8-2022

Lentil (Lens culinaris Medik.) Prebiotic Carbohydrates and Protein Quality: Uncovering Genomic Associations and Developing Rapid FTIR Phenotyping Methods

Nathan Johnson
njohns9@g.clemson.edu

LENTIL (*LENS CULINARIS* MEDIK.) PREBIOTIC CARBOHYDRATES AND PROTEIN QUALITY: UNCOVERING GENOMIC ASSOCIATIONS AND DEVELOPING RAPID FTIR PHENOTYPING METHODS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Plant and Environmental Sciences

---

by
Nathan Jay Johnson
August 2022

Accepted by:
Dr. Dil Thavarajah, Committee Chair
Dr. Shiv Kumar
Dr. J. Lucas Boatwright
Dr. William Bridges
Dr. Stephen Kresovich

ABSTRACT

Lentil (*Lens culinaris* Medik.) is a cool-season food legume cultivated around the globe. This pulse crop boasts a rich nutrient profile including high concentrations of prebiotic carbohydrates, protein, essential amino acids, and micronutrients, such as folate, iron, zinc, and selenium. Prebiotic carbohydrates promote a healthy gut microbiome, which, in turn, is associated with reduced risk of numerous pathologies including obesity/overweight, type II diabetes, irritable bowel disease, and colon cancer. Known as "poor man's meat," lentil also provides high quality plant-based protein at a low cost. As the world increasingly looks to crops to supplement and replace animal-based protein, lentil protein offers an excellent alternative. To fully take advantage of lentil's unique nutrient profile and promote global food security, breeding programs may wish to add prebiotic carbohydrates and protein quality to their breeding target traits. Additionally, with the advance of genomics-assisted breeding approaches, genetic markers could significantly accelerate breeding efforts through marker-assisted selection and genomic selection. However, crucial lentil population data, genetic resources, and high-throughput phenotyping methods are lacking. To help address this gap, the present research quantifies seed prebiotic carbohydrates (sugar alcohols, raffinose-family oligosaccharides, fructooligosaccharides, and resistant starch) and protein quality traits (amino acids and *in vitro* protein digestibility) and calculates trait heritability estimates in a lentil diversity panel. Genome-wide association studies identify significantly associated SNP markers and candidate genes, while admixture analysis elucidates lentil ancestral subpopulations and their global distribution. Finally, the development of high-throughput Fourier-Transform infrared spectroscopy (FTIR) phenotyping methods promises to significantly reduce breeding operation costs in developed and developing countries alike. Thus, this research advances lentil nutritional breeding to aid in the

development of new germplasm and varieties targeted for unique growing environments and consumer populations.

DEDICATION

To Hannah, my wife and my love,

who has supported me most and sacrificed even more than I know.

I am forever grateful.

And to Jesus, the Messiah—my Rock and Morning Light—for giving me

a field of lentils to stand upon (II Samuel 23).

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Dil Thavarajah. From the offer to join her program to final manuscript edits, her encouragement, patience, and skillful mentoring know no bounds. Her passion for science and global food security is contagious. I would also like to thank my committee members—Dr. Shiv Kumar for his aid and insight as a leading global lentil breeder, Dr. William Bridges for his skilled teaching of statistics and acclaimed encouragement of grad students like myself, and Dr. Stephen Kresovich for his guiding insights and counsel into my research and academic pursuits. I would like to acknowledge and thank Dr. Lucas Boatwright for his extensive investment into my program through computational assistance, teaching, and encouragement.

Funding support for this research was provided by the USDA National Institute of Food and Agriculture including: the Plant Health and Production and Plant Products: Plant Breeding for Agricultural Production program area (grant no. 2018-67014-27621/project accession no. 1015284), Hatch project [1022664], and the Organic Agriculture Research and Extension Initiative (OREI) (award no. 2018-51300-28431/proposal no. 2018-02799). Funding was also provided by the International Center for Agricultural Research in the Dry Areas (ICARDA, Morocco), the Good Food Institute, and the Feed the Future Innovation Lab for Crop Improvement through the United States Agency for International Development (USAID) under Cooperative Agreement No 7200AA19LE00005/Subaward no 89915-11295.

TABLE OF CONTENTS

Page

Table of Contents (Continued) Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

THE ROLES AND POTENTIAL OF LENTIL PREBIOTIC CARBOHYDRATES IN HUMAN AND PLANT HEALTH

Abstract

Diet-related ailments, such as obesity and micronutrient deficiencies, have global adverse impacts on society. Lentil is an important staple crop, especially in South Asia and Africa, and has been associated with the prevention of chronic illnesses, including type II diabetes, obesity, and cancer. Lentil, a cool-season food legume, is rich in protein and micronutrients while also containing a range of prebiotic carbohydrates, such as raffinose family oligosaccharides (RFOs), fructooligosaccharides, sugar alcohols (SAs), and resistant starch (RS), which contribute to lentil's health benefits. Prebiotic carbohydrates are fermented by beneficial microorganisms in the colon, which impart health benefits to the consumer. Prebiotic carbohydrates are also vital to lentil plant health, being associated with carbon transport/storage and abiotic stress tolerance. Important to both human and plant health, prebiotic carbohydrates in lentil are a prominent candidate for nutrigenomic breeding efforts. New lentil cultivars could help to combat global health problems, while also proving resilient to climate change. The objectives of this review are to: (a) discuss the benefits lentil prebiotic carbohydrates confer to human and plant health; (b) describe the biosynthesis pathways of two prominent prebiotic carbohydrate families in lentil, RFOs and SAs; and (c) consider the potential of prebiotic carbohydrates in terms of future nutritional breeding efforts.

Introduction

Lentil (*Lens culinaris* Medikus) is an ancient crop. Cultivated lentil dates to before 7000 BCE with likely origin and domestication in southern Turkey and northern Syria (Cubero, Perez de la Vega, & Fratini, 2009). The genus *Lens* contains four species: *L. culinaris* (ssp. *culinaris*, *orientalis*, *tomentosus*, and *odemensis*), *L. ervoides*, *L. lamottei*, and *L. nigricans* (Wong et al., 2015). Lentil is a diploid with seven chromosome pairs ($2n = 14$), with an estimated genome size of 4,063 Mb (Rizvi, Aski, Sarker, Dikshit, & Yadav, 2019). Lentil is a staple crop in much of the world, consumed particularly in South Asia and Africa. World lentil production, led by Canada, India, Turkey, and the United States, exceeded 7.5 million tons in 2017 (FAOSTAT, 2017).

Lentil is commonly consumed as a soup or "dahl," a Southeast Asian dish typical in India, Nepal, Bangladesh, and Sri Lanka. Lentil has been referred to colloquially as "Poor man's meat," as it is a rich source of nutrients, composed of 60%–67% carbohydrate, 20%–36% protein, <4% lipid, and 2%–3% ash on a dry basis (Bhatty, 1988). Its nutritional values compare favorably to other significant legumes and cereals, such as chickpea, soybean, rice, and wheat (Table 1.1). Lentil is an excellent source of energy; it is high in protein (typical of legumes), low in lipids, compared to chickpea and soybean, and rich in minerals and vitamins, compared to rice and wheat (Table 1.1). Consequently, a diet rich in lentil and other legumes has many health benefits. For example, substituting a half serving of legumes for eggs, bread, rice, or baked potato reduces the risk of developing diabetes (Becerra-Tomás et al., 2018). This effect is in part attributed to the low glycemic index of lentil and other legumes. Red lentil glycemic index (21%) compares favorably to other grain carbohydrate sources, such as multigrain bread (62%), basmati rice (69%), and whole-wheat pasta (55%; Henry, Lightowler, Strik, Renton, & Hails, 2005). A lentil-based diet reduces total and low-density lipoprotein cholesterol and the risk of

cardiovascular disease (Abeysekara, Chilibeck, Vatanparast, & Zello, 2012), increases satiety (McCrory, Hamaker, Lovejoy, & Eichelsdoerfer, 2010), and is considered a potential solution to help combat obesity (Siva, Johnson, et al., 2018). Many of lentil's health benefits are likely due to the type and concentration of prebiotic carbohydrates present in the seed and how these change during cooking, cooling, and reheating (Johnson, Thavarajah, Combs, & Thavarajah, 2013).

Prebiotic carbohydrates are specific colonic nutrients that act as biosynthetic precursors for human microbiota activity, which in turn leads to possible health benefits related to combating type II diabetes and obesity. In addition to human health benefits, prebiotic carbohydrates also benefit plant health by increasing leaf raffinose family oligosaccharides (RFOs) to enhance drought (Bartels & Sunkar, 2005), chilling (Nishizawa, Yabuta, & Shigeoka, 2008), and freezing tolerance (Pennycooke, Jones, & Stushnoff, 2003). Sugar alcohols (SAs) also increase chilling (Chiang, Stushnoff, McSay, Jones, & Bohnert, 2005), drought (Pujni, Chaudhary, & Rajam, 2007), and salinity tolerance in a range of plants (Zhifang & Loescher, 2003). These RFOs and SAs generally act as signaling compounds for both biotic and abiotic stresses (Valluru & Van den Ende, 2011). With climate conditions changing globally, future lentil production might be limited due to increased incidence of drought and higher temperatures. The significance of prebiotic carbohydrates to human and plant health means the type and concentration thereof in lentil are essential traits for nutrigenomic breeding efforts. Nutritionally improved lentil cultivars could help to combat global health problems, while simultaneously enhancing resilience to the effects of climate change (Muehlbauer et al., 2006).

Lentil Prebiotic Carbohydrates

Lentil contains a range of prebiotic carbohydrates including average concentrations of 4,071 mg of RFOs, 1,423 mg of SAs, 62 mg of FOSs, and 7,500 mg of RS per 100 g (Johnson et al., 2013). A recent study reported the prebiotic carbohydrate profile after removing protein and fat from lentil seeds (Table 1.2: Siva, Thavarajah, Kumar, & Thavarajah, 2019). Among simple sugars, sucrose was the most abundant (1,174–2,288 mg/100 g) followed by glucose (21–61 mg/100 g), fructose (0.2–21.9 mg/100 g), mannose (1.2–7.9 mg/100 g), and rhamnose (0.5–1.0 mg/100 g). For SAs, sorbitol concentrations (606–733 mg/100 g) were the highest followed by mannitol (9–31 mg/100 g) and xylitol (14–31 mg/100 g) regardless of the lentil market class. Among RFOs, stachyose (2,236–2,348 mg/100 g) was more abundant than raffinose (403–646 mg/100 g) and verbascose (581–1,769 mg/100 g). Considering lentil FOSs, kestose levels were higher than nystose levels. Other prebiotic carbohydrates present were arabinose (2,419–2,630 mg/100 g), xylose (1,912–1,936 mg/100 g), and cellulose (611–640 mg/100 g).

Lentil prebiotic carbohydrate concentrations vary by growing location. Johnson, Thavarajah, Thavarajah, Fenlason, et al. (2015) analyzed lentil samples from six countries (Table 1.3). They observed that total low-molecular weight carbohydrate concentrations were generally the highest in regions with less rainfall, higher temperatures, and higher estimated stress index. This suggests a mechanism of abiotic stress tolerance correlated with the type and level of prebiotic carbohydrates in lentil seeds. Total RFO concentrations ranged from 5,225 mg/100 g in Syria to 7,149 mg/100 g in Morocco. Total SA concentrations ranged from 1,385 mg/100 g in Washington State to 2,019 mg/100 g in Morocco. Further to variability due to location, they noted variation among the nine genotypes analyzed as well as a genotype × location interaction. The significant genotype × growing location interaction supports

the hypothesis that increasing the nutritional value of lentil prebiotic carbohydrates can be achieved by selecting ideal growing areas and suitable cultivars for developing nutritionally superior varieties (Johnson, Thavarajah, Thavarajah, Fenlason, et al., 2015).

Concentration of prebiotic carbohydrate can also vary by location and genotype, or by method of food processing (Johnson, Thavarajah, Thavarajah, Payne, et al., 2015; Siva, Thavarajah, & Thavarajah, 2018). Lentils are often cooked, cooled, and reheated before consumption; hence these processes are important considerations in terms of their impact on the prebiotic carbohydrates undergoing these processes prior to consumption. Johnson, Thavarajah, Thavarajah, Payne, et al. (2015) measured prebiotic carbohydrate concentrations in whole and dehulled red and green lentil when raw and after cooking, cooling, and reheating. RFO concentrations decreased with processing (Figure 1.1), although the differences between raw and reheated lentil were only significant in whole lentil products. Differences in RS concentrations between raw/cooked and cooled/reheated were significant, indicating RS increases when food products are cooled after cooking, likely due to annealing. Siva, Thavarajah, et al. (2018) also showed this trend in RS. Additionally, they measured SA concentrations and found that sorbitol and mannitol concentrations significantly increase from cooked to cooled lentil in most market classes and then decrease again with reheating (Figure 1.2). These studies show that cooking/cooling/reheating processes can increase the health benefits of lentil via modulation of prebiotic carbohydrate concentrations.

Prebiotic Carbohydrates

The definition of a *prebiotic* has evolved since its coining in 1995. Complementary to the probiotic concept, Gibson and Roberfroid (1995) originally defined a prebiotic as a "*non-digestible food ingredient that beneficially affects the host by selectively stimulating the growth*

and/or activity of one or a limited number of bacteria already resident in the colon." This definition was revised in 2004 to three criteria that restricted prebiotic foods to ingredients that are (a) resistant to mammalian digestion; (b) fermented by intestinal microflora; and (c) selectively stimulate the growth and/or activity of intestinal bacteria associated with health and well-being (Gibson, Probert, Van, Rastall, & Roberfroid, 2004). The definition was further broadened in 2008 by the Food and Agricultural Organization of the United Nations to allow the possibility of extraintestinal sites and eliminate the requirement of selective fermentation (Pineiro et al., 2008). The definition was critiqued by Gibson et al. (2010) for this latter omission and also for not adequately excluding antibiotics. Reaffirming selective fermentation and establishing "*a niche*," Gibson et al. (2010) defined a *dietary prebiotic* as "*a selectively fermented ingredient that results in specific changes in the composition and/or activity of the gastrointestinal microbiota, thus conferring benefit(s) upon host health*." Selective fermentation was again challenged by Bindels, Delzenne, Cani, and Walter (2015), who eliminated this requirement from their definition and again restricted *prebiotic* to the gastrointestinal tract. In 2016, the International Scientific Association for Probiotics and Prebiotics (ISAPP) came to the current consensus definition: "*a substrate that is selectively utilized by host microorganisms conferring a health benefit*" (Gibson et al., 2017). This current definition has broadened the scope of prebiotics beyond carbohydrate substrates in the gastrointestinal tract by acknowledging the potential for non-gastrointestinal sites and non-carbohydrate substances. However, the definition has retained the selective fermentation component, which the ISAPP sees as vital to the concept of prebiotics (Gibson et al., 2017). While the definition has broadened beyond dietary carbohydrates, research on prebiotics has primarily focused on dietary prebiotic carbohydrates, and, consequently, these are our focus here regarding lentil.

Prebiotic carbohydrates can be categorized based on their degree of polymerization, sugar subunits, and linkage configuration. Naturally occurring prebiotic carbohydrates are divided into two major groups: dietary fiber and SAs (Roberfroid, 2007). Dietary fiber is comprised of starch polysaccharides (RS) and non-starch polysaccharides (RFOs, fructooligosaccharide [FOSs], galactooligosaccharides, xylooligosaccharides, hemicellulose, cellulose, pectin, and inulin; Roberfroid, 2007). These prebiotic carbohydrates are associated with many human health benefits, because they promote satiety, lower high cholesterol, and regulate postprandial blood glucose levels (Beserra et al., 2015). Most naturally occurring prebiotic carbohydrates are found in fresh vegetables, legumes, and fruits at concentrations ranging from trace amounts in wheat, to moderate levels in onion and green bananas, to relatively high concentrations (35.7–47.6 g/100 g) in chicory root (Van Loo et al., 1999).

As a staple part of many diets, legumes, such as lentil and chickpea, provide an excellent source of prebiotic carbohydrates (Table 1.2). Legumes tend to have higher concentrations of SA, RFO, fiber, and RS than prebiotic-rich fruits and vegetables, which tend to be higher in simple sugars and fructooligosaccharides (Table 1.2). For example, lentil and chickpea contain mean sorbitol concentrations of 0.66 and 0.52 g/100 g, respectively, compared to not detected and 1.09 g/100 g in onion and nectarine, respectively. With the exception of 0.23 g/100 g of raffinose in onion, nectarine and onion are void of detectable concentrations of RFO. Lentil and chickpea, however, have total RFO concentrations of 4.14 and 1.09 g/100 g, respectively. Although all legumes have merit as prebiotic-rich foods, our focus here is lentil, which is one of the most studied cool-season food legumes.

Lentil Prebiotic Carbohydrates and Gut Health

The human gastrointestinal tract, with a surface area of over 300 m^2, hosts more than 100 trillion microorganisms (Savage, 1977). These microbes, collectively termed “the microbiome”, comprise 10 times more cells than human cells and over 100 times more genetic information than the human genome (Bäckhed, Ley, Sonnenburg, Peterson, & Gordon, 2005). The microbiome is a dynamic ecosystem, with a myriad of interactions between microbes and human tissues that change throughout the course of human growth and development. Increasingly, the microbiome is recognized as an extra-human organ, capable of protecting the host from invading pathogens, stimulating the immune system, increasing the availability of nutrients, stimulating bowel motility, and improving lipid levels in the body (Holzapfel & Schillinger, 2002). However, gut microbiota are also involved with a host of disease processes, including obesity, diabetes, infections, inflammatory bowel disease, cancer, and many others (Lynch & Pedersen, 2016). Primary determinants of microbiota composition and function include age, environment, genetic factors, diet, health status, and medical interventions, such as the use of antimicrobial agents (Lozupone, Stombaugh, Gordon, Jansson, & Knight, 2012). The concept of modulating the gut microbiome's composition and function through diet, primarily through prebiotics, has gained enormous attention (Bindels et al., 2015). Prebiotics are fermented by hindgut microflora into active metabolites—short-chain fatty acids, branched-chain fatty acids, vitamins, and bile acid derivatives—that bathe the lumen of the intestinal tract. These compounds, in turn, produce a wide range of important physiological benefits, including anti-inflammatory and immune cell regulation (Arpaia et al., 2013), antineoplastic properties (Furusawa et al., 2013), and metabolic regulation (Gao et al., 2009).

We are now discovering the importance of the microbiome in early childhood growth and development. Moderate acute malnutrition in Bangladeshi children has been related to premature microbiota composition (Subramanian et al., 2014). Supplementation with gut microbial flora from healthy children and with foods rich in several prebiotic ingredients alleviated acute malnutrition with an associated normalization of age-appropriate hindgut microflora (Gehrig et al., 2019). Moreover, an altered gut microbiome has also been implicated in autism spectrum disorder, although this interaction is not yet thoroughly understood (Li, Hu, Ou, & Xia, 2019). Prospective studies with prebiotics in autistic children, when combined with exclusion of a dietary component, have revealed modest improvements in behavioral symptoms; however, randomized controlled trials have not been able to demonstrate these effects (Ng et al., 2019). These discoveries highlight opportunities for further research toward how novel dietary approaches can improve early childhood growth and development. As lentils provide significant levels of prebiotic carbohydrate, we propose they are an ideal food source for increasing prebiotic carbohydrates in people's diets and for imparting the health benefits these may provide. Indeed, the results from a recent study in rats further support the notion that a lentil-rich diet may have significant health benefits because of the superior nutritional value of its prebiotic carbohydrates and the concomitant increase in the activity of hindgut bacteria (Siva, Johnson, et al., 2018). Specifically, rats fed on a lentil diet had a significantly lower mean body weight (443 ± 47 g/rat) than those fed on control (511 ± 51 g/rat) or corn (502 ± 38 g/rat) diets; in addition, mean percent body fat and triglyceride concentration were lower and lean body mass was higher in rats fed on the lentil diet. Moreover, the fecal abundance of *Actinobacteria* and *Bacteroidetes* (beneficial bacteria) was significantly higher and the abundance of *Firmicutes* (pathogenic bacteria) was lower in rats fed the lentil diet versus the control diet.

When considering the impact of diet on the microbiome and chronic disease, we recommend a diet with satisfactory levels of prebiotics. Legumes, such as lentil, are a rich and affordable source of prebiotic carbohydrates with 100 g of lentil providing 12 g of prebiotic carbohydrates (Siva et al., 2019). This recommendation is especially applicable to countries where legumes are often neglected in people's diets. Creativity in processing methods and marketing approaches, such as the recent advance of plant-based burgers, could help to popularize lentil and other legumes in countries where they are not generally widely consumed.

Prebiotic Carbohydrates and Plant Health

As would be expected due to their high concentrations in lentil seed, prebiotic carbohydrates are vital to lentil plant health. Several functions of these carbohydrates have been elucidated. Here we discuss two of the most abundant families of prebiotic carbohydrates in lentil, RFOs and SAs, and their roles as (a) primary photosynthetic products and carbon transport molecules; (b) carbon stores; and (c) aids of abiotic stress tolerance, namely temperature, drought, and salinity stress.

Raffinose family oligosaccharides and SAs are primary photosynthetic products and carbon transport molecules in many higher plants. Labeled $^{14}CO_2$ studies have revealed that the primary soluble carbon products synthesized through photosynthesis in higher plants are sucrose (ubiquitous), RFOs, and SAs (Loescher & Everard, 2000). The orders of plants that utilize RFOs as a photosynthetic product and storage molecule include Lamiales, Cucurbitales, Cornales, and some Celastrales (Sengupta & Majumder, 2015). *Ajuga reptans* L. is the premier example of this type of plant, which uses stachyose as its primary carbon transport molecule. To store carbon, it synthesizes RFO of higher degrees of polymerization (DP), which become trapped for storage purposes (Bachman, Matile, & Keller, 1994). Lentil is not known to synthesize RFOs in leaves

as a primary photosynthetic product and, consequently, also does not transport carbon via RFOs (Obendorf & Gorecki, 2012). Instead, sucrose and SAs function as the transport molecules to the seed during seed filling. RFOs are formed in maturing lentil seeds at high concentrations (Obendorf & Gorecki, 2012). Likewise, for SAs, Grant and ap Rees (1981) showed that approximately 70% of fixed carbon in apple leaves was made into sucrose and sorbitol. Similarly, Loescher, Tyson, Everard, Redgwell, and Bieleski (1992) found that 80%–90% of the fixed carbon was transformed into mannitol and sucrose in celery. Similar patterns of SA accumulation have been shown in lilac and apricot (Loescher & Everard, 2000). Although sucrose is the primary photosynthetic product and carbon transport molecule in legumes, SAs may also function passively in this capacity, being found in both the leaf and seed (Amede, Schubert, & Stahr, 2011; Johnson et al., 2013).

Raffinose family oligosaccharides and SAs also serve as a carbon store. As noted, some plants (i.e., *A. reptans*) store RFOs in their leaves by increasing DP. RFOs are primarily known for their accumulation in seeds during late development (Sengupta & Majumder, 2015) and are especially prevalent in legumes (Obendorf & Gorecki, 2012). RFOs protect the embryo during desiccation. During germination, RFOs are rapidly hydrolyzed by α-galactosidases but do not appear to be necessary for germination (Peterbauer & Richter, 2001). The use of SAs as a carbon store is largely dependent on tissue type, developmental stage, and environment. For example, apple leaves contain 0.9% sorbitol (dry weight) in June but 4.8% in late July (Loescher & Everard, 2000). Physiologically mature lentil seeds contain significant concentrations of both sorbitol and mannitol (Johnson et al., 2013).

