

4.7 ACKNOWLEDGEMENTS

The authors wish to acknowledge Sarah Turner-Hissong for her assistance in early pipeline development.

4.8 REFERENCES

Alonso-Blanco, Carlos, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M Borgwardt, Jun Cao, Eunyoung Chae, Todd M Dezwaan, and Wei Ding. 2016. '1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana', *Cell*, 166: 481-91.

Barrett, Jeffrey C, B Fry, JDMJ Maller, and Mark J Daly. 2004. 'Haploview: analysis and visualization of LD and haplotype maps', *Bioinformatics*, 21: 263-65.

Box, George EP, and David R Cox. 1964. 'An analysis of transformations', *Journal of the Royal Statistical Society: Series B (Methodological)*, 26: 211-43.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler. 2007. 'TASSEL: software for association mapping of complex traits in diverse samples', *Bioinformatics*, 23: 2633-5.

Cook, R.D. 1977. 'Detection of influential observation in linear regression', *Technometrics*, 19: 5-19.

Flint-Garcia, Sherry A., Anne-Celine Thuillet, Jianming Yu, Gael Pressoir, Susan M Romero, Sharon E. Mitchell, John Doebley, Stephen Kresovich, Major M. Goodman, and Edward S. Buckler. 2005. 'Maize association population: a high-resolution platform for quantitative trait locus dissection', *The Plant Journal*, 44: 1054-64.

Fox, John, Sanford Weisberg, Daniel Adler, Douglas Bates, Gabriel Baud-Bovy, Steve Ellison, David Firth, Michael Friendly, Gregor Gorjanc, and Spencer Graves. 2012. 'Package ‘car’', *Vienna: R Foundation for Statistical Computing*.

Grimm, D. G., D. Roqueiro, P. A. Salomé, S. Kleeberger, B. Greshake, W. Zhu, C. Liu, C. Lippert, O. Stegle, B. Schölkopf, D. Weigel, and K. M. Borgwardt. 2017. 'easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies', *Plant Cell*, 29: 5-19.

Horton, Matthew W, Angela M Hancock, Yu S Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N Wayan Muliyati, Alexander Platt, F Gianluca Sperone, and Bjarni J Vilhjálmsson. 2012. 'Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel', *Nature genetics*, 44: 212.

Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. 2005. *Applied linear statistical models* (McGraw-Hill Irwin Boston).

Lipka, Alexander E, Feng Tian, Qishan Wang, Jason Peiffer, Meng Li, Peter J Bradbury, Michael A Gore, Edward S Buckler, and Zhiwu Zhang. 2012. 'GAPIT: genome association and prediction integrated tool', *Bioinformatics*, 28: 2397-99.

Liu, Xiaolei, Meng Huang, Bin Fan, Edward S Buckler, and Zhiwu Zhang. 2016. 'Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies', *PLoS genetics*, 12: e1005767.

Nordborg, Magnus, Tina T Hu, Yoko Ishino, Jinal Jhaveri, Christopher Toomajian, Honggang Zheng, Erica Bakker, Peter Calabrese, Jean Gladstone, and Rana Goyal. 2005. 'The pattern of polymorphism in Arabidopsis thaliana', *PLoS biology*, 3: e196.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. 'Inference of population structure using multilocus genotype data', *Genetics*, 155: 945-59.

Segura, V., B. J. Vilhjálmsson, A. Platt, A. Korte, Ü Seren, Q. Long, and M. Nordborg. 2012. 'An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations', *Nat Genet*, 44: 825-30.

Seren, Ü, B. J. Vilhjálmsson, M. W. Horton, D. Meng, P. Forai, Y. S. Huang, Q. Long, V. Segura, and M. Nordborg. 2012. 'GWAPP: a web application for genome-wide association mapping in Arabidopsis', *Plant Cell*, 24: 4793-805.

Wang, J, and Z Zhang. 2018. 'GAPIT Version 3:An Interactive Analytical Tool for Genomic Association and Prediction', *Bioinformatics. Draft.*

Wang, SS, F Wang, SJ Tian, MX Wang, N Sui, and XS Zhang. 2014. 'Transcript profiles of maize embryo sacs and preliminary identification of genes involved in the embryo sac-pollen tube interaction', *Frontiers in Plant Science*, 5: 1-15.

Ziegler, G., A. Terauchi, A. Becker, P. Armstrong, K. Hudson, and I. Baxter. 2013. 'Ionomic Screening of Field-Grown Soybean Identifies Mutants with Altered Seed Elemental Composition', *Plant Genome*, 6: 1-9.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

Seeds are integral to life as we know it. They are vital sources of protein and calories in the diets of people and livestock around the world. However, there is much room for improvement in the biofortification of protein composition of seeds. Despite previous attempts at altering the FAA and PBAA pools in the seed, drastically altering protein content by reprogramming the proteome has proven to be challenging. This is primarily due to the rebalancing phenomenon and a lack of understanding thereof. The exact mechanism(s) regulating amino acid composition and their interplay remain largely elusive.

To that end, the objective of this dissertation was to advance knowledge of the regulatory mechanisms of seed amino acid regulation using the model plant species *Arabidopsis*. Crop species such as maize and soybean have much more complicated genomes and are more cumbersome to grow and harvest. Since it is believed that amino acid regulation is a conserved mechanism across all plant species, a commonsensical first step is to identify the underpinnings of genetic regulation in a model crop with a vast number of resources. *Arabidopsis* was an optimal choice.

In Chapter 2, using the *Arabidopsis* 360 population, I find a novel connection between FAA primary metabolism in the Glutamate Family with aliphatic glucosinolate biosynthesis. Although glutamine is not an essential amino acid, it is a key FAA in plant metabolism, namely nitrogen metabolism. Thus, understanding glutamine natural variation and regulation in the seed is of particular interest. In this study, I validated putative genes of interest with additional experimentation including QTL and mutant analysis. An independent follow-up experiment found that elimination of aliphatic glucosinolates in the seed disrupted sulfur and nitrogen homeostasis, resulting in the elevated glutamine response in the seed. These results were fascinating as it uncovered a seed specific role of glutamine in nitrogen homeostasis, linked primary and secondary metabolism, and confirmed changes occurring throughout development can be identified in the dry seed.

This work identified fascinating connection among glutamine, glucosinolates, and sulfur content in seeds; however, the exact mechanism by which this regulation is sensed in the plant was not revealed. To uncover the direct mechanism, I propose a sulfur feeding experimentation of the *myb28/29* KO mutant lines. Although there is no guarantee that sulfur feeding will increase sulfur in the seeds, it has been previously reported that plants can absorb amino acids through their roots (Svennerstam et al., 2011). Thus, an easy and cheap experiment would be to supplement amino acids containing sulfur such as methionine or cysteine, or supplement with sulfur rich compounds during weekly watering throughout development. If the high glutamine phenotype is eliminated in the mutants through sulfur supplementation, this suggests that indeed the elevated glutamine phenotype in the mutant is due to sulfur starvation in the seed. Furthermore, it would be interesting to collect additional omics data, namely transcriptomics and proteomics throughout development in the feeding study to further elucidate gene expression and protein abundance alterations across development.

In Chapter 3, I harnessed phenotypic variation in the *Arabidopsis* 1001 population in a GWAS to identify candidate genes. I compare GWAS results with multi-omics data from two SSP mutants. By harnessing protein homeostasis occurring at the population level coupled with active rebalancing occurring in the SSP mutants, I was able to uncover putative regulatory mechanisms associated with PBAA composition in the seed. Chapter 3 improved upon the work in Chapter 2 by increasing the power to identify smaller metabolic changes by using a much larger *Arabidopsis* population with many more SNPs, as well as, incorporating a much larger number of derived ratio-traits to identify the transitional machinery and the cell cycle as putative regulatory mechanisms.

Chapter 3 would, however, benefit from additional proteomic and metabolomic data across development. The developmental transcriptomic data for the two SSP mutant and WT contributed unexpected and informative information regarding the developmental gene expression landscape; however, to complement such data, I also measured PBAA and FAA levels in the genotypes across the seven developmental timepoints. Now that the dry seed has been characterized in this study, the next step is to integrate the remaining metabolomics data. Additionally, proteomics data for the genotypes across development is crucial data that needs to be incorporated. Proteomics data is at beginning stages of quality control and cleaning but should be later incorporated to create a truly comprehensive developmental omics data set for the CRU- and napin-RNAi mutants.

In Chapter 3, I uncovered a unique connection of the PBAA rebalancing with ribosomal transcription and potentially translational machinery. However, as I mentioned above, this analysis is currently lacking proteomics and metabolomics data across seed development. To validate the putative results and further define regulatory mechanisms, in addition to determining what biological level and stage the regulation takes place, additional network analyses with development time-series transcriptome, proteome, and metabolome data are required. For example, the additional developmental metabolome and proteome data across development could be integrated with current developmental transcriptomic data to highlight putative pathways of interest. Use of publicly available tools made specifically for multi-omics integration, such as IMPRes (Jiang et al., 2020), should be used to better integrate data. Additionally, metabolite-gene networks could be created per genotype (CRU-, napin-RNAi, and WT) to elucidate biological pathways and interesting interactions, such as metabolite – gene interactions. Publicly available tools such as WGCNA (Langfelder & Horvath, 2008) could be used to correlate FAA and PBAA traits with genes in the network. Further analysis of network rewiring could also be completed between genotypes to highlight putative mechanistic differences and similarities. Identification of shared pathways in the SSP that differ from the WT could further elucidate regulatory pathways that are being turned on and off in response to PBAA rebalancing.

Lastly, despite being unable to pinpoint a mechanism of regulation from the DEGs, it is clear that the mutants are indeed responding to the SSP perturbations through transcriptomic changes. Such results are exciting as it means that future transgenic approaches in altering PBAA may work through alteration of transcription. Further work needs to be completed to address specific contributions of each gene(s) of interests identified in this study. Using promising CRISPR techniques, identified genes can be knocked-out, knocked-down and/or overexpressed in a seed specific manner. Follow-up changes after gene alteration could be assessed in the proteome and metabolome. Of the most promising genes in this study that could be first analyzed included HCCGs that are also found as DEPs such as AT3G52930, AT3G54050, AT1G42970, AT3G26060, AT1G76550. Also, ribosomal genes identified in the CRU- red HCCG cluster that were also DEGs in the napin-RNAi mutant would be exciting avenues for future analysis: AT3G53430, AT2G18110, AT5G48760, and AT2G35040.

Lastly in Chapter 4 to further the knowledge base of amino acid composition in the seed, I present HAPPI GWAS, an integrative GWAS tool that can be utilized in future work with additional crop species. The tool addressed a need in the quantitative genetics community to have the capability to run many traits through not only GWAS, but also pre-GWAS and post-GWAS analysis. HAPPI GWAS seamlessly integrates a comprehensive GWAS pipeline to output a list of candidate genes and tables and figures that are ready for biological interpretation. The tool is written in R and can be easily run on the command line.

One of the advantages of HAPPI GWAS is the tool's ability to run through the entire GWAS pipeline after the user changes only select parameters and introduces input files. However, the tools still requires running on the command line, which can be unappealing to select potential users. To that end, future improvements for HAPPI GWAS include further advancing user experience through the creation of a GUI or website, thus eliminating the command line interface. Furthermore, genomic data from additional plant species could be housed within the tool to further increase usability.

Collectively, this dissertation demonstrates the benefit of harnessing natural variation in large populations across several omics platforms and using additional datasets to confirm and further pinpoint findings. Although it has only scraped the surface to what can be mined from the large, comprehensive datasets, this work contributes novel and interesting findings to the area of seed amino acid regulation.

5.1 REFERENCES

Jiang, Y., D. Wang, D. Xu, and T. Joshi. 2020. 'IMPRes-Pro: A high dimensional multiomics integration method for in silico hypothesis generation', *Methods*, 173: 16-23.

Langfelder, Peter, and Steve Horvath. 2008. 'WGCNA: an R package for weighted correlation network analysis', *BMC bioinformatics*, 9.

Svennerstam, H., S. Jämtgård, I. Ahmad, K. Huss-Danell, T. Näsholm, and U. Ganeteg. 2011. 'Transporters in Arabidopsis roots mediating uptake of amino acids at naturally occurring concentrations', *New Phytol*, 191: 459-67.

APPENDIX A: CHPATER 1 PREFERNTIAL RETENTION OF GENES FROM ONE PARENTAL GENOME AFTER POLYPLOIDY ILLUSTRATES THE NATRUE AND SCOPE OF THE GENOMIC CONFLICTS INDUCED BY HYBRIDIZATION

Marianne Emery[1], M. Madeline S. Willis[2], Yue Hao[3], Kerrie Barry[4], Khouanchy Oakgrove[4], Yi Peng[4], Jeremy Schmutz[4,5], Eric Lyons[6], J. Chris Pires[1,7,8], Patrick P. Edger[9,10], and Gavin C. Conant[3,11-13]

[1]Division of Biological Sciences, University of Missouri-Columbia, MO, USA
[2]Department of Biochemistry, University of Missouri-Columbia, MO, USA
[3]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA
[4]Department of Energy Joint Genome Institute, Walnut Creek CA, USA
[5]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA.
[6]School of Plant Sciences, University of Arizona, Tucson AZ, USA.
[7]Informatics Institute, University of Missouri-Columbia, MO, USA
[8]Bond Life Sciences Center, University of Missouri-Columbia, MO, USA
[9]Department of Horticulture, Michigan State University, East Lansing, MI, USA
[10]Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI, USA
[11]Division of Animal Sciences, University of Missouri-Columbia, MO, USA
[12]Program in Genetics, North Carolina State University, Raleigh, NC, USA
[13]Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA

Emery, M., Willis, M. M. S., Hao, Y., Barry, K., Oakgrove, K., Peng, Y., ... & Conant, G. C. (2018). Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. *PLoS genetics*, *14*(3), e1007267.

A.1.1 ABSTRACT

Polyploidy is increasingly seen as a driver of both evolutionary innovation and ecological success. One source of polyploid organisms' successes may be their origins in the merging and mixing of genomes from two different species (e.g., allopolyploidy). Using POInT (the Polyploid Orthology Inference Tool), we model the resolution of three allopolyploidy events, one from the bakers' yeast (*Saccharomyces cerevisiae*), one from the thale cress (*Arabidopsis thaliana)* and one from grasses including *Sorghum bicolor*. Analyzing a total of 21 genomes, we assign to every gene a probability for having come from each parental subgenome (i.e., derived from the diploid progenitor species), yielding orthologous segments across all genomes. Our model detects statistically robust evidence for the existence of *biased fractionation* in all three lineages, whereby genes from one of the two subgenomes were more likely to be lost than those from the other subgenome. We further find that a driver of this pattern of biased losses is the co-retention of genes from the same parental genome that share functional interactions. The pattern of biased fractionation after the *Arabidopsis* and grass allopolyploid events was surprisingly constant in time, with the same parental genome favored throughout the lineages' history. In strong contrast, the yeast allopolyploid event shows evidence of biased fractionation only immediately after the event, with balanced gene losses more recently. The rapid loss of functionally associated genes from a single subgenome is difficult to reconcile with the action of genetic drift and suggests that selection may favor the removal of specific duplicates. Coupled to the evidence for continuing, functionally-associated biased fractionation after the *A. thaliana* At-α event, we suggest that, after allopolyploidy, there are functional conflicts between interacting genes encoded in different subgenomes that are ultimately resolved through preferential duplicate loss.

A.1.2 INTRODUCTION

Polyploidy events (also known as whole-genome duplications or WGDs) are widespread across the eukaryotic tree of life (Van de Peer et al., 2009) and have long interested geneticists and evolutionary biologists for reasons varying from the nature of interspecific crosses to the organismal effects of changes in gene copy number to the origins of novel functions in evolution (Birchler & Veitia, 2012; Clausen & Goodspeed, 1925; Ohno, 1970; Taylor & Raes, 2004). Recent work has associated genome duplications with evolutionary innovations (Conant & Wolfe, 2007; Edger et al., 2015; Merico et al., 2007; van Hoek & Hogeweg, 2009) and with shifts in net diversification rates (Kellogg, 2016; Mayrose et al., 2011; Schranz et al., 2012; Zhan et al., 2016).

Understanding how polyploidy contributes to these biologically important processes requires coming to grips with three key patterns in the evolution of polyploid genomes. The first is the rapid loss of genetic redundancy after polyploidy. Most WGD-created duplicate genes, termed “ohnologs” (Wolfe, 2000), do not survive: their losses start very soon after WGD (Gaeta et al., 2007; Scannell et al., 2007; Soltis et al., 2004) and may be governed epigenetically in this period (Edger et al., 2016). The net result of such losses can be dramatic: only 551 of an estimated 5000 duplicate gene pairs produced by the WGD in yeast survive in the *Saccharomyces cerevisiae* genome (Byrne & Wolfe, 2005). Nonetheless, the footprint of WGD is clear in the extant patterns of double-conserved synteny (DCS; Kellis et al., 2004; Wolfe & Shields, 1997): homologs of genes from a single genomic region in an non-polyploid relative will be split between two regions in the polyploid genomes (upper and lower blocks of Figure A.1).