Lastly, RFOs and SAs aid plants experiencing abiotic stress. During abiotic stress, several compounds accumulate, including RFOs and SAs. These compounds aid the plant in

survival through these extreme conditions by balancing osmotic pressures and have, therefore, been called “osmoprotectants” (Bohnert & Jensen, 1996). RFOs and SAs substitute for water as compatible solutes; they may provide a medium for enzyme function and protect enzymes from free radicals and consequent denaturing (Smirnoff & Cumbes, 1989). Studies using transgenic plants with upregulated RFOs and SAs have shown increased drought, cold/freezing, and salinity tolerance (Gangola & Ramadoss, 2018; Loescher & Everard, 2000; Sengupta & Majumder, 2015).

Biochemical synthesis pathways have been elucidated for both RFOs and SAs and are detailed separately below (Figure 1.3). Understanding these pathways will help to identify and exploit molecular and genetic markers that can be used in lentil breeding programs. RFOs represent a series of increasing DP formed through the addition of galactose monomers to sucrose via 1,6-α glycosidic linkage, building raffinose (DP3), stachyose (DP4), and verbascose (DP5). Higher DP (DP15 or greater) exist in some plants, such as lupin seeds (Kannan, Sharma, Gangola, Sari, & Chibbar, 2018), but are not detected in lentil. The primary RFO biosynthesis pathway uses galactinol as the galactosyl donor. Galactinol is formed via galactinol synthase from UDP-galactose and L-myo-inositol (Figure 1.3). Raffinose synthase binds the galactosyl from galactinol to a sucrose molecule to form raffinose. Stachyose synthase binds galactosyl to raffinose to form stachyose. In addition, verbascose synthase binds galactosyl to stachyose to form verbascose. RFO synthesis takes place primarily in the cytosol. A secondary RFO biosynthesis pathway exists in *A. reptans* (Bachmann et al., 1994). This pathway is independent of galactinol, using a galactosyltransferase enzyme to transfer a galactosyl unit from one RFO to another to create higher DP oligosaccharides (Sengupta & Majumder, 2015).

The most abundant and well-studied SAs in higher plants are sorbitol (glucitol) and mannitol. Both have reduced forms of hexose sugars (glucose and mannose) and share similar pathways (Figure 1.3). Sorbitol biosynthesis has been characterized in the Rosaceae family (Williamson, Jennings, Guo, Pharr, & Ehrenshaft, 2002). Glucose-6P is converted into sorbitol-6-P via sorbitol-6-P dehydrogenase, which is subsequently dephosphorylated by a phosphatase, yielding sorbitol. Mannitol biosynthesis has been characterized in celery (Williamson et al., 2002). Parallel to sorbitol biosynthesis, mannose-6-P is converted into mannitol-1-P via mannose-6-P reductase, which is then dephosphorylated by a phosphatase, yielding mannitol (Figure 1.3).

Breeding Approaches for Lentil Prebiotic Carbohydrates

Due to lentil's excellent overall nutritional makeup, it has already been targeted for biofortification (Kumar, Sen, Kumar, Gupta, & Singh, 2016). However, efforts have primarily been directed toward combatting micronutrient deficiency or "hidden hunger" (Kumar et al., 2016). Micronutrient biofortification seeks to increase concentrations of essential micronutrients, such as iron, zinc, and selenium, while decreasing levels of antinutrients, such as phytic acid, which lowers mineral bioavailability (Thavarajah et al., 2011). Prebiotic carbohydrates, such as RFOs and SAs, now show potential for biofortification. Johnson, Thavarajah, Thavarajah, Fenlason, et al. (2015) showed that lentil RFO concentration varies by genotype, while SA concentration varies both by variety and location. This finding suggests that prebiotic carbohydrate biofortification efforts are likely to succeed in producing lentil varieties with optimized prebiotic carbohydrate levels for human health, which may be increased or decreased based on the target population. Many people suffer from flatulence and bloating upon ingestion of high levels of RFOs, such as those in most legumes. This adverse effect prevents susceptible

populations from eating legumes, such as lentil, thus depriving them of associated nutritional benefits. This potential tradeoff between high RFO content and flatulence may make breeding for higher RFO content unacceptable to some consumers.

Significant genetic variability exists for lentil prebiotic carbohydrates (Frias, Vidal-Valverde, Bakhsh, Arthur, & Hedley, 1994; Johnson, Thavarajah, Thavarajah, Fenlason, et al., 2015; Tahir, Vandenberg, & Chibbar, 2011), indicating the possibility for genetic manipulation through conventional or molecular breeding approaches. Recent advances in genomic tools and techniques have great potential to accelerate current breeding efforts toward lentil varieties with moderate to high levels of prebiotic carbohydrates (Kumar, Rajendran, Kumar, Hamwieh, & Baum, 2015). Additionally, genome-wide association studies may reveal other genes/QTLs that affect the levels of prebiotic carbohydrates in lentil.

Conclusion

Lentil is a rich source of prebiotic carbohydrates including SAs, RFOs, FOSs, and other polysaccharides such as cellulose, hemicellulose, and amylose. In addition to the human nutritional benefits, prebiotic carbohydrates have a significant influence on plant health, a feature that will significantly benefit the breeding of pulse crops for climate resilience. Consequently, lentil prebiotic carbohydrates are an important breeding target, requiring further characterization and evaluation of germplasm. Phenotyping diverse lentil mapping populations could identify future genetic markers associated with high levels of prebiotic carbohydrates and thus significantly accelerate nutritional breeding for different growing environments and consumer preference (Varshney et al., 2013). These genetic markers could then be used to screen locally grown varieties as well as to develop new cultivars with special consumer requirements; for example, breeder-friendly genetic markers can be used to develop new varieties with moderate

RFOs and increased levels of FOSs and RS to reduce flatulence in populations sensitive to RFOs. Globally, the development and selection of lentil genotypes with enhanced levels of prebiotic carbohydrates could not only provide significant health benefits to society but could also provide economic benefits through improved crop sustainability and production.

Acknowledgments

Funding support for this project was provided by the Plant Health and Production and Plant Products: Plant Breeding for Agricultural Production program area (grant no. 2018-67014-27621/project accession no. 1015284) of the USDA National Institute of Food and Agriculture and the International Center for Agricultural Research in the Dry Areas (ICARDA, Morocco).

Tables and Figures

Table 1.1: Nutritional values per 100 g of raw lentil, chickpea, soybean, rice, and wheat

Nutrient	Lentil	Chickpea	Soybean	Rice	Wheat
Proximate analysis					
Water (g)	8.3	7.7	8.5	11.6	12.4
Energy (kcal)	352	378	446	365	332
Protein (g)	25	20	36	7	10
Total lipid (g)	1.1	6.0	20	0.7	2.0
Carbohydrate (by difference, g)	63	63	30	80	74
Fiber (g)	11	12	9	1	13
Sugars (g)	2.0	11	7	0.1	1.0
Minerals (mg)					
Calcium (Ca)	35	57	277	28	33
Iron (Fe)	6.5	4.3	16	0.8	3.7
Magnesium (Mg)	47	79	280	25	117
Phosphorus (P)	281	252	704	115	323
Potassium (K)	677	718	1797	115	394
Sodium (Na)	6	24	2	5	3
Zinc (Zn)	3.3	2.8	4.9	1.1	3.0
Vitamins					
Vitamin C (mg)	4.5	4.0	6.0	0.0	0.0
Thiamin (mg)	0.87	0.48	0.87	0.07	0.3
Riboflavin (mg)	0.21	0.21	0.87	0.05	0.19
Niacin (mg)	2.61	1.54	1.62	1.60	5.35
Vitamin B-6 (mg)	0.54	0.54	0.38	0.16	0.19
Folate, DFE (µg)	479	557	375	8	28
Vitamin A, RAE (µg)	2	3	1	0	0
Vitamin E (mg)	0.49	0.82	0.85	0.11	0.53
Vitamin K (µg)	5.0	9.0	47.0	0.1	1.9

Source: Original data obtained from the USDA Nutrient Database for Standard Reference (2018).

Table 1.2: Mean carbohydrate concentrations in raw prebiotic-rich foods (lentil, chickpea, onion, and nectarine)

Carbohydrates (g/100g)	**Lentil**	**Chickpea**	**Onion**	**Nectarine**
Sugar alcohols				
Sorbitol	0.66±0.056	0.52±0.048	nd	1.08 ±0.079
Mannitol	0.02±0.008	0.02±0.006	nd	nd
Xylitol	0.02±0.006	0.02±0.002	nd	0.28±0.026
Simple sugars (SAs)				
Glucose	0.03±0.016	0.03±0.004	0.42±0.01	1.50±0.083
Fructose	0.01±0.009	tr	1.21±0.34	1.15±0.052
Sucrose	1.71±0.435	2.15±0.433	0.43±0.02	3.50±0.198
Raffinose family oligosaccharides (RFOs)				
Raffinose	0.50±0.116	0.44±0.120	0.23±0.011	nd
Stachyose	2.29±0.100	0.53±0.112	nd	nd
Verbascose	1.35±0.437	0.12±0.030	--	--
Fructooligosaccharides (FOSs)				
Kestose	0.33±0.080	0.04±0.018	1.15±0.046	0.18±0.011
Nystose	tr	0.01±0.006	0.77±0.028	0.65±0.015
Soluble Fiber	1.44±0.11	tr	--	--
Insoluble Fiber	19.0±1.27	13.9±0.09	--	--
Resistant Starch	3.25±0.42	3.39±0.96	--	--

Note: Data are expressed as mean ± *SD*. Abbreviations: nd, not detected; tr, trace amount. Sugar alcohol, simple sugar, and oligosaccharide data were obtained from Siva et al. (2019) and Jovanovic-Malinovska, Kuzmanova, and Winkelhausen (2014) for lentil/chickpea and onion/nectarine, respectively. Fiber and resistant starch data were obtained from de Almeida Costa, Silva, Pissini Machado Reis, and Oliveira (2006).

Table 1.3: Prebiotic carbohydrate concentrations vary by growing location

Country	Sugar Alcohols (mg/100 g)			Raffinose Family Oligosaccharides (mg/100 g)	
	Sorbitol	Mannitol	Galactinol	Raffinose+Stachyose	Verbascose
Washington, USA[†]	1259	57	69	3956	2453
Terbol, Lebanon	1528	117	52	3314	1926
Morocco[‡]	1824	132	63	4802	2347
Breda, Syria	1419	87	46	3318	1907
Sanliurfa, Turkey	1328	111	53	3494	2273
Akaki, Ethiopia	1611	118	89	3774	2272
Mean	1509	106	63	3847	2266

[a] Mean values of three locations in Washington, USA (Garfield, Fairfield, and Pullman) are reported. [b] Mean values of three locations in Morocco (Jemaat, Shaim, and Marchouche) are reported. Original data obtained from Johnson, Thavarajah, Thavarajah, Fenlason, et al. (2015).

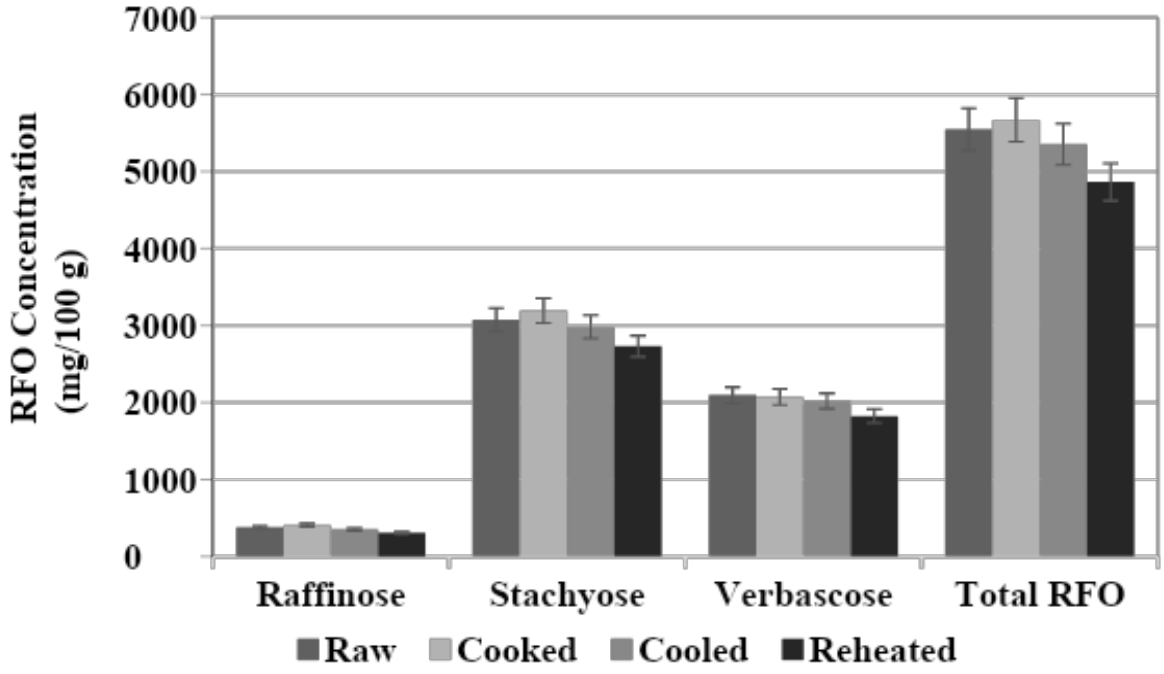

Figure 1.1: Mean raffinose family oligosaccharide (RFO) concentrations of raw, cooked, cooled, and reheated lentil. Original data obtained from Johnson, Thavarajah, Thavarajah, Payne, et al. (2015)

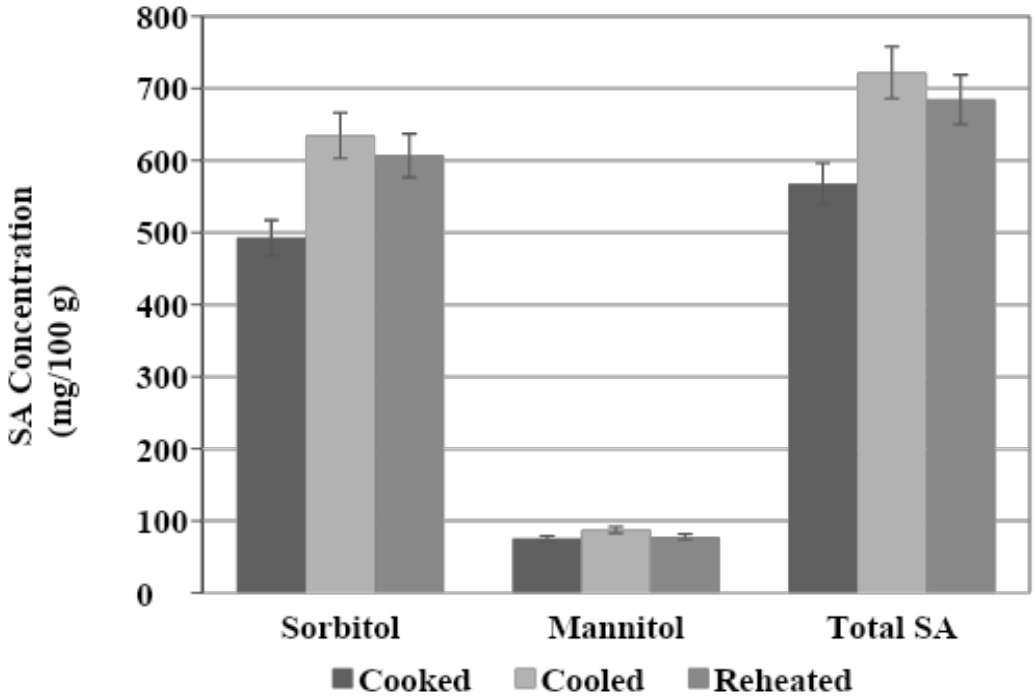

Figure 1.2: Mean sugar alcohol (SA) concentrations of cooked, cooled, and reheated lentil. Original data obtained from Siva, Thavarajah, et al. (2018)

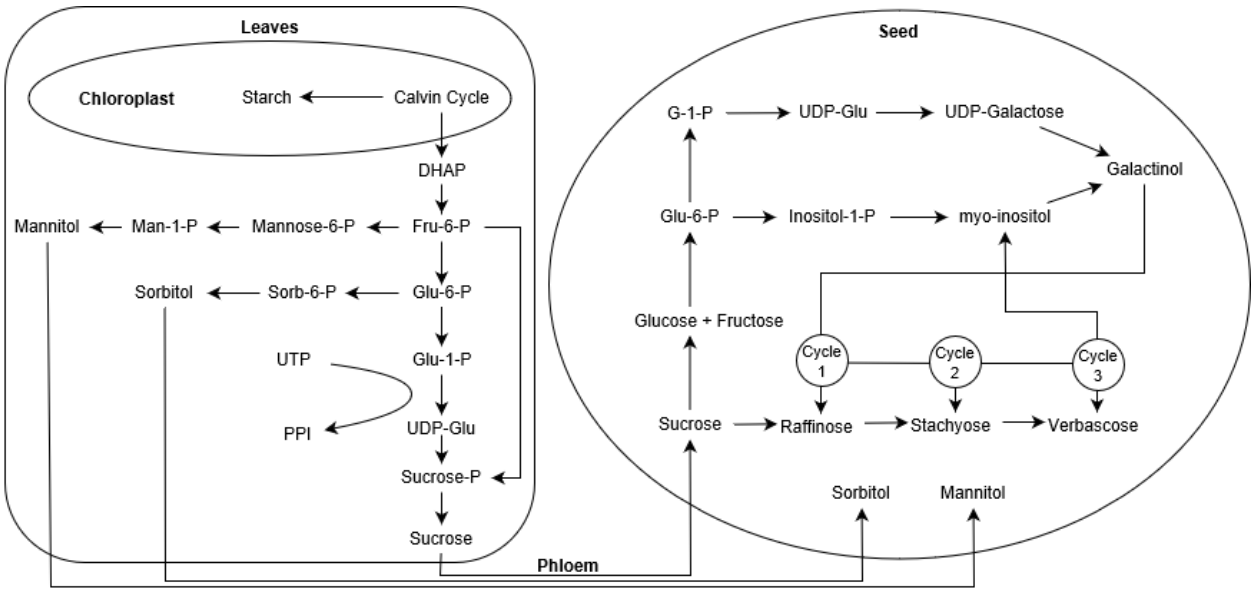

Figure 1.3: Biosynthetic pathway of raffinose family oligosaccharides and sugar alcohols from leaves to seed. Figure created from Gangola and Ramadoss (2018), Loescher and Everard (2000), and Dumschott, Richter, Loescher, and Merchant (2017)

References

Abeysekara, S., Chilibeck, P. D., Vatanparast, H., & Zello, G. A. (2012). A pulse-based diet is effective for reducing total and LDL-cholesterol in older adults. *British Journal of Nutrition*, 108, S103– S110. https://doi.org/10.1017/S0007114512000748

Amede, T., Schubert, S., & Stahr, K. (2011). Mechanisms of drought resistance in grain legumes I: Osmotic adjustment. *SINET: Ethiopian Journal of Science*, 26(1), 37– 46. https://doi.org/10.4314/sinet.v26i1.18198

Arpaia, N., Campbell, C., Fan, X., Dikiy, S., van der Veeken, J., DeRoos, P., … Rudensky, A. Y. (2013). Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature*, 504, 451– 455. https://doi.org/10.1038/nature12726

Bachmann, M., Matile, P., & Keller, F. (1994). Metabolism of the raffinose family oligosaccharides in leaves of *Ajuga reptans* L. (cold acclimation, translocation, and sink to source transition: Discovery of chain elongation enzyme). *Plant Physiology*, 105, 1335– 1345.

Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., & Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science*, 307, 1915– 1920. https://doi.org/10.1126/science.1104816

Bartels, D., & Sunkar, R. (2005). Drought and salt tolerance in plants. *Critical Reviews in Plant Sciences*, 24, 23– 58. https://doi.org/10.1080/07352680590910410

Becerra-Tomás, N., Díaz-López, A., Rosique-Esteban, N., Ros, E., Buil-Cosiales, P., Corella, D., … Lamuela-Raventós, R. M. (2018). Legume consumption is inversely associated with type 2 diabetes incidence in adults: A prospective assessment from the PREDIMED study. *Clinical Nutrition*, 37, 906– 913. https://doi.org/10.1016/j.clnu.2017.03.015

Beserra, B. T. S., Fernandes, R., Rosario, V. A., Mocellin, M. C., Kuntz, M. G. F., & Trindade, E. B. S. M. (2015). A systematic review and meta-analysis of the prebiotics and synbiotics effects on glycaemia, insulin concentrations and lipid parameters in adult patients with overweight or obesity. *Clinical Nutrition*, 34, 845– 858. https://doi.org/10.1016/j.clnu.2014.10.004

Bhatty, R. S. (1988). Composition and quality of lentil (*Lens culinaris* Medik): A review. *Canadian Institute of Food Science and Technology Journal*, 21, 144– 160. https://doi.org/10.1016/S0315-5463(88)70770-1

Bindels, L. B., Delzenne, N. M., Cani, P. D., & Walter, J. (2015). Opinion: Towards a more comprehensive concept for prebiotics. *Nature Reviews Gastroenterology and Hepatology*, 12, 303– 310. https://doi.org/10.1038/nrgastro.2015.47

Bohnert, H. J., & Jensen, R. G. (1996). Strategies for engineering water-stress tolerance in plants. *Trends in Biotechnology*, 14, 89– 97. https://doi.org/10.1016/0167-7799(96)80929-2

Chiang, Y.-J., Stushnoff, C., McSay, A. E., Jones, M. L., & Bohnert, H. J. (2005). Overexpression of mannitol-1-phosphate dehydrogenase increases mannitol accumulation and adds protection against chilling injury in petunia. *Journal of the American Society for Horticultural Science*, 130, 605– 610. https://doi.org/10.21273/JASHS.130.4.605

Cubero, J. I., Perez de la Vega, M., & Fratini, R. (2009). Origin, phylogeny, domestication and spread. In W. Erskine, F. J. Muehlbauer, A. Sarker, & B. Sharma (Eds.), The lentil: Botany, production and uses (pp. 13– 33). Wallingford, UK: CABI.

De Almeida Costa, G. E., Da Silva, Q.-M., Pissini Machado Reis, S. M., & De Oliveira, A. C. (2006). Chemical composition, dietary fibre and resistant starch contents of raw and

cooked pea, common bean, chickpea and lentil legumes. *Food Chemistry*, 94, 327– 330. https://doi.org/10.1016/j.foodchem.2004.11.020

Dumschott, K., Richter, A., Loescher, W., & Merchant, A. (2017). Post photosynthetic carbon partitioning to sugar alcohols and consequences for plant growth. *Phytochemistry*, 144, 243– 252. https://doi.org/10.1016/j.phytochem.2017.09.019

FAOSTAT. (2017). Food and Agriculture Organization of the United Nations Statistics Division Portal. Retrieved from http://www.fao.org/faostat/en/#home. Accessed 13 September 2019.

Frias, J., Vidal-Valverde, C., Bakhsh, A., Arthur, A. E., & Hedley, C. (1994). An assessment of variation for nutritional and non-nutritional carbohydrates in lentil seeds (*Lens culinaris*). *Plant Breeding*, 113, 170– 173. https://doi.org/10.1111/j.1439-0523.1994.tb00719.x

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T. A., Nakato, G., Takahashi, D., … Ohno, H. (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature*, 504, 446– 450. https://doi.org/10.1038/nature12721

Gangola, M. P., & Ramadoss, B. R. (2018). Sugars play a critical role in abiotic stress tolerance in plants. In S. H. Wani (Ed.), Biochemical, physiological and molecular avenues for combating abiotic stress tolerance in plants (pp. 17– 38). Academic Press.

Gao, Z., Yin, J., Zhang, J., Ward, R. E., Martin, R. J., Lefevre, M., … Ye, J. (2009). Butyrate improves insulin sensitivity and increases energy expenditure in mice. *Diabetes*, 58, 1509– 1517. https://doi.org/10.2337/db08-1637

Gehrig, J. L., Venkatesh, S., Chang, H.-W., Hibberd, M. C., Kung, V. L., Cheng, J., … Gordon, J. I. (2019). Effects of microbiota-directed foods in gnotobiotic animals and

undernourished children. *Science*, 365, eaau4732. https://doi.org/10.1126/science.aau4732

Gibson, G. R., Hutkins, R., Sanders, M. E., Prescott, S. L., Reimer, R. A., Salminen, S. J., … Reid, G. (2017). Expert consensus document: The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics. *Nature Reviews Gastroenterology and Hepatology*, 14, 491– 502. https://doi.org/10.1038/nrgastro.2017.75

Gibson, G. R., Probert, H. M., Loo, J. V., Rastall, R. A., & Roberfroid, M. B. (2004). Dietary modulation of the human colonic microbiota: Updating the concept of prebiotics. *Nutrition Research Reviews*, 17, 259– 275. https://doi.org/10.1079/NRR200479

Gibson, G. R., & Roberfroid, M. B. (1995). Dietary modulation of the human colonic microbiota: Introducing the concept of prebiotics. *The Journal of Nutrition*, 125, 1401–1412. https://doi.org/10.1093/jn/125.6.1401

Gibson, G. R., Scott, K. P., Rastall, R. A., Tuohy, K. M., Hotchkiss, A., Dubert-Ferrandon, A., … Buddington, R. (2010). Dietary prebiotics: Current status and new definition. *Food Science & Technology Bulletin: Functional Foods*, 7, 1– 19. https://doi.org/10.1616/1476-2137.15880

Grant, C. R., & ap Rees T. (1981). Sorbitol metabolism by apple seedlings. *Phytochemistry*, 20, 1505– 1511. https://doi.org/10.1016/S0031-9422(00)98521-2

Henry, C. J. K., Lightowler, H. J., Strik, C. M., Renton, H., & Hails, S. (2005). Glycaemic index and glycaemic load values of commercially available products in the UK. *British Journal of Nutrition*, 94, 922– 930. https://doi.org/10.1079/BJN20051594

Holzapfel, W. H., & Schillinger, U. (2002). Introduction to prebiotics and probiotics. *Food Research International*, 35, 109– 116.

Johnson, C. R., Thavarajah, D., Combs, G. F., & Thavarajah, P. (2013). Lentil (*Lens culinaris* L.): A prebiotic-rich whole food legume. *Food Research International*, 51, 107– 113. https://doi.org/10.1016/j.foodres.2012.11.025

Johnson, C. R., Thavarajah, D., Thavarajah, P., Fenlason, A., McGee, R., Kumar, S., & Combs, G. F. (2015). A global survey of low-molecular weight carbohydrates in lentils. *Journal of Food Composition and Analysis*, 44, 178– 185. https://doi.org/10.1016/j.jfca.2015.08.005

Johnson, C. R., Thavarajah, D., Thavarajah, P., Payne, S., Moore, J., & Ohm, J. B. (2015). Processing, cooking, and cooling affect prebiotic concentrations in lentil (*Lens culinaris* Medikus). *Journal of Food Composition and Analysis*, 38, 106– 111. https://doi.org/10.1016/j.jfca.2014.10.008

Jovanovic-Malinovska, R., Kuzmanova, S., & Winkelhausen, E. (2014). Oligosaccharide profile in fruits and vegetables as sources of prebiotics and functional foods. *International Journal of Food Properties*, 17, 949– 965. https://doi.org/10.1080/10942912.2012.680221

Kannan, U., Sharma, R., Gangola, M. P., Sari, N., & Chibbar, R. N. (2018). Improving grain quality in pulses: Strategies to reduce raffinose family oligosaccharides in seeds. *Ekin Journal of Crop Breeding and Genetics*, 4, 70– 88.

Kumar, J., Sen, G. D., Kumar, S., Gupta, S., & Singh, N. P. (2016). Current knowledge on genetic biofortification in lentil. *Journal of Agricultural and Food Chemistry*, 64, 6383– 6396. https://doi.org/10.1021/acs.jafc.6b02171

Kumar, S., Rajendran, K., Kumar, J., Hamwieh, A., & Baum, M. (2015). Current knowledge in lentil genomics and its application for crop improvement. *Frontiers in Plant Science*, 6, 1– 13. https://doi.org/10.3389/fpls.2015.00078

Li, K., Hu, Z., Ou, J., & Xia, K. (2019). Altered gut microbiome in autism spectrum disorder: Potential mechanism and implications for clinical intervention. *Global Clinical and Translational Research*, 1, 45– 52. https://doi.org/10.36316/gcatr.01.0006

Loescher, W., & Everard, J. (2000). Regulation of sugar alcohol biosynthesis. *Photosynthesis: Physiology and Metabolism*, 9, 275– 299. https://doi.org/10.1007/0-306-48137-5_12

Loescher, W. H., Tyson, R. H., Everard, J. D., Redgwell, R. J., & Bieleski, R. L. (1992). Mannitol synthesis in higher plants. *Plant Physiology*, 98, 1396– 1402. https://doi.org/10.1104/pp.98.4.1396

Loo, J. V., Cummings, J., Delzenne, N., Englyst, H., Franck, A., Hopkins, M., … van den Heuvel, E. (1999). Functional food properties of non-digestible oligosaccharides: A consensus report from the ENDO project (DGXII AIRII-CT94-1095). *British Journal of Nutrition*, 81, 121– 132. https://doi.org/10.1017/S0007114599000252

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, 489, 220– 230. https://doi.org/10.1038/nature11550

Lynch, S. V., & Pedersen, O. (2016). The human intestinal microbiome in health and disease. *New England Journal of Medicine*, 375, 2369– 2379. https://doi.org/10.1056/NEJMra1600266

McCrory, M. A., Hamaker, B. R., Lovejoy, J. C., & Eichelsdoerfer, P. E. (2010). Pulse consumption, satiety, and weight management. *Advances in Nutrition*, 1, 17– 30. https://doi.org/10.3945/an.110.1006

Muehlbauer, F. J., Cho, S., Sarker, A., McPhee, K. E., Coyne, C. J., Rajesh, P. N., & Ford, R. (2006). Application of biotechnology in breeding lentil for resistance to biotic and abiotic stress. *Euphytica*, 147, 149– 165. https://doi.org/10.1007/s10681-006-7108-0

Ng, Q. X., Loke, W., Venkatanarayanan, N., Lim, D. Y., Sen, S. A. Y., & Yeo, W. S. (2019). A systematic review of the role of prebiotics and probiotics in autism spectrum disorders. *Medicina*, 55, 1– 10. https://doi.org/10.3390/medicina55050129

Nishizawa, A., Yabuta, Y., & Shigeoka, S. (2008). Galactinol and raffinose constitute a novel function to protect plants from oxidative damage. *Plant Physiology*, 147, 1251– 1263. https://doi.org/10.1104/pp.108.122465

Obendorf, R. L., & Gorecki, R. J. (2012). Soluble carbohydrates in legume seeds. *Seed Science Research*, 22, 219– 242. https://doi.org/10.1017/S0960258512000104

Pennycooke, J. C., Jones, M. L., & Stushnoff, C. (2003). Down-regulating α-galactosidase enhances freezing tolerance in transgenic petunia. *Plant Physiology*, 133, 901– 909. https://doi.org/10.1104/pp.103.024554

Peterbauer, T., & Richter, A. (2001). Biochemistry and physiology of raffinose family oligosaccharides and galactosyl cyclitols in seeds. *Seed Science Research*, 11, 185– 197. https://doi.org/10.1079/SSR200175

Pineiro, M., Asp, N.-G., Reid, G., Macfarlane, S., Morelli, L., Brunser, O., & Tuohy, K. (2008). FAO technical meeting on prebiotics. *Journal of Clinical Gastroenterology*, 42, S156– S159. https://doi.org/10.1097/MCG.0b013e31817f184e

Pujni, D., Chaudhary, A., & Rajam, M. V. (2007). Increased tolerance to salinity and drought in transgenic indica rice by mannitol accumulation. *Journal of Plant Biochemistry and Biotechnology*, 16, 1– 7. https://doi.org/10.1007/BF03321921

Rizvi, A. H., Aski, M., Sarker, A., Dikshit, H. K., & Yadav, P. (2019). Origin, distribution, and gene pools. In M. Singh (Ed.), Lentils (pp. 7– 19). Academic Press.

Roberfroid, M. (2007). Prebiotics: The concept revisited. *The Journal of Nutrition*, 137, 830S– 837S. https://doi.org/10.1093/jn/137.3.830S

Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology*, 31, 107– 133. https://doi.org/10.1146/annurev.mi.31.100177.000543

Sengupta, S., & Majumder, A. L. (2015). Significance of galactinol and raffinose family oligosaccharide synthesis in plants. *Frontiers in Plant Science*, 6, 1– 11. https://doi.org/10.3389/fpls.2015.00656

Siva, N., Johnson, C. R., Richard, V., Jesch, E. D., Whiteside, W., Abood, A. A., … Thavarajah, D. (2018). Lentil (*Lens culinaris* Medikus) diet affects the gut microbiome and obesity markers in rat. *Journal of Agricultural and Food Chemistry*, 66, 8805– 8813. https://doi.org/10.1021/acs.jafc.8b03254

Siva, N., Thavarajah, P., Kumar, S., & Thavarajah, D. (2019). Variability in prebiotic carbohydrates in different market classes of chickpea, common bean, and lentil collected from the American local market. *Frontiers in Nutrition*, 6, 1– 11. https://doi.org/10.3389/fnut.2019.00038

Siva, N., Thavarajah, P., & Thavarajah, D. (2018). The impact of processing and cooking on prebiotic carbohydrates in lentil. *Journal of Food Composition and Analysis*, 70, 72– 77. https://doi.org/10.1016/j.jfca.2018.04.006

Smirnoff, N., & Cumbes, Q. J. (1989). Hydroxyl radical scavenging activity of compatible solutes. *Phytochemistry*, 28, 1057– 1060. https://doi.org/10.1016/0031-9422(89)80182-7

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M. A., … Gordon, J. I. (2014). Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature*, 510, 417– 421. https://doi.org/10.1038/nature13421

Tahir, M., Vandenberg, A., & Chibbar, R. N. (2011). Influence of environment on seed soluble carbohydrates in selected lentil cultivars. *Journal of Food Composition and Analysis*, 24, 596– 602. https://doi.org/10.1016/j.jfca.2010.04.007

Thavarajah, D., Thavarajah, P., Wejesuriya, A., Rutzke, M., Glahn, R. P., Combs, G. F., & Vandenberg, A. (2011). The potential of lentil (*Lens culinaris* L.) as a whole food for increased selenium, iron, and zinc intake: Preliminary results from a 3 year study. *Euphytica*, 180, 123– 128. https://doi.org/10.1007/s10681-011-0365-6

USDA. (2018). National nutrient database for standard reference release legacy reports 16069, 16056, 16108, 20444, & 20649. Retrieved from https://ndb.nal.usda.gov/ndb/search/list. Accessed 12 February 2019.

Valluru, R., & Van den Ende, W. (2011). Myo-inositol and beyond – Emerging networks under stress. *Plant Science*, 181, 387– 400. https://doi.org/10.1016/j.plantsci.2011.07.009

Varshney, R. K., Mohan, S. M., Gaur, P. M., Gangarao, N., Pandey, M. K., Bohra, A., … Gowda, C. (2013). Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnology Advances*, 31, 1120– 1134. https://doi.org/10.1016/j.biotechadv.2013.01.001

Williamson, J. J. D., Jennings, D. B., Guo, W.-W., Pharr, D. M., & Ehrenshaft, M. (2002). Sugar alcohols, salt stress, and fungal resistance: Polyols—Multifunctional plant protection?

Journal of the American Society for Horticultural Science, 127, 467– 473. https://doi.org/10.21273/jashs.127.4.467

Wong, M. M. L., Gujaria-Verma, N., Ramsay, L., Yuan, H. Y., Caron, C., Diapari, M., … Bett, K. E. (2015). Classification and characterization of species within the genus lens using genotyping-by-sequencing (GBS). *PLoS ONE*, 10, 1– 16. https://doi.org/10.1371/journal.pone.0122025

Zhifang, G., & Loescher, W. H. (2003). Expression of a celery mannose 6-phosphate reductase in Arabidopsis thaliana enhances salt tolerance and induces biosynthesis of both mannitol and a glucosyl-mannitol dimer. *Plant, Cell and Environment*, 26, 275– 283. https://doi.org/10.1046/j.1365-3040.2003.00958.x

CHAPTER TWO

GENOME-WIDE ASSOCIATION MAPPING OF LENTIL (*LENS CULINARIS* MEDIKUS) PREBIOTIC CARBOHYDRATES TOWARD IMPROVED HUMAN HEALTH AND CROP STRESS TOLERANCE

Abstract

Lentil, a cool-season food legume, is rich in protein and micronutrients with a range of prebiotic carbohydrates, such as raffinose-family oligosaccharides (RFOs), fructooligosaccharides (FOSs), sugar alcohols (SAs), and resistant starch (RS), which contribute to lentil's health benefits. Beneficial microorganisms ferment prebiotic carbohydrates in the colon, which impart health benefits to the consumer. In addition, these carbohydrates are vital to lentil plant health associated with carbon transport, storage, and abiotic stress tolerance. Thus, lentil prebiotic carbohydrates are a potential nutritional breeding target for increasing crop resilience to climate change with increased global nutritional security. This study phenotyped a total of 143 accessions for prebiotic carbohydrates. A genome-wide association study (GWAS) was then performed to identify associated variants and neighboring candidate genes. All carbohydrates analyzed had broad-sense heritability estimates (H^2) ranging from 0.22 to 0.44, comparable to those reported in the literature. Concentration ranges corresponded to percent recommended daily allowances of 2–9% SAs, 7–31% RFOs, 51–111% RS, and 57–116% total prebiotic carbohydrates. Significant SNPs and associated genes were identified for numerous traits, including a galactosyltransferase (*Lcu.2RBY.1g019390*) known to aid in RFO synthesis. Further studies in multiple field locations are necessary. Yet, these findings suggest the potential for molecular-assisted breeding for prebiotic carbohydrates in lentil to support human health and crop resilience to increase global food security.

Introduction

The World Health Organization estimates that non-communicable diseases (NCDs), such as cardiovascular disease and diabetes, cause 71% of global deaths[1]. The United Nations Sustainable Goals by 2030 include the reduction of NCD mortality by one-third as a primary health goal[1]. NCD risk factors are diverse; however, some, such as obesity, overweight, and malnutrition, clearly have a dietary link. Consequently, food security and consumer acceptance of nutritious foods are vital to lowering NCD risk. Compounding the problem is the threat of climate change to global food security[2]. Anticipated increases in temperature and drought will have harmful effects on crop yields and the people dependent upon them. Thus, ensuring the production of nutritionally dense staple food crops, such as pulses, is essential to address these global food security challenges. Amid the complexity of these issues, we put forward lentil (*Lens culinaris* Medikus), a staple food crop rich in prebiotic carbohydrates, as one piece in the broader solution. Lentil prebiotic carbohydrates are an ideal target for genomic-assisted breeding approaches to combat NCD and ensure global food security.

Lentil is a nutritionally dense cool-season pulse crop with notable concentrations of protein (20–30%), low-digestible carbohydrates (20%), fat (1%), iron (Fe), zinc (Zn), and a range of vitamins[3]. A study in rats shows a lentil diet can significantly lower mean body weight, percent body fat, and blood plasma triglyceride levels and increase lean body mass than control or corn diet[4]. Lentil's health benefits are in part due to its high concentrations of prebiotic or low-digestible carbohydrates, including raffinose-family oligosaccharides (RFOs; 4071 mg/100 g), sugar alcohols (SAs; 1423 mg/100 g), fructooligosaccharides (FOSs; 62 mg/100 g), and resistant starch (RS; 7500 mg/100 g)[5]. A prebiotic is *"a substrate that is selectively utilized by host microorganisms conferring a health benefit"*[6]. When consumed, prebiotics pass through the

upper gastrointestinal tract and are fermented by beneficial microorganisms in the colon, which benefits their human host. The human gastrointestinal tract is lined with trillions of microorganisms, composing the microbiome[7]. These microbes are vital to colon health and function, aiding in immune system stimulation, nutrient breakdown and absorption, and bowel motility[8]. Adverse microbiome compositions have been associated with various ailments, such as obesity, diabetes, infection, and colon cancer[9]. Modulation of the microbiome, primarily through prebiotic consumption, can improve health outcomes. For example, a prebiotic-rich diet restored the microbiome composition and plasma biomarkers of malnourished Bangladeshi children to levels similar to healthy children[10].

Lentil prebiotic carbohydrates also serve a vital role in plant health. Lentil accumulates RFOs in its seeds at high concentrations. Although few studies have been done on lentil RFOs, soybean seedlings have been shown to use this carbon store for energy; however, RFOs do not appear necessary for successful germination[11]. Abiotic stress studies in *Arabidopsis thaliana* show upregulation of RFOs under drought, salinity, cold, and heat stress[12,13]. Further, a transgenic *A. thaliana* line overexpressing three genes essential in RFO synthesis demonstrated increased drought, salinity, and cold tolerance[12]. Similar results are reported for SAs[14]. These carbohydrates function as osmoregulants, cell signals, free radical scavengers, and compatible solutes for enzyme function[15].

As a staple food crop, lentil may be ideal for marker-assisted breeding efforts to alter prebiotic carbohydrate concentrations to reduce NCDs and advance global food security, now threatened by climate change. However, traditional breeding techniques are particularly challenging for quantitative nutritional traits in mature seeds. Analysis by high-performance anion-exchange chromatography is time-consuming and expensive; therefore, molecular

techniques have been explored as a way to significantly accelerate the breeding process[16,17]. Genome-wide association studies (GWAS) can detect quantitative trait loci (QTL) associated with prebiotic carbohydrate concentrations and help identify genetic markers needed for molecular breeding techniques. Lentil is a diploid (2n = 14) with a large ~ 4 Gb genome[18]. This allows for the use of numerous tools developed for diploid crops and simplifies some analysis. However, the large repetitive genome poses some additional challenges, such as generating a reference genome (yet unpublished) and sequencing new lines. One of the advantages of using genotyping-by-sequencing (GBS) methods is eliminating some of this complexity by reducing repetitive DNA sequencing[19]. GWAS using genotyping-by-sequencing (GBS) data have identified markers for *Aphanomyces* root rot resistance[20] and abiotic stress tolerance[21] in lentil. However, this is the first comprehensive study to report GWAS findings for prebiotic carbohydrates in lentil. Two lentil mapping populations were obtained from the International Centre for Agricultural Research in the Dry Areas (ICARDA), Rabat Institute, Rabat, Morocco. The heat tolerance population (150 accessions) and the global mapping population (128 accessions) were grown in a completely randomized design with two replicates at the Clemson University Greenhouse Complex, Clemson, SC, USA. The objectives of this study were to (1) identify and quantify prebiotic carbohydrates in a lentil association mapping population grown under greenhouse conditions, (2) identify SNP markers and candidate genes for lentil prebiotic carbohydrates through GWAS, and (3) identify lentil prebiotic carbohydrate breeding targets for human nutrition and climate resilience.

Results

Population Composition

The two lentil mapping populations were combined for statistical analysis, and an additional 14 lines were added for which data was available. Due to population overlap and poor grain yields, the total number of unique accessions with low-molecular-weight carbohydrate data was 143 with 1–5 replicates per accession. The lentil population included 60 from Asia, 40 from Europe, 16 from Africa, 13 from North America, eight from ICARDA, and six from South America (Table 2.1).

Prebiotic Carbohydrates

Low-molecular-weight carbohydrate analysis was conducted on 143 accessions with 1–5 replicates (Table 2.2). Starch data were only collected from the heat tolerance population and included 102 accessions with 1–2 replicates (Table 2.2). Mean carbohydrate concentrations (used in the GWAS) were approximately normally distributed, as indicated by the normal red curves fitted to the concentration histograms (Fig. 2.1). For SAs, sorbitol (sor) had a mean concentration of approximately 4.5 times that of mannitol (man), at 206.8 and 46.8 mg/100 g, respectively. Simple sugars glucose (glu), fructose (fru), and sucrose (suc) had mean concentrations of 93, 69, and 496 mg/100 g, respectively. RFOs stachyose + raffinose (sta + raf) and verbascose + kestose (ver + kes) had mean concentrations of 578 and 318 mg/100 g, respectively (Table 2.2). Sta + raf and suc had the highest concentrations of all low-molecular-weight carbohydrates measured. Polysaccharides included RS, non-resistant starch (NRS), and total starch (TS) and had mean concentrations of 16.4, 39.6, and 56.0 g/100 g, respectively. All carbohydrates analyzed had modest broad-sense heritability estimates (H^2) ranging from 0.22 (TS) to 0.45 (man).

Concentration ranges corresponding to 2–9%, 7–31%, 51–111%, and 57–116% of the recommended daily allowance (RDA) for SAs, RFOs, RS, and total prebiotic carbohydrates, respectively.

Significant differences in carbohydrate concentrations by continent of origin were evident for sor, suc, ver + kes, NRS, and TS (Fig. 2.2). SA concentrations were highest in accessions from South America (sor) and North America (man) and lowest in the ICARDA accessions. Simple sugar concentrations were highest in accessions from Europe (glu, fru) and North America (suc) and lowest in accessions from Africa (glu, fru) and ICARDA (suc). RFO concentrations were highest in accessions from Europe (sta + raf) and North America (ver + kes) and lowest in accessions from ICARDA. Finally, starch concentrations were highest in accessions from Africa (RS) and ICARDA (NRS, TS) and lowest in accessions from South America (RS) and North America (NRS, TS).

Significant single nucleotide polymorphisms (SNPs) were identified for fru, sta + raf, RS, and TS (Fig. 2.3, Table 2.3). Significant SNPs tended not to be in linkage disequilibrium with adjacent SNPs, likely due to the low coverage of GBS data and the large genome size. Three SNPs were significantly associated with man (chromosomes 2–4), with one (CHR2_558954064) identified by both software programs employed (GAPIT and GEMMA) and having a minor allele frequency (MAF) of 5.9%. One SNP was significantly associated with glu (chromosome 6). Ten SNPs were significantly associated with fru (chromosomes 1–5), two of which (CHR1_153779147, CHR5_316719059) were identified by both software programs with MAFs of 7.3 and 5.2%, respectively. One SNP was significantly associated with suc (chromosome 6) and was identified by both software programs with an MAF of 5.2%. Twenty-two SNPs were significantly associated with sta + raf (chromosomes 1, 4–7), with one (CHR6_371563912)

identified by both software programs with an MAF of 9.8%. Ten SNPs were significantly associated with RS (chromosomes 1–3, 6–7), and one was significantly associated with TS (chromosome 7). Linkage blocks containing significant SNPs largely exceeded 100 kb and contained genes too numerous to include here. Genes within 100 kb flanking regions can be accessed in Supplemental Information (https://doi.org/10.1038/s41598-021-93475-3).