The second key trend is that, despite the rapidity of these duplicate losses, they are nonrandom, with certain functional classes of genes being overly frequent among surviving ohnologs and other being overly rare. In both yeasts and angiosperms, genes involved in DNA repair and those targeted to the organelles were rapidly returned to single copy after WGD (Conant, 2014; De Smet et al., 2013). On the other hand, genes coding for transcription factors, ribosomal proteins and kinases were over-retained in duplicate after independent WGD events across a phylogenetically wide range of organisms from amoebae and plants to vertebrates and yeasts (Aury et al., 2006; G. Blanc & K. H. Wolfe, 2004; Maere et al., 2005; Makino & McLysaght, 2010; Seoighe & Wolfe, 1998). The force underlying these convergent patterns of loss/retention is most likely selection to maintain dosage balance among interacting gene products (Birchler et al., 2005). The *dosage balance hypothesis* explains a variety of observations about the evolution of both polyploid and non-polyploid genomes, including the pattern of post-WGD duplicate retentions (Birchler & Veitia, 2007; Conant et al., 2014; Edger & Pires, 2009; Freeling, 2009; Freeling & Thomas, 2006; Makino & McLysaght, 2010) and the tendency of these same gene families not to undergo single gene duplications, where balance would be perturbed (Dopman & Hartl, 2007; Maere et al., 2005; Wapinski et al., 2007). Similarly, genes in central network positions or whose products are parts of protein complexes are likely to show dosage phenotypes (Papp et al., 2003) and are over-retained after WGD (Bekaert et al., 2011; Conant, 2014).

The third and final trend in post-WGD evolution is that when genes are lost, they are apparently not always lost equally from the paired DCS regions. This pattern of *biased fractionation* has been observed across a range of WGD events, primarily in angiosperms

(Schnable et al., 2011; Tang et al., 2012; Thomas et al., 2006) but also from other taxa (Sankoff et al., 2010). The most plausible current hypothesis for why biased fractionation occurs is that the events in question were allopolyploidies (Garsmeur et al., 2013; Thomas et al., 2006). In the alternative case of autopolyploidy, the paired genomic regions created by polyploidy are identical, and we know of no mechanism by which these identical regions could be stably marked over evolutionary time so as to differ strongly in their duplicate retention patterns. However, the converse is *not* true: the absence of biased fractionation cannot be taken as evidence for autopolyploidy. If the genomes that merged were from closely related taxa, bias is not necessarily expected.

As for the genetic mechanism behind the bias in ohnolog losses, biases in gene expression between the two subgenomes in recent allopolyploids appear to be common (Buggs et al., 2010; Wang et al., 2006) and the chromosomal regions with lowered expression also appear more prone to ohnolog loss (Cheng et al., 2012; Schnable et al., 2011), leading to the suggestion that biased fractionation might result from a tendency for the ohnolog with lower expression to be less likely to show a fitness defect when lost. One potential source of these initial differences in expression might then be the difference in transposon load between the subgenomes of an allopolyploid, with the transposon-rich genome facing greater silencing and hence higher rates of gene loss (Garsmeur et al., 2013; Schnable et al., 2011).

A difficulty that arises in the analysis of biased fractionation (BF) is that there has been a degree of circularity in its detection. Because rearrangements occur after WGD events, the duplicated regions in a paleopolyploid genome, which are identified by shared gene order or *synteny*, will be separated from each other by breakpoints. Within each syntenic block, the identification of the homeologous region with more retained genes is straight forward. However, when comparing a single polyploid genome to a diploid outgroup, it is difficult to formally refute the possibility that the parent-of-origin of the highly retained subgenome in one block might be the same as that of the lowly retained subgenome in another (but see; Sankoff et al., 2010). This difficulty in fact motivates the phylogenetic approach to studying polyploidy that we use below. There are also other potential factors that might be involved in driving BF that remain to be investigated. For instance, the convergent pattern of rapid losses in gene coding for the DNA repair enzymes (Conant, 2014; De Smet et al., 2013) suggests that there may be incompatibilities between the versions of these genes contributed by the two allopolyploid parents. If such incompatibilities were common, they could contribute to BF by favoring retention from a single subgenome once the symmetry of a particular genetic module has been broken by the first loss.

Using POInT, the Polyploid Orthology Inference Tool, we analyzed the resolution of three WGD events, one in yeasts (Wolfe & Shields, 1997), one in the grasses (the ρ event; Paterson et al., 2004; Van de Peer et al., 2017) and the most recent event (At-α) in *Arabidopsis thaliana* and its relatives. Using POInT's synteny-based estimates of post-WGD gene losses, we show that BF was a genome-wide evolutionary pattern after the At-α and ρ WGD events and persisted over long periods. In contrast, in yeasts we find evidence for BF only in a very short time interval post-WGD. In *Arabidopsis*, we also find that there is preferential co-retention of genes from the same subgenome whose products interact, as opposed to interactions involving proteins from different parents. Collectively, these results suggest that biased fractionation is at least in part a relic of conflicts between the paralogous genes contributed by the two parents at the time of the allopolyploidy.

A.1.3 METHODS

Identifying double-conserved synteny blocks in polyploid genomes

Our previous POInT analyses in yeast were based on human curated datasets (Byrne & Wolfe, 2005; Gordon et al., 2009). We do not have such inferences for either the At-α or the grass ρ event. Instead, using experience from previous projects (Joyce et al., 2017; Tang et al., 2012), we developed a new pipeline for inferring the paralogous genomic regions created by a WGD in the genomes sharing that event. We then merged these regions of DCS (Kellis et al., 2004; Wolfe & Shields, 1997) across all polyploid genomes and sought an ancestral gene order that minimized the number of synteny breaks. Figure A.1 shows examples of such DCS blocks for At- α.

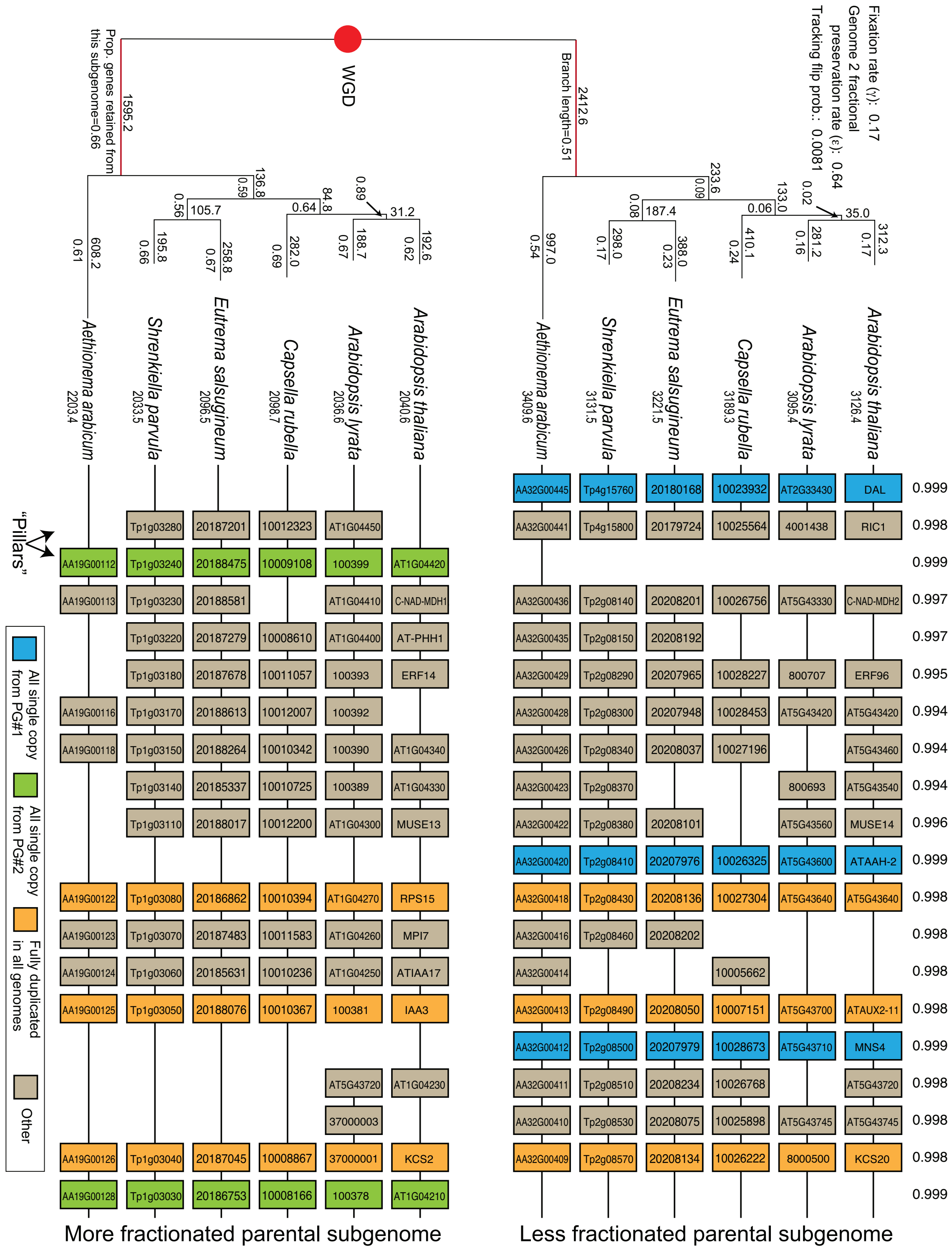


Figure A. 1. POInT's inferences regarding the loss of genes post-WGD. The At-α duplication produced two sets of homoeologous regions, one from the parental subgenome with more surviving genes ("Less fractionated subgenome," upper track) and one with fewer ("More fractionated subgenome," lower track). Genes in these tracks may have surviving duplicates in at least some taxa (orange/tan), or they may be single-copy in all species (blue if derived from the less fractionated subgenome and green if from the more fractionated one). Under each taxon name is the number of single-copy genes predicted to have been retained from that parental subgenome in that taxon. The branch length (numbers under the branches of the upper tree) gives the value of α×time in the model of Figure A.2B: larger values correspond to a relatively higher chance that a position with a ohnolog pair present at the start of a branch will be single-copy by its end. Numbers above the branches give POInT's estimate of the number of genes returned to single copy deriving from the less fractionated (upper panel) and more fractioned (lower panel) subgenomes, respectively. Under the branches of the lower tree are the branch-specific ratio of genes retained from subgenome #2 relative to subgenome #1: these values can be compared to the overall estimate of this parameter, which is 0.64, shown in the upper left. POInT's estimates of the other global parameters for this model are also given here. Above each pillar of genes is POInT's estimate of the posterior probability of the set of subgenome assignments depicted, relative to the other 2^n-1 possible assignments (where n is the number of genomes). The two root branches are shown in red: these correspond to branches where the biased fractionation parameter ε was allowed to differ from the rest of the tree in our analyses of temporal patterns of biased fractionation (Methods). Similar trees depicting loss events for the grass and yeast WGDs are given as S1 Figure.

The goal of the pipeline is to find a common set of DCS blocks shared by the genomes of the six Brassicaceae species that possess At-α: *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000)*, Arabidopsis lyrata* (Hu et al., 2011)*, Capsella rubella* (Slotte et al., 2013)*, Shrenkiella parvula* (Dassanayake et al., 2011), formerly known as *Thellungiella parvula* or erroneously as *Thellungiella halophila* (Koch & German, 2013), *Eutrema salsugineum* (Yang et al., 2013)*,* and *Aethionema arabicum* (Haudry et al., 2013) and for the four grasses with ρ: *Brachypodium distachyon* (Initiative, 2010), *Oropetium thomaeum* (VanBuren et al., 2015), *Setaria italica* (Zhang et al., 2012) and *Sorghum bicolor* (Paterson et al., 2009). To do so, we used outgroup genomes that lacked the WGD in question. For the At-α event, we used the draft genome of the outgroup plant *Cleome violacea*, which split from the six taxa studied prior to that event (Schranz et al., 2012): it likewise lacks the WGD found in other taxa in the Cleomaceae (Edger et al., 2015). The *C. violacea* genome is available from the CoGe comparative genomics portal (https://genomevolution.org/coge/) under accession number 23822. For the grass ρ event, we used the genome of the pineapple *Ananas comosus* as an outgroup (Ming et al., 2015). CoGe accession numbers for all plant genomes used are listed in S1 Data.

The product of a WGD is a set of duplicated genes in a genome that each originate from a single ancestral gene. Here, the *C. violacea* and pineapple genomes give us an estimate of these ancestral loci, and we seek to place either one (e.g., a duplicate loss has happened) or two genes (the ohnologs survive) from the duplicated genome in a "pillar" with each such ancestral gene (see Figure A.1). Genome annotation files for these 12 plant genomes were obtained from CoGe (Lyons & Freeling, 2008). With these data in hand, the inference of the shared DCS blocks that serve as POInT's input is a three step process: 1) a homology search of each polyploid genome against the diploid outgroup, 2) inference of species-specific DCS blocks and 3) inference of a common set of DCS blocks across all genomes along with an estimate of their ancestral order at the time of the polyploidy.

Step 1: Homology search. For At-α, we used a fast homology search program based on the SeqAn package (Doring et al., 2008; Taxis et al., 2015) to identify pairs of homologous genes, one from a genome with At-α and one from *C. violacea*. We defined a pair of genes as being homologous for the purposes of DCS inference if their protein sequences: 1) share two 7 amino acid residue exact matches, 2) have the shorter sequence having 80% of the length of the longer, and 3) show 70% amino acid identity overall. Because of the greater evolutionary distances involved in the grass ρ event, we used a slower but more sensitive BLAST-based search, employing our tool GenomeHistory to do so (Altschul et al., 1997; Conant & Wagner, 2002). In this case, we required a maximal BLAST E-value of 10^{-8} to identify matches between the four duplicated grass and pineapple: we then used the same 70% identity and 80% aligned length cutoffs as used with At-α to select homologs.

Step 2: Genome-specific DCS inference. Sequence homology alone is insufficient to identify the DCS blocks given the angiosperms' history of nested polyploidy (Van de Peer et al., 2009). Instead, for the second step of the pipeline, we used gene order information (synteny) to identify which of the potentially many homologs in each polyploid genome are the WGD-produced ohnologs. We frame this problem as follows. First, we define a set A of n DCS blocks that consists of ancestral pillars A_i such that $A_i \in A | 1 \leq i \leq n$. Each pillar is linked to a unique gene from *C. violacea* or pineapple and has elements $A_i(p_1)$ and $A_i(p_2)$, which represent the potential homologous genes created by WGD. Each pillar A_i also has associated a set of genes $\{h_1 ... h_h\}$ from the polyploid genome that are homologous to the pillar's ancestral gene. A maximum of two of these homologs can be assigned to $A_i(p_1)$ and $A_i(p_2)$. We next define $O(A_1...A_n)$ to be the order of the pillars in A for our analysis. Hence, $A_{O(i)}$ represents the i^{th} pillar in this ordering. For a given $A_{O(i)}(p_k)|1 \leq k \leq 2$, define $A_{O(i+j)}(p_k)$ such that j=min(x; i+1≤x≤n) where $A_{O(i+x)}(p_k) \neq \emptyset$: in other words, $i+j$ is the next pillar after i in $O(A_1...A_n)$ with an assigned gene for parental genome k. We define the score s of such a combination of homolog assignments and pillar orders:

$$s = \sum_{i=1}^{n}\sum_{k=1}^{2}\left|\begin{matrix}1 & A_{O(i)}(p_k)\ and\ A_{O(i+j)}(p_k)\ are\ neighbors \\ 0 & otherwise\end{matrix}\right. \quad (1)$$

In other words, the score is the sum of the number of positions in $O(A_1...A_n)$ where the genes in each pillar are the genomic neighbors of the genes in the next non-empty position. We cannot simply use the pillar order seen in the outgroup, because neither *C. violacea* nor pineapple is the true ancestor of the WGD events in question: both have evolved independently for many millions of years. Instead we must optimize $O(A_1...A_n)$. Note that, throughout this pipeline, neighbor is understood to exclude any genes that are not part of the current analysis set. For instance, a gene in *Arabidopsis thaliana* with no identified *C. violacea* homolog is ignored in the neighbor computation because it could never appear in an ancestral pillar. By the same logic, any position for which $A_{O(i)}(p_k)$ and $A_{O(i+j)}(p_k)$ are not neighbors is defined as a synteny break, and, if this situation is true for both $k=1$ and $k=2$, we refer to position i as having a double synteny break.

To infer the combination of the homolog assignments $A_i(p_k)|1 \leq i \leq n, 1 \leq k \leq 2$ and the ordering $O(A_1...A_n)$, we used simulated annealing (Conant & Wolfe, 2006; Kirkpatrick et al., 1983). This algorithm proposes random changes to either $O(A_1...A_n)$ or to the $A_i(p_k)$ assignments with the goal of maximizing *s*, which recomputed after each such change. We used the extant *C. violacea* and pineapple gene orders as our initial orders and made increasingly long runs until longer run times no longer produced meaningfully higher values of *s*.