Discussion

This study estimated the concentrations of 10 different carbohydrates in a lentil mapping population to understand underlying genetic mechanisms. To our knowledge, it is the first publication to identify associated SNPs and candidate genes for lentil prebiotic carbohydrates via GWAS. Furthermore, it stands as one of the few GWAS for lentils irrespective of the trait. The findings are essential for developing markers for molecular-assisted breeding approaches for nutritional and climate-change resilience breeding objectives in lentils. Prebiotic carbohydrates are important traits relevant both to human health and crop climate-change resilience. Specifically, a healthy gastrointestinal microbiome is sustained mainly by consuming prebiotic carbohydrates in the human diet, which promote the growth of beneficial microorganisms, such as *Lactobacilli* and *Bifidobacteria*[22]. A healthy microbiome has been associated with numerous health benefits, including increased mineral absorption and reduced risk of colon cancer, diabetes, irritable bowel disease, and others[9]. In addition, these carbohydrates play an essential role in increasing the plant's abiotic stress tolerance, being associated with tolerance to salinity, heat, cold, and freezing stresses[12,13,14,15].

Low-molecular-weight carbohydrate concentrations were generally consistent with values found in the literature for lentils; however, mean concentrations of sor, suc, sta + raf, and ver + kes were on the low end of normal[5,23,24]. Typical lentil SA concentration ranges are 1000–

2000 mg/100 g (sor) and 50–300 mg/100 g (man); values measured here are notably lower for sor (113–328 mg/100 g) and similar for man (2–357 mg/100 g). Typical simple sugar concentration ranges are 20–300 mg/100 g glu, 0.2–50 mg/100 g fru, and 1000–2500 mg/100 g suc; values measured here are similar for glu (36–315 mg/100 g), higher for fru (7–325 mg/100 g), and lower for suc (208–1010 mg/100 g). Typical RFO concentrations are 1500–5000 mg/100 g sta + raf and 500–2500 mg/100 g ver + kes; values measured here are both notably lower at 344–1748 mg/100 g sta + raf and 164–647 mg/100 g ver + kes. Total starch concentrations were consistent with the literature[5,23]; however, RS concentrations were higher than expected based on literature values, at 10–22 g/100 g compared to 5–10 g/100 g. This also corresponded to lower NRS values than expected. Overall, significant variation was evident within this population grown under greenhouse conditions. Larger variation in concentrations would be expected in field trials in addition to genotype × environment effects.

Heritability estimates showed cautious potential for breeding for these traits. Sugar alcohols' broad-sense heritability estimates are not commonly calculated in grain crops. Sorbitol heritability estimate in peach was reported as 0.7–0.8[25], which is higher than noted for lentil in the present study (0.34). Estimates for simple sugar and RFO heritability are consistent with other literature on pulse crops. H^2 values for glucose and sucrose (0.20 and 0.34) are compatible with other pulse crops, ranging from 0.2–0.4 and 0.2–0.5, respectively[26,27]. The H^2 value for fructose is high compared to 0.05–0.07 in chickpea[26]. The H^2 value for stachyose + raffinose of 0.41 is comparable to heritability of 0.2–0.5 in common bean and desi and Kabuli chickpea[26,27]. Resistant starch ($H^2 = 0.31$) is a novel phenotype for which heritability estimates are limited; however, total starch heritability of 0.3–0.4 has been reported in barley[28], which is slightly higher than the value of 0.22 for lentil in the present study. This study indicates low to medium

heritability estimates for lentil prebiotic carbohydrates, suggesting that the environment may play a more significant role than genotype in determining these concentrations; this may challenge breeding for these traits. However, this is the first study to measure heritability in these traits for lentils and was performed in a controlled greenhouse environment, so it is too early to make any definitive statements for or against breeding prospects. Field trials with multiple locations will be vital toward estimating heritability more accurately and determining genotype × environment effects. In addition, increasing the lentil population size to encompass broader genetic diversity will potentially increase heritability estimates.

Based on %RDA values, there is significant potential within the *Lens culinaris* species for selecting lentil lines of high or low prebiotic carbohydrate content. Our results also suggest the potential for incorporating prebiotic carbohydrates as a nutritional trait in breeding programs. From a dietary perspective, specific lentil accessions may be selected based on their prebiotic concentration, potentially providing up to 100% of the RDA. Human populations with obesity would benefit from varieties with increased prebiotic carbohydrate levels; these varieties may also increase climate resilience for global food security. For populations where specific prebiotics in lentil may cause undesirable side effects, including bloating, flatulence, indigestion, need lentil cultivars with lower total prebiotic concentrations, or particular carbohydrates could be targeted, such as RFOs, which are the carbohydrate family primarily implicated in indigestion[29]. Target concentrations may vary depending on the desired outcome and population; nevertheless, RS, which makes up most prebiotic content in lentils, may prioritize the most significant trait of interest. Whereas non-resistant starch is digested and absorbed in the upper digestive tract, RS is not broken down by digestive enzymes and consequently enters the colon, fermented by microorganisms[30].

Prebiotic carbohydrate concentrations vary by growing location[24]. The present study showed that some prebiotic carbohydrate concentrations also vary by continent of origin, although this difference is not significant in most cases. This result can be interpreted with contrasting ramifications. In the cases where little difference is detected (man, sta + raf, and RS), this may suggest that the trait is highly conserved. If so, the lentil plant must tightly regulate these concentrations to produce viable seed; manipulating these concentrations through breeding would then be challenging and, if successful, may have a detrimental effect on the plant and agronomic traits, including yield.

In contrast, where concentrations differ by continent of origin (sor, ver + kes) may suggest that prebiotic carbohydrate concentrations have been under selective pressure in the lentil's evolutionary development[31]. During lentil's introduction to new regions, differences in climate would have been a prominent source of pressure driving variation alongside historical agronomic breeding. If prebiotic carbohydrate concentrations played a role in these historical adaptations, exploring their potential in developing varieties resilient under various environmental conditions is warranted. Namely, the warmer, dryer climates feared to result from climate change. More studies, including a larger population and multiple field trials, are needed to support these hypotheses with heritability.

GWAS has been successfully used in other crops to identify significant SNPs and candidate genes for simple sugars and RFOs[32,33]. Few GWAS on lentil have been reported in the literature, likely due to the lack of genetic resources. The development of genetic resources for lentil and other legumes has lagged behind other crops, such as maize and sorghum. For example, the lentil genome remains unpublished, in part due to its size and repetitive nature. In addition, the quality of the genome available through the University of Saskatchewan was

relatively poor until the recent release of version 2.0, which incorporated multiple sequencing platforms as well as long and short reads (presentation and communication with Kirsten Bett of University of Saskatchewan at North American Pulse Improvement Association, Fargo, ND, Nov 6–8, 2019).

This GWAS on lentil prebiotic carbohydrates uncovered several significantly associated SNPs. SNP markers were identified for the prebiotic carbohydrates man, sta + raf, RS, and the non-prebiotic carbohydrates glu, fru, suc, and TS. Due to the ubiquity of SNPs in the genome, they are convenient markers for GWAS. Though a significant SNP is often not the causative mutation, it may be in linkage with the causative mutation. Genes within 100 kb of each significant SNP are shown in Supplemental Information (https://doi.org/10.1038/s41598-021-93475-3). A number of significant SNPs were identified within genes. For example, CHR1_143888359 was located within Lcu.2RBY.1g019390, homologous to a galacturonosyltransferase in *Arabidopsis thaliana*. Generally, this gene class is known for the synthesis of pectin in cell walls[34]; however, the transfer of galactose is the primary step in RFO synthesis carried out by galactosyltransferases[35]. Thus, this discovery offers a potential gene target for altering RFO concentration in lentil.

Conclusion

Lentil prebiotic carbohydrates play a vital role in plant physiology and should be further explored as a means of breeding lentil varieties for changing climates. Additionally, prebiotic carbohydrates are important for human health, specifically for their role in regulating and modulating the gut microbiome. Thus, increased consumption of lentil and other pulse crops could have a beneficial effect on many people's health. Future studies should validate identified candidate genes to verify their function and uncover causative mutations. Once confirmed,

markers can be confidently developed for molecular-assisted breeding for prebiotic carbohydrates. Markers, such as microsatellites, could be used in molecular-assisted breeding approaches to incorporate the desired alleles and then recover the elite cultivar genotype through backcrossing aided by markers scattered across the genome[36].

Materials and Methods

Materials

Standards, chemicals, and high-purity solvents used for prebiotic carbohydrate analysis were purchased from Sigma Aldrich Co. (St. Louis, MO), Fisher Scientific (Waltham, MA), VWR International (Radnor, PA), and Tokyo Chemical Industry (Portland, OR) and used without further purification. Water, distilled, and deionized (ddH_2O) to the resistance of ≥ 18.2 $M\Omega \times cm$ (PURELAB flex 2 system, ELGA LabWater North America, Woodridge, IL) was used for sample and reagent preparation.

Greenhouse

Two lentil mapping populations were obtained from the International Centre for Agricultural Research in the Dry Areas (ICARDA), Rabat-Institute, Rabat, Morocco. The heat tolerance population (150 accessions) and the global mapping population (128 accessions) were grown in a completely randomized design with two replicates ($n = 558$) at the Clemson University Greenhouse Complex, Clemson, SC, USA (Table 2.1). The soil in each pot was saturated with ddH_2O and allowed to drain overnight. At seeding, pots were at 80% pot capacity. Greenhouse conditions were day and night temperatures of 22/20 °C. Photosynthetically active radiation levels were 300 $\mu mol/m^2/s$ using a 16-h photoperiod and 50–60% relative humidity. All

pots were watered to approximately 70% of free-draining moisture content every day, and 250 mL of the nutrient solution were added to all pots every 2 weeks, as per standard procedures for lentils at the Clemson University Pulse Quality and Nutrition program. Nutrient concentrations of the all-purpose 20-20-20 fertilizer solution (Plant Products Co. Ltd., Brampton, ON, Canada) were 20% total N, 20% total P, 20% soluble K, 0.02% B, 0.05% chelated Cu, 0.1% chelated Fe, 0.05% Mo, 0.05% Zn, and 1% EDTA. All plants were hand-harvested at physiological maturity, air-dried (40 °C), and hand-threshed. The total seed weight per pot was recorded, and the seeds were stored at − 40 °C until analysis.

Low Molecular Weight Carbohydrates Prebiotic Carbohydrate Analysis

Lentil seeds were ground (Blade Coffee Grinder, KitchenAid, St. Joseph, MI, USA) and sieved to 0.5-mm particle size. Carbohydrates were extracted following Muir et al.[37] with modification. Each flour sample was weighed (150 mg) into a centrifugal polypropylene tube (VWR International, Radnor, PA, USA). After adding 10 mL of water, each tube was mixed on a vortex mixer and placed in a water bath for 1 h at 80 ℃. Tubes were then centrifuged at 3000*g* for 10 min. The supernatant was filtered through a 13 mm × 0.45 μm nylon syringe filter (Thermo Fisher Scientific, MA, USA) into an HPLC vial for analysis.

Low molecular weight carbohydrate analysis was performed following Feinberg et al.[38] on a Dionex ICS-5000^{+} system (Thermo Scientific, Waltham, MA, USA) equipped with a pulsed amperometric detector (PAD) with a working gold electrode and a silver-silver chloride reference electrode. The separation was achieved using a Dionex CarboPac PA1 analytical column (250 × 4 mm) in series with a Dionex CarboPac PA1 guard column (50 × 4 mm). Pure standards were used to identify peaks, generate calibration curves, and monitor detector sensitivity. A lentil lab reference sample was used to monitor extraction consistency.

Concentrations were quantified within a linear range of 0.1–500 ppm with a minimum detection limit of 0.1 ppm. Concentrations in samples were calculated following $X = (C \times V)/m$, where X is the moisture-corrected analyte concentration in the sample, C is the concentration in the filtrate, V is the sample volume, and *m* is the mass of the dry lentil flour.

Starch Analysis

Resistant, non-resistant, and total starch were measured using the AOAC approved Megazyme resistant starch assay method[39]. Each sample was weighed (100 mg) into a centrifugal polypropylene tube. Enzyme solution was added (2 mL), which contained amyloglucosidase (3 U/mL) and α-amylase (10 mg/mL) in sodium maleate buffer (100 mM, pH 6.0). Tubes were incubated with constant circular shaking (200 strokes/min) for 16 h at 37 °C. Ethanol (4 mL; 99%) was added, followed by vortex mixing centrifugation (1500*g* for 10 min) and decanting into 100-mL volumetric flasks. Two additional washings of the sample were performed, adding 2 mL of ethanol (50%) and vortex mixing to suspend the pellet, followed by an additional 6 mL of ethanol (50%), vortex mixing, centrifugation, and decanting. Pooled non-resistant starch washings were brought to 100 mL volume with water. Pellets containing resistant starch were dissolved in 2 mL of 2 M KOH with a magnetic stir bar for 20 min in an ice water bath. Sodium acetate buffer (8 mL, 1.2 M, pH 3.8) was added, followed immediately by 0.1 mL of amyloglucosidase (AMG; 3300 U/mL). Samples were incubated at 50 °C in a water bath for 30 min. Tubes were then centrifuged (1500*g* for 10 min). Resistant starch and non-resistant starch fractions were quantified via spectrophotometry as follows. Starch solution (0.1 mL) and glucose oxidase/peroxidase (GOPOD) reagent (3 mL) were added to a glass tube and incubated for 20 min at 50 °C. A glucose standard (1 mg/mL in 0.2% benzoic acid) was included in each batch. Absorbance was measured at 510 nm against a reagent blank. Non-resistant starch was

calculated by the formula NRS (g/100 g sample) = ΔE × F/W × 90, where ΔE is the absorbance of the sample, F is the absorbance to microgram conversion factor (100/absorbance of glucose standard), W is the sample dry weight, and 90 includes adjustments for volume, unit conversions, and free to anhydrous glucose. Resistant starch was calculated by a similar formula: RS (g/100 g sample) = ΔE × F/W × 9.27, where 9.27 includes adjustments for volume, unit conversions, and free to anhydrous glucose. Total starch was calculated as TS = RS + NRS.

Statistical Analysis

Carbohydrate concentration means, standard errors, and ranges were averaged across replications for each accession. Carbohydrate distributions were displayed as histograms, and normal curves were fit to the histograms to determine how closely the values followed a normal distribution. To compare each carbohydrate concentration among a continent of origin, a statistical model was developed with the mean concentration of each carbohydrate as the response variable and continent as a fixed effect. The model was estimated using standard least squares. ANOVA was used to determine if the continent effect was significant. Fisher's Protected Least Significant Difference Test was used to compare mean concentrations by continent of origin for each carbohydrate. P-value < 0.05 was considered evidence of statistical significance. To estimate broad-sense heritability (H^2), a statistical model was developed with the mean concentration of each carbohydrate as the response variable and genotype as a random effect. The model was estimated using the restricted maximum likelihood (REML) method. H^2 was identified as the proportion of variance due to genotype. Percent recommended daily allowances (%RDA) were calculated for total SA, total RFO, and RS, and total prebiotic carbohydrate concentrations based on 7 g/day for sugar alcohols, 7 g/day for RFOs, 20 g/day for RS, and 20 g/day for total prebiotic content[40,41,42]. All calculations were performed using JMP 14.0.0.

Genome-Wide Association Studies

Previously sequenced genotyping-by-sequencing (GBS) data were used for genome-wide association analysis[21]. The TASSEL-GBS pipeline[43] with default parameters was used for aligning reads to the reference genome (Lens culinaris v2.0) and for single nucleotide polymorphism (SNP) calling. Beagle 5.0 with default settings was used for imputation[44]. VCFtools was used for filtering the VCF file to include only the 143 lentil lines included in the study (102 for starch) and to exclude sites with less than 5% minor allele frequency (MAF) and more than 20% missing data, leaving 22,222 high-quality SNPs for analysis[45]. Association analyses were conducted with two software programs and models: the Genome Association and Prediction Integrated Tool (GAPIT) in R using the FarmCPU model[46] and the Genome-wide Efficient Mixed Model Association Algorithm (GEMMA) using a linear mixed model for univariate analyses[47]. Least square means from the JMP analysis were used. The population structure was estimated with the VanRaden kinship matrix algorithm in GAPIT. PLINK[48] was used to calculate linkage disequilibrium decay around significant SNPs to determine linkage blocks and identify candidate genes from a GFF3 file.

Acknowledgements

Funding support for this project was provided by the Plant Health and Production and Plant Products: Plant Breeding for Agricultural Production program area [grant no. 2018-67014-27621/project accession no. 1015284], the Organic Agriculture Research and Extension Initiative (OREI) (award no. 2018-51300-28431/proposal no. 2018-02799), and the support of the American People provided to the Feed the Future Innovation Lab for Crop Improvement through the United States Agency for International Development (USAID) under Cooperative Agreement

No 7200AA19LE00005/Subaward no 89915-11295, and the International Center for Dry Land Agriculture (ICARDA, Morocco). The authors like to thank Drs. Kirstin Bett (University of Saskatchewan, Canada), Rebecca McGee (USDA-ARS, Washington State University, WA, USA), Jodi Humamm, and Dorrie Main (Washington State University, WA, USA) for giving access to the lentil reference genome and genotyping files. Finally, we are grateful to Dr. Stephan Kresovich for his guidance on obtaining funds and developing the project proposal with Dil Thavarajah.

Tables and Figures

Table 2.1: *Lens culinaris* ssp. *culinaris* population origin information.

Continent	Country	Accessions
Africa (16)	Algeria (2)	ILL858, ILL4781
	Egypt (2)	ILL702, ILL1046
	Ethiopia (4)	ILL1706, ILL1959, ILL5639, ILL5956
	Libya (1)	ILL4804
	Morocco (2)	ILL6493, ILL7727
	Sudan (2)	ILL1861, ILL5505
	Tunisia (3)	ILL918, ILL1890, ILL6272
Asia (60)	Afghanistan (2)	ILL213, ILL2217
	Armenia (2)	ILL86, ILL619
	Azerbaijan (1)	ILL1671
	Bangladesh (3)	ILL7774, ILL7789, ILL8007
	India (6)	ILL931, ILL3517, ILL4152, ILL4164, ILL4900, ILL5151
	Iran (8)	ILL223, ILL257, ILL769, ILL1013, ILL1097, ILL2406, ILL4791, ILL4886
	Iraq (1)	ILL4899
	Jordan (4)	ILL2150, ILL5261, ILL5384, ILL6925
	Lebanon (3)	ILL191, ILL840, ILL5626
	Nepal (4)	ILL3485, ILL3487, ILL7437, ILL8010
	Pakistan (3)	ILL2297, ILL7650, ILL8114
	Palestine (1)	ILL4606
	Russia (3)	ILL597, ILL4819, ILL4830
	Saudi Arabia (1)	ILL7745
	Syria (6)	ILL158, ILL490, ILL4471, ILL5509, ILL5595, ILL6644
	Tajikistan (2)	ILL598, ILL6126
	Turkey (7)	ILL71, ILL129, ILL550, ILL556, ILL635, ILL2181, ILL6149
	Uzbekistan (1)	ILL4875
	Yemen (2)	ILL950, ILL6281
Europe (40)	Albania (1)	ILL4841
	Belgium (1)	ILL224, ILL6185, ILL7495
	Croatia (1)	ILL4915
	Cyprus (2)	ILL890, ILL5968
	Czech Republic (1)	ILL4409
	France (1)	ILL6528
	Germany (2)	ILL4831, ILL4881
	Greece (4)	ILL304, ILL4857, ILL5519, ILL5533
	Hungary (1)	ILL719
	Italy (4)	ILL343, ILL5416, ILL5418, ILL7084
	North Macedonia (2)	ILL623, ILL624

	Norway (1)	ILL4782
	Poland (2)	ILL705, ILL5424
	Portugal (1)	ILL4956
	Romania (1)	ILL4774
	Serbia and Montenegro (1)	ILL1949
	Spain (4)	ILL4926, ILL5028, ILL5653, Pardina
	Ukraine (3)	ILL82, ILL595, ILL7090
	United Kingdom (3)	ILL348, ILL4345, ILL6415
	Yugoslavia (2)	ILL2230, ILL2231
ICARDA (8)	ICARDA (8)	ILL6994, ILL7012, ILL7978, ILL7979, ILL7981, ILL9888, ILL10053, ILL10281
North America (13)	Canada (4)	ILL4738, Eston, Richlea, Viceroy
	Guatemala (1)	ILL494
	Mexico (3)	ILL502, ILL5553, ILL5645
	United States (5)	ILL4671, Brewer, Crimson, Merrit, Redchief
South America (6)	Argentina (2)	ILL268, ILL4605
	Chile (2)	ILL956, ILL1005
	Columbia (1)	ILL1649
	Uruguay (1)	ILL4778

Note: Numbers in parentheses are accession counts per location.

Table 2.2: Carbohydrate analysis with the number of accessions (N), range, overall mean with standard error (SE), and heritability estimates (H^2)

Carbohydrate type	N	Range	Mean ± (SE)	Genotype	H^2	%RDA[a]
Sugar alcohols (mg/100 g)						
Sorbitol	143	113d–328	207 ± 3	***	0.34	
Mannitol	143	2–357	46 ± 3	***	0.45	
Total sugar alcohols	143	126–609	253 ± 5			2–9
Simple sugars (mg/100 g)						
Glucose	143	36–315	93 ± 3	***	0.20	
Fructose	143	7–325	69 ± 4	***	0.23	
Sucrose	143	208–1010	496 ± 9	***	0.34	
Total simple sugars	143	275–1326	658 ± 12			
Raffinose-family oligosaccharides (mg/100 g)						
Stachyose + Raffinose	143	344–1748	578 ± 12	***	0.41	
Verbascose + Kestose	143	164–647	318 ± 7	***	0.29	
Total RFOs	143	508–2167	896 ± 16			7–31
Starch polysaccharides (g/100 g)						
Resistant starch	102	10.1–22.1	16.4 ± 0.2	***	0.31	51–111
Non-resistant starch	102	27.1–48.3	39.6 ± 0.4	***	0.37	
Total starch polysaccharides	102	44.7–68.2	56.0 ± 0.4	**	0.22	
Total prebiotic carbohydrates	102	11.4–23.2	17.5 ± 0.2			57–116

[a] %RDA is based on a recommended daily allowance of 7 g/day for sugar alcohol[41], 7 g/day for raffinose-family oligosaccharides[42], and 20 g/day for resistant starch and total prebiotic carbohydrates[43]. A genotype is noted as significant at ** $P<0.05$ and *** $P<0.01$. H^2 broad-sense heritability estimate. Total prebiotic carbohydrates include resistant starch, raffinose-family oligosaccharides, and sugar alcohols. N: number of samples.

Table 2.3: Significant SNPs identified using GAPIT and GEMMA software.

Carbohydrate	**Significant SNP**	**p-value (GAPIT)**	**p-wald[a] (GEMMA)**	**MAF (%)**
Mannitol	CHR2_558954064	1.6E-06	7.4E-07	5.9
	CHR3_346516487	NS	1.3E-06	6.6
	CHR4_179223602*	NS	6.0E-08	8.0
Glucose	CHR6_290592280	NS	2.2E-06	14.7
Fructose	CHR1_153779145	NS	7.2E-07	7.3
	CHR1_153779147	1.3E-12	7.2E-07	7.3
	CHR1_537293922*	NS	7.4E-08	7.0
	CHR1_537449765	NS	7.4E-08	7.0
	CHR2_352993379	1.1E-06	NS	8.4
	CHR3_167167171	1.8E-06	NS	19.2
	CHR3_39693339	2.5E-07	NS	7.7
	CHR4_268011619*	8.1E-09	NS	16.8
	CHR4_316841184	1.6E-07	NS	5.2
	CHR5_316719059*	1.0E-18	4.3E-07	5.2
Sucrose	*CHR6_116880302*	1.8E-06	9.0E-07	5.2
Stachyose+Raffinose	CHR1_35234757*	1.7E-07	NS	9.8
	CHR1_46070961*	6.0E-08	NS	7.0
	CHR1_143888359*	NS	4.1E-08	5.6
	CHR4_32265499	2.5E-07	NS	8.4
	CHR4_87430341	4.2E-07	NS	6.6
	CHR5_235283678	1.5E-07	NS	12.6
	CHR6_116870957	NS	4.2E-08	5.2
	CHR6_116880302	NS	1.6E-07	5.2
	CHR6_117946950	NS	3.8E-07	5.6
	CHR6_121916516	NS	7.0E-07	6.3
	CHR6_126221589	NS	2.0E-06	6.3
	CHR6_126869085	NS	2.0E-06	6.3
	CHR6_128918912	NS	7.0E-07	6.3
	CHR6_128918961	NS	7.0E-07	6.3
	CHR6_130768080	NS	7.0E-07	6.3
	CHR6_137205602	NS	1.1E-06	5.9
	CHR6_137205644	NS	1.1E-06	5.9
	CHR6_137214200	NS	1.9E-06	5.9
	CHR6_137214204	NS	1.9E-06	5.9
	CHR6_371563912	8.5E-12	2.6E-08	9.8
	CHR7_371621305	NS	3.6E-07	5.9
	CHR7_371621330	NS	3.6E-07	5.9
Resistant Starch	CHR1_181806369*	3.8E-07	NS	6.9
	CHR1_505079023	9.0E-11	NS	8.8
	CHR2_137384845*	NS	1.0E-06	22.1

	CHR2_137480326	NS	1.1E-07	22.1
	CHR2_137480370	NS	1.1E-07	22.1
	CHR2_239215652	1.8E-07	NS	8.3
	CHR2_451413537	3.1E-11	NS	13.7
	CHR3_45534258*	6.0E-07	NS	27.0
	CHR6_116906535	2.3E-07	NS	5.2
	CHR6_387488515	2.2E-07	NS	6.9
Total Starch	CHR7_84111711	1.4E-11	3.2E-07	4.9

* Located within a gene. Italicized SNP (*CHR6_116880302)* is associated with both suc and sta+raf. NS = not identified as significant by the software. [a]GEMMA p-wald values were from the Wald test.

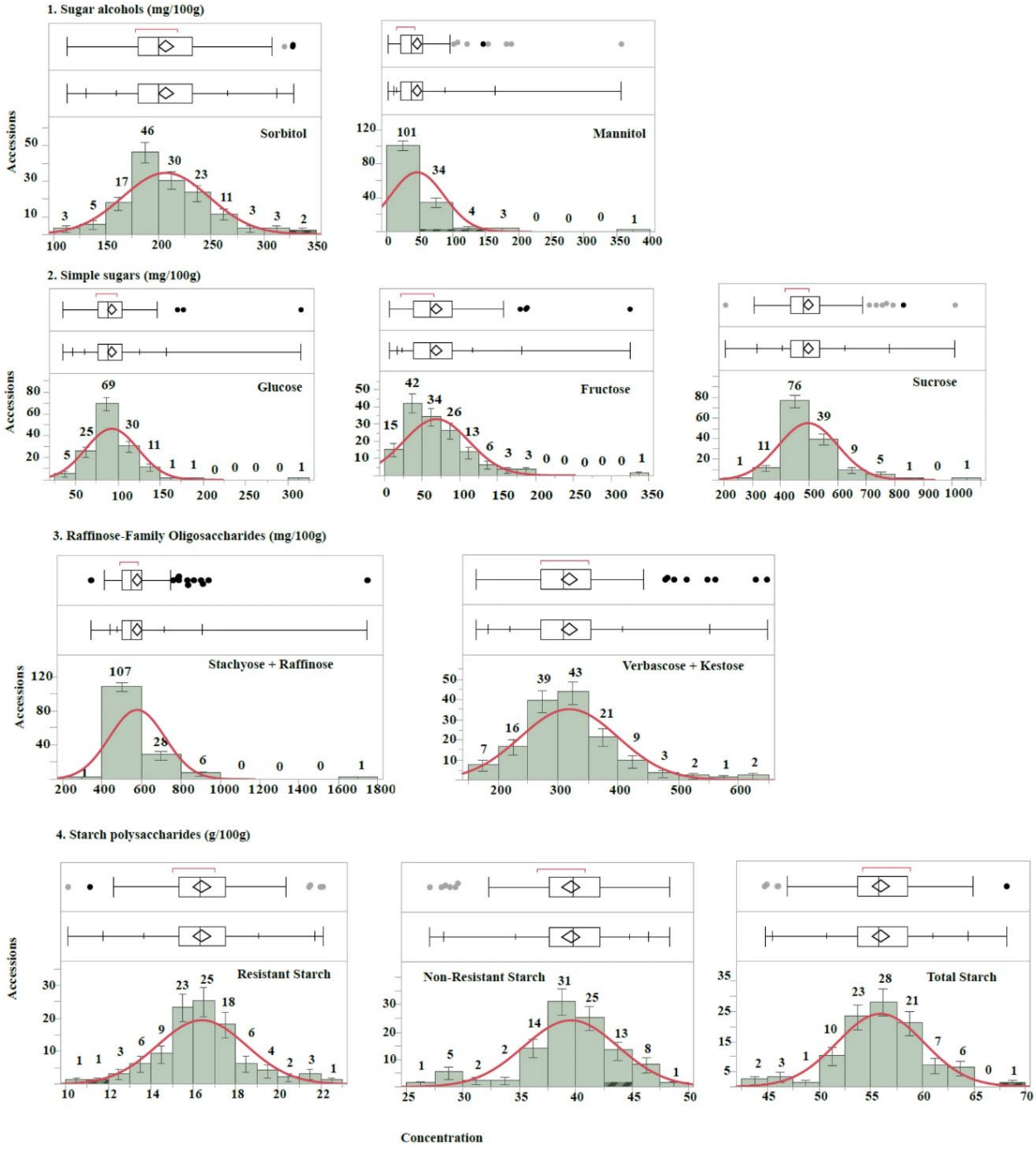

Figure 2.1: Histograms of accession means with normal curve fits. 1. Sugar alcohols (mg/100 g); 2. Simple sugars (mg/100 g); 3. Raffinose-family oligosaccharides (mg/100 g); 4. Starch polysaccharides (g/100 g). The first box plot (Tukey outlier) shows possible outliers as points, while the second box plot (normal quantile) includes all data in estimates. Red normal curves were fitted to the data based on the mean, standard deviation, and sample size.

Figure 2.2: Comparison of carbohydrate concentrations by continent of origin. Bars separated by different letters have significantly different means ($p < 0.05$). Bars labeled as ICARDA originated as part of the ICARDA breeding program.

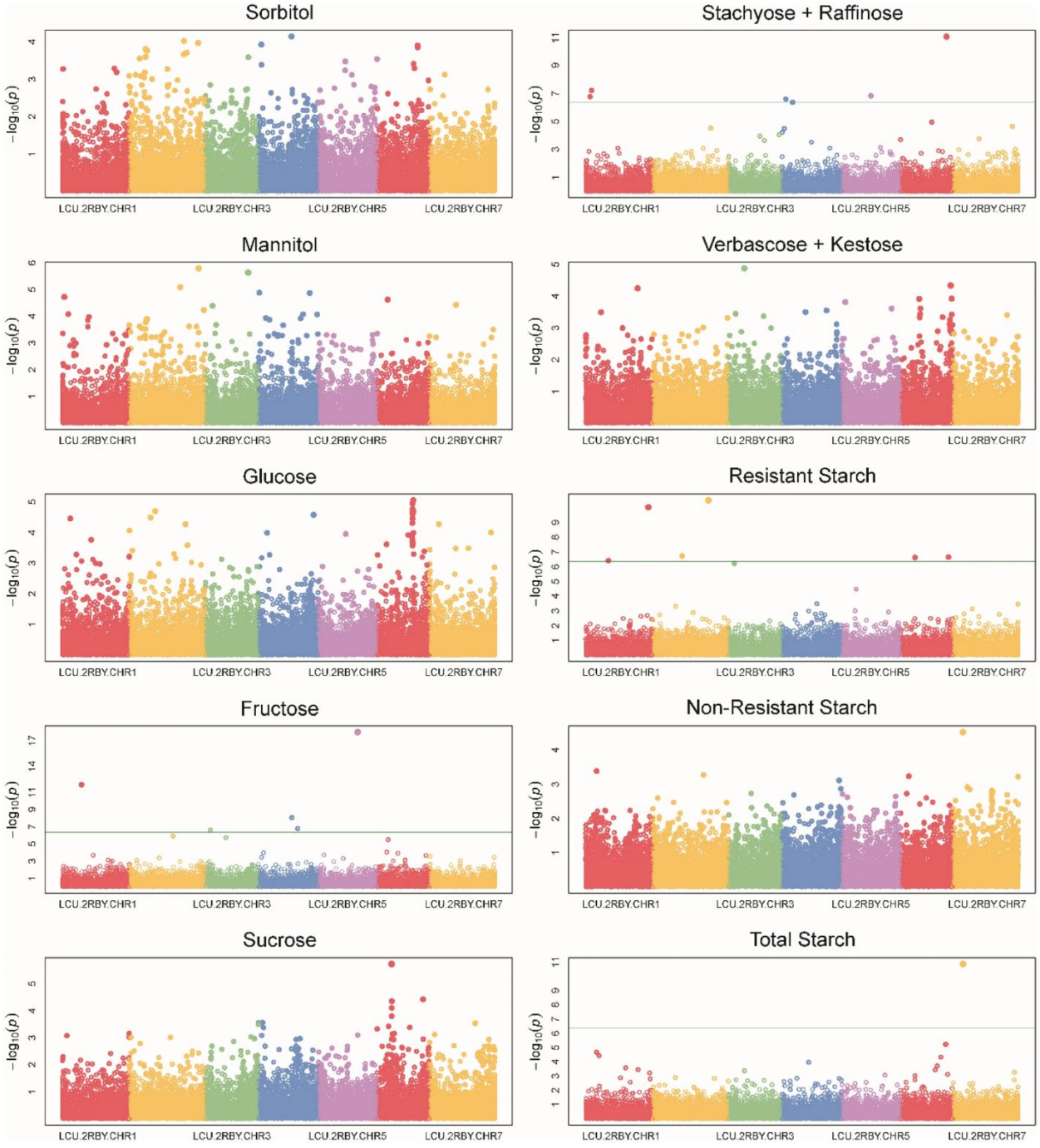

Figure 2.3: Genome-wide association study Manhattan plots from GAPIT. The green line represents the Bonferroni significance threshold ($p < 0.01/22{,}222$).

References

1. Nugent, R. *et al.* Investing in non-communicable disease prevention and management to advance the Sustainable Development Goals. *Lancet* **391**, 2029–2035 (2018).
2. Tietjen, B. *et al.* Climate change-induced vegetation shifts lead to more ecological droughts despite projected rainfall increases in many global temperate drylands. *Glob. Chang. Biol.* **23**, 2743–2754 (2017).
3. Thavarajah, D., Johnson, C. R., McGee, R. & Thavarajah, P. Phenotyping Nutritional and Antinutritional Traits. In *Phenomics Crop Plants Trends, Options Limitations* 223–233 https:// doi. org/ 10. 1007/ 978- 81- 322- 2226-2 (2015).
4. Siva, N. *et al.* Lentil (*Lens culinaris* Medikus) diet affects the gut microbiome and obesity markers in rat. *J. Agric. Food Chem.* **66**, 8805–8813 (2018).
5. Johnson, C. R., Thavarajah, D., Combs, G. F. & Thavarajah, P. Lentil (*Lens culinaris* L.): A prebiotic-rich whole food legume. *Food Res. Int.* **51**, 107–113 (2013).
6. Gibson, G. R. *et al.* Expert consensus document: The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics. *Nat. Rev. Gastroenterol. Hepatol.* **14**, 491–502 (2017).
7. Savage, D. C. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.* **31**, 107–133 (1977).
8. Holzapfel, W. H. & Schillinger, U. Introduction to prebiotics and probiotics. *Food Res. Int.* **35**, 109–116 (2002).
9. Lynch, S. V. & Pedersen, O. The human intestinal microbiome in health and disease. *N. Engl. J. Med.* **375**, 2369–2379 (2016).

10. Gehrig, J. L. *et al.* Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. *Science* **365**, eaau4732 (2019).

11. Dierking, E. C. & Bilyeu, K. D. Raffinose and stachyose metabolism are not required for efficient soybean seed germination. *J. Plant Physiol.* **166**, 1329–1335 (2009).

12. Taji, T. *et al.* Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *Plant J.* **29**, 417–426 (2002).

13. Panikulangara, T. J., Eggers-Schumacher, G., Wunderlich, M., Stransky, H. & Schoffl, F. Galactinol synthase1. A novel heat shock factor target gene responsible for heat-induced synthesis of raffinose family oligosaccharides in arabidopsis. *Plant Physiol.* **136**, 3148–3158 (2004).

14. Loescher, W. & Everard, J. Regulation of sugar alcohol biosynthesis. *Photosynth. Physiol. Metab.* **9**, 275–299 (2000).

15. Gangola, M. P. & Ramadoss, B. R. Sugars play a critical role in abiotic stress tolerance in plants. In *Biochemical, Physiological and Molecular Avenues for Combating Abiotic Stress Tolerance in Plants* 17–38 https:// doi. org/ 10. 1016/ B978-0- 12- 813066- 7. 00002-4 (Elsevier Inc., 2018).

16. Abberton, M. *et al.* Global agricultural intensification during climate change: A role for genomics. *Plant Biotechnol. J.* **14**, 1095–1098 (2016).

17. Varshney, R. K. *et al.* Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnol. Adv.* **31**, 1120–1134 (2013).

18. Arumuganathan, K. & Earle, E. D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–218 (1991).

19. Scheben, A., Batley, J. & Edwards, D. Genotyping-by-sequencing approaches to characterize crop genomes: Choosing the right tool for the right application. *Plant Biotechnol. J.* **15**, 149–161 (2017).

20. Ma, Y. *et al.* Dissecting the genetic architecture of aphanomyces root rot resistance in lentil by QTL mapping and genome-wide association study. *Int. J. Mol. Sci.* **21**, 2129 (2020).

21. Amin, M. N. *Molecular Analysis of Abiotic Stress in Lentil (Lens culinaris. Medik)* (Washington State University, 2018).

22. Gibson, G. R. & Roberfroid, M. B. Dietary modulation of the human colonic microbiota: Introducing the concept of prebiotics. *J. Nutr.* **125**, 1401–1412 (1995).

23. Bhatty, R. S. & Slinkard, A. E. Composition, starch properties and protein quality of lentils. *Can. Inst. Food Sci. Technol.* **12**, 88–92 (1979).

24. Johnson, C. R. *et al.* A global survey of low-molecular weight carbohydrates in lentils. *J. Food Compos. Anal.* **44**, 178–185 (2015).

25. Wu, B. H. *et al.* Maternal inheritance of sugars and acids in peach (*P. persica* (L.) Batsch) fruit. *Euphytica* **188**, 333–345 (2012).

26. Gangola, M. P., Khedikar, Y. P., Gaur, P. M., Baišga, M. & Chibbar, R. N. Genotype and growing environment interaction shows a positive correlation between substrates of raffinose family oligosaccharides (RFO) biosynthesis and their accumulation in chickpea (*Cicer arietinum* L.) seeds. *J. Agric. Food Chem.* **61**, 4943–4952 (2013).

27. McPhee, K. E., Zemetra, R. S., Brown, J. & Myers, J. R. Genetic analysis of the raffinose family oligosaccharides in common bean. *J. Am. Soc. Hortic. Sci.* **127**, 376–382 (2002).

28. Fox, G. *et al.* Is malting barley better feed for cattle than feed barley?. *J. Inst. Brew.* **115**, 95–104 (2009).

29. Marteau, P. & Seksik, P. Tolerance of probiotics and prebiotics. *J. Clin. Gastroenterol.* **38**, S67–S69 (2004).

30. McCleary, B. V. & Monaghan, D. A. Measurement of resistant starch. *J. AOAC Int.* **85**, 665–675 (2002).

31. Becklin, K. M. *et al.* Examining plant physiological responses to climate change through an evolutionary lens. *Plant Physiol.* **172**, 635–649 (2016).

32. Matros, A. *et al.* Genome-wide association study reveals the genetic complexity of fructan accumulation patterns in barley grain. *J. Exp. Bot.* https:// doi. org/ 10. 1093/ jxb/ erab0 02 (2021).

33. Sui, M. *et al.* Genome-wide association analysis of sucrose concentration in soybean (*Glycine max* L.) seed based on high-throughput sequencing. *Plant Genome* **13**, 1–18 (2020).

34. Madrid Liwanag, A. J. *et al.* Pectin biosynthesis: GALS1 in *Arabidopsis thaliana* is a β-1,4-galactan β-1,4-galactosyltransferase. *Plant Cell* **24**, 5024–5036 (2013).

35. Tapernoux-Luthi, E. M., Bohm, A. & Keller, F. Cloning, functional expression, and characterization of the raffinose oligosaccharide chain elongation enzyme, galactan:galactan galactosyltransferase, from common bugle leaves. *Plant Physiol.* **134**, 1377–1387 (2004).

36. Bailey-Serres, J. *et al.* Submergence tolerant rice: SUB1's journey from landrace to modern cultivar. *Rice* **3**, 138–147 (2010).

37. Muir, J. G. *et al.* Measurement of short-chain carbohydrates in common Australian vegetables and fruits by high-performance liquid chromatography (HPLC). *J. Agric. Food Chem.* **57**, 554–565 (2009).

38. Feinberg, M., San-redon, J. & Assie, A. Determination of complex polysaccharides by HPAE-PAD in foods: Validation using accuracy profile. *J. Chromatogr. B* **877**, 2388–2395 (2009).

39. Megazyme. Resistant Starch Assay Procedure (AOAC). (2019).

CHAPTER THREE

FOURIER-TRANSFORM INFRARED SPECTROSCOPY (FTIR) AS A HIGH-THROUGHPUT PHENOTYPING TOOL FOR QUANTIFYING PROTEIN QUALITY IN PULSE CROPS

Abstract

Fourier-transform mid-infrared (FT-MIR) spectroscopy is a high-throughput, cost-effective method to quantify nutritional traits, such as total protein and sulfur-containing amino acid (SAA) concentrations, in plant matter. This study used the spectroscopic technique FT-MIR coupled with attenuated total internal reflectance sampling interface to develop multivariate models for total protein concentration in chickpea (*Cicer arietinum* L.), dry pea (*Pisum sativum* L.), and lentil (*Lens culinaris* Medik.), in addition to SAA concentration in lentil. Total nitrogen data from combustion analysis and SAA data from high-performance liquid chromatography analysis following acid hydrolysis were used for model calibration and validation. Models for the total protein concentration of chickpea (calibration root mean square error [RMSE] = 0.093, R^2 = 0.948, prediction RMSE = 0.10), dry pea (calibration RMSE = 0.096, R^2 = 0.845, prediction RMSE = 0.093), and lentil (calibration RMSE = 0.13, R^2 = 0.845, prediction RMSE = 0.11) utilized infrared regions associated with protein structures, namely amide bands A, I, and II. In sulfur-related models for lentil total SAA (calibration RMSE = 0.014, R^2 = 0.827, prediction RMSE = 0.022) and methionine (calibration RMSE = 0.0075, R^2 = 0.815, prediction RMSE = 0.014) models utilized the C-S and S-CH_3 stretching and bending bands. Study findings support the conclusion that FT-MIR spectroscopy is a promising high-throughput and cost-effective phenotyping technique that will allow quantifying protein traits quickly and easily in pulse crops.

Abbreviations

AA amino acids

ATR attenuated total reflectance

FIR far-infrared

FTIR Fourier-transform infrared spectroscopy

FT-MIR Fourier-transform mid-infrared

HPLC high-performance liquid chromatography

IR infrared

MIR mid-infrared

NIR near-infrared

PLS partial least squares

QTL quantitative trait loci

RMSE root means square error

SAA sulfur-containing amino acids

ZFF zero-fill factor

Introduction

Pulse crops, such as chickpea (*Cicer arietinum* L.), dry pea (*Pisum sativum* L.), and lentil (*Lens culinaris* Medik.), are an essential part of the global food system to provide plant-based protein, low digestible carbohydrates, and a range of micronutrients (Foyer et al., 2016; Johnson et al., 2020). These staple crops are increasing in popularity as plant-based protein sources—a trend expected to continue based on many factors such as health benefits and climate change (Graça et al., 2019; Kim et al., 2019; Pimentel & Pimentel, 2003). Pulses tend to be low in sulfur-containing amino acids (SAA) (Boye et al., 2012), so varieties high in methionine and cystine are a vital breeding objective to increase the protein quality in plant-based diets. However, measuring the concentration of amino acids (AA), particularly SAA, is challenging, as they are susceptible to acid degradation and thus require an additional protective oxidation step. A typical method takes two to three days for sample digestion before AA quantification. Instruments to measure AA concentrations, such as high-performance liquid chromatography (HPLC), are generally low-throughput, expensive, time-consuming, and require highly skilled operators. Quantitative Fourier-transform mid-infrared (FT-MIR) spectroscopy methods offer a promising alternative to conventional methods for analyzing protein and SAA. Samples can be analyzed in seconds without the chemicals and consumables required by traditional techniques.

Infrared (IR) is a low-energy region in the electromagnetic spectrum extending from 12,800 to 10 cm^{-1} (Skoog et al., 2016) and consists of the near-infrared (NIR; 12,800–4,000 cm^{-1}), mid-infrared (MIR; 4,000–200 cm^{-1}), and far-infrared (FIR; 200–10 cm^{-1}) spectrums (Skoog et al., 2016). Infrared spectroscopy using interferometers coupled with Fourier-transform (FT) algorithms are termed Fourier-transform infrared spectroscopy (FTIR) instruments and have several advantages over previous dispersive spectroscopy instruments, including (a) greater

energy intensity due to the lack of slits and fewer optics to attenuate the source radiation (mechanically simpler), known as Jacquinot's (throughput) advantage; (b) simultaneous collection of multiple wavelengths (without the need for scanning), resulting in a shorter collection time and consequent increases in the signal-to-noise ratio, known as Fellgett's (multiplex) advantage; and (c) increased wavenumber accuracy inherent to the internal laser calibration and interferometer, enabling multiple scans to be collected *and* averaged, known as Connes' advantage (Perkins, 1987; Skoog et al., 2016). Fourier-transform instruments in the near, mid, and far regions probe high-frequency oscillations (vibrational overtones), fundamental vibrational modes, and low energy vibrations (Berthomieu & Hienerwadel, 2009; Capuano & van Ruth, 2015; El Khoury & Hellwig, 2017). However, the fundamental oscillations in MIR spectroscopy provide quantitative data from unique functional group oscillations (Leong et al., 2018). The overtones arising in the NIR range lack a robust quantitative background due to the complexity of unresolved bands (Capuano & van Ruth, 2015). Thus, chemometric models underlying NIR spectroscopy may not produce consistent quantitative results across diverse samples, such as grain flours from different regions or years, despite success in training sets. NIR spectroscopy was first reported for the evaluation of protein in pulses in Williams et al. (1978), yet the method has been little reported since, with even less work reported using MIR spectroscopy. The stronger absorption bands of MIR spectra provide a superior platform for consistent chemometrics with greater selectivity and sensitivity, which will not change with crop genotype, growing location, or year. Therefore, FT-MIR can be used to simultaneously identify and quantify molecules (i.e., proteins, carbohydrates, etc.) based on their distinct functional groups without further sample preparation.