A. thaliana and its relatives share a history of WGD (Maere et al., 2005): prior to the WGD-α event modeled here there was another WGD, termed WGD- β which is shared with *C. violacea.* One might wonder if our simulated annealing algorithm has mistaken synteny blocks surviving from WGD-α for the more recent products of WGD-α. We suspect that any such errors are quite rare for two reasons. First, *C. violacea* also experienced WGD- β and hence also possesses the corresponding synteny blocks, meaning that they are accounted for in the inputs to our simulated annealing routines. Second, we only considered homology relationships between genes in *C. violacea* and in *A. thaliana, A.lyrata, C. rubella, S. parvula* and *E. salsugineum* with nonsynonymous divergence (K_a) less than 0.1 and between *C. violacea* and *A. arabicum* with $K_a \leq 0.2$. As a result, between 41% and 45% of the genes from *C. violacea* have only a single homolog identified in the other 6 genomes and hence cannot represent ambiguous surviving blocks from WGD- β in *C. violacea*. Hence, it is difficult to see how ancestral WGD- β blocks would have infiltrated our inferences in significant numbers.

Step 3: Inferring a global ancestral ordering for POInT analyses. Using the four/six individually optimized set of ancestral pillars (for ρ and At-α, respectively) with assigned genes (the $A_i(p_k)$ values for each genome), we extracted, for each genome, only ancestral pillars for which each gene in the pillar had synteny support (i.e., each gene was a neighbor of at least one other gene in that pillar set). Using the outgroup gene from each ancestral pillar as an index, we then merged all of these inferences. Because we required that at least one gene from each genome be in each pillar, the effect of this merging was to limit our analyses to a set of m=7243 and =3091 ancestral pillars for At-α and ρ, respectively. However, those pillars have shared syntenic support across all genomes. The optimal ancestral order for each extant genome differs, so once the ancestral pillars were assembled, we inferred a globally-optimal ancestral order $O(AG_1..AG_m)$, again using simulated annealing. The optimality criterion here was to maximize the number of neighbor relationships, but in this case the $A_i(p_k)$ assignments were held constant and only $O(AG_1..AG_m)$ was changed.

To assess the influence of the ancestral ordering on POInT's estimates, we fit the WGD-*bf* model (Figure A.2B) to both the initial *C. violacea* order and to the 10 inferences of $O(AG_1..AG_m)$ with the largest simulated annealing scores, using the order with the highest likelihood for further analyses (S1 Table). We similarly used the ancestral ordering of highest likelihood for our ρ analyses.

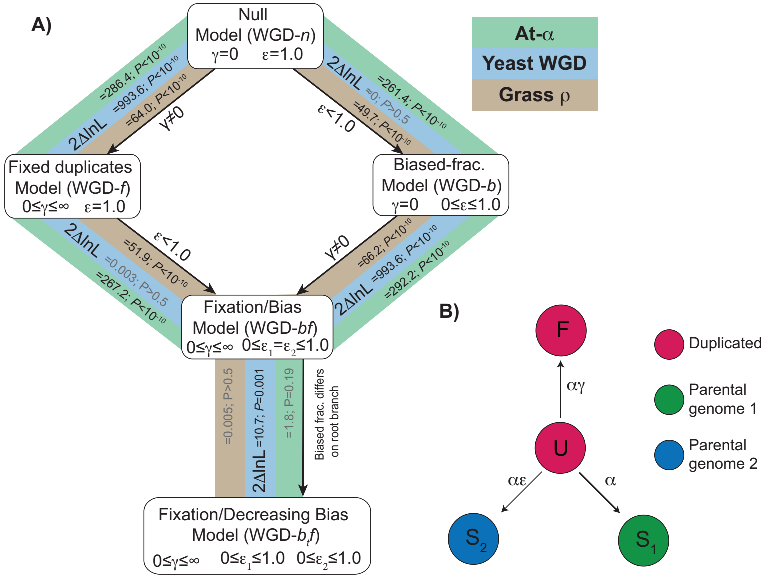


Figure A. 2. Modeling WGD resolution with POInT. We employed a number of models of the fates of the duplicates produced by WGD. **A)** Statistical relationships between the various models for the yeast WGD (blue), At-α (green) and ρ (brown) events. The simplest model (WGD-n) considers only a balanced process of gene loss. From this model, we can either allow duplicate genes to become fixed (for instance by neo- or sub-functionalization, WGD-f) or for one of the two parental subgenomes to lose more genes than the other (WGD-b). Using a likelihood ratio test (LRT), we find that, for all three WGD events, allowing duplicate fixation significantly improves the fit of the data to the models ($P<10^{-10}$, LRT, *Methods*). However, for the yeast dataset, there is no significant evidence for biased fractionation (P>0.5, LRT), while for two plant WGDs, adding it significantly improves the fit ($P<10^{-10}$; LRT). From these two models, we can then allow the other process. Again, for yeast, there is significant evidence for fixation but not biased fractionation ($P<10^{-10}$ and P>0.5, respectively, LRT) while for At-α and ρ, there is significant evidence for both ($P<10^{-10}$ in each case, LRT). We also tested a model where the biased fractionation parameter ε (see panel **B**) was allowed to differ on the shared root branch of the tree (WGD-b_tf) compared to all of the other branches. For the two plant WGD events,there is no significant evidence that the level of biased fractionation differed early in history of the WGD relative to later in time (P≥0.19, *Results*). On the other hand, for the yeast WGD, biased fractionation was much more intense soon after the polyploidy event and weakened later (P=0.001; *Results*). **B)** Model states and parameters. Our model has four states, two duplicated ones (**U**=undifferentiated duplicates and **F**=fixed duplicates) and two single copy states (**S_1** and **S_2**, corresponding to the two parental subgenomes). The base loss rate (α) is compounded with the estimated time to give the branch lengths of Figure A.1. The relative fixation rate γ (0≤ γ <∞) gives the rate of duplicate fixation relative to the loss rate α. Likewise, the fractionation bias parameter ε (0≤ ε ≤1) gives the excess of preservations from subgenome 1 relative to subgenome 2 (assumed to be the more fractionated subgenome).

Extracting a "high synteny" subset of ancestral pillars.

To assess if the fragmentation of synteny blocks was artificially leading us to invoke BF, we also extracted from our full At-α dataset a smaller set of ancestral loci with strong syntenic support, including only pillars with full syntenic support in at least one direction (e.g., two links per pillar per genome). The result was a dataset of m_h=4556 ancestral loci for which we also inferred an optimal ancestral ordering. No such analysis was performed for α due to the small total number of ancestral pillars found. Table 1 gives the parameter estimates from all four datasets for various ancestral orders.

Table A. 1. POInT estimates for different datasets and ancestral orders.

Description	**Ancestral loci**[a]	**#breaks**[b]	**#double breaks**[c]	**WGD-*bf* lnL**[d]	**Fixation rate (γ)**[e]	**Bias strength (ε)**[f]
At-α, Full: *C. violacea* order	7243	6614	3021	-25357.46	0.160	0.538
At-α, Full: Optimized order	7243	5468	1129	-24497.04	0.169	0.645
At-α, High-synteny: *C. violacea* order	4556	3544	1039	-12837.67	0.205	0.718
At-α, High-synteny: Optimized order	4556	2266	252	-12442.51	0.220	0.786
Grass ρ, Pineapple order	3091	4387	2299	-8822.89	0.049	0.400
Grass ρ, Optimized order	3091	2457	434	-8199.10	0.061	0.730
Yeast WGD	4065	4346	796	-19374.10	0.137	0.955[g]

a: Number of ancestral loci studied.
b: Number of synteny breaks across the polyploid genomes.
c: Number of cases where both parental subgenomes showed a synteny break after an ancestral locus (see *Methods*).
d: ln-likelihood from fitting WGD-*bf* to this ancestral order.
e: Maximum likelihood estimate of the relative duplicate fixation rate for this ancestral order (see Figure A.2).
f: Maximum likelihood estimate of the relative rate of retention from the more fractionated subgenome for this ancestral order (see Figure A.2).
g: γ not significantly different from 1.0; see Figure A.2.

Modeling the evolution of WGD events with POInT

We have previously described POInT (Conant, 2014; Conant & Wolfe, 2008), which fits a Markov model to duplicate loci created by WGD. The model has four states (Figure A.2B), namely **U** (undifferentiated duplicated genes), **F** (fixed duplicate genes) and $\mathbf{S_1}$ and $\mathbf{S_2}$ (the single copy states): it is a generalization of a model proposed by Lewis (2001). Note that once the genes of each post-WGD genome have been assembled into ancestral pillars using the simulated annealing approach above, the *sequences* of the genes of the post-WGD genomes are never used again: all of POInT's inferences are based on shared DCS information. Since our prior work, we have completely re-written POInT to allow for user-defined evolutionary models, computing the resulting transition probabilities by exponentiating the user-supplied instantaneous rate matrix (Muse & Gaut, 1994). Using this new version of POInT, we fit five models to our four datasets (two from At-α and one each from the yeast and grass WGD events, Figure A.2). We used likelihood ratio tests to assess whether more complex models better fit the loss data than did simpler models (Sokal & Rohlf, 1995).

POInT's focus on WGD has advantages over applying more general gene birth-death models to polyploid species (De Bie et al., 2006; Rabier et al., 2014). POInT models the process of duplicate loss and retention jointly across all genomes and along a phylogeny. Hence, the probability of a particular model state at a given ancestral locus is conditioned on all other loci and all other genomes. This conditioning is performed by analogy to the linkage analysis model of Lander and Green (1987) using the hidden-Markov approach of Felsenstein and Churchill (1996). The states the Markov model considers are the set of 2^n possible orthology relationships between the $2n$ different loci (e.g., 2 duplicated loci in each of *n* genomes). The likelihood of site *i+1* having orthology state *j* given that site *i* has that orthology assignment is (1-θ), where θ is a small constant estimated from data ($0.0004 \leq \theta \leq 0.0081$ across these analyses). In cases where there is a double break in gene order in a particular genome, θ=0.5.

From this model structure, we can infer orthologous chromosomal regions produced by WGD between the genomes studied, along with confidence estimates in these inferences (Figure A.1). The previous version of POInT did not distinguish between states $\mathbf{S_1}$ and $\mathbf{S_2}$. The result was degeneracy in the inferences of orthologous regions. In other words, assigning the first member of each DCS pair to subgenome 1 and the second to subgenome 2 produced orthology assignment 111111 across the six genomes, which was identical in likelihood to assignment 222222. (The computation is completely analogous for the other two WGD events studied.) Effectively, this degeneracy corresponds to flipping the upper and lower panels of Figure A.1, because each of the 2^n possible orthology assignments has an equivalent assignment with all 1s converted to 2s and *vice versa*.

To model the process of BF, we relaxed this assumption by introducing parameter ε (Figure A.2B). This parameter makes losses to state $\mathbf{S_2}$ potentially less common than to $\mathbf{S_1}$. If BF is present in the data, the maximum likelihood estimate of ε will be less than 1.0, and the likelihood of orthology assignment 111111 will no longer be the same as 222222. We can then use the POInT model to estimate the posterior probability of the subgenome assignments (the numbers shown above every column in Figure A.1) at every pillar. For convenience we refer to the resulting two regions as deriving from allopolyploid parents 1 and 2 (Garsmeur et al., 2013), respectively, defining parent 1 as containing genes in state $\mathbf{S_1}$ (e.g., it is potentially less fractionated), similar to Thomas et al., (Thomas et al., 2006).

In previous work in yeast (Conant, 2014; Conant & Wolfe, 2008; Scannell et al., 2007), we found evidence for "convergent" gene losses that were phylogenetically independent and yet more often from the same subgenome than could be explained by chance. We modeled these events by adding two duplicated converging states to our model, $\mathbf{C_1}$ and $\mathbf{C_2}$. Gene losses from $\mathbf{C_1}$ were always to $\mathbf{S_1}$ and similarly for $\mathbf{C_2}$. We fit versions of this model both with ($0 \leq \varepsilon \leq 1.0$) and without ($\varepsilon = 1$) BF to our yeast, grass and At-α data: while these models improved the fit relative to the WGD-*bf* model used here, we present our results in terms of the WGD-*bf* model because both model classes give similar parameter estimates (S2 Table), and the more complex models do not add insight for the questions considered here.

Dependence of POInT parameter estimates on the assumed phylogeny

Because we analyzed only four genomes sharing the grass ρ event, it was possible to use POInT to test all 15 possible rooted phylogenetic trees for these taxa to assess the dependence of our inferences on the inferred phylogeny. We present our results in terms of the optimal tree, but the global parameter estimates for the WGD-*bf* model were very similar for all topologies ($0.061 \leq \gamma \leq 0.067$; $0.719 \leq \varepsilon \leq 0.739$; $0.0061 \leq \theta \leq 0068$; Figure A.2).

Network analyses of biased losses

We asked whether genes surviving from one or the other of the subgenomes showed patterns of interconnection in the networks of *Arabidopsis thaliana*. We use the BioGrid database (Stark et al., 2011) to extract known protein-protein interactions (*Arabidopsis* Interactome Mapping Consortium, 2011). We tested for paucity of interactions between the products of genes from different subgenomes with a randomization approach. We thus compared the number of interactions between gene products from alternative subgenomes in the actual data to this value computed after 1000 randomizations of the subgenome assignments. To assess the degree to which our conclusions were potentially affected by errors in the assignment of genes to subgenomes, we conducted our tests at a range of confidences in subgenome assignment (Figure A.3).

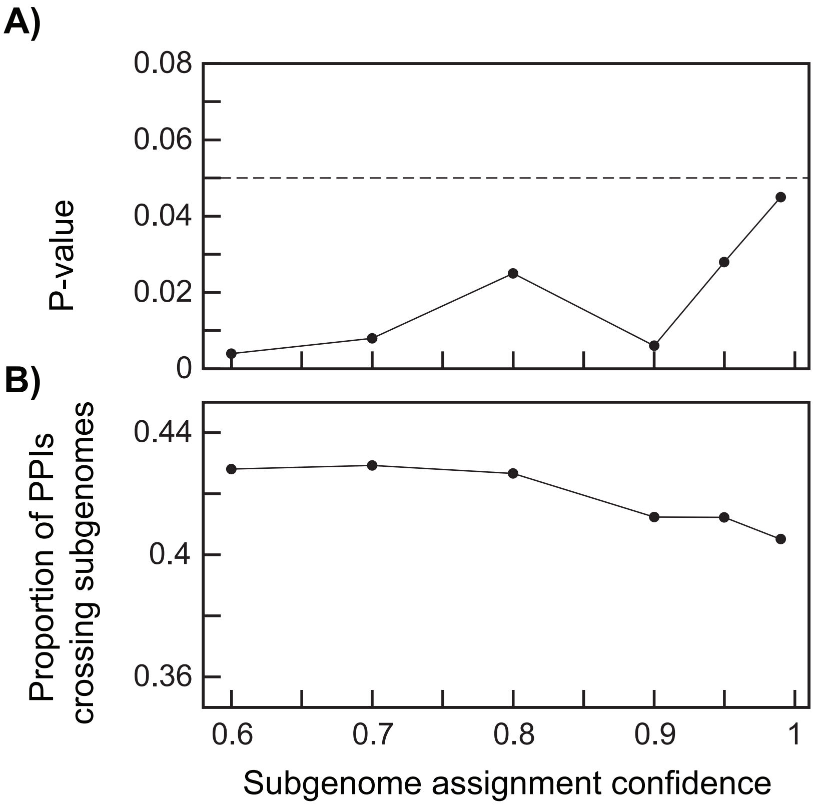


Figure A. 3. Protein interactions between single-copy genes from alternative subgenomes are rarer than expected. We extracted single-copy genes for a range of values of POInT's overall confidence in pillar assignments to subgenomes (x-axis) and computed the P-value for the test of the null hypothesis of no fewer protein-protein interactions between products of genes from alternative subgenomes than expected (y-axis; panel **A**: see *Methods*). We also computed the frequency of such "crossing" interactions relative to interactions between products of the same subgenome (y-axis, panel **B**).

GO analyses of biased losses

We used the Gene List Analysis tool from the PANTHER classification system (Mi et al., 2017) to perform statistical overrepresentation tests to find over/under-represented Gene Ontology (GO) terms associated with biological processes, molecular functions, or cellular components. The input of our analysis consists of two sets of genes: the target list to analyze, and a reference list. The expected number of genes for a GO term in the target list was calculated based on the number of genes with that term in the reference list: binomial statistics for each GO term associated with genes in the target list were then computed from these expectations (Mi et al., 2013).

We first performed an overrepresentation test for 4,086 single copy genes from both subgenomes against the reference set of 4,152 surviving duplicated genes. The over/under represented GO terms in the analysis were filtered with a threshold *P*-value ≤ 0.01 after Bonferroni correction, and only terms with a fold-enrichment larger than 1.5 (overrepresented) or smaller than 0.67 (underrepresented) are reported. We next compared 2,552 single copy genes from subgenome 1 (dominant) relative to the terms for the 1,534 genes from subgenome 2 (more fractionated) with a similar approach. To compensate for the smaller number of terms found to be enriched in this second analysis, we used an FDR-corrected *P*-value of 0.05 as a threshold. Full lists of all significantly enriched terms for any comparison with associated GO identifiers are given as S3-5 Tables.

A.1.4 RESULTS

Modeling WGD evolution with POInT

Using POInT, we analyzed the resolution of three phylogenetically widely-spaced polyploidy events: the WGD in the ancestor of *Saccharomyces cerevisiae* and relatives (Marcet-Houben & Gabaldon, 2015; Wolfe & Shields, 1997), the ρ event found in the ancestor of the grasses (Paterson et al., 2004; Van de Peer et al., 2017) and the At-α event shared by the model plant *Arabidopsis thaliana* and its relatives (G. Blanc & K. H. Wolfe, 2004; Maere et al., 2005). Previous work has suggested that all of these WGDs were allopolyploid events (Garsmeur et al., 2013; Marcet-Houben & Gabaldon, 2015), meaning the duplicated regions in the extant polyploid genomes (hereafter subgenomes) derive from parental genomes from differing species. Whatever their origins, however, these subgenomes produced by polyploidy are now distinct due to their individual histories of gene loss. In order to assign the extant genes to one of the two subgenomes, we applied new duplicate resolution models that distinguished between a less fractionated genome (more surviving genes) and the more fractionated genome (fewer surviving genes; Garsmeur et al., 2013; Thomas et al., 2006).

As previously described (Conant, 2014; Conant & Wolfe, 2008; Scannell et al., 2007), we used ohnologs from the Yeast Genome Order Browser project and an inferred ancestral genome order as POInT's inputs for the yeast analyses (Byrne & Wolfe, 2005; Gordon et al., 2009). For the At-α and ρ events, no such data exist, so we developed a new pipeline that uses sequence homology and shared gene order (synteny) to assign genes from the polyploid genomes to a "pseudo-ancestral" gene from the extant outgroups *Cleome violacea* (for At-α) and pineapple (for ρ). First, we used simulated annealing to assign genes from each of the polyploid genomes to double-conserved synteny (DCS) blocks. These assignments were made forcing pairs of regions in the polyploid genomes to possess one or two homologous genes to one gene from a single region in outgroup genome: the simulated annealing algorithm then sought such assignments that maximized the shared gene order (see *Methods* for additional details). We then merged these single-genome inferences into a set of 7243 and 3091 (for At-α and ρ, respectively) ancestral gene pillars, each consisting of at least one gene from every genome that shared synteny with at least one other gene (see Figure A.1). We then again used simulated annealing to optimize our estimate of ancestral genome order of these loci by maximizing the synteny among the pillars. Figure A.1 gives an example of the estimates made by POInT based on these inferred pillars: from the inferred pillar order, POInT is able to estimate the probability associated with assigning each genome segment from each species to either of the two subgenomes (numbers above the columns in that figure).

Using these data, we tested the hypothesis that biased fractionation (BF) was observed after the three WGD, explored its temporal characteristics and sought to associate it with functional properties of the genes in question.

Biased fractionation was common after At-αand ρ.

By fitting nested models of evolution to these datasets, we tested for the presence of ohnolog fixation and biased fractionation after the three WGD events. Fixation (WGD-*f*, Figure A.2A) is inferred when a WGD-produced duplicate pair has persisted across the tree longer than would be expected given the loss rates. There is evidence of such fixation events after all three WGDs ($P<10^{-10}$, likelihood ratio test, Figure A.2 and *Methods*). We model biased fractionation (BF, WGD-*b*, Figure A.2A) as a preference for losses of genes from subgenome 2 ($0\leq \varepsilon \leq1$, Figure A.2B) over subgenome 1. Note that the identity of subgenome 2 is inferred from the data and bespeaks no lack of generality in our model.

At-α and ρ show strong evidence of BF ($P<10^{-10}$, likelihood ratio test, Figure A.2A and *Methods*). However, similar to previous analyses of the yeast WGD (Kellis et al., 2004), we find no statistical evidence for a *general* BF process after the yeast WGD ($P>0.5$, likelihood ratio test, LRT, Figure A.2A). Our estimate of the strength of BF after At-α is nearly identical to that found by Thomas and coauthors when considering only the *A. thaliana* genome (Thomas et al., 2006), with the more fractionated subgenome showing approximately 2 single copy genes deriving from it for every 3 from the less fractionated subgenome. The bias estimated for the ρ event was slightly weaker: 3 genes from the more fractionated genome retained for every 4 from the other subgenome. We note that these estimates vary somewhat depending on the quality of the syntenic data used as the input for POInT: when we used the highly non-optimal *C. violacea* gene order (which has many more syntenic breaks), the estimated ratio of single copy genes from the more and less fractionated genomes was closer to 1:2 (S1 Table). However, it is unlikely that further order optimization would raise the estimates of the BF parameter ε (e.g., imply less fractionation): all of the estimated ancestral orders gave similar estimates of ε with no trend of increasing ε with smaller numbers of breaks (S1 Table). Likewise, we inferred a "highly syntenic" dataset of 4556 ancestral pillars for At-α that included only pillars with fully syntenic connections to at least one other pillar (*Methods*). While the estimate of ε for this dataset is higher than that for the full dataset (Table 1), it is still significantly different from 1.0 ($P<10^{-10}$). Moreover, some of the increase in ε here may be attributable to the greater number of surviving duplicates (larger γ, see Table 1).

Biased fractionation occurred in a brief interval after the yeast WGD but has been a continuous process after At-α and ρ.

The process of duplicate loss immediately post-WGD differs from that observed later (Conant, 2014; De Smet et al., 2013). We hence fit a model where the strength of BF was allowed to differ on the shared root branch (Figure A.1) relative to the remaining branches. For At-α and ρ there is no significant evidence for such a difference (ε_{early}=0.67/0.74, ε_{late}=0.63/0.73, for At-α and ρ, respectively; *P*≥ 0.19). However, the strength of biased fractionation immediately after the yeast WGD was much higher than that seen later (ε_{early}=0.47, ε_{late}=0.99; *P*=0.001), showing that our initial conclusion of no BF in yeast was an artifact of low temporal resolution in the WGD-*bf* model. Approximately 277 single-copy genes from the less fractionated parent, and only 135 from the more fractionated one, were returned to single-copy along the shared root branch following the yeast WGD (S1 Fig.). We note that it is difficult to directly compare the yeast and plant results because of the differing shape of the post-WGD phylogenies for the datasets. The yeast WGD was characterized by very rapid post-WGD speciation (Scannell et al., 2006; Scannell et al., 2007): thus only 412/4099 (10%) of the ohnolog pairs had lost a gene before the first speciation (S1 Fig.). On the other hand, the taxa sharing At-α had undergone ohnolog losses at 4008/7243 (55%) of the ancestral positions before the speciation event that split *Aethionemae arabicum* from the other Brassicaceae (Figure A.1), with a similar proportion of losses on the root branch after ρ (S1 Fig.). The phylogenies reflect this difference, with POInT's estimate of the length of the root branch in the yeast analysis being 0.063 verses 0.55 for At-α and 0.63 for ρ (recall that branch lengths are proportional to the probability of an ohnolog loss along that branch). The tribe Aethionemae is sister to the remainder of all extant Brassicaceae species (Huang et al., 2016). Hence, at least for At-α, there might have been short period of more intense biased fractionation that we cannot detect due to the lack of an extant early diverging lineage such as those we have studied in the yeasts.

Biased fractionation is a genome-wide phenomenon

As mentioned, it is not guaranteed that two genomic regions each showing a higher retention rate than their homeologous partners necessarily originate from the same parental subgenome (the circularity problem in measuring BF). We used the high-synteny subset of the At-α data to assess the degree of this problem. From it, we produced visual representation of the set of ancestral synteny blocks POInT was using for its inferences. In Figure A.4B, we show how often 5, 4, or 3 genomes agree from pillar to pillar in their subgenome assignments. Notably, when only 3 of 6 genomes are required to agree at high probability, the model infers a relatively small number of ancestral syntenic blocks, consistent with a set of ancestral chromosomes prior to At-α. Moreover, these blocks are identifiable without the assumption of BF (e.g., they are also inferable from the WGD-*f* model, Figure A.4B) and, at least for most of the larger blocks, give estimates of BF similar to the dataset as a whole (Figure A.4C). Hence, it is clear that biased fractionation is not an artifact of synteny-block inference. Similar diagrams for the full At-α dataset, the ρ dataset and yeast are given in S2 Fig.

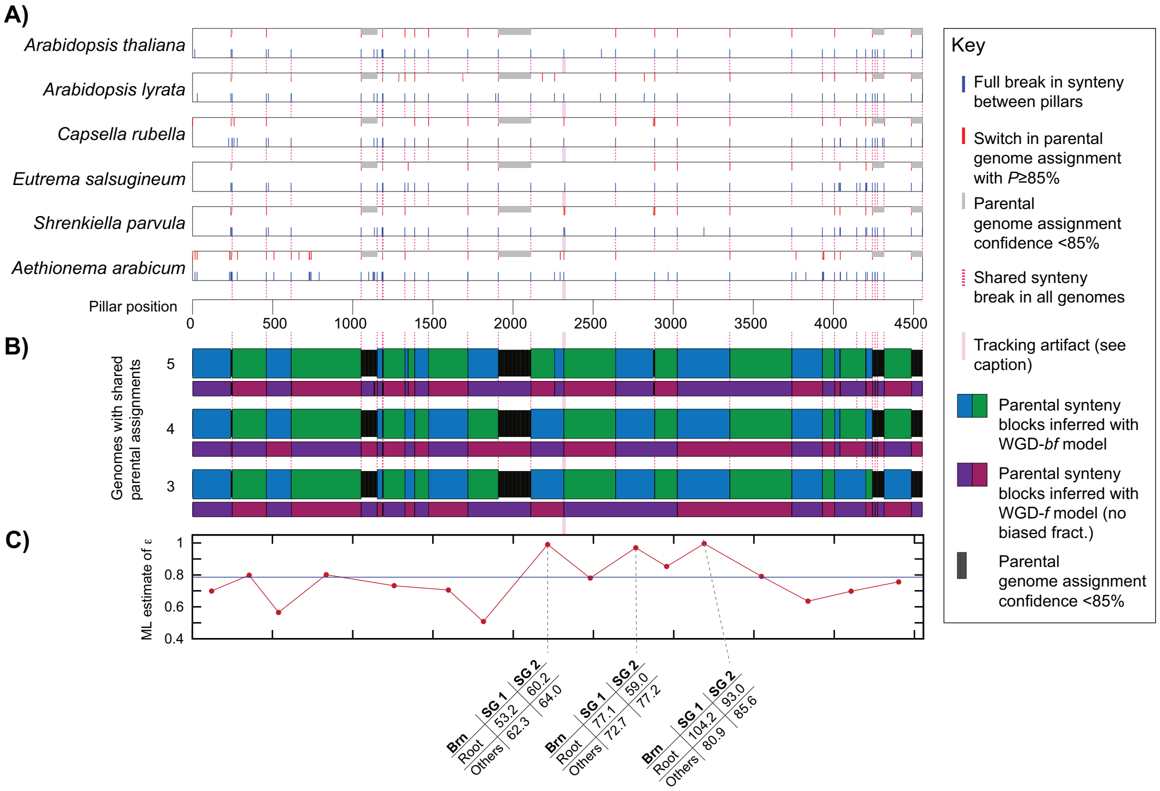


Figure A. 4. Consistency across the ancestral genome of POInT's estimates of the subparental genome of origin. **A)** In the six panels, we illustrate how often POInT's assignment of parental subgenome of origin for At-α changes between two successive pillars when considering the "high synteny" dataset. A red tick at position i corresponds to a situation where POInT assigned parents-of-origin to two chromosomal regions at position i-1 with probability of ≥85% and either the *opposite* combination of parents at position i or with the same assignment but with confidence less than 85%. Gray ticks, in turn, correspond to those positions immediately after a red tick where the confidence in the parental assignments is less than 85%. The blue ticks in the lower half of each block indicate positions where there is a double synteny break after position i-1 (see *Methods*). At these positions, the parental inferences at position i are independent of those at *i-1*. Locations where all 6 genomes have such breaks are shown with the pink dotted lines. **B)** Estimates of shared parental blocks across genomes. With very few exceptions, locations where POInT finds a change in subgenome assignments correspond to these six-fold synteny breaks from **A**. Each blue/green colored block corresponds to a situation where at least 5, 4, or 3 genomes (top, middle and bottom, respectively) agree between every neighbor as to the subgenome assignment at a confidence of 85% or more. Narrower black regions are regions where there is no position-to-position agreement in assignment for any number of genomes (e.g., these are regions where our confidence in subgenome assignments is low overall). Any shared loss of synteny can induce a new block: such synteny breaks might, for instance, reflect a shift to new ancestral chromosome. For reference, we also show the set of blocks inferred with the WGD-f model as the smaller set of red/purple blocks. This model does not include BF, making it degenerate, so that subgenome 1 and 2 can be swapped. We therefore define one region of one genome as being subgenome #1 and make the block assignments correspondingly. Almost all of the phasing of blocks can be done without the assumption of BF, as is seen with the similarity between the blue/green and red/purple blocks. The implication of this fact is that the blocks are defined by the pattern of shared gene losses and that including BF in the model serves only to allow us to assign unlinked blocks to the same subgenomes based on their BF patterns. **C)** For the 16 blocks with more than 100 pillars, we show the estimates of the strength of BF (maximum likelihood estimate of ε; y-axis) judged solely from that block (block mid-point on the x-axis). These values indicate strong BF in all but three cases: in most of the larger blocks the estimated strength of BF is nearly identical to that for the full dataset (blue line). For the three blocks with weak evidence for BF ($\varepsilon \approx 1.0$), we further interrogated the patterns of gene loss (tables at bottom). In two of three cases, the signal of BF is relatively strong along the shared root branch where most losses occurred, with conflicting patterns on other branches. We attribute these differences to sampling effects among the relatively small number of losses along each branch. For the final block, with coordinates from pillars 2113 to 2318, the inferred pattern of losses contradicts the subgenome assignment, with more inferred losses from subgenome 1. When we examined the pattern of synteny breaks in this region, we discovered an anomaly: all of the genomes except *Eutrema salsugineum* had a synteny break at the end of this block: E. salsugineum instead had a break six pillars later (the pink shaded region). Hence, this synteny pattern caused the block to be linked to the next, larger, block, giving rise to the incongruous gene loss inferences. Equivalent figures for the full At-α dataset, the yeasts and the grasses are given as S2 Fig.

Protein products of single copy genes from different subgenomes rarely physically interact

Using data from BioGrid (*Arabidopsis* Interactome Mapping Consortium, 2011; Stark et al., 2011), we asked whether protein-protein interactions between the products of *A. thaliana* single-copy genes from alternate subgenomes were rarer than would be expected by chance. Across a large range of subgenome confidence estimates from POInT, there were fewer such "crossing" interactions than expected (Figure A.3A), and the frequency of such interactions decreases as our confidence in the subgenome assignments increases (Figure A.3B). Similar analyses were not performed for the ρ and yeast WGD events due to the lack of large-scale interaction data and the lack of substantial fractionation, respectively.

BF has retained genes of distinct functions from each subgenome

As seen in previous analyses (G. Blanc & K. H. Wolfe, 2004; De Smet et al., 2013; Freeling, 2009; Maere et al., 2005), the surviving At-α ohnologs are enriched or depleted for a number of GO ontology categories (Figure A.5 and S3 Fig.). We had anticipated that those categories that were depleted for ohnolog pairs might represent a set of single-copy genes drawn preferentially from the dominant subgenome. However, such was not the case: even at a quite liberal FDR-corrected significant threshold ($P \leq 0.05$), there are relatively few GO terms significantly differentially retained between the single copy genes of the two subgenomes. Moreover, these terms do not overlap with the ohnolog-depleted terms: instead the single copy genes operating in the endoplasmic reticulum more often derive from the less-fractionated subgenome (Figure A.5). Similarly, genes involved in the cell cycle and circadian rhythm are preferentially drawn from the more fractionated subgenome and those for developmental genes in phloem or xylem from the less-fractionated subgenome (S3 Fig.).