The functional groups of proteins (N-H and C = O) and SAA (C-S and C-H of S-CH_3) have permanent dipole moments, and such groups can be readily probed with FT-MIR spectroscopy (Barth, 2007; Berthomieu & Hienerwadel, 2009). Total protein and SAA offer a helpful picture of protein quality in pulses since pulses are high in protein but limited by SAA (Bhatty, 1988; Sarwar & Peace, 1986). Standard laboratory approaches for measuring protein and SAA include the Dumas method (nitrogen analysis through combustion), Kjeldahl method, UV-visible spectroscopy (Chang & Yan, 2019), and various chromatography techniques, such as HPLC with diode array detection. Most of the above approaches are destructive to the sample, require extensive analysis time, chemicals, and skills and are thus expensive. Amino acid analysis, for example, costs over $100 per sample. Total protein analysis is less expensive at ~$6 but remains a constraint when analyzing thousands of samples. Consequently, these methods do not qualify as high-throughput workflows desired in nutritional breeding programs. In contrast, FT-MIR spectroscopy is a nondestructive, high-throughput approach requiring little operating costs or training. Therefore, the objectives of this paper are two-fold: (a) demonstrate FT-MIR as a potential high-throughput, nondestructive, and cost-effective phenotyping technique for pulse nutritional traits, and (b) present multivariate models for the quantification of protein and SAA in pulse crops based on FT-MIR spectra.

Materials and Methods

FTIR Instrumentation and Data Analysis Software

A Cary 630 FTIR spectrometer with a diamond attenuated total reflectance (ATR) module (Agilent Technologies) was used to acquire all MIR spectroscopic data. The data acquisition was performed within a spectral range of 650–4,000 cm^{-1} under Happ-Genzel apodization. The instrument acquisition parameters were optimized for each trait to enable the

collection of spectral data with sufficient selectivity and sensitivity for quantitative analysis (Table 3.2). The data were analyzed with MicroLab Expert software (version 1.1) developed by Agilent Technologies for multivariate statistical modeling (chemometric modeling). Scatter plots were generated, and pooled *t*-tests were performed in JMP Pro (14.0.0).

Chickpea, Dry Pea, and Lentil Seed Samples

All pulse seed samples were collected from U.S. breeding programs, specifically the USDA-ARS chickpea breeding program at Washington State University and the organic pulse nutritional breeding program at Clemson University. For chickpea and dry pea, a total of 100–150 dry seeds were selected from each breeding line and ground to a maximum particle size of 0.5 mm, using a cyclone sample grinder (UDY Corporation). Likewise, 10–50 seeds were selected from each lentil line and ground using a blade coffee grinder (KitchenAid) and sieved to a maximum particle size of 0.5 mm. The powdered subsamples were stored before analysis in a cold room maintained at 10 °C with a humidity level of ~50%.

Total Nitrogen Analysis

The total nitrogen content of all pulse flours was analyzed on a combustion nitrogen analyzer at the Clemson Agricultural Service Laboratory (Clemson, SC). The final protein concentration was determined by multiplying total nitrogen by a factor of 6.25 (Salo-väänänen & Koivistoinen, 1996).

Sulfur-Containing Amino Acid Analysis

Lentil SAA concentrations were determined using an acid hydrolysis method with a pre-oxidation step, followed by HPLC analysis. The hydrolysis method was adapted from Gehrke et al. (1985) and Manneberg et al. (1995). In brief, 40 mg of lentil flour was weighed into glass culture tubes (16 × 125 mm, polytetrafluoroethylene [PTFE] lined cap). A lentil lab reference

standard was included in each batch to monitor batch-to-batch variation. Five mL of chilled performic acid (9:1 ratio of formic acid and hydrogen peroxide) was added to each tube to convert the SAA to stable derivatives, methionine sulfone, and cysteic acid. The tubes were gently swirled on a vortex mixer and refrigerated in an ice bath overnight (16 h). Caps were removed, and PTFE boiling rods (1/8 in. × tube length) were added. Samples were evaporated to dryness in an oil bath under vacuum (~70–80 °C, ~610 mmHg; 3 gal. resin trap; BACOENG). The tube rack was elevated with a stir bar underneath to improve consistent evaporation across the batch. The pressure was slowly lowered to prevent bumping. Tubes were removed, and residual oil was wiped off. Caps were removed, and 4.9 mL of 6 M HCl (hydrochloric acid) was added, along with 0.1 mL internal standard mix (25 mM norvaline and sarcosine each). Tubes were tightly capped and gently swirled. Proteins were hydrolyzed in an oven at 110 °C for 24 hr. Tubes were then allowed to cool to room temperature and vortex mixed. Samples were filtered (0.22 μm polypropylene syringe filter), and 1 mL was added to a clean glass tube to be evaporated to dryness as before. Samples were reconstituted with 1 mL mobile phase A and loaded into HPLC vials for analysis.

Amino acid concentrations were measured using an HPLC method adapted from Agilent application notes (Agilent Application Note, 2010; Long, 2015). An Agilent 1100 series system (Agilent Technologies) was used for analysis. A diode array detector (DAD) collected spectra at 338 nm, 10 nm bandwidth (reference 390 nm, 20 nm bandwidth) and 262 nm, 10 nm bandwidth (reference 390 nm, 20 nm bandwidth). Mobile phase A consisted of 10 mM Na_2HPO_4 (sodium phosphate), 10 mM $Na_2B_4O_7{\bullet}10H_2O$ (sodium tetraborate decahydrate), and 5 mM NaN_3 (sodium azide) and was adjusted to pH 8.2 with concentrated HCl and subsequently filtered through 0.2 μm regenerated cellulose membrane. Solution B consisted of acetonitrile/methanol/water

(45:45:10, v/v/v). Separation was achieved on an Agilent Poroshell HPH-C18 3 × 100 mm analytical column (Part Number 695975-502; Agilent Technologies) with the corresponding Poroshell HPH-C18 3 × 5 mm guard column (Part Number 823750–928). The G1329A autosampler derivatized AAs with OPA (*o*-phthalaldehyde) and FMOC (9-fluorenylmethyl chloroformate). Vials of borate buffer (Part Number 5061-3339), H_2O (water) needle wash, and injection diluent (100 mL solution A, 0.4 mL H_3PO_4 conc.) were also required. The injection method was as follows (default speed and offset were used except where noted): (a) draw 2.5 µL from borate buffer, (b) draw 0.5 µL from a sample, (c) mix 3 µL from the air for five times, (d) wait 0.2 min, (e) draw 0 µL from needle wash, (f) draw 0.5 µL from OPA (vial insert) using 2 mm offset, (g) mix 3.5 µL from the air for six times, (h) draw 0 µL from needle wash, (i) draw 0.4 µL from FMOC (vial insert) using 2 mm offset, (j) mix 3.9 µL from the air for 10 times, (k) draw 32 µL from injection diluent, (l) mix 20 µL from the air for eight times, and (m) inject. See Table 3.1 for instrument method and conditions. Dilution series were made for calibration standard curves from 9 to 900 pmol/µL with norvaline (primary AA) and sarcosine (secondary AA) as internal standards at 500 pmol/µl. Calibration curves were generated for each AA from the ratio of AA/internal standards. Standards included cysteic acid, aspartic acid, glutamic acid, asparagine, serine, glutamine, histidine, glycine, threonine, methionine sulfone, arginine, alanine, tyrosine, cystine, valine, methionine, tryptophan, phenylalanine, isoleucine, leucine, lysine, hydroxyproline, and proline.

Chickpea Total Nitrogen Model

The diamond ATR surface was cleaned with HPLC grade methanol (Fisher Scientific) before spectra of the ground chickpea samples (fully homogenized by mixing) were collected. Instrument and model parameters are available in Table 3.2. The instrument acquisition

parameters were set to absorbance mode with 64 scans (~30 s) per spectrum (Table S3.1 [Appendix A]), 4 cm^{-1} resolution, and no zero-fill factor (ZFF). Each breeding line was analyzed seven times. The most stable spectra with constant intensity were selected without averaging for calibration. Background corrections (36 scans) were performed between each spectral collection. Protein is a macronutrient with easily resolved IR bands, requiring less stringent acquisition parameters than SAA, as discussed below. The calibration set included 55 breeding lines (154 spectra) from the 2018 chickpea population, and the validation set included 22 breeding lines (84 spectra) from the 2020 chickpea population for the partial least squares (PLS-1) model (Tobias, 1995). The Savitzky-Golay first-order derivative and smoothing algorithm (smoothing window of 21) was applied to all spectra. The model was calibrated with nitrogen values obtained from a nitrogen analyzer. The PLS-1 model was developed based on the regions sensitive to the total protein concentration (3,682.61–3,006.98 cm^{-1}, N-H stretch; 1,718.30–1,487.21 cm^{-1}, amide bands I and II), and eight PLS model factors were included in the model. The model was run with full cross-validation.

Dry Pea Total Nitrogen Model

The same background correction and data acquisition steps as for chickpea were followed (Table 3.2). However, the calibration set included 40 breeding lines (135 spectra) from the 2019 dry pea population, and the validation set included 22 breeding lines (59 spectra) from the 2020 dry pea population. The spectra were initially normalized to a scale of 0 to 1, and the Savitzky-Golay first-order derivative and smoothing algorithm (smoothing window of 21) was applied. The model was calibrated with total nitrogen values, as done for the chickpea model. The PLS-1 model was developed based on the same spectral ranges as the total nitrogen model above;

however, 11 PLS model factors were included in the model. The model was run with full cross-validation.

Lentil Total Nitrogen and SAA Models

The diamond ATR window was cleaned with HPLC grade methanol and allowed to dry before each spectrum was collected. The background was collected every 30 min or less for convenience. Fourier-transform mid-infrared spectra were collected for 50 lentil breeding lines, and six spectra were collected per breeding line. Acquisition parameters included 200 scans per background and 100 scans (~75 s) per spectrum at a resolution of 2 cm^{-1} and a ZFF of 2 (Table 3.2). All spectra were normalized to a scale of 0 to 1. Unlike the previous models, the spectra were not derivatized by the Savitzky-Golay algorithm because the spectra were highly structured and informative at a resolution of 2 cm^{-1} and with a ZFF of 2. The increased scan number and resolution generated detailed spectra and allowed for the quantification of SAA, which are at low concentrations in lentil. For ease, the same spectra were used for the protein model. Additionally, this allows for the models to be combined into a single method for generating protein and SAA data simultaneously.

A PLS-1 model for total nitrogen in lentil flour was developed using Agilent MicroLab Expert software. The most stable spectra were applied in calibration without averaging. The calibration set included 32 breeding lines (57 spectra), and the validation set included 18 breeding lines (25 spectra). The model utilized the same spectral regions as in chickpea and dry pea and included five PLS model factors. PLS-1 models for total SAA and methionine were similarly attempted. In the model for total SAA, the calibration set included 37 breeding lines (53 spectra), and the validation set included 24 breeding lines (34 spectra). The model utilized 721.24–867.07, 1,231.88–1,469.96, 1,904.20–2,241.99 and 2,825.78–2,994.91 cm^{-1} spectral

regions and included eight PLS model factors. Furthermore, the methionine model included 26 breeding lines (39 spectra) and 22 breeding lines (31 spectra) for calibration and validation, respectively. The model utilized 674.65–808.37, 1,182.03–1,484.41, 1,975.49–2,158.59, and 2,658.52–2,991.19 cm^{-1} spectral regions with eight PLS model factors. All lentil models were run with full cross-validation.

Results and Discussion

This study successfully demonstrated that FT-MIR is a robust, nondestructive tool for measuring protein and SAA in pulse crops. Proteins and SAA have polar functional groups sensitive to MIR energy. The functional groups of proteins (N-H and C = O) in chickpea, dry pea, and lentil flour were analyzed through FT-MIR spectroscopy. Associated IR bands were identified at ~1,550 cm^{-1} (amide II bands), ~1,650 cm^{-1} (amide I band), and between 3,310 and 3,270 cm^{-1} (amide A band) (Tiwari & Singh, 2012). Multivariate models (PLS-1) were developed associating these regions with total nitrogen content. In chickpea, predicted protein concentrations of the validation set ranged from 18.3 to 23.9%, with a mean of 20.9% (Table 3.3). The chickpea total nitrogen model achieved an R^2 of 0.948, a calibration root means square error (RMSE) of 0.093, and a prediction RMSE of 0.10 (Figure 3.1b and Table 3.4). For dry pea, the predicted total protein concentration of the validation set ranged from 18.1 to 23.1%, with a mean of 21.2%. The dry pea total nitrogen model achieved a calibration RMSE of 0.096, an R^2 of 0.845, and a prediction RMSE of 0.093 (Figure S3.1b [Appendix A]). For lentil, predicted protein concentrations ranged from 25.4 to 33.3%, with a mean of 28.3%. The lentil total nitrogen model achieved an R^2 of 0.845, a calibration RMSE of 0.13, and a prediction RMSE of 0.11 (Figure S3.2b [Appendix A]). These models predicted mean protein concentrations in chickpea, dry pea, and lentil within the cited ranges in the literature (chickpea:

15.6–22.4%, dry pea: 20–25%, and lentil: 20.6–31.4%), demonstrating the applicability of the method in the field (Jarpa-Parra, 2018; Khan et al., 2016; Upadhyaya et al., 2016). Furthermore, pooled two-tailed *t*-tests performed on each crop (chickpea: $P > |t| = 0.93$; dry pea: $P > |t| = 0.97$; lentil: $P > |t| = 0.82$) targeting the means of actual and predicted protein concentrations of validation data showed no significant difference.

The functional groups of SAA (C-S and C-H of S-CH_3) in lentil flour were similarly analyzed. SAA is a valuable nutritional breeding trait because lentil (and other pulse crops) is nutritionally limited by SAA, methionine, and cysteine, despite being high in total protein. These low concentrations present a challenge for IR band resolution and consequent quantification. However, this study successfully identified bands in the lentil MIR spectrum (~751–685,~2,493–2,157, and ~2,977–2,861 cm^{-1}) associated with C-H stretching of methyl mercaptan (S-CH_3) and C-S stretching in pure methionine were recognized (Figures 3.2a & S3.3a–S3.4 [Appendix A]). The bands apparent at ~2,991–2,659 cm^{-1} and 1,470–1,232 cm^{-1} represent the total C-H, C-CH_2, and C-CH_3 oscillations in lentil flour. The region between ~2,159–1,975 cm^{-1} (the phonon band arising due to the oscillations of the carbon lattice of ATR- diamond) strengthened the prediction of the multivariate regression models for total SAA and methionine. The lentil SAA model achieved an R^2 of 0.827, and the predicted validation data ranged from 0.207 to 0.326%, with a mean of 0.258%. In this model, the calibration RMSE was 0.014, and the prediction RMSE was 0.022 (Figure 3.2b). Further, the methionine model achieved an R^2 of 0.815 and predicted the validation results between 0.194–0.294%, with a mean of 0.222%. The methionine model had the calibration and prediction RMSEs at 0.0075 and 0.014, respectively (Figure S3.3b [Appendix A]). The lines of best fit for the validation data (Figures 3.1b–3.2b & S3.1b–S3.3b [Appendix A]; blue lines) have deviated slightly from that of the calibration data (Figures 3.1b–3.2b &

S3.1b–S3.3 [Appendix A]; black lines). The *t*-tests performed for total SAA and methionine ($P > |t| = 0.35$ and $P > |t| = 0.76$, respectively) returned no significant differences between actual and predicted means. The predicted lentil methionine mean, 0.22%, agrees well with the literature (0.22%, USDA ARS, 2019). Total SAA makes up ~2% of the total protein content of lentils, whereas SAA comprise ~4% of beef and chicken protein and ~8% of chicken egg protein (USDA ARS, 2019). Lentil and other pulse crops are not a good source of SAA; however, genetic selection and breeding may help increase their SAA concentrations. Developing lentil varieties with high SAA concentrations could help improve the dietary intake of better-quality protein and develop food products, such as protein powder, that contain high-quality protein without adding another high-SAA source.

Chemometric models with well-recognized and consistent underlying bands will aid in the development of analytical methods and accurate, consistent modeling regardless of differing sample origins. While the prediction RMSEs indicate these models have high predictive ability for each sample, the *t*-tests indicate they also accurately predict the population means. The calibration data were not used in model validation, and the purpose of calibration data was to build the model, whereas validation data was to test the model. Thus, these total protein, total SAA, and methionine chemometric models have consistent applicability over these pulse crops regardless of sample origin. Accordingly, FT-MIR spectroscopy provides added advantages for stable and straightforward chemometric modeling compared with methods associated with the NIR range, which lacks a strong quantitative foundation (Guo et al., 2016).

Traditional univariate statistical regression modeling based on Beer-Lambert was unsuitable for complex sample systems like lentil and chickpea. Partial least squares regression (a multivariate statistical regression algorithm) was applied with chemometric modeling

throughout this study, where the best predictive use of spectral variables can be enhanced. The use of PLS regression reduced the dimensionality of the multivariate space in a supervised manner, maintaining a good correlation between dependents (absorbance values) and independent (analyte concentrations) variables (Saikat et al., 2008). Therefore, PLS-1 proved to be an excellent choice for correlating nutrient data with the spectral regions associated with protein functional groups. Fourier-transform mid-infrared spectroscopic data were utilized with minimal mathematical pre-processing (averaging, normalization, and the Savitzky-Golay derivative and smoothing algorithm). In FT-MIR spectroscopy, the spectra are always associated with functional groups and molecular skeletal structures (Yadav, 2005). Fully resolved functional group bands act as fingerprints for traits (analytes). In proteins, the A, I, and II amide bands (Figures 3.1a & S3.1a–S3.2a [Appendix A]) were significantly associated with protein content in our models. The C-H stretching bands of methyl mercaptans and C-S stretching bands in methionine are mainly associated with our total SAA and methionine models. Other spectral regions common to both lentil flour and the standard compounds (Figure S3.4) were also selected to enhance the regression in the chemometric models. Notably, different spectral acquisition parameters were followed in the lentil models than the chickpea and dry pea models during spectral sampling. This was to ensure sufficiently high resolution and scan number in the lentil spectra to observe the minor bands associated with methionine at low concentrations in the sample matrix. Once highly resolved spectra were employed, the number of spectra required for a consistent model in the lentil models was lower than for the chickpea and dry pea models, which employed lower resolution parameters and had fewer spectral details (data points) in each spectrum. However, high resolution is not required for a bulk trait such as total protein because the associated amide bands are distinct and quickly resolved. The use of first derivatives in the

chickpea and dry pea spectra further strengthened the predictive ability of the two respective chemometric models related to total proteins.

Breeding programs require the generation of large amounts of phenotypic data. Nutritional traits are no exception, yet higher costs are associated with collecting these data than traditional agronomic traits such as yield. With the great promise of molecular-based breeding approaches, such as marker-assisted backcrossing and genomic selection along with genome-wide association studies, large datasets are needed to discover quantitative trait loci (QTL) and elucidate underlying gene pathways associated with traits (Liu et al., 2020; Roorkiwal et al., 2016; Sab et al., 2020; Upadhyaya et al., 2016). The application of conventional protocols in quantifying nutrients (nutritional phenotyping) is not suitable for the large volume of samples from the field. Significant challenges with traditional quantitative analysis techniques include long analysis times, highly trained workers, chemical costs, chemical disposal, and instrument maintenance. Fourier-transform infrared spectroscopy analysis time is short (i.e., less than a minute), and the method does not require a skilled operator (Capuano & van Ruth, 2015). It also requires minimal sample preparation, minimizing the risks of hazardous chemical usage and chemical cost. Compared with the complex compartmentalization typical of liquid and gas chromatography systems, the compact instrumentation occupies little space and is relatively simple in construction. Maintenance costs are also considerably lower than other analytical instruments (Minali & Rein, 2015). Therefore, FT-MIR spectroscopy can support a high-throughput and efficient workflow for the quantitative analysis of nutritional traits.

Accordingly, the chemometric regression (PLS-1) models for total protein and methionine could be an essential part of this high-throughput phenotyping workflow. This analytical technique could lower costs in breeding programs globally and open possibilities for

developing and under-resourced countries to adopt the technique in their breeding programs. The methods and models presented in this study can accelerate nutritional breeding programs by reducing the time and cost of analysis and by being incorporated into QTL discovery pipelines. Rapid, low-cost data generation is advantageous for efficiently increasing sample size and power in genome-wide association studies. Once QTLs are detected, flanking markers can be used in marker-assisted selection (MAS) to verify the presence or absence of favorable alleles in progeny. MAS could be an effective technique for nutritional traits because the phenotype can be predicted without processing and analyzing the seed. Seedlings could be genetically tested and selected or discarded before flowering, allowing for same-generation hybridization, essentially cutting generation time in half.

Conclusions

Fourier-transform mid-infrared spectroscopy is conveniently applicable with simple chemometric modeling to predict the concentrations of total proteins and SAA in chickpea, dry pea, and lentil. Well-recognized functional groups (bands) associated with total protein content and SAA content in the MIR range make multivariate modeling relatively simple. Therefore, the present work on FT-MIR spectroscopy creates a platform for high-throughput and nondestructive phenotyping with minimal costs and chemical hazards. Further, these techniques can reduce breeding program expenses globally and allow under-resourced countries to expand into nutritional phenotypes, such as those with improved protein content. Future studies may benefit from exploration of different modeling techniques and larger sample sizes for calibration and validation.

Acknowledgements

This project was supported by the American people via the Feed the Future Innovation Lab for Crop Improvement through the United States Agency for International Development (USAID, award no 7200AA19LE00005/subaward no 89915-11295 awarded to DT); the Organic Agriculture Research and Extension Initiative (OREI) (award no. 2018-51300-28431/proposal no. 2018–02799) of the United States Department of Agriculture, National Institute of Food and Agriculture, the Pulse Health Initiative (USDA-ARS awarded to DT); the Good Food Institute, and the USDA National Institute of Food and Agriculture, [Hatch] project [1022664] awarded to DT. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the USDA, USAID, or the United States Government. The authors thank Drs. George Vandermark, USDA-ARS, Washington State, and Siv Kumar, ICARDA, Morocco, providing pulse seed samples for model development.

Tables and Figures

Table 3.1: HPLC gradient method and conditions (max pressure: 400 bar; column temp: 40 °C)

Time	**A**	**B**	**Flow rate**
min	—% MP—		mL/min
0.00	100.0	0.0	0.25
3.00	100.0	0.0	0.25
10.40	81.5	18.5	0.62
23.00	43.0	57.0	0.62
23.10	0.0	100.0	0.62
27.00	0.0	100.0	0.62
27.10	100.0	0.0	0.62
27.90	100.0	0.0	0.62
28.00	100.0	0.0	0.25
33.00	100.0	0.0	0.25

Note. MP, mobile phase.