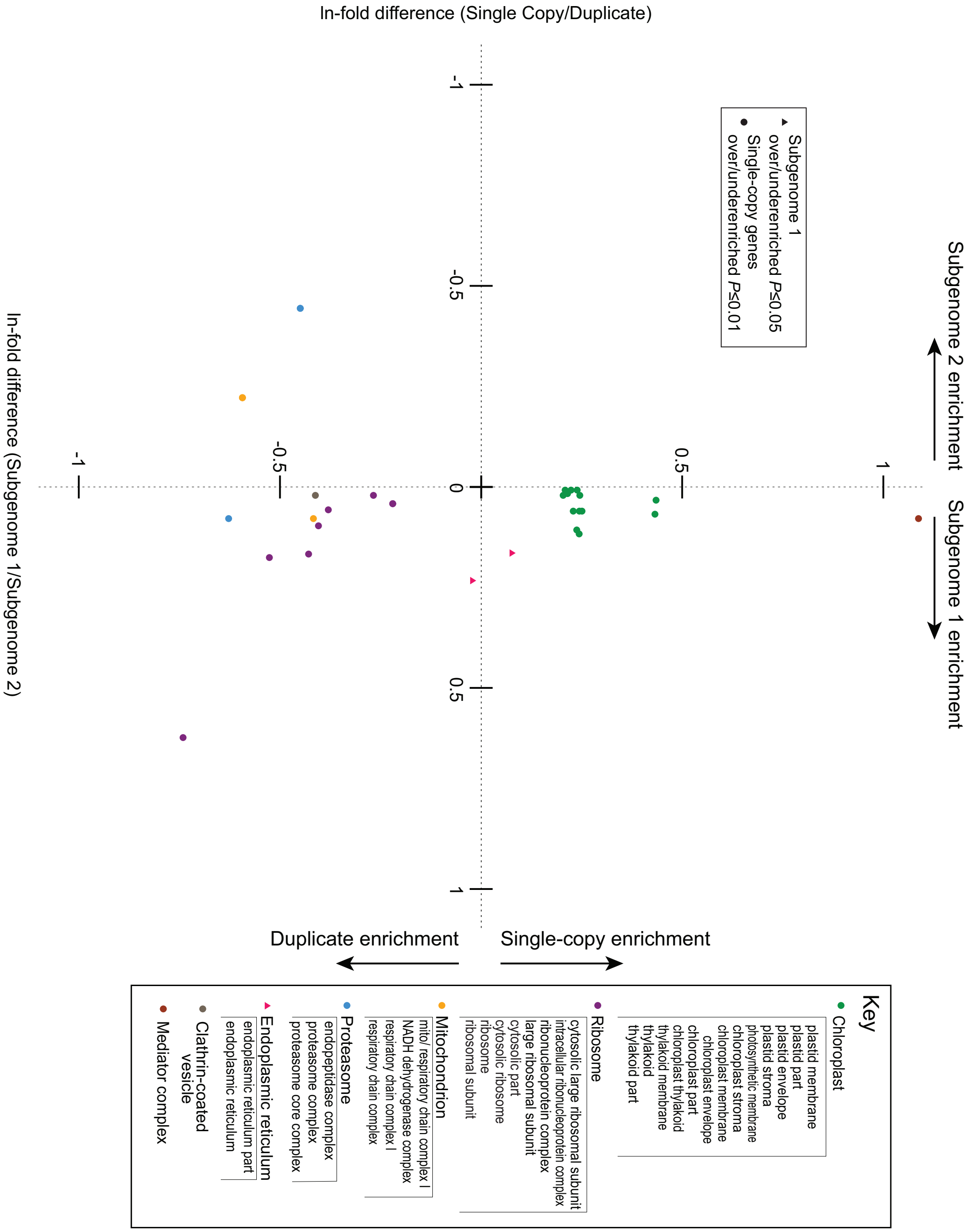


Figure A. 5. Statistically overrepresented GO terms from the cellular component hierarchy associated with At-α duplication status and parental subgenome of origin (see *Methods*). On the y axis is the ln(fold-enrichment) of each GO terms among the single copy genes relative to the surviving duplicates from At-α. Dots represent cellular component terms that are significantly over (positive values) or underrepresented (negative values) among single copy genes relative to duplicates (Bonferroni corrected *P-value* ≤ 0.01 and a fold-enrichment of > ± 1.5). On the x axis is the ln(fold-enrichment) of GO terms of genes from subgenome 1 (the less fractionated genome) relative to those from subgenome 2 (the more fractionated one). GO terms that are overrepresented in genes from subgenome 1 with a *P-value* ≤ 0.05 after Bonferroni correction are shown as triangles. Points are colored based on the compartment in question, as indicated in the key at right. The patterns seen for the “Molecular Function” and “Biological Process” categories of terms are presented in S3 Fig.

A.1.5 DISCUSSION

There is considerable and accumulating evidence for the actions of biased fractionation (BF) after WGD in angiosperms (Sankoff et al., 2010; Schnable et al., 2011; Tang et al., 2012; Thomas et al., 2006) and strong suggestions that allopolyploidy is more likely to produce such biases than autopolyploidy (Garsmeur et al., 2013). Nonetheless, there remains at least a theoretical danger that analyses of BF that consider only a single polyploid genome at a time (often by comparison to a diploid outgroup; Cheng et al., 2012; Schnable et al., 2012; Schnable et al., 2011; Tang et al., 2012; Woodhouse et al., 2010) could mistake the random variation in preservation in small synteny blocks for biases in fractionation.

The results presented here refute this concern, and indicate that, at a minimum, BF acts consistently across regions at the chromosome scale. Our confidence in this conclusion is driven by the concordance of multiple lines of evidence as to the presence and strength of BF. At a methodological level, POInT integrates across multiple genomes, such that lineage-specific synteny breaks are passed through using data from genomes without such breaks (subject to limitations in genome assemblies and in the degree of shared history in the genomes). This approach dramatically increases synteny block size (see Figure A.4). Moreover, POInT employs a very strict and transparent definition of synteny: only genomic neighbors are considered to be in synteny, meaning that POInT employs no parameters such as a window size that need to be tuned by the user and that could confound inferences. POInT also employs a robust modeling framework similar to those used in sequence evolution studies (Liò & Goldman, 1998) and allows for explicit statistical tests for the presence of BF. Using this framework, we have shown very strong statistical support for

BF after two independent WGD events: At-α and the grass ρ event, with a ratio of single copy genes from the less and more fractionated subgenomes somewhere between 3:2 and 5:4, in line with previous estimates (Thomas et al., 2006). This modeling approach has the further advantage of avoiding the circularity in block estimation: POInT infers parental genome assignments on the basis of shared gene losses, a point we have exploited previously (Casola et al., 2012; Conant, 2014; Evangelisti & Conant, 2010; Scienski et al., 2015). As a result, POInT effectively recovers the same shared parental genome assignments under a model without biased fractionation (red/purple blocks in Figure A.4) as it does under the BF model. Moreover, the simultaneous consideration of multiple genomes allows us to assess if the evidence for BF is consistent across those genomes: our loss estimates for each branch of the phylogeny all show BF of roughly similar magnitude, despite the fact that losses on the different tip branches of the phylogeny in Figure A.1 are necessarily independent (an estimate of the BF ratio is given under each branch of the lower tree in that Figure). Finally, the absence of evidence for BF on most branches of the post-WGD yeast phylogeny (which was recently conclusively found to be an allopolyploidy; Marcet-Houben & Gabaldon, 2015) illustrates that POInT is fully capable of rejecting the hypothesis of BF when evidence for it is weak (or temporally variable in this case).

One might argue instead that BF favored some chromosomes from one parental genome and some from another. However, this position is inconsistent with the results of our interaction data and GO term analyses, since such interactions more often occur between products of genes from the same subgenome than between products of genes encoded on different subgenomes, and genes assigned to the same subgenome show consistency in low-level GO term associations. Likewise, there is a good accordance between the estimates of the strength of BF in three of the four largest synteny blocks of Figure A.4 and the overall estimate: were BF a chromosome-by-chromosome phenomenon, it is difficult to understand why its strength would be so consistent across blocks.

While POInT represents a significant improvement over analyses of single polyploid genomes, there are always limitations to any modeling framework. From a practical point of view, our inferences are limited by the quality of the genomic data used as inputs: the more fragmented these genome assemblies, the less power POInT has to infer parental genomes of origin. The inference of DCS blocks by simulated annealing is a costly and computationally difficult problem, and while our scoring functions are reasonable, they may not be the optimal method for inferring ancestral genome orders (Gordon et al., 2009). As mentioned in the *Methods* section, there is also a potential for older polyploidies that are shared by the outgroup genome to mislead our scaffolding, although we do not believe this problem was significant here. Finally, POInT itself is imperfect in how it treats uncertainty in parental genome assignments: the error parameter θ estimates the degree to which the input data fails to conform to POInT's underlying model. While our results above appear to be robust to these various sources of error, future studies of polyploid genomes with improved approaches could give more refined estimates of parental genomes of origin and fine-scale temporal patterns of post-polyploidy gene losses.

Having reaffirmed that BF is a robustly detectable phenomenon in the evolution of polyploid genomes, it is reasonable to try to better understand its origins. In this vein, several of our observations, which arise from POInT's unique capacity to probe polyploidy phylogenetically, serve to again suggest a link between BF and the hypothesized effects of allopolyploidy. The association of genes that physically interact with the same parental genome is one example of such an observation. Another is the conclusion that, after the At-α and ρ events, the strength of BF was uniform in time, but in yeast, BF was associated only with the very earliest stages of WGD resolution. We have previously found that a very particular group of genes, involved in DNA repair and mitochondrial function, were returned to single copy immediately after the yeast WGD (Conant, 2014). Given the biases in those losses found here, it appears likely that BF in yeast was a result of selection for the removal of some ohnolog copies in order to prevent the mixing of genes for these two functions from the two diploid progenitor species. It is likely that the DNA repair enzymes and nuclear-encoded proteins targeted to the mitochondria have co-evolved separately in each parental genome (and that only one of the two parents contributed a mitochondrial genome to the hybridization). If true, these hypotheses would suggest that BF in yeast resulted from selection to maintain co-adapted genes after hybridization. Because these losses, in addition to being biased towards one subgenome and a limited set of functions, occurred very rapidly after the WGD event (Scannell et al., 2007), it is difficult to believe they occurred through purely neutral processes: the proposal by De Smet et al., (2013) that forces such as dominant negative interactions may have driven selection to favor certain losses seems increasingly plausible. These results also reinforce a point we have made several times before: one's understanding of the forces acting on a polyploid genome may depend on *when* in its history you look (Conant, 2014; Conant et al., 2014; Mayfield-Jones et al., 2013).

Our analyses are compatible with differences in gene expression driving BF (Chang et al., 2010; Schnable et al., 2011). However, the BF process does not appear to be solely a product of expression: the presence of co-evolved modules in the two parental genomes also apparently plays a role. Not only do we see a strong bias in the retention of DNA repair enzymes and mitochondrially-targeted proteins in yeast, but we also see a relative absence of protein-protein interactions between proteins encoded by different subgenomes in *A. thaliana*. This hypothesis would also explain our previous observation that both ribosomal proteins and histones underwent post-WGD gene conversions in yeasts (Evangelisti & Conant, 2010; Scienski et al., 2015), as gene conversion represents a second mechanism for resolving parent-of-origin conflicts induced by polyploidy.

Returning to our point about the timing of post-WGD events, we propose that the process of BF and the selection that retains some ohnologs to preserve dosage balance are linked. In this view, some genetic modules (a vague but still useful concept; Wisecaver et al., 2017) do not tolerate being duplicated and are quickly returned to single-copy (De Smet et al., 2013). Others remain duplicated as predicted by the DBH (Birchler & Veitia, 2012; Freeling, 2009). However, these duplications are not necessarily stable over long timescales (Conant, 2014; Conant et al., 2014): any incompatibilities between the subgenomes will favor one subgenome when duplicates are in the end lost. The origins of these conflicts most likely arise through co-evolution between genes in individual genomes (Codoner & Fares, 2008). From our GO analyses, it appears that the effects of this co-evolution decay quickly as one moves away from directly interacting genes: hence many biological processes have "mixed and matched" set of genes from the two subgenomes.

The three WGD events considered here cannot completely resolve these questions: the yeast WGD mostly lacks prolonged BF, while the early events after At-α and ρ are difficult to identify because of the long shared post-WGD branch. In the future, we will perform similar analyses with the recent *Brassica* hexaploidy to further refine our understanding of post-WGD functional evolution. So doing will not only improve our understanding of polyploidy but also of the nature of the functional links and the degree of co-evolution inherent in the interacting macromolecules that make up the cell.

A.1.6 SUPPLMENTARY INFORMATION

Supplemental Figures S1-S3, Supplemental Tables S1-S5 and Supplemental Data S1-3 are available with the original *PLOS Genetics* publication downloadable by accessing the publication https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007267 or downloading files directly using the following links:

S1 Fig: https://doi.org/10.1371/journal.pgen.1007267.s001

S2 Fig: https://doi.org/10.1371/journal.pgen.1007267.s002

S3 Fig: https://doi.org/10.1371/journal.pgen.1007267.s003

S1 Table: https://doi.org/10.1371/journal.pgen.1007267.s004

S2 Table: https://doi.org/10.1371/journal.pgen.1007267.s005

S3 Table: https://doi.org/10.1371/journal.pgen.1007267.s006

S4 Table: https://doi.org/10.1371/journal.pgen.1007267.s007

S5 Table: https://doi.org/10.1371/journal.pgen.1007267.s008

S1 Data: https://doi.org/10.1371/journal.pgen.1007267.s009

S2 Data: https://doi.org/10.1371/journal.pgen.1007267.s010

S3 Data: https://doi.org/10.1371/journal.pgen.1007267.s011

A.1.7 ACKNOWLEDGEMENTS

We would like to thank J. Washburn for assistance with the grass phylogeny, J. Birchler, J. Thorne, X. Ji and members of our fall 2016 Molecular and Network Evolution course for helpful discussions and M. Becchi, D. Li and H. Wu for computational assistance.

A.1.8 REFERENCES

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., & Ding, W. (2016). 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell, 166*(2), 481-491.

Altenbach, S. B., Kuo, C.-C., Staraci, L. C., Pearson, K. W., Wainwright, C., Georgescu, A., & Townsend, J. (1992). Accumulation of a Brazil nut albumin in seeds of transgenic canola results in enhanced levels of seed protein methionine. *Plant molecular biology, 18*(2), 235-245.

Altenbach, S. B., Pearson, K. W., Meeker, G., Staraci, L. C., & Sun, S. S. (1989). Enhancement of the methionine content of seed proteins by the expression of a chimeric gene encoding a methionine-rich protein in transgenic plants. *Plant molecular biology, 13*(5), 513-522.

Altenbach, S. B., & Simpson, R. B. (1990). Manipulation of methionine-rich protein genes in plant seeds. *Trends in Biotechnology, 8*, 156-160.

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. *59*(2), 118-125.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped Blast and Psi-Blast : A new-generation of protein database search programs. *Nucleic Acids Research, 25*(#17), 3389-3402.

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

Amir, R., & Tabe, L. (2006). Molecular approaches to improving plant methionine content.

Angelovici, R., Batushansky, A., Deason, N., Gonzalez-Jorge, S., Gore, M. A., Fait, A., & DellaPenna, D. (2017). Network-guided GWAS improves identification of genes affecting free amino acids. *Plant physiology, 173*(1), 872-886.

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 in plant science, 15*(4), 211-218.

Angelovici, R., Lipka, A. E., Deason, N., Gonzalez-Jorge, S., Lin, H., Cepela, J., Buell, R., Gore, M. A., & Dellapenna, D. (2013). Genome-wide analysis of branched-chain amino acid levels in Arabidopsis seeds. *Plant Cell, 25*(12), 4827-4843. https://doi.org/10.1105/tpc.113.119370

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

Arabidopsis Interactome Mapping Consortium. (2011). Evidence for network evolution in an Arabidopsis interactome map. *Science, 333*(6042), 601-607. https://doi.org/333/6042/601 [pii] 10.1126/science.1203877

Arends, D., Prins, P., Jansen, R. C., & Broman, K. W. (2010). R/qtl: high-throughput multiple QTL mapping. *Bioinformatics, 26*(23), 2990-2992. https://doi.org/10.1093/bioinformatics/btq565

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., & Hu, T. T. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature, 465*(7298), 627.

Aury, J. M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Camara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A. M., Kissmehl, R., Klotz, C., Koll, F., Le Mouel, A., Lepere, G., Malinsky, S., Nowacki, M., Nowak, J. K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Betermier, M., Weissenbach, J., Scarpelli, C., Schachter, V., Sperling, L., Meyer, E., Cohen, J., & Wincker, P. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature, 444*(7116), 171-178.

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2004). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics, 21*(2), 263-265.

Batushansky, A., Toubiana, D., & Fait, A. (2016). Correlation-based network generation, visualization, and analysis as a powerful tool in biological studies: a case study in cancer cell metabolism. *Biomed Res Int, 2016*, 1-9. https://doi.org/10.1155/2016/8313272

Baud, S., Boutin, J.-P., Miquel, M., Lepiniec, L., & Rochat, C. (2002). An integrated overview of seed development in Arabidopsis thaliana ecotype WS. *Plant Physiology and Biochemistry, 40*(2), 151-160. https://doi.org/https://doi.org/10.1016/S0981-9428(01)01350-X

Bekaert, M., Edger, P. P., Pires, J. C., & Conant, G. C. (2011). Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative followed by absolute dosage constraints. *The Plant Cell, 23*, 1719-1728.

Benderoth, M., Textor, S., Windsor, A. J., Mitchell-Olds, T., Gershenzon, J., & Kroymann, J. (2006). Positive selection driving diversification in plant secondary metabolism. *Proceedings of the National Academy of Sciences of the United States of America, 103*(24), 9118-9123. https://doi.org/10.1073/pnas.0601738103

Besnard, J., Pratelli, R., Zhao, C., Sonawala, U., Collakova, E., Pilot, G., & Okumoto, S. (2016). UMAMIT14 is an amino acid exporter involved in phloem unloading in Arabidopsis roots. *Journal of Experimental Botany, 67*, 6385-6397.

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

Birchler, J. A., Riddle, N. C., Auger, D. L., & Veitia, R. A. (2005). Dosage balance in gene regulation: biological implications. *Trends Genet, 21*(4), 219-226. https://doi.org/S0168-9525(05)00050-8 [pii] 10.1016/j.tig.2005.02.010

Birchler, J. A., & Veitia, R. A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell, 19*(2), 395-402. https://doi.org/tpc.106.049338 [pii] 10.1105/tpc.106.049338

Birchler, J. A., & Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc Natl Acad Sci U S A, 109*(37), 14746-14753. https://doi.org/1207726109 [pii] 10.1073/pnas.1207726109

Birtić, S., & Kranner, I. (2006). Isolation of high-quality RNA from polyphenol-, polysaccharide-and lipid-rich seeds. *Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques, 17*(3), 144-148.

Blanc, G., & Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell, 16*(7), 1679-1691.

Blanc, G., & Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell, 16*, 1679-1691.