Table 3.2: Instrument acquisition and model parameters

Model name	Instrument scans # (background/sample)	Resolution cm^{-1}	Zero-fill factor	Preprocessing	Calibration breeding lines #	Validation breeding lines #	Calibration spectra #	Validation spectra #
Chickpea total protein	36/64[a]	4	None	D+S	55	22	154	84
Dry pea Total Protein	36/64[a]	4	None	N, D+S	40	22	135	59
Lentil total protein	200/100[b]	2	2	N	32	18	57	25
Lentil SAA	200/100[b]	2	2	N	37	24	53	34
Lentil methionine	200/100[b]	2	2	N	26	22	39	31

Note. D+S = Savitzky-Golay first-order derivative and smoothing algorithm (smoothing window of 21), N = Normalization (0 to 1).
[a]64 scans ≈ 30 s at 4 cm^{-1} resolution. [b]100 scans ≈ 75 s at 2 cm^{-1} resolution.

Table 3.3: Actual vs. model predicted data

Model name	**Actual calibration set range)**	**Actual calibration set true mean**	**Actual validation set range**	**Actual validation set true mean**	**Predicted validation set range**	**Predicted validation set true mean**	***t*-test**
	—% protein—						
Chickpea total protein	15.4–24.6	20.0	18.1–24.6	20.3	18.3–23.9	20.9	NS
Dry pea total protein	18.3–23.9	21.1	18.4–23.6	21.0	18.1–23.1	21.2	NS
Lentil total protein	25.7–33.7	29.7	24.7–31.1	29.6	25.4–33.3	28.3	NS
Lentil SAA	0.211–0.348	0.279	0.197–0.321	0.265	0.207–0.326	0.258	NS
Lentil methionine	0.185–0.264	0.224	0.200–0.251	0.221	0.194–0.294	0.222	NS

Note. NS = actual and predicted means of validation data were not significant at $P < .05$; SAA, sulfur-containing amino acid.

Table 3.4: Chemometric model statistics

Model name	R^2	**RMSEC**	**RMSECV**	**RMSEP**	**SEP**	**Bias**
Chickpea total protein	0.948	0.093	0.093	0.10	0.10	–0.0057
Dry pea total protein	0.845	0.096	0.096	0.093	0.091	0.0039
Lentil total protein	0.845	0.13	0.13	0.11	0.11	0.016
Lentil SAA	0.827	0.014	0.014	0.022	0.021	–0.0066
Lentil methionine	0.815	0.0075	0.0075	0.014	0.014	0.0011

Note. RMSEC, root mean square error of calibration; RMSECV, root mean square error of cross validation; RMSEP, root mean square error of prediction; SEP, standard error of prediction.

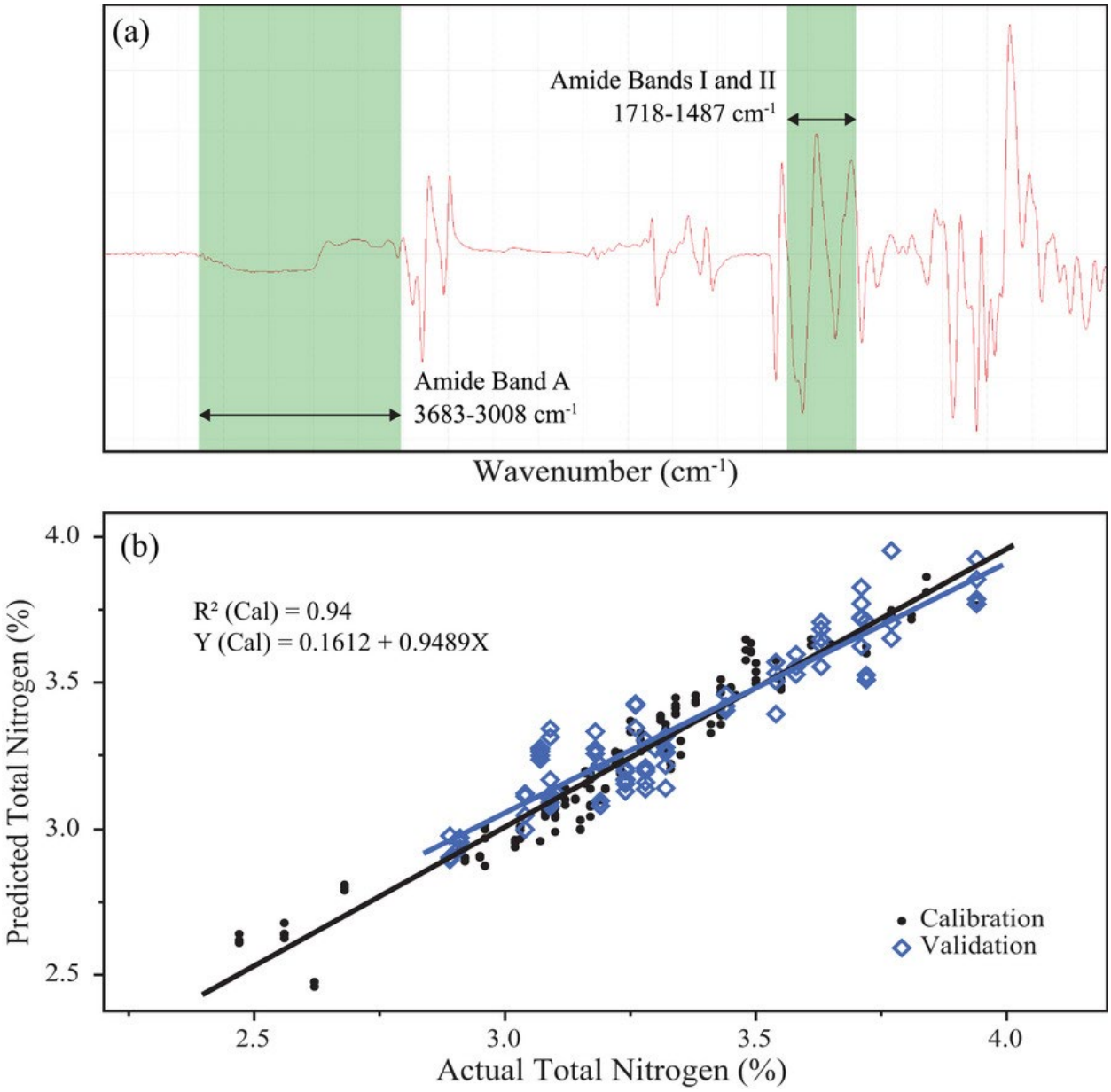

Figure 3.1: Chickpea N model. (a) Average chickpea mid-infrared first-derivative absorbance spectrum. Regions in green were selected for the total nitrogen model in chickpea. (b) Scatter plot of actual vs. predicted total nitrogen (%) of calibration and validation data with lines of best fit

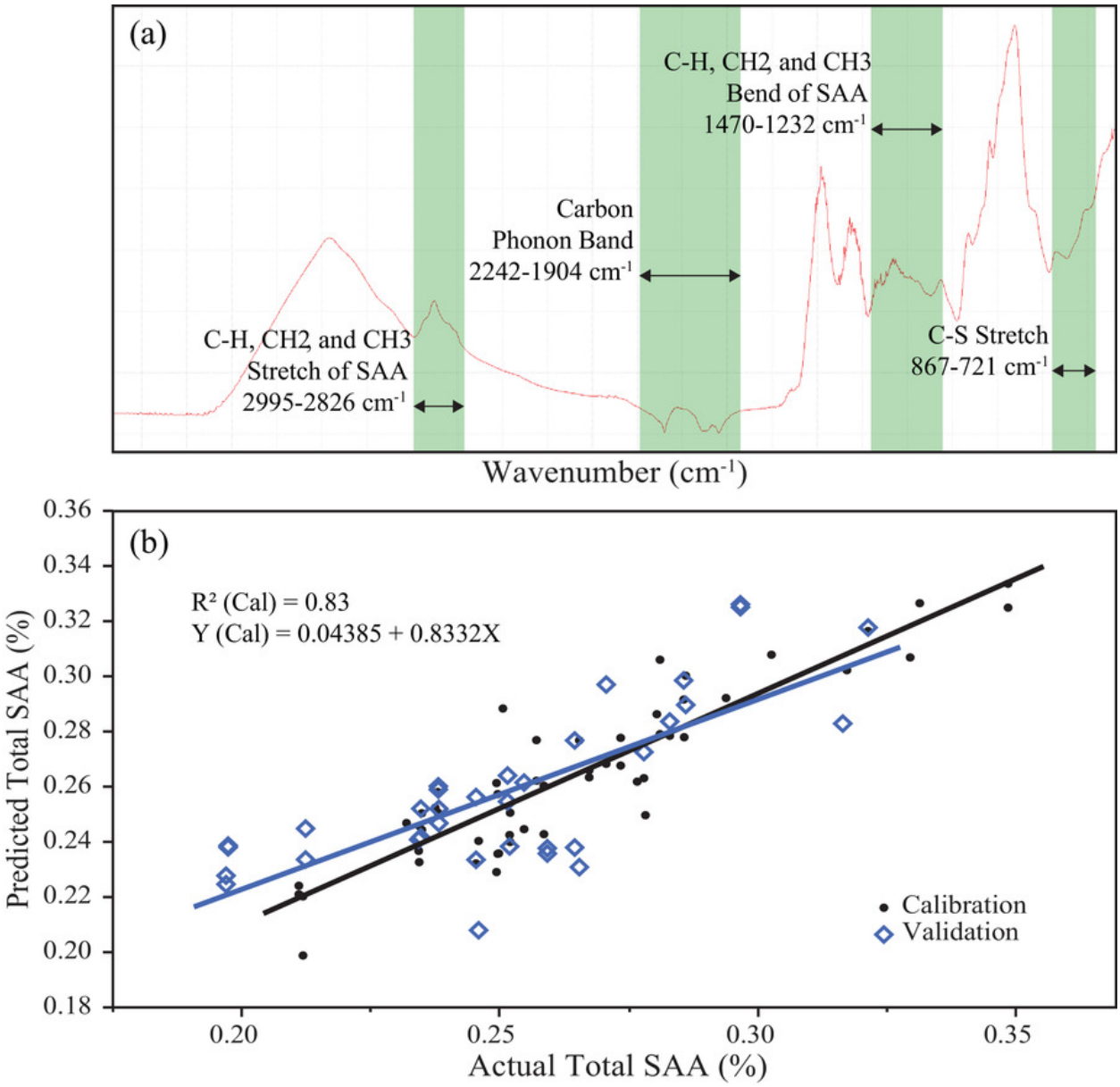

Figure 3.2: Lentil SAA model. (a) Average lentil MIR absorbance spectrum. Regions in green were selected for the total sulfur-containing amino acid (SAA) model in lentil. (b) Scatter plot of actual vs. predicted total SAA (%) of calibration and validation data with lines of best fit.

References

Agilent Application Note. (2010). *Separation of two sulfurated amino acids with other seventeen amino acids by HPLC with pre-column derivatization*. Agilent Technologies.

Barth, A. (2007). Infrared spectroscopy of proteins. *Biochimica et Biophysica Acta - Bioenergetics*, *1767*(9), 1073–1101. https://doi.org/10.1016/j.bbabio.2007.06.004

Berthomieu, C., & Hienerwadel, R. (2009). Fourier transform infrared (FTIR) spectroscopy. *Photosynthesis Research*, *101*(2–3), 157–170. https://doi.org/10.1007/s11120-009-9439-x

Bhatty, R. S. (1988). Composition and quality of lentil (*Lens culinaris* Medik): A review. *Canadian Institute of Food Science and Technology Journal*, *21*(2), 144–160. https://doi.org/10.1016/S0315-5463(88)70770-1

Boye, J., Wijesinha-Bettoni, R., & Burlingame, B. (2012). Protein quality evaluation twenty years after the introduction of the protein digestibility corrected amino acid score method. *British Journal of Nutrition*, *108*(Suppl), (2). https://doi.org/10.1017/S0007114512002309

Capuano, E., & van Ruth, S. M. (2015). *Infrared spectroscopy: Applications*. *Encyclopedia of food and health* (1st ed.). Elsevier Ltd. https://doi.org/10.1016/B978-0-12-384947-2.00644-9

Chang, S. K. C., & Yan, Z. (2019). Protein analysis. In S. S. Nielsen (Ed.), *Food analysis* (5th ed., pp. 315–327). Springer. https://doi.org/10.1007/978-3-319-45776-5

El Khoury, Y., & Hellwig, P. (2017). Far infrared spectroscopy of hydrogen bonding collective motions in complex molecular systems. *Chemical Communications*, *53*(60), 8389–8399. https://doi.org/10.1039/C7CC03496B

Foyer, C. H., Lam, H.-M., Nguyen, H. T., Siddique, K. H. M., Varshney, R. K., Colmer, T. D., Cowling, W., Bramley, H., Mori, T. A., Hodgson, J. M., Cooper, J. W., Miller, A. J., Kunert, K., Vorster, J., Cullis, C., Ozga, J. A., Wahlqvist, M. L., Liang, Y., Shou, H., … Considine, M. J. (2016). Neglecting legumes has compromised human health and sustainable food production. *Nature Plants*, *2*, 16112. https://doi.org/10.1038/nplants.2016.112

Gehrke, C. W., Wall, L. L., Sr., Absheer, J. S., Kaiser, F. E., & Zumwalt, R. W. (1985). Sample preparation for chromatography of amino acids: Acid hydrolysis of proteins. *Journal of Association of Official Analytical Chemists*, *68*(5), 811–821. https://doi.org/10.1093/jaoac/68.5.811

Graça, J., Godinho, C. A., & Truninger, M. (2019). Reducing meat consumption and following plant-based diets: Current evidence and future directions to inform integrated transitions. *Trends in Food Science and Technology*, *91*(July), 380–390. https://doi.org/10.1016/j.tifs.2019.07.046

Guo, T., Feng, W. H., Liu, X. Q., Gao, H. M., Wang, Z. M., & Gao, L. L. (2016). Fourier transform mid-infrared spectroscopy (FT-MIR) combined with chemometrics for quantitative analysis of dextrin in Danshen (Salvia miltiorrhiza) granule. *Journal of Pharmaceutical and Biomedical Analysis*, *123*, 16–23. https://doi.org/10.1016/j.jpba.2015.11.021

Jarpa-Parra, M. (2018). Lentil protein: A review of functional properties and food application. An overview of lentil protein functionality. *International Journal of Food Science and Technology*, *53*(4), 892–903. https://doi.org/10.1111/ijfs.13685

Johnson, N., Johnson, C. R., Thavarajah, P., Kumar, S., & Thavarajah, D. (2020). The roles and potential of lentil prebiotic carbohydrates in human and plant health. *Plants, People, Planet*, *2*, 310–319. https://doi.org/10.1002/ppp3.10103

Khan, T. N., Meldrum, A., & Croser, J. S. (2016). Pea: Overview. In C. Wrigley, H. Corke, K. Seetharaman, & J. Faubion (Eds.), *Encyclopedia of food grains* (2nd ed.; pp. 324–333). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-394437-5.00037-1

Kim, H., Caulfield, L. E., Garcia-Larsen, V., Steffen, L. M., Coresh, J., & Rebholz, C. M. (2019). Plant-based diets are associated with a lower risk of incident cardiovascular disease, cardiovascular disease mortality, and all-cause mortality in a general population of middle-aged adults. *Journal of the American Heart Association*, *8*(16) e012865. https://doi.org/10.1161/JAHA.119.012865

Leong, S. S., Ng, W. M., Lim, J. K., & Yeap, S. P. (2018). Dynamic light scattering: Effective sizing technique for characterization. of magnetic nanoparticles. In S. Sharma (Ed.), *Handbook of materials characterization* (pp. 77–111). Springer. https://doi.org/10.1007/978-3-319-92955-2_3

Liu, X., Qin, D., Piersanti, A., Zhang, Q., Miceli, C., & Wang, P. (2020). Genome-wide association study identifies candidate genes related to oleic acid content in soybean seeds. *BMC Plant Biology*, *20*(1), 1–14. https://doi.org/10.1186/s12870-020-02607-w

Long, W. (2015). Automated amino acid analysis using an *Agilent Poroshell HPH-C18 Column*. Agilent Technologies.

Manneberg, M., Lahm, H. W., & Fountoulakis, M. (1995). Quantification of cysteine residues following oxidation to cysteic acid in the presence of sodium azide. *Analytical Biochemistry*, *231*(2), 349–353. https://doi.org/10.1006/abio.1995.9988

Minali, D., & Rein, A. (2015). *The Agilent Cary 630 FTIR Spectrometer quickly identifies and qualifies pharmaceuticals*. Agilent Technologies.

Perkins, W. D. (1987). Fourier transform infrared spectroscopy: II. Advantages of FT-IR. *Journal of Chemical Education*, *64*(11), A269. https://doi.org/10.1021/ed064pa269

Pimentel, D., & Pimentel, M. (2003). Sustainability of meat-based and plant-based diets and the environment. *American Journal of Clinical Nutrition*, *78*(3, Suppl.), 660–663. https://doi.org/10.1093/ajcn/78.3.660s

Roorkiwal, M., Rathore, A., Das, R. R., Singh, M. K., Jain, A., Srinivasan, S., Gaur, P. M., Chellapilla, B., Tripathi, S., Li, Y., Hickey, J. M., Lorenz, A., Sutton, T., Crossa, J., Jannink, J. L., & Varshney, R. K. (2016). Genome-enabled prediction models for yield related traits in chickpea. *Frontiers in Plant Science*, *7*, 1666. https://doi.org/10.3389/fpls.2016.01666

Sab, S., Lokesha, R., Mannur, D. M., Somasekhar, J., K., Mallikarjuna, B. P., Laxuman, C., Yeri, S., Valluri, V., Bajaj, P., Chitikineni, A., Vemula, A., Rathore, A., Varshney, R. K., Shankergoud, I., & Thudi, M. (2020). Genome-wide SNP discovery and mapping QTLs for Seed iron and zinc concentrations in chickpea (*Cicer arietinum* L.). *Frontiers in Nutrition*, *7*, 559120. https://doi.org/10.3389/fnut.2020.559120

Saikat, M., Jun, Y., Maitra, S., & Yan, J. (2008). Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Casualty Actuarial Society*, 79–90.

Salo-väänänen, P. P., & Koivistoinen, P. E. (1996). Determination of protein in foods: Comparison of net protein and crude protein (N × 6.25) values. *Food Chemistry*, *57*(1), 27–31. https://doi.org/https://doi.org/10.1016/0308-8146(96)00157-4

Sarwar, G., & Peace, R. W. (1986). Comparisons between true digestibility of total nitrogen and limiting amino acids in vegetable proteins fed to rats. *The Journal of Nutrition*, *116*(7), 1172–1184. https://doi.org/10.1093/jn/116.7.1172

Skoog, D. A., Hanlan, J., & West, D. M. (2016). *Principles of instrumental analysis* (7th ed.). Cengage.

Tiwari, B., & Singh, N. (2012). *Pulse chemistry and technology*. The Royal Society of Chemistry.

Tobias, R. D. (1995). An introduction to partial least squares regression. In *Proceedings of the 20th Annual SAS Users Group International Conference*. SAS Institute Inc.

USDA ARS. (2019). *FoodData Central* (NDB# 16069). https://fdc.nal.usda.gov/fdc-app.html#/food-details/172420/nutrients

Upadhyaya, H. D., Bajaj, D., Narnoliya, L., Das, S., Kumar, V., Gowda, C. L. L., Sharma, S., Tyagi, A. K., & Parida, S. K. (2016). Genome-wide scans for delineation of candidate genes regulating seed-protein content in chickpea. *Frontiers in Plant Science*, *7*, 302. https://doi.org/10.3389/fpls.2016.00302

Williams, P. C., Stevenson, S. G., Starkey, P. M., & Hawtin, G. C. (1978). The application of near infrared reflectance spectroscopy to protein-testing in pulse breeding programmes. *Journal of the Science of Food and Agriculture*, *29*(3), 285–292. https://doi.org/https://doi.org/10.1002/jsfa.2740290315

Yadav, L. D. S. (2005). *Organic spectroscopy*. Springer. https://doi.org/10.1007/978-1-4020-2575-4

CHAPTER FOUR

GENOME-WIDE ASSOCIATION MAPPING OF LENTIL (*LENS CULINARIS* MEDIK.) PROTEIN QUALITY TRAITS

Abstract

Lentil (*Lens culinaris* Medik.) contains ~25% high-quality plant-based protein in addition to high concentrations of prebiotic carbohydrates and micronutrients, such as folate, iron, zinc, and selenium. As the economic and environmental costs of animal-based protein rise, plant-based proteins, such as lentil, will become increasingly important to global food systems. Consequently, evaluating and targeting protein quality traits for genomic-assisted breeding is a valuable objective for lentil breeding programs. To this end, this study measured protein quality traits (amino acids and protein digestibility) in a lentil diversity panel grown under greenhouse conditions. Repeatability estimates were calculated, indicating low to moderate heritability in protein quality traits. Twelve traits were strongly correlated with each other ($r > 0.70$; Ala, Arg, Asp, Glu, Gly, Ile, Leu, Met, Ser, Thr, Val and total amino acid concentration [TA]). Admixture analysis was performed and subpopulations were evaluated based on their global distributions and effect on protein quality traits. Finally, genome-wide association studies were performed to identify SNP markers significantly association with protein quality traits. Candidate genes in local linkage disequilibrium with significant SNPs were identified and evaluated for physiological importance.

Abbreviations

AA	amino acid
Arg	arginine
Asp	aspartate
Ala	alanine
Cys	cystine
Glu	glutamate
Gly	glycine
His	histidine
H-Pro	hydroxyproline
Ile	isoleucine
Leu	leucine
Lys	lysine
Met	methionine
PDg	protein digestibility
Phe	phenylalanine
Pro	proline
S-AA	sulfur-containing amino acid
Ser	serine
TA	total amino acid
Thr	threonine
Val	valine

Introduction

Protein quality, in addition to protein content, is an important consideration when evaluating plant-based protein sources, such as lentil (*Lens culinaris* Medik.). Lentil and other legumes are a staple part of many traditional diets around the world and feature prominently in the Mediterranean diet, which has been shown to lower the risk of all-cause, cardiovascular, and cancer mortality (Papandreou *et al.*, 2019). A 100 g serving of lentil provides 25 g of protein or 50% of the recommended dietary allowance (National Research Council Subcommittee on the Tenth Edition of the Recommended Dietary Allowances, 1989; U.S. Department of Agriculture, 2019) and is also a rich source of prebiotic carbohydrates, vitamins, and minerals (Johnson *et al.*, 2020). As many people begin to transition away from animal-based protein sources, lentil is an excellent alternative nutritionally, economically, and environmentally. Lentil is less expensive to purchase than meat, has a reduced emissions impact on the environment, and lowers nitrogen fertilizer requirements by fixing atmospheric nitrogen in root nodules (Foyer *et al.*, 2016; Semba *et al.*, 2021). However, a challenge that lentil protein, along with other plant-based proteins, faces is a lower protein quality than animal protein.

Proteins are macromolecules composed of amino acids bound together by peptide bonds. A healthy diet not only requires a sufficient quantity of protein but also sufficient quantities of the essential amino acids Phe, Val, Trp, Thr, Ile, Met, His, Leu, and Lys. These are amino acids that humans cannot synthesize and, therefore, must consume in their diets or suffer malnutrition. Lentil is a good source of Asn, Ala, Asp, and Glu; however, lentil's first limiting amino acids are the sulfur-containing amino acids, Met and Cys (Salaria *et al.*, 2022). The body can synthesize Cys from Met and, consequently, they are combined for intake requirements. Cys is vital for its role in protein folding due to its ability to form disulfide bonds, while Met is significant for its

role as the amino acid that begins translation as well for its derivatives, glutathione and S-adenosyl methionine (SAM), which are important in oxidative protection and DNA methylation, respectively, and have been examined in such diverse pathologies as obesity and Parkinson's disease (Barbosa *et al.*, 2021; Jalgaonkar *et al.*, 2022). In addition to amino acid composition, protein quality also depends on protein digestibility (PDg), which determines how well the protein can be catabolized during digestion and utilized by the body. Lentil PDg is ~84% which is excellent relative to other crops (cf. oat 72%, wheat 77%, soybean 78%); however, plant proteins tend to have lower digestibility than animal proteins (cf. meat/poultry/fish 94%, milk 95%, egg 97%) (Gilbert *et al.*, 2011).