Bose, U., Broadbent, J. A., Byrne, K., Blundell, M. J., Howitt, C. A., & Colgrave, M. L. (2020). Proteome Analysis of Hordein-Null Barley Lines Reveals Storage Protein Synthesis and Compensation Mechanisms. *J Agric Food Chem, 68*(20), 5763-5775. https://doi.org/10.1021/acs.jafc.0c01410

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological), 26*(2), 211-243.

Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R., & Görlach, J. (2001). Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. *Plant Cell, 13*(7), 1499-1510. https://doi.org/10.1105/tpc.010011

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics, 23*(19), 2633-2635. https://doi.org/10.1093/bioinformatics/btm308

Bright, S. W., Kueh, J. S., & Rognes, S. E. (1983). Lysine transport in two barley mutants with altered uptake of basic amino acids in the root. *Plant Physiology, 72*(3), 821-824.

Broman, K. W., Gatti, D. M., Simecek, P., Furlotte, N. A., Prins, P., Sen, Ś., Yandell, B. S., & Churchill, G. A. (2019). R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations. *Genetics, 211*(2), 495-502.

Buggs, R. J., Chamala, S., Wu, W., Gao, L., May, G. D., Schnable, P. S., Soltis, D. E., Soltis, P. S., & Barbazuk, W. (2010). Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular ecology, 19*(s1), 132-146.

Burow, M., Atwell, S., Francisco, M., Kerwin, R. E., Halkier, B. A., & Kliebenstein, D. J. (2015). The glucosinolate biosynthetic gene AOP2 mediates feed-back regulation of jasmonic acid signaling in Arabidopsis. *Molecular plant, 8*(8), 1201-1212.

Burow, M., Halkier, B. A., & Kliebenstein, D. J. (2010). Regulatory networks of glucosinolates shape Arabidopsis thaliana fitness. *Curr Opin Plant Biol, 13*(3), 348-353. https://doi.org/10.1016/j.pbi.2010.02.002

Byrne, K. P., & Wolfe, K. H. (2005). The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research, 15*(10), 1456-1461.

Casola, C., Conant, G. C., & Hahn, M. W. (2012). Very low rate of gene conversion in the yeast genome. *Molecular Biology and Evolution, 29*(12), 3817-3826. https://doi.org/mss192 [pii] 10.1093/molbev/mss192

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

Chan, E. K., Rowe, H. C., Corwin, J. A., Joseph, B., & Kliebenstein, D. J. (2011). Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana. *PLoS biology, 9*(8), e1001125.

Chang, P. L., Dilkes, B. P., McMahon, M., Comai, L., & Nuzhdin, S. V. (2010). Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. *Genome Biology, 11*(12), R125. https://doi.org/10.1186/gb-2010-11-12-r125

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H., Dong, H., Zhang, W., Zhang, L., Yu, S., Wang, G., Lian, X., & Luo, J. (2014).

Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics, 46*, 714-721.

Chen, W., Wang, W., Peng, M., Gong, L., Gao, Y., Wan, J., Wang, S., Shi, L., Zhou, B., & Li, Z. (2016). Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nature communications, 7*, 12767.

Chen, Y., Lun, A. T., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res, 5*, 1438. https://doi.org/10.12688/f1000research.8987.2

Chen, Y. Z., Pang, Q. Y., He, Y., Zhu, N., Branstrom, I., Yan, X. F., & Chen, S. (2012). Proteomics and metabolomics of Arabidopsis responses to perturbation of glucosinolate biosynthesis. *Mol Plant, 5*(5), 1138-1150. https://doi.org/10.1093/mp/sss034

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., & Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PLoS ONE, 7*(5), e36442.

Churchill, G. A., & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics, 138*(3), 963-971.

Clausen, R., & Goodspeed, T. (1925). Interspecific hybridization in Nicotiana. II. A tetraploid glutinosa-tabacum hybrid, an experimental verification of Winge's hypothesis. *Genetics, 10*(3), 278.

Codoner, F. M., & Fares, M. A. (2008). Why should we care about molecular coevolution? *Evolutionary Bioinformatics, 4*, 29-38.

Cohen, H., Pajak, A., Pandurangan, S., Amir, R., & Marsolais, F. (2016). Higher endogenous methionine in transgenic Arabidopsis seeds affects the composition of storage proteins and lipids. *Amino acids, 48*(6), 1413-1422.

Coleman, C. E., & Larkins, B. A. (1999). The prolamins of maize. In *seed Proteins* (pp. 109-139). Springer.

Conant, G. C. (2014). Comparative genomics as a time machine: How relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. *Molecular Biology and Evolution, 31*(12), 3184-3193. https://doi.org/10.1093/molbev/msu250

Conant, G. C., Birchler, J. A., & Pires, J. C. (2014). Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current opinion in plant biology, 19*, 91-98. https://doi.org/10.1016/j.pbi.2014.05.008

Conant, G. C., & Wagner, A. (2002). GenomeHistory: A software tool and its application to fully sequenced genomes. *Nucleic Acids Research, 30*(15), 3378-3386.

Conant, G. C., & Wolfe, K. H. (2006). Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biology, 4*, e109.

Conant, G. C., & Wolfe, K. H. (2007). Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology, 3*, 129.

Conant, G. C., & Wolfe, K. H. (2008). Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics, 179*, 1681-1692.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics, 19*(1), 5-19.

Dassanayake, M., Oh, D. H., Haas, J. S., Hernandez, A., Hong, H., Ali, S., Yun, D. J., Bressan, R. A., Zhu, J. K., Bohnert, H. J., & Cheeseman, J. M. (2011). The genome of the extremophile crucifer Thellungiella parvula. *Nature Genetics, 43*(9), 913-918. https://doi.org/10.1038/ng.889

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics, 22*(10), 1269-1271.

De Clercq, A., Vandewiele, M., Van Damme, J., Guerche, P., Van Montagu, M., Vandekerckhove, J., & Krebbers, E. (1990). Stable Accumulation of Modified 2S Albumin Seed Storage Proteins with Higher Methionine Contents in Transgenic Plants. *Plant Physiology*, 970-979.

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C., Maere, S., & Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A, 110*(8), 2898-2903. https://doi.org/1300127110 [pii] 10.1073/pnas.1300127110

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.

Diepenbrock, C. H., Kandianis, C. B., Lipka, A. E., Magallanes-Lundback, M., Vaillancourt, B., Gongora-Castillo, E., Wallace, J. G., Cepela, J., Mesberg, A., Bradbury, P. J., Ilut, D. C., Mateos-Hernandez, M., Hamilton, J., Owens, B. F., Tiede, T., Buckler, E. S., Rocheford, T., Buell, C. R., Gore, M. A., & DellaPenna,

D. (2017). Novel Loci Underlie Natural Variation in Vitamin E Levels in Maize Grain. *Plant Cell, 29*(10), 2374-2392. https://doi.org/10.1105/tpc.17.00475

DiLeo, M. V., Strahan, G. D., den Bakker, M., & Hoekenga, O. A. (2011). Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. *PLoS One, 6*(10), e26683. https://doi.org/10.1371/journal.pone.0026683

Dinkins, R. D., Reddy, M. S., Meurer, C. A., Yan, B., Trick, H., Thibaud-Nissen, F., Finer, J. J., Parrott, W. A., & Collins, G. B. (2001). Increased sulfur amino acids in soybean plants overexpressing the maize 15 kDa zein protein. *In Vitro Cellular & Developmental Biology-Plant, 37*(6), 742-747.

[Record #395 is using a reference type undefined in this output style.]

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. https://doi.org/10.1093/bioinformatics/bts635

Dopman, E. B., & Hartl, D. L. (2007). A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A, 104*(50), 19920-19925. https://doi.org/10.1073/pnas.0709888104

Doring, A., Weese, D., Rausch, T., & Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics, 9*, 11. https://doi.org/1471-2105-9-11 [pii] 10.1186/1471-2105-9-11

Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res, 38*(Web Server issue), W64-70. https://doi.org/10.1093/nar/gkq310

Edger, P. P., Heidel-Fischer, H. M., Bekaert, M., Rota, J., Glöckner, G., Platts, A. E., Heckel, D. G., Der, J. P., Wafula, E. K., & Tang, M. (2015). The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences, 112*(27), 8362-8366.

Edger, P. P., & Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res, 17*(5), 699-717. https://doi.org/10.1007/s10577-009-9055-9

Edger, P. P., Smith, R. D., McKain, M. R., Cooley, A. M., Vallejo-Marin, M., Yuan, Y., Bewick, A. J., Ji, L., Platts, A. E., & Bowman, M. J. (2016). Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140 year old naturally established neo-allopolyploid monkeyflower. *bioRxiv*, 094797.

elBaradi, T. T., van der Sande, C. A., Mager, W. H., Raué, H. A., & Planta, R. J. (1986). The cellular level of yeast ribosomal protein L25 is controlled principally by rapid degradation of excess protein. *Curr Genet, 10*(10), 733-739. https://doi.org/10.1007/BF00405095

Evangelisti, A. M., & Conant, G. C. (2010). Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. *Genome Biology and Evolution, 2*, 826-834.

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 Physiol, 142*(3), 839-854. https://doi.org/10.1104/pp.106.086694

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. https://doi.org/10.1104/pp.106.086694

Falco, S., Guida, T., Locke, M., Mauvais, J., Sanders, C., Ward, R., & Webber, P. (1995). Transgenic canola and soybean seeds with increased lysine. *Bio/technology, 13*(6), 577.

FAO. *Staple foods: What do people eat?* . Retrieved June 11 from (http://www.fao.org/3/u8480e/u8480e07.htm

Felsenstein, J., & Churchill, G. A. (1996). A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution, 13*(1), 93-104.

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.

[Record #32 is using a reference type undefined in this output style.]

Flint-Garcia, S. A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., Kresovich, S., Goodman, M. M., & Buckler, E. S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal, 44*(6), 1054-1064. https://doi.org/https://doi.org/10.1111/j.1365-313X.2005.02591.x

Forsyth, J. L., Beaudoin, F., Halford, N. G., Sessions, R. B., Clarke, A. R., & Shewry, P. R. (2005). Design, expression and characterisation of lysine-rich forms of the barley seed protein CI-2. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1747*(2), 221-227.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., & Graves, S. (2012). Package 'car'. *Vienna: R Foundation for Statistical Computing*.

Foyer, C. H., Bloom, A. J., Queval, G., & Noctor, G. (2009). Photorespiratory metabolism: genes, mutants, energetics, and redox signaling. *Annu Rev Plant Biol, 60*, 455-484. https://doi.org/10.1146/annurev.arplant.043008.091948

Francisco, M., Joseph, B., Caligagan, H., Li, B., Corwin, J. A., Lin, C., Kerwin, R., Burow, M., & Kliebenstein, D. J. (2016). The Defense Metabolite, Allyl Glucosinolate, Modulates Arabidopsis thaliana Biomass Dependent upon the Endogenous Glucosinolate Pathway. *Frontiers in Plant Science, 7*, 774. https://doi.org/10.3389/fpls.2016.00774

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology, 60*, 433-453. https://doi.org/10.1146/annurev.arplant.043008.092122

Freeling, M., & Thomas, B. C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research, 16*, 805-814.

Frerigmann, H., & Gigolashvili, T. (2014). MYB34, MYB51, and MYB122 distinctly regulate indolic glucosinolate biosynthesis in Arabidopsis thaliana. *Molecular plant, 7*(5), 814-828.

Frerigmann, H., & Gigolashvili, T. (2014). MYB34, MYB51, and MYB122 distinctly regulate indolic glucosinolate biosynthesis in Arabidopsis thaliana. *Mol Plant, 7*(5), 814-828. https://doi.org/10.1093/mp/ssu004

Gaeta, R. T., Pires, J. C., Iniguez-Luy, F., Leon, E., & Osborn, T. C. (2007). Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. *Plant Cell, 19*(11), 3403-3417.

Galili, G. (2011). The aspartate-family pathway of plants: linking production of essential amino acids with energy and stress regulation. *Plant signaling & behavior, 6*(2), 192-195.

Galili, G., & Amir, R. (2013). Fortifying plants with the essential amino acids lysine and methionine to improve nutritional quality. *Plant Biotechnol J, 11*(2), 211-222. https://doi.org/10.1111/pbi.12025

Galili, G., & Höfgen, R. (2002). Metabolic engineering of amino acids and storage proteins in plants. *Metabolic engineering, 4*(1), 3-11.

Gao, J., Yu, X., Ma, F., & Li, J. (2014). RNA-seq analysis of transcriptome and glucosinolate metabolism in seeds and sprouts of broccoli (Brassica oleracea var. italic). *Plos One, 9*(2), e88804. https://doi.org/10.1371/journal.pone.0088804

Garsmeur, O., Schnable, J. C., Almeida, A., Jourda, C., D'Hont, A., & Freeling, M. (2013). Two Evolutionarily Distinct Classes of Paleopolyploidy. *Molecular Biology and Evolution, 31*(2), 448-454. https://doi.org/mst230 [pii] 10.1093/molbev/mst230

Ghislain, M., Frankard, V., Vandenbossche, D., Matthews, B. F., & Jacobs, M. (1994). Molecular analysis of the aspartate kinase-homoserine dehydrogenase gene from Arabidopsis thaliana. *Plant molecular biology, 24*(6), 835-851.

Gonzalez-Jorge, S., Ha, S. H., Magallanes-Lundback, M., Gilliland, L. U., Zhou, A., Lipka, A. E., Nguyen, Y. N., Angelovici, R., Lin, H., Cepela, J., Little, H., Buell, C. R., Gore, M. A., & Dellapenna, D. (2013). Carotenoid cleavage dioxygenase4 is a negative regulator of β-carotene content in Arabidopsis seeds. *Plant Cell, 25*(12), 4812-4826. https://doi.org/10.1105/tpc.113.119677

Gordon, J. L., Byrne, K. P., & Wolfe, K. H. (2009). Additions, losses and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genetics, 5*(5), e1000485.

Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., & Aebersold, R. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol Cell Proteomics, 1*(4), 323-333. https://doi.org/10.1074/mcp.m200001-mcp200

Grimm, D. G., Roqueiro, D., Salomé, P. A., Kleeberger, S., Greshake, B., Zhu, W., Liu, C., Lippert, C., Stegle, O., Schölkopf, B., Weigel, D., & Borgwardt, K. M. (2017). easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies. *Plant Cell, 29*(1), 5-19. https://doi.org/10.1105/tpc.16.00551

Grover, Z., & Ee, L. C. (2009). Protein energy malnutrition. *Pediatric Clinics, 56*(5), 1055-1068.

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.

Hagan, N., Upadhyaya, N., Tabe, L., & Higgins, T. (2003). The redistribution of protein sulfur in transgenic rice expressing a gene for a foreign, sulfur-rich protein. *The Plant Journal, 34*(1), 1-11.

Hajduch, M., Hearne, L. B., Miernyk, J. A., Casteel, J. E., Joshi, T., Agrawal, G. K., Song, Z., Zhou, M., Xu, D., & Thelen, J. J. (2010). Systems analysis of seed filling in Arabidopsis: using general linear modeling to assess concordance of transcript and protein expression. *Plant Physiol, 152*(4), 2078-2087. https://doi.org/10.1104/pp.109.152413

Halkier, B. A., & Gershenzon, J. (2006). Biology and biochemistry of glucosinolates. *Annu Rev Plant Biol, 57*, 303-333. https://doi.org/10.1146/annurev.arplant.57.032905.105228

Haudry, A., Platts, A. E., Vello, E., Hoen, D. R., Leclercq, M., Williamson, R. J., Forczek, E., Joly-Lopez, Z., Steffen, J. G., & Hazzouri, K. M. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics, 45*(8), 891-898.

Haughn, G. W., Davin, L., Giblin, M., & Underhill, E. W. (1991). Biochemical genetics of plant secondary metabolites in Arabidopsis thaliana: the glucosinolates. *Plant Physiology, 97*(1), 217-226.

Henriques, R., Bogre, L., Horvath, B., & Magyar, Z. (2014). Balancing act: matching growth with environment by the TOR signalling pathway. *Journal of Experimental Botany, 65*(10), 2691-2701. https://doi.org/10.1093/jxb/eru049

Heremans, B., & Jacobs, M. (1997). A Mutant of Arabidopsis thaliana lpar; L.) Heynh. with Modified Control of Aspartate Kinase by Threonine. *Biochemical genetics, 35*(3-4), 139-153.

Herman, E. M. (2014). Soybean seed proteome rebalancing. *Frontiers in plant science, 5*, 437.

Hoffman, L. M., Donaldson, D. D., & Herman, E. M. (1988). A modified storage protein is synthesized, processed, and degraded in the seeds of transgenic plants. *Plant molecular biology, 11*(6), 717-729.

Holding, D., & Messing, J. (2013). Evolution, structure, and function of prolamin storage proteins. *Seed genomics*, 138-158.

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N. W., Platt, A., Sperone, F. G., & Vilhjálmsson, B. J. (2012). Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature genetics, 44*(2), 212.

Hou, A., Liu, K., Catawatcharakul, N., Tang, X., Nguyen, V., Keller, W. A., Tsang, E. W., & Cui, Y. (2005). Two naturally occurring deletion mutants of 12S seed storage proteins in Arabidopsis thaliana. *Planta, 222*(3), 512-520.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F., Van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., & Guo, Y. L. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics, 43*(5), 476-481. https://doi.org/10.1038/ng.807

Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M. A., Al-Shehbaz, I., & Edger, P. P. (2016). Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution, 33*(2), 394-412.

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. https://doi.org/10.1105/tpc.003905

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.

Ingle, R. A. (2011). Histidine biosynthesis. *The Arabidopsis book/American Society of Plant Biologists, 9*.

Initiative, I. B. (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature, 463*(7282), 763.

Jensen, L. M., Kliebenstein, D. J., & Burow, M. (2015). Investigation of the multifunctional gene AOP3 expands the regulatory network fine-tuning glucosinolate production in Arabidopsis. *Front Plant Sci, 6*, 762. https://doi.org/10.3389/fpls.2015.00762

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 opaque2 mutant by transcriptional and proteomic analysis. *BMC plant biology, 13*(1), 60.

Jiang, Y., Wang, D., Xu, D., & Joshi, T. (2020). IMPRes-Pro: A high dimensional multiomics integration method for in silico hypothesis generation. *Methods, 173*, 16-23. https://doi.org/10.1016/j.ymeth.2019.06.013

Johnson, S. D., Griffiths, M. E., Peter, C. I., & Lawes, M. J. (2009). Pollinators, "mustard oil" volatiles, and fruit production in flowers of the dioecious tree Drypetes natalensis (Putranjivaceae). *American Journal of Botany, 96*, 2080–2086.

Jones, C. G., & Firn, R. D. (1991). On the Evolution of Plant Secondary Chemical Diversity. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 333*(1267), 273-280. https://doi.org/DOI 10.1098/rstb.1991.0077

Joyce, B. L., Haug-Baltzell, A., Davey, S., Bomhoff, M., Schnable, J. C., & Lyons, E. (2017). FractBias: a graphical tool for assessing fractionation bias following polyploidy. *Bioinformatics, 33*(4), 552-554. https://doi.org/10.1093/bioinformatics/btw666

Karchi, H., Shaul, O., & Galili, G. (1993). Seed-specific expression of a bacterial desensitized aspartate kinase increases the production of seed threonine and methionine in transgenic tobacco. *The Plant Journal, 3*(5), 721-727.

Katz, E., Nisani, S., Yadav, B. S., Woldemariam, M. G., Shai, B., Obolski, U., Ehrlich, M., Shani, E., Jander, G., & Chamovitz, D. A. (2015). The glucosinolate breakdown product indole-3-carbinol acts as an auxin antagonist in roots of Arabidopsis thaliana. *Plant J, 82*(4), 547-555. https://doi.org/10.1111/tpj.12824

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature, 428*, 617-624.

Kellogg, E. A. (2016). Has the connection between polyploidy and diversification actually been tested? *Current opinion in plant biology, 30*, 25-32.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., & Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature genetics, 39*(9), 1151.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., & Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nat Genet, 39*. https://doi.org/10.1038/ng2115

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. https://doi.org/10.1007/s00425-011-1487-8

Kirkpatrick, S., Gelatt, C. D. J., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671-680.

Kliebenstein, D. J., D'Auria, J. C., Behere, A. S., Kim, J. H., Gunderson, K. L., Breen, J. N., Lee, G., Gershenzon, J., Last, R. L., & Jander, G. (2007). Characterization of seed-specific benzoyloxyglucosinolate mutations in Arabidopsis thaliana. *Plant J, 51*(6), 1062-1076. https://doi.org/10.1111/j.1365-313X.2007.03205.x

Kliebenstein, D. J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J., & Mitchell-Olds, T. (2001). Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiol, 126*(2), 811-825. https://doi.org/10.1104/pp.126.2.811

Koch, M. A., & German, D. (2013). Taxonomy and systematics are key to biological information: Arabidopsis, Eutrema (Thellungiella), Noccaea and Schrenkiella (Brassicaceae) as examples. *Frontiers in plant science, 4*, 267.

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. https://doi.org/10.1105/tpc.1.1.105

Kroymann, J., Donnerhacke, S., Schnabelrauch, D., & Mitchell-Olds, T. (2003). Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. *Proc Natl Acad Sci U S A, 100 Suppl 2*, 14587-14592. https://doi.org/10.1073/pnas.1734046100

Kroymann, J., Textor, S., Tokuhisa, J. G., Falk, K. L., Bartram, S., Gershenzon, J., & Mitchell-Olds, T. (2001). A gene controlling variation in Arabidopsis glucosinolate composition is part of the methionine chain elongation pathway. *Plant Physiol, 127*(3), 1077-1088.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (Vol. 5). McGraw-Hill Irwin Boston.

La, T., Large, E., Taliercio, E., Song, Q., Gillman, J. D., Xu, D., Nguyen, H. T., Shannon, G., & 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.

Lambert, R., Alexander, D., & Dudley, J. (1969). Relative Performance of Normal and Modified Protein (Opaque-2) Maize Hybrids 1. *Crop Science, 9*(2), 242-243.

Lander, E. S., & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences, U.S.A., 84*, 2363-2367.

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

Le Roch, K. G., Johnson, J. R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., Yan, S. F., Williamson, K. C., Holder, A. A., Carucci, D. J., Yates, J. R., & Winzeler, E. A. (2004). Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res, 14*(11), 2308-2318. https://doi.org/10.1101/gr.2523904

[Record #431 is using a reference type undefined in this output style.]

Lea, P. J., & Miflin, B. J. (1974). Alternative route for nitrogen assimilation in higher plants. *Nature, 251*(5476), 614-616. https://doi.org/10.1038/251614a0

Lee, S. I., Kim, H. U., Lee, Y.-H., Suh, S.-C., Lim, Y. P., Lee, H.-Y., & Kim, H.-I. (2001). Constitutive and seed-specific expression of a maize lysine-feedback-insensitive dihydrodipicolinate synthase gene leads to increased free lysine levels in rice seeds. *Molecular Breeding, 8*(1), 75-84.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology, 50*, 913-925.

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

Lipka, A. E., Gore, M. A., Magallanes-Lundback, M., Mesberg, A., Lin, H., Tiede, T., Chen, C., Buell, C. R., Buckler, E. S., Rocheford, T., & Dellapenna, D. (2013). Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. *G3, 3*(8), 1287-1299. https://doi.org/10.1534/g3.113.006148

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

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.

Liò, P., & Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome Research, 8*, 1233-1244.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Tohge, T., Fernie, A. R., 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.

Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D., & Daniel-Vedele, F. (2002). Bay-0× Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theoretical and Applied Genetics, 104*(6-7), 1173-1184.

[Record #396 is using a reference type undefined in this output style.]

Lund, S. P., Nettleton, D., McCarthy, D. J., & Smyth, G. K. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical applications in genetics and molecular biology, 11*(5).

Lyons, E., & Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal, 53*(4), 661-673.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., & Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences, U.S.A., 102*(15), 5454-5459.

Magrath, R., Bano, F., Morgner, M. . (1994). Genetics of aliphatic glucosinolates I. Side chain elongation in Brassica napus and Arabidopsis thaliana. *Heredity 72*, 290-299.

Magrath, R., & Mithen, R. (1993). Maternal Effects on the Expression of Individual Aliphatic Glucosinolates in Seeds and Seedlings of Brassica-Napus. *Plant Breeding, 111*(3), 249-252.

Majumdar, R., Barchi, B., Turlapati, S. A., Gagne, M., Minocha, R., Long, S., & Minocha, S. C. (2016). Glutamate, Ornithine, Arginine, Proline, and Polyamine

Metabolic Interactions: The Pathway Is Regulated at the Post-Transcriptional Level. *Frontiers in Plant Science, 7*, 78.

Makino, T., & McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A, 107*(20), 9270-9274. https://doi.org/0914697107 [pii] 10.1073/pnas.0914697107

Malinovsky, F. G., Thomsen, M. F., Nintemann, S. J., Jagd, L. M., Bourgine, B., Burow, M., & Kliebenstein, D. J. (2017). An evolutionarily young defense metabolite influences the root growth of plants via the ancient TOR signaling pathway. *Elife, 6*. https://doi.org/10.7554/eLife.29353

Marcet-Houben, M., & Gabaldon, T. (2015). Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLoS Biology, 13*(8), e1002220. https://doi.org/10.1371/journal.pbio.1002220

[Record #393 is using a reference type undefined in this output style.]

Mayfield-Jones, D., Washburn, J. D., Arias, T., Edger, P. P., Pires, J. C., & Conant, G. C. (2013). Watching the grin fade: Tracing the effects of polyploidy on different evolutionary time scales. *Seminars in Cellular and Developmental Biology, 24*, 320-331.

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., & Otto, S. P. (2011). Recently formed polyploid plants diversify at lower rates. *Science, 333*(6047), 1257. https://doi.org/science.1207205 [pii] 10.1126/science.1207205

Mazur, B., Krebbers, E., & Tingey, S. (1999). Gene discovery and product development for grain quality traits. *Science, 285*(5426), 372-375.

[Record #394 is using a reference type undefined in this output style.]

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., Samaras, P., Richter, S., Shikata, H., Messerer, M., Lang, D., Altmann, S., Cyprys, P., Zolg, D. P., Mathieson, T., Bantscheff, M., Hazarika, R. R., Schmidt, T., Dawid, C., Dunkel, A., Hofmann, T., Sprunck, S., Falter-Braun, P., Johannes, F., Mayer, K. F. X., Jürgens, G., Wilhelm, M., Baumbach, J., Grill, E., Schneitz, K., Schwechheimer, C., & Kuster, B. (2020). Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature, 579*(7799), 409-414. https://doi.org/10.1038/s41586-020-2094-2

Merico, A., Sulo, P., Piškur, J., & Compagno, C. (2007). Fermentative lifestyle in yeasts belonging to the *Saccharomyces* complex. *FEBS Journal, 274*, 976-989.

Mertz, E. T., Bates, L. S., & Nelson, O. E. (1964). Mutant gene that changes protein composition and increases lysine content of maize endosperm. *Science, 145*(3629), 279-280.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research, 45*(D1), D183-D189. https://doi.org/10.1093/nar/gkw1138

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc, 8*(8), 1551-1566. https://doi.org/10.1038/nprot.2013.092

Ming, R., VanBuren, R., Wai, C. M., Tang, H., Schatz, M. C., Bowers, J. E., Lyons, E., Wang, M.-L., Chen, J., & Biggers, E. (2015). The pineapple genome and the evolution of CAM photosynthesis. *Nature Genetics, 47*(12), 1435.

Molvig, L., Tabe, L. M., Eggum, B. O., Moore, A. E., Craig, S., Spencer, D., & Higgins, T. J. (1997). Enhanced methionine levels and increased nutritive value of seeds of transgenic lupins (Lupinus angustifolius L.) expressing a sunflower seed albumin gene. *Proceedings of the National Academy of Sciences, 94*(16), 8393-8398.

Morton, K. J., Jia, S., Zhang, C., & Holding, D. R. (2016). Proteomic profiling of maize opaque endosperm mutants reveals selective accumulation of lysine-enriched proteins. *J Exp Bot, 67*(5), 1381-1396. https://doi.org/10.1093/jxb/erv532

Muse, S. V., & Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution, 11*(5), 715-724.

Nguyen, T.-P., Cueff, G., Hegedus, D. D., Rajjou, L., & Bentsink, L. (2015). A role for seed storage proteins in Arabidopsis seed longevity. *Journal of Experimental Botany, 66*(20), 6399-6413.

Nikiforova, V. J., Bielecka, M., Gakiere, B., Krueger, S., Rinder, J., Kempa, S., Morcuende, R., Scheible, W. R., Hesse, H., & Hoefgen, R. (2006). Effect of sulfur availability on the integrity of amino acid biosynthesis in plants. *Amino Acids, 30*(2), 173-183. https://doi.org/10.1007/s00726-005-0251-4

Nikiforova, V. J., Kopka, J., Tolstikov, V., Fiehn, O., Hopkins, L., Hawkesford, M. J., Hesse, H., & Hoefgen, R. (2005). Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of Arabidopsis plants. *Plant Physiol, 138*(1), 304-318. https://doi.org/10.1104/pp.104.053793

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., & Goyal, R. (2005). The pattern of polymorphism in Arabidopsis thaliana. *PLoS biology, 3*(7), e196.

Nour-Eldin, H. H., Andersen, T. G., Burow, M., Madsen, S. R., Jorgensen, M. E., Olsen, C. E., Dreyer, I., Hedrich, R., Geiger, D., & Halkier, B. A. (2012). NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. *Nature, 488*(7412), 531-534.

Nour-Eldin, H. H., Andersen, T. G., Burow, M., Madsen, S. R., Jørgensen, M. E., Olsen, C. E., Dreyer, I., Hedrich, R., Geiger, D., & Halkier, B. A. (2012). NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. *nature, 488*(7412), 531.

Ohno, S. (1970). *Evolution by gene duplication*. Springer.

Okumoto, S., Funck, D., Trovato, M., & Forlani, G. (2016). Editorial: Amino Acids of the Glutamate Family: Functions beyond Primary Metabolism. *Frontiers in Plant Science, 7*, 318.

Papp, B., Pal, C., & Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature, 424*(6945), 194-197.

Parkin, I., Magrath, R., Keith, D., Sharpe, A., Mithen, R., & Lydiate, D. (1994). Genetics of Aliphatic Glucosinolates .2. Hydroxylation of Alkenyl Glucosinolates in Brassica-Napus. *Heredity, 72*, 594-598. https://doi.org/DOI 10.1038/hdy.1994.82

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., & Poliakov, A. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature, 457*(7229), 551.

Paterson, A. H., Bowers, J. E., & Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A, 101*(26), 9903-9908. https://doi.org/10.1073/pnas.0307901101

Pei, G., Chen, L., & Zhang, W. (2017). WGCNA Application to Proteomic and Metabolomic Data Analysis. *Methods Enzymol, 585*, 135-158. https://doi.org/10.1016/bs.mie.2016.09.016

Petersen, B. L., Chen, S., Hansen, C. H., Olsen, C. E., & Halkier, B. A. (2002). Composition and content of glucosinolates in developing Arabidopsis thaliana. *Planta, 214*(4), 562-571. https://doi.org/10.1007/s004250100659

Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., Ågren, J., Bossdorf, O., Byers, D., & Donohue, K. (2010). The scale of population structure in Arabidopsis thaliana. *PLoS genetics, 6*(2), e1000843.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

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.

Qi, Z., Zhang, Z., Wang, Z., Yu, J., Qin, H., Mao, X., Jiang, H., Xin, D., Yin, Z., Zhu, R., Liu, C., Yu, W., Hu, Z., Wu, X., Liu, J., & Chen, Q. (2018). Meta-analysis and transcriptome profiling reveal hub genes for soybean seed storage composition during seed development. *Plant Cell Environ, 41*(9), 2109-2127. https://doi.org/10.1111/pce.13175

Rabier, C. E., Ta, T., & Ane, C. (2014). Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Molecular Biology and Evolution, 31*(3), 750-762. https://doi.org/10.1093/molbev/mst263

Rajjou, L., Duval, M., Gallardo, K., Catusse, J., Bally, J., Job, C., & Job, D. (2012). Seed germination and vigor. *Annu Rev Plant Biol, 63*, 507-533. https://doi.org/10.1146/annurev-arplant-042811-105550

Research, I. G. (2017). *Amino Acid Market: Global Industry Analysis, Trends, Market Size & Forecast to 2023* Retrieved June 9 from ttps://www.researchandmarkets.com/research/xsjjhs/amino_acid

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., Altmann, T., Stitt, M., Willmitzer, L., & Melchinger, A. E. (2012). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proceedings of the National Academy of Sciences, 109*(23), 8872-8877.

Rohr, F., Ulrichs, C., Mucha-Pelzer, T., & Mewis, I. (2006). Variability of aliphatic glucosinolates in Arabidopsis and their influence on insect resistance. *Commun Agric Appl Biol Sci, 71*(2 Pt B), 507-515.

Sankoff, D., Zheng, C., & Zhu, Q. (2010). The collapse of gene complement following whole genome duplication. *BMC Genomics, 11*(1), 313.

Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S., & Wolfe, K. H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature, 440*, 341-345.

Scannell, D. R., Frank, A. C., Conant, G. C., Byrne, K. P., Woolfit, M., & Wolfe, K. H. (2007). Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proceedings of the National Academy of Sciences, U.S.A., 104*, 8397-8402.

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

Scheible, W. R., Morcuende, R., Czechowski, T., Fritz, C., Osuna, D., Palacios-Rojas, N., Schindelasch, D., Thimm, O., Udvardi, M. K., & Stitt, M. (2004). Genome-wide reprogramming of primary and secondary metabolism, protein synthesis, cellular growth processes, and the regulatory infrastructure of Arabidopsis in response to nitrogen. *Plant Physiol, 136*(1), 2483-2499. https://doi.org/10.1104/pp.104.047019

Schmidt, M. A., Barbazuk, W. B., Sandford, M., May, G., Song, Z., Zhou, W., Nikolau, B. J., & 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, 156*(1), 330-345.

[Record #46 is using a reference type undefined in this output style.]

Schnable, J. C., Freeling, M., & Lyons, E. (2012). Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol Evol, 4*(3), 265-277. https://doi.org/evs009 [pii] 10.1093/gbe/evs009

Schnable, J. C., Springer, N. M., & Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences, 108*(10), 4069-4074.

Schranz, M. E., Mohammadin, S., & Edger, P. P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current opinion in plant biology, 15*(2), 147-153. https://doi.org/S1369-5266(12)00046-5 [pii] 10.1016/j.pbi.2012.03.011

Schwender, J., König, C., Klapperstück, M., Heinzel, N., Munz, E., Hebbelmann, I., Hay, J. O., Denolf, P., De Bodt, S., & Redestig, H. (2014). Transcript abundance on its own cannot be used to infer fluxes in central metabolism. *Frontiers in plant science, 5*, 668.

Scienski, K., Fay, J. C., & Conant, G. C. (2015). Patterns of Gene Conversion in Duplicated Yeast Histones Suggest Strong Selection on a Coadapted Macromolecular Complex. *Genome Biology and Evolution, 7*(12), 3249-3258.

Scossa, F., Laudencia-Chingcuanco, D., Anderson, O. D., Vensel, W. H., Lafiandra, D., D'Ovidio, R., & Masci, S. (2008). Comparative proteomic and transcriptional profiling of a bread wheat cultivar and its derived transgenic line overexpressing a low molecular weight glutenin subunit gene in the endosperm. *Proteomics, 8*(14), 2948-2966. https://doi.org/10.1002/pmic.200700861

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet, 44*(7), 825-830. https://doi.org/10.1038/ng.2314

Seoighe, C., & Wolfe, K. H. (1998). Extent of genomic rearrangement after genome duplication in yeast. *Proceedings of the National Academy of Sciences, U.S.A., 95*(#8), 4447-4452.

Seren, Ü., Vilhjálmsson, B. J., Horton, M. W., Meng, D., Forai, P., Huang, Y. S., Long, Q., Segura, V., & Nordborg, M. (2012). GWAPP: a web application for genome-wide association mapping in Arabidopsis. *Plant Cell, 24*(12), 4793-4805. https://doi.org/10.1105/tpc.112.108068

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. *Funct Integr Genomics, 14*(4), 683-696. https://doi.org/10.1007/s10142-014-0388-x

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res, 13*(11), 2498-2504. https://doi.org/10.1101/gr.1239303

Shaul, O., & Galili, G. (1992). Increased lysine synthesis in tobacco plants that express high levels of bacterial dihydrodipicolinate synthase in their chloroplasts. *The Plant Journal, 2*(2), 203-209.

Shrestha, V. (2020). *UNCOVERING THE GENETIC ARCHITECTURE AND METABOLIC BASIS OF AMINO ACID COMPOSITION IN MAIZE KERNELS USING MULTI-OMICS INTEGRATION* University of Missouri, Columbia].

Skokut, T. A., Wolk, C. P., Thomas, J., Meeks, J. C., & Shaffer, P. W. (1978). Initial organic products of assimilation of [N]ammonium and [N]nitrate by tobacco cells cultured on different sources of nitrogen. *Plant Physiology, 62*, 299-304.

[Record #392 is using a reference type undefined in this output style.]

Slaten, M. L., Yobi, A., Bagaza, C., Chan, Y. O., Shrestha, V., Holden, S., Katz, E., Kanstrup, C., Lipka, A. E., Kliebenstein, D. J., Nour-Eldin, H. H., & Angelovici, R. (2020). mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis. *Plant Physiol, 183*(2), 483-500. https://doi.org/10.1104/pp.20.00039

Slotte, T., Hazzouri, K. M., Agren, J. A., Koenig, D., Maumus, F., Guo, Y. L., Steige, K., Platts, A. E., Escobar, J. S., Newman, L. K., Wang, W., Mandakova, T., Vello, E., Smith, L. M., Henz, S. R., Steffen, J., Takuno, S., Brandvain, Y., Coop, G., Andolfatto, P., Hu, T. T., Blanchette, M., Clark, R. M., Quesneville, H., Nordborg, M., Gaut, B. S., Lysak, M. A., Jenkins, J., Grimwood, J., Chapman, J., Prochnik, S., Shu, S., Rokhsar, D., Schmutz, J., Weigel, D., & Wright, S. I. (2013). The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nature Genetics, 45*(7), 831-835. https://doi.org/10.1038/ng.2669

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry: 3rd Edition*. W. H. Freeman and Company.

Soltis, D. E., Soltis, P. S., PIRES, J. C., Kovarik, A., Tate, J. A., & Mavrodiev, E. (2004). Recent and recurrent polyploidy in Tragopogon (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biological Journal of the Linnean Society, 82*(4), 485-501.

Sonderby, I. E., Burow, M., Rowe, H. C., Kliebenstein, D. J., & Halkier, B. A. (2010). A Complex Interplay of Three R2R3 MYB Transcription Factors Determines the Profile of Aliphatic Glucosinolates in Arabidopsis1[C][W][OA]. *Plant Physiology, 153*(1), 348-363.

Sonderby, I. E., Hansen, B. G., Bjarnholt, N., Ticconi, C., Halkier, B. A., & Kliebenstein, D. J. (2007). A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. *Plos One, 2*(12), e1322. https://doi.org/10.1371/journal.pone.0001322

Stark, C., Breitkreutz, B. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K., & Tyers, M. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research, 39*(Database issue), D698-704. https://doi.org/gkq1116 [pii] 10.1093/nar/gkq1116

Svennerstam, H., Jämtgård, S., Ahmad, I., Huss-Danell, K., Näsholm, T., & Ganeteg, U. (2011). Transporters in Arabidopsis roots mediating uptake of amino acids at naturally occurring concentrations. *New Phytol, 191*(2), 459-467. https://doi.org/10.1111/j.1469-8137.2011.03699.x

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., & 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.

Tabe, L., & Higgins, T. (1998). Engineering plant protein composition for improved nutrition. *Trends in plant science, 3*(7), 282-286.

Tabe, L. M., & Droux, M. (2002). Limits to sulfur accumulation in transgenic lupin seeds expressing a foreign sulfur-rich protein. *Plant Physiology, 128*(3), 1137-1148.

Tan-Wilson, A. L., & Wilson, K. A. (2012). Mobilization of seed protein reserves. *Physiol Plant, 145*(1), 140-153. https://doi.org/10.1111/j.1399-3054.2011.01535.x

Tang, H., Woodhouse, M. R., Cheng, F., Schnable, J. C., Pedersen, B. S., Conant, G., Wang, X., Freeling, M., & Pires, J. C. (2012). Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. *Genetics, 190*(4), 1563-1574. https://doi.org/genetics.111.137349 [pii] 10.1534/genetics.111.137349

Tan-Wilson, A. L., & Wilson, K. A. (2012). Mobilization of seed protein reserves. *Physiologia Plantarum, 145*(1), 140-153.

Taxis, T. M., Wolff, S., Gregg, S. J., Minton, N. O., Zhang, C., Dai, J., Schnabel, R. D., Taylor, J. F., Kerley, M. S., Pires, J. C., Lamberson, W. R., & Conant, G. C. (2015). The players may change but the game remains: Network analyses of ruminal microbiomes suggest taxonomic differences mask functional similarity. *Nucleic Acids Research, 43*(20), 9600-9612.

Taylor, J. S., & Raes, J. (2004). Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics, 38*, 615-643.

Team, R. (2014). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria2014. *URL:(https://www. R-project. org)*.

Textor, S., Bartram, S., Kroymann, J., Falk, K. L., Hick, A., Pickett, J. A., & Gershenzon, J. (2004). Biosynthesis of methionine-derived glucosinolates in Arabidopsis thaliana: recombinant expression and characterization of methylthioalkylmalate synthase, the condensing enzyme of the chain-elongation cycle. *Planta, 218*(6), 1026-1035. https://doi.org/10.1007/s00425-003-1184-3

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature, 408*(6814), 796-815. https://doi.org/10.1038/35048692

Thomas, B. C., Pedersen, B., & Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research, 16*(7), 934-946. https://doi.org/10.1101/gr.4708406

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 Res, 45*(W1), W122-W129. https://doi.org/10.1093/nar/gkx382

Torrent, M., Alvarez, I., Geli, M. I., Dalcol, I., & Ludevid, D. (1997). Lysine-rich modified γ-zeins accumulate in protein bodies of transiently transformed maize endosperms. *Plant molecular biology, 34*(1), 139-149.

Toubiana, D., Semel, Y., Tohge, T., Beleggia, R., Cattivelli, L., Rosental, L., Nikoloski, Z., Zamir, D., Fernie, A. R., & Fait, A. (2012). Metabolic profiling of a mapping population exposes new insights in the regulation of seed metabolism and seed, fruit, and plant relations. *Plos genetics, 8*(3), e1002612.

Tzin, V., & Galili, G. (2010). New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. *Molecular plant, 3*(6), 956-972.

Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics, 10*(10), 725-732.

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics, 18*(7), 411-424.

van Hoek, M. J., & Hogeweg, P. (2009). Metabolic adaptation after whole genome duplication. *Molecular Biology and Evolution, 26*(11), 2441-2453. https://doi.org/msp160 [pii] 10.1093/molbev/msp160

VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., Spittle, K., Hall, R., Gu, J., & Lyons, E. (2015). Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature, 527*(7579), 508.

Vaughn, J. N., Nelson, R. L., Song, Q., Cregan, P. B., & Li, Z. (2014). The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. *G3: Genes, Genomes, Genetics, 4*(11), 2283-2294.

Velasco, P., Soengas, P., Vilar, M., Cartea, M. E., & del Rio, M. (2008). Comparison of glucosinolate profiles in leaf and seed tissues of different Brassica napus crops. *Journal of the American Society for Horticultural Science, 133*(4), 551-558. https://doi.org/Doi 10.21273/Jashs.133.4.551

Verslues, P. E., Lasky, J. R., Juenger, T. E., Liu, T.-W., & Kumar, M. N. (2014). Genome-wide association mapping combined with reverse genetics identifies new effectors of low water potential-induced proline accumulation in Arabidopsis. *Plant physiology, 164*(1), 144-159.

Wanasundara, J. P. (2011). Proteins of Brassicaceae oilseeds and their potential as a plant protein source. *Critical reviews in food science and nutrition, 51*(7), 635-677.

Wang, J., Tian, L., Lee, H.-S., Wei, N. E., Jiang, H., Watson, B., Madlung, A., Osborn, T. C., Doerge, R., & Comai, L. (2006). Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics, 172*(1), 507-517.

Wang, J., & Zhang, Z. (2018). GAPIT Version 3:An Interactive Analytical Tool for Genomic Association and Prediction. *Bioinformatics. Draft.*(7).

Wang, S., Wang, F., Tian, S., Wang, M., Sui, N., & Zhang, X. (2014). Transcript profiles of maize embryo sacs and preliminary identification of genes involved in the embryo sac-pollen tube interaction. *Frontiers in Plant Science, 5*, 1-15.

Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature, 449*, 54-61.

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., & Yan, J. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat Commun, 5*, 3438. https://doi.org/10.1038/ncomms4438

Wenefrida, I., Utomo, H. S., Blanche, S. B., & Linscombe, S. D. (2009). Enhancing essential amino acids and health benefit components in grain crops for improved nutritional values. *Recent patents on DNA & gene sequences, 3*(3), 219-225.

Wentzell, A. M., Rowe, H. C., Hansen, B. G., Ticconi, C., Halkier, B. A., & Kliebenstein, D. J. (2007). Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *Plos genetics, 3*(9), e162.

Wentzell, A. M., Rowe, H. C., Hansen, B. G., Ticconi, C., Halkier, B. A., & Kliebenstein, D. J. (2007). Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genet, 3*(9), 1687-1701. https://doi.org/10.1371/journal.pgen.0030162

WHO. *Global Database on Child Growth and Malnutrition* Retrieved June 9 from http://www.who.int/ nutgrowthdb/about/introduction/en/

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G. V., & Provart, N. J. (2007). An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS One, 2*(8), e718. https://doi.org/10.1371/journal.pone.0000718

Wisecaver, J. H., Borowsky, A. T., Tzin, V., Jander, G., Kliebenstein, D. J., & Rokas, A. (2017). A global co-expression network approach for connecting genes to specialized metabolic pathways in plants. *The Plant Cell*, tpc. 00009.02017.

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.

Wolfe, K. H. (2000). Robustness: It's not where you think it is. *Nature Genetics, 25*, 3-4.

Wolfe, K. H., & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature, 387*(#6634), 708-713.

Woodhouse, M. R., Schnable, J. C., Pedersen, B. S., Lyons, E., Lisch, D., Subramaniam, S., & Freeling, M. (2010). Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biology, 8*(6), e1000409.

Wu, S., Alseekh, S., Cuadros-Inostroza, Á., Fusari, C. M., Mutwil, M., Kooke, R., Keurentjes, J. B., Fernie, A. R., 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.

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.

Yang, R., Jarvis, D. E., Chen, H., Beilstein, M. A., Grimwood, J., Jenkins, J., Shu, S., Prochnik, S., Xin, M., Ma, C., Schmutz, J., Wing, R. A., Mitchell-Olds, T., Schumaker, K. S., & Wang, X. (2013). The Reference Genome of the Halophytic Plant Eutrema salsugineum. *Front Plant Sci, 4*, 46. https://doi.org/10.3389/fpls.2013.00046

Yao, M., Guan, M., Zhang, Z., Zhang, Q., Cui, Y., Chen, H., Liu, W., Jan, H. U., Voss-Fels, K. P., & 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.

Yao, M., Guan, M., Zhang, Z., Zhang, Q., Cui, Y., Chen, H., Liu, W., Jan, H. U., Voss-Fels, K. P., Werner, C. R., He, X., Liu, Z., Guan, C., Snowdon, R. J., Hua, W., & Qian, L. (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), 320. https://doi.org/10.1186/s12864-020-6711-0

Yobi, A., & Angelovici, R. (2018). A High-throughput absolute-level quantification of protein-bound amino acids in seeds. *Curr Protoc Plant Biol*, e20084. https://doi.org/10.1002/cppb.20084

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., Batushansky, A., Oliver, M. J., & Angelovici, R. (2019). Adaptive responses of amino acid metabolism to the combination of desiccation and low nitrogen availability in Sporobolus stapfianus [journal article]. *Planta, 249*, 1535-1549. https://doi.org/10.1007/s00425-019-03105-6

Zhan, S. H., Drori, M., Goldberg, E. E., Otto, S. P., & Mayrose, I. (2016). Phylogenetic evidence for cladogenetic polyploidization in land plants. *American Journal of Botany, 103*(7), 1252-1258.

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z., & Wang, W. (2012). Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nature Biotechnology, 30*(6), 549.

Zhang, H., Wang, M. L., Schaefer, R., Dang, P., Jiang, T., & Chen, C. (2019). GWAS and Coexpression Network Reveal Ionomic Variation in Cultivated Peanut. *J Agric Food Chem, 67*(43), 12026-12036. https://doi.org/10.1021/acs.jafc.9b04939

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, L., Tan, Q., Lee, R., Trethewy, A., Lee, Y. H., & Tegeder, M. (2010). Altered xylem-phloem transfer of amino acids affects metabolism and leads to increased seed yield and oil content in Arabidopsis. *Plant Cell, 22*(11), 3603-3620. https://doi.org/10.1105/tpc.110.073833

Zhang, Y., Li, B., Huai, D., Zhou, Y., & Kliebenstein, D. J. (2015). The conserved transcription factors, MYB115 and MYB118, control expression of the newly evolved benzoyloxy glucosinolate pathway in Arabidopsis thaliana. *Frontiers in Plant Science, 6*, 343.

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.

Zhu, X., & Galili, G. (2004). Lysine metabolism is concurrently regulated by synthesis and catabolism in both reproductive and vegetative tissues. *Plant Physiology, 135*(1), 129-136.

Ziegler, G., Terauchi, A., Becker, A., Armstrong, P., Hudson, K., & Baxter, I. (2013a). Ionomic Screening of Field-Grown Soybean Identifies Mutants with Altered Seed Elemental Composition. *Plant Genome, 6*(2), 1-9.

Ziegler, G., Terauchi, A., Becker, A., Armstrong, P., Hudson, K., & Baxter, I. (2013b). Ionomic Screening of Field-Grown Soybean Identifies Mutants with Altered Seed Elemental Composition. *Plant Genome, 6*(2). https://doi.org/10.3835/plantgenome2012.07.0012

VITA

Marianne Slaten grew-up in rural Missouri. Her childhood experiences led to a curiosity and desire for a greater understanding of agricultural systems which led her to attend Truman State University where she earned a Bachelor's of Science Degree in Agricultural Sciences. During her time at Truman, she worked in the lab of Dr. Mark Campbell researching the diversity of starchy mutant alleles in maize. Inspired by the worldly notion that her research may contribute to the nutritional needs of others, Marianne attended Iowa State University where she earned her Master's Degree in Plant Breeding and Genetics. Marianne worked in the USDA lab of Dr. Paul Scott where she continued work with mutant starchy alleles, introgressing novel starch phenotypes into maize varieties commonly utilized in the snack industry. In addition, Marianne led a de novo assembly of a genomic region responsible for maize gametophytic incompatibility. Realizing the impact that bioinformatics had on the field of plant genetics, Marianne moved to Columbia, Missouri to attend the University of Missouri. Marianne started her PhD with a brief stop in Dr. Chris Elsik's lab gaining valuable bioinformatic skills before landing in the lab of Dr. Ruthie Angelovici where she would finish her degree. After completion of her degree, Marianne will start as a Postdoctoral researcher with Dr. Bob Sharp at the University of Missouri.