In order to ensure nutritional food security, lentil biofortification is being pursued for protein quality and other nutritional traits (Kumar *et al.*, 2016). Significant genetic variation has been observed for protein traits in lentil, including protein content, storage protein structure and weight, and amino acid concentrations (Alghamdi *et al.*, 2014; Ghumman *et al.*, 2019; Hang *et al.*, 2022). Accelerating the rate of genetic gain by reducing breeding cycle time is a primary objective in breeding programs, and the rise of genomics has allowed for significant gains through the use of genetic markers and molecular breeding approaches (Cobb *et al.*, 2019; Kumar *et al.*, 2021). Genome-wide association studies (GWAS) have been used extensively in plant science to identify genetic markers and candidate genes associated with a range of traits (Tibbs Cortes *et al.*, 2021). In lentil, GWAS has been used to identify markers associated with days to flower, seeds per pod, and 100 seed weight (Rajendran *et al.*, 2021); salinity tolerance (Dissanayake *et al.*, 2021); Aphanomyces root rot resistance (Ma *et al.*, 2020); Fe and Zn concentrations (Kumar *et al.*, 2019); prebiotic carbohydrate concentrations (Johnson *et al.*, 2021); as well as amino acid concentrations quantified by near-infrared spectroscopy (Hang,

2021; Hang *et al.*, 2022). Protein quality traits have also been explored via GWAS in cereal crops (Chen *et al.*, 2016) and legumes such as chickpea (Karaca *et al.*, 2019) and common bean (Katuuramu *et al.*, 2018). Genomic markers can be used to accelerate breeding efforts. One prominent method is marker-assisted selection which leverages genetic markers associated with a causal allele or gene to select for a trait, such as through introgression and backcrossing. This strategy has been used with much success in developing maize with high beta-carotene content (Muthusamy *et al.*, 2014) and rice with increased flood resistance (Bailey-Serres *et al.*, 2010).

Biofortification of lentil protein quality is a desirable breeding objective; however, genetic markers and genes associated with these traits are limited. Therefore, the objectives of this study were to (1) quantify protein quality traits (amino acids and PDg) in a lentil association mapping population grown under greenhouse conditions, (2) evaluate ancestral subpopulation global distribution and subpopulation effects on protein quality traits, and (3) identify SNP markers and candidate genes associated with these traits.

Results

Summary Statistics and Correlations

AA concentration means ranged from 0.21 to 4.45 % (Table 4.1). S-AAs, Met and Cys, had mean concentrations of 0.21 and 0.22%, respectively. The highest four mean AA concentrations were Arg (2.72%), H-Pro (3.16%), Asp (3.87%), and Glu (4.45%). The mean total AA concentration was 29.0%. Mean protein digestibility was 88% (Table 4.1). Figure S4.1 (Appendix B) displays histograms representing protein quality trait distributions as they compare to normal density curves of the same mean and standard deviation. However, Cys and, to a lesser degree, Ile and Lys appear bimodal. A 100 g serving of lentil, as estimated from trait means, would provide 100% of the recommended dietary allowance (RDA) of Ile, Leu, Lys, Phe, and Thr (Table

4.1). The same serving would provide 25 and 43 % of the RDA of Val and total AA, respectively. The lowest %RDA values were for lentil's limiting AAs, Met and Cys, at 19 and 20%, respectively. These are the S-AAs. Repeatability estimates for protein quality traits ranged from 3.8 to 34.9% (Table 4.1). The seven highest estimates were PDg (34.9%), Met (26.4%), Total AA (23.5%), Thr (22.4%), Asp (22.2%), Ile (21.5%), and Leu (21.4%). Repeatability estimates for ratios of individual AAs to total AA ranged from 0 to 27% (Table S4.1 [Appendix B]). These estimates were 4.8% lower on average from their respective non-ratio AAs. Met:TA was the exception with a slightly higher estimate (27%) than Met (26.4%).

The following traits were strongly correlated with one another ($r > 0.7$): Gly, Ser, Ala, Leu, Val, Asp, Ile, Met, Thr, Arg, Glu, and TA (Table 4.2). Lys had a moderate correlation ($r = 0.56–0.67$) with most of the strongly correlated traits, except a low correlation with Met ($r = 0.44$). Leu and Ile had the strongest correlation of 0.99. The sulfur-containing amino acids, cystine and methionine, had a moderate correlation of 0.68. Pro and H-Pro also had a moderate correlation of 0.66. PDg had only three correlations with $r > 0.23$; these were Asp ($r = 0.40$), Arg ($r = 0.39$), and TA ($r = 0.28$). Most trait correlations were significant at $p < 0.05$; however, H-Pro, His, and PDg were noteworthy for having seven, four, and four insignificant correlations, respectively.

Population Structure and Subpopulation Trait Differences

The lentil diversity panel was determined to have six ancestral subpopulations by ADMIXTURE analysis (Figure 4.1b). Subpopulations 1 through 6 were composed of 32, 13, 25, 46, 15, and 27 accessions, respectively. Mostly subtle visual associations were seen between subpopulations and regions of origin (Figure 4.1a). Subpopulation six was mostly absent from countries in North and South America. Accessions from the United States and Canada were composed mostly of clusters four and five, while Syrian accessions saw substantial representation

of all six subpopulation clusters. Subpopulation two showed little admixture with the other subpopulations (Figure 4.1b) and was seen almost exclusively represented by the accessions originating from Syria (Figure 4.1a), which was the center for the ICARDA lentil breeding program. The first three principal components (PCs) of the principal component analysis (PCA) accounted for 13.9, 10.1, and 5.1% of the total variance. Clear separation of ADMIXTURE clusters was seen in the PCA scatter plots (Figure 4.1c & Figure 4.1d). Accessions within subpopulation two were tightly clustered on the PCA scatterplots and were clearly delineated from the other clusters by PC3 (Figure 4.1d).

When analysis of variance was performed between ADMIXTURE clusters for each protein quality trait, significantly different ($p < 0.05$) means were identified for Ala, Arg, Cys, and His (Figure 4.2). Pair-wise comparision showed that cluster two had a mean in the highest letter category across all four traits, while cluster four had a mean in the lowest letter category across all four traits. Thus, across these four traits, clusters two and four always had significantly different means. Ala clusters two and six had significantly higher means than clusters three and four. Arg clusters one, two, and six had significantly higher means than cluster four, while clusters three and four had significantly lower means than cluster two. Cys clusters two and six had significantly higher means than clusters one and four. His clusters two and three had significantly higher means than cluster six, while clusters four and six had significantly lower means than cluster two.

Genome-Wide Association Studies

Fifty significantly associated SNPs were identified across 17 protein quality traits (Table 4.3; Figures 4.3 & S4.2 [Appendix B]). These SNPs were distributed across 46 linkage disequilibrium (LD) blocks, which contained a total of 157 genes (Tables 4.3 & 4.4; Supplemental Data: GWAS Exhaustive [https://github.com/njohns4/LentilProteinQualityGWAS]). Minor allele

frequencies of these SNPs in the final variant call file ranged from 0.3 to 47%. QQ plots showed that some models controlled for false positives better than others for certain traits (Figures S4.3, S4.4, & S4.5 [Appendix B]). For instance, the model SUPER showed pronounced deviation from the null hypothesis (red dotted line) for the traits Ala, Leu, Met, Asp, Arg, Cys, Phe, and Glu:TA. This indicates that the model poorly fit the trait data (Figures S4.3, S4.4, & S4.5 [Appendix B]). This inflation of *p*-values was noticeable on the Met Manhattan plot as well (Figure 4.3, yellow points). The model FarmCPU also had several traits where *p*-values deviated from the line early at approximately –log(3) (Figures S4.3, S4.4, & S4.5 [Appendix B]). Consequently, LD blocks associated with traits solely by one of these two models were excluded from Table 4.4. Thirteen traits were associated with SNPs identified by multiple GWAS models and include: Ala, Val, Leu, Ile, Thr, Met, Lys, Asp, Asp:TA, Met:TA, Gly:TA, His:TA, and PDg (Figure 4.3). Two LD blocks were associated with multiple traits (Figure 4.3, grey dashed boxes). The first of these blocks (Chr3_115394955–116212912) was associated with PDg and two aspartate family traits (Asp and Asp:TA). The block was 818 kb and contained 15 genes including four glutathione S-transferase genes and three protease family protein genes (Table 4.4). The second LD block (Chr3_424696277–424813245) was associated with three pyruvate family amino acids (Ala, Val, and Leu) and three aspartate family amino acids (Ile, Thr, and Met). The block was 117 kb and contained four genes: Lcu.2RBY.3g073770 (gibberellin-2-beta-dioxygenase), Lcu.2RBY.3g073780 (gibberellin-2-beta-dioxygenase), Lcu.2RBY.3g073790 (stem 28 kDa glycoprotein), and Lcu.2RBY.3g073800 (plant receptor-like kinase). SNP densities were visualized by the green to red scaled plots above chromosome numbers in Figures 4.3 and S4.2 (Appendix B). Densities were relatively low (0–21 SNPs per Mb) across most of the genome with regions near the end of chromosomes showing higher SNP densities (>30 SNPs per Mb).

Discussion

Amino acid concentrations agreed well with the U.S. Department of Agriculture's reference values (2019). The reference mean fell within the ranges reported here, with the exceptions of Cys (0.32% reference vs. 0.15–0.29%) and Val (1.22% vs. 0.14–0.26%), where the reference was higher, and Pro (1.03% vs. 1.44–3.96%), where the reference was lower (Table 4.1). Cys is a sulfur-containing amino acid known to degrade during acid hydrolysis. The method used included a pre-oxidative step using performic acid. This was intended to convert all Cys to cysteic acid, which is a more stable derivative. Nonetheless, some Cys is expected to be degraded, which may be the case here. Val has long been noted as being resistant to digestion when bonded to Val or Ile, so this may help explain the low Val concentration (Nair *et al.*, 1976). Although Pro is high compared to the reference, this concentration agrees with the literature (Salaria *et al.*, 2022). Notably, many of the standard reference values are low compared to the literature. Means and ranges also agreed well with values estimated by near-infrared spectroscopy, with the exceptions of high Arg and Pro and low Lys and Val (Hang, 2021). The mean total amino acid concentration (29%; Table 4.1) agreed well with ranges found in the literature for protein content (U.S. Department of Agriculture, 2019). The protein digestibility estimate (88%, Table 4.1) accords well with the range of 50–95% found in the literature (Shekib *et al.*, 1986; Monsoor & Yusuf, 2002; Martín-Cabrejas *et al.*, 2009). Percent recommended dietary allowance values reinforce that lentil is a good source of protein and essential amino acids, except for the limiting amino acids Met and Cys.

Repeatability is considered the upper bound of broad-sense heritability (Kruijer *et al.*, 2014). Protein content in food legumes is significantly affected by environment and genotype–environment effects (Pratap & Kumar, 2011). Heritability estimates (broad-sense) of amino acid

and protein concentrations vary across legume species and studies. The total amino acid repeatability estimate presented here (23.5%, Table 4.1) are low within this range. Heritability estimates of individual amino acids are not widely reported in legumes. However, in soybean, heritability estimates of individual amino acids ranged from 40.9% (Trp) to 81.8% (Asp) (Jiang & Katuuramu, 2021), while Met was high (99.7%) in chickpea (Desai *et al.*, 2015). Repeatability estimates (Table 4.1) ranged from 3.8% (Phe) to 26.4% (Met). The heritability of protein content in food legumes is moderate to high, ranging from 20 to 85% (Baudoin & Maquet, 1999; Pratap & Kumar, 2011; Patil *et al.*, 2020). Estimates in lentil fall within this range (Gautam *et al.*, 2018). The PDg repeatability estimate was 34.9% (Table 4.1), which was the highest of the protein quality traits. Protein digestibility heritability estimates are not widely reported for legumes. However, heritability estimates in sorghum range from 91 to 96% (Pfeiffer, 2017; Abdelhalim *et al.*, 2019). Repeatability estimates for lentil protein quality traits were low to moderate and comparable to literature values. This indicates that some of these traits, such as Met and PDg, are good breeding target traits for increased protein quality.

Strong correlations were seen between several protein quality traits, which is consistent with the literature (Wang & Daun, 2006; Hang, 2021). The present study found strong correlations ($r > 0.7$) between Gly, Ser, Ala, Leu, Val, Asp, Ile, Met, Thr, Arg, Glu, and total amino acid concentration (Table 4.2). Cys is also seen to have low to moderate correlations ($r = 0.2–0.7$) with other amino acids here and by Hang (2021). Interestingly, the trait that correlated strongest with Cys was Met, the other sulfur-containing amino acid. PDg was only weakly correlated ($r < 0.39$) with other protein quality traits. However, its strongest correlation was with Arg ($r = 0.39$), which is one of the amino acids incorporated into the *in vitro* PDg calculation. As

will be discussed below, several strongly correlated amino acids share significantly associated SNPs (Table 4.4).

ADMIXTURE analysis determined the optimal number of ancestral subpopulations to be k = 6 (Figure 4.1b). This is comparable but higher than other admixture analyses using the software STRUCTURE, which found between 3 and 5 ancestral subpopulations for *Lens culinaris* Medik. (Khazaei *et al.*, 2016; Kumar *et al.*, 2019; Pavan *et al.*, 2019; Liber *et al.*, 2021; Rajendran *et al.*, 2021). The absence of subpopulation six from most of the accessions in North and South America suggests either that germplasm from this subpopulation was not widely incorporated into these regions or that North and South American accessions from subpopulation six were simply not included in the present study. North and South American accessions were primarily from subpopulations four and five. In contrast, Syria, which includes the largest number of accessions (n = 35), contains substantial representation of all six subpopulations. This is not surprising since ICARDA, the source of the populations used in this study, was located in Syria. Subpopulation two is distinct because it shows relatively little admixture compared to the other subpopulations (Figure 4.1b), which suggests a highly related (inbred) group of accessions that is relatively distant genetically from the other subpopulations. This is further confirmed by the PCA (Figures 4.1c & 4.1d), which shows that accessions classified as subpopulation two are tightly clustered and clearly delineated from the other clusters by PC 3 (Figure 4.1d). Interestingly, subpopulation two is also primarily represented in the accessions originating from Syria.

Analysis of variance showed significant differences between the means of subpopulations across accessions for Ala, Arg, Cys, and His (Figure 4.2). Subpopulations two and four had the highest and lowest respective means for each trait. This suggests that genetically and

phenotypically divergent accessions could be selected from these subpopulations for recombinant population development. Additionally, high Ala and Arg accessions could be selected from subpopulations one, two, five, or six. High Cys accessions could be selected from subpopulations two, three, five, or six. And high His accessions could be selected from subpopulations one, two, three, or five.

Genome-wide association studies resulted in identification of fifty significantly associated SNPs and 157 genes across 46 linkage disequilibrium (LD) blocks and 17 protein quality traits (Table 4.3, Table 4.4, Supplemental Data: GWAS Exhaustive [https://github.com/njohns4/LentilProteinQualityGWAS]). Most of the significant SNPs had minor allele frequencies below 0.10 (Table 4.3). The number of false positives in GWAS does increase at lower minor allele frequencies (Tabangin *et al.*, 2009); however, many causative mutations are expected to occur at low frequencies in a population due to purifying selection (Tibbs Cortes *et al.*, 2021). Consequently, these SNPs and the genes in local LD with them should be investigated for their effect on protein quality traits but with informed caution. All significant SNPs with minor allele frequencies above 0.32 were detected exclusively by either the model SUPER or FarmCPU. These models deviated significantly from the null hypothesis for many traits as can be seen by the elevated SNP *p*-values in QQ plots (Figures S4.3, S4.4, & S4.5 [Appendix B]) and even some Manhattan plots (Met, Figure 4.3). SNPs associated with traits exclusively by one of these two models were excluded from Table 4.4 because they have a higher chance of being false positives. These SNPs and the genes in local LD with them should be pursued only with great caution. LD blocks contained only one or two SNPs per block. This is not surprising due to the low SNP density observed here (22,280 SNPs / 3.69 Gb reference genome = ~ 6 SNPs / Mb; Figure 4.3). However, this could also be indicative of false positives.

Two LD blocks are noteworthy because they were associated with multiple traits by multiple models (Table 4.4, Figure 4.3). Chr3_115394955–116212912 was associated with Asp, Asp:TA, and PDg. Interestingly, Asp was only weakly correlated with PDg (r = 0.40, Table 4.2). Although amino acids and their corresponding ratio with TA (such as Asp and Asp:TA) might be expected to share significant associations, this was the only pair that was identified. (It might also be expected that ratios of TA may share associations with TA; however, since TA was not significantly associated with any marker, this was not the case.) This LD block was 818 kb and contained 15 genes. Four genes were identified as glutathione S-transferase genes by homology. Glutathione S-transferase genes are a supergene family whose products aid in neutralizing toxins by helping facilitate the anti-oxidative activity of glutathione (Gullner *et al.*, 2018). These genes are upregulated during stress, as has been shown in lentil under arsenic stress (Talukdar, 2016). A glutathione S-transferase gene has also been proposed as a candidate gene for Verticillium wilt disease resistance in Arabidopsis (Gong *et al.*, 2018). The role of glutathione S-transferase in protein quality is unclear. Chr3_115394955–116212912 also contained three protease family genes. Proteases degrade unwanted proteins and maintain protein quality in plant cells (García-Lorenzo *et al.*, 2006). Protease inhibitors in grain reduce the activity of protease enzymes during animal digestion; consequently, seed with higher concentrations of protease inhibitors lower protein digestibility (Singh & Jambunathan, 1981). It is hypothesized that regulation of protease genes and protease inhibitor genes are synchronized within the plant leading to this association with protein digestibility.

The second LD block associated with multiple traits was Chr3_424696277–424813245, which was associated with Ala, Ile, Leu, Met, Thr, and Val (Figure 4.3). These traits were highly correlated (Table 4.2). The block contained four genes, two of which were gibberellin 2-beta-

dioxygenase genes. This gene family is involved in numerous developmental processes in plants, such as seed germination, leaf expansion, shoot/stem lengthening, and reproductive structures and processes (Wang *et al.*, 2014). It is likely, therefore, that these genes would affect protein quality traits; however, altering expression may affect numerous traits besides protein quality traits. The other two genes were a stem 28 kDa glycoprotein gene and a plant receptor-like kinase, both of which are broad descriptors, requiring functional analysis for further investigation.

Conclusion

Lentil's high concentration of high-quality plant-based protein makes it a prime candidate for protein biofortification. To that end, this study measured protein quality traits in a lentil diversity panel. These traits included 17 amino acids, total amino acid content, and protein digestibility. The ratios of individual amino acids to total amino acid content were also evaluated. Correlations between traits were measured. Admixture analysis revealed six lentil ancestral subpopulations represented in the population, and global distribution of these subpopulations revealed subpopulations to differ by mean concentrations of Ala, Arg, Cys, and His. Subpopulation two was found to be unique to accessions originating from Syria. Genome-wide association studies associated 50 SNPs with 17 protein quality traits, 42 LD blocks, and 157 genes. Future studies are needed to evaluate candidate genes, especially glutathione S-transferase, protease family, and gibberellin 2-beta-dioxygenase genes.

Materials and Methods

Diversity Panel Composition

Two mapping populations obtained from the International Center for Agricultural Research in the Dry Areas (ICARDA) were grown with two replicate pots per accession in the Clemson greenhouse complex in 2018 (Johnson *et al.*, 2021). Samples from the two populations were combined for analysis. After accounting for population overlap and low yields from some accessions and replicates, 183 unique accessions with 1–4 replicates each were analyzed for protein quality traits.

Amino Acid Analysis

Reagents, solvents, and high-purity standards for amino acid analysis were purchased from Sigma Aldrich Co. (St. Louis, MO), Fisher Scientific (Waltham, MA), and VWR International (Radnor, PA). Ultrapure water was used in all analyses (PURELAB flex 2 system, ELGA LabWater North America, Woodridge, IL). The amino acid analysis is reported elsewhere (Madurapperumage *et al.,* 2022) as an adaptation from the literature (Gehrke *et al.*, 1985; Manneberg *et al.*, 1995). In brief, 40 mg of lentil flour (particle size ≤ 0.5 mm) was weighed into glass culture tubes (16 x 125 mm, PTFE lined cap). Performic acid was synthesized from formic acid and hydrogen peroxide (9:1 ratio). Once chilled in an ice bath, 5 mL of performic acid was added to each tube and gently swirled on a vortex mixer before being capped and refrigerated for 16 hr to convert Cys and Met to derivatives, methionine sulfone and cysteic acid, which are more stable under acid hydrolysis. A 1/8 in. x tube length PTFE boiling rod was inserted into each tube before being evaporated to dryness in a vacuum oil bath (3 gal. resin trap, BACOENG, Suzhou, China) at ~70–80 ˚C and ~610 mmHg. Once cooled, tubes were removed and 4.9 mL of 6 M HCl and 0.1 mL of internal standard mix (25 mM norvaline, 25 mM sarcosine) was added to each tube

before being capped and gently swirled. Tubes were then placed in a gravity convection oven at 110 ˚C for 24 hr to hydrolyze peptide bonds. Samples were cooled to room temperature, vortex mixed, and filtered through a 0.22 µm polypropylene syringe filter. One mL of sample was added to a clean culture tube and evaporated to dryness as before. Samples were rehydrated with 1 mL of HPLC mobile phase A and pipetted into HPLC vials for analysis.

Amino acid analysis was performed via high-performance reverse phase chromatography on a 1100 series Agilent system (Agilent Technologies, Santa Clara, CA, USA) according to a method adapted from Agilent application notes (Agilent Application Note, 2010; Long, 2015). Amino acids were detected on a diode array detector at two wavelengths (338 nm, 10 nm bandwidth, reference 390 nm, 20 nm bandwidth and 262 nm, 10 nm bandwidth, reference 390 nm, 20 nm bandwidth). An aqueous and an organic solvent were used for mobile phase A and B, respectively. Mobile phase A contained 10mM sodium phosphate, 10 mM sodium tetraborate decahydrate, and 5 mM sodium azide with a pH adjusted to 8.2 with 12 M HCl. The solution was then filtered through 0.2 µm regenerated cellulose. Mobile phase B consisted of 45% methanol, 45% acetonitrile, and 10% water (v/v/v). Injection diluent was prepared by adding 0.4 mL concentrated phosphoric acid to 100 mL mobile phase A. An Agilent Poroshell HPH-C18 analytical column (3 x 100 mm) in series with the corresponding guard column (3 x 5 mm) were used for separation of amino acids. A gradient method was employed with linear adjustment between the following times (concentration mobile phase A, flow rate): 0.0 min (100%, 0.25 mL/min), 3.0 min (100%, 0.25 mL/min), 10.4 min (81.5%, 0.62 mL/min), 23.0 min (43%, 0.62 mL/min), 23.1 min (0%, 0.62 mL/min), 27 min (0%, 0.62 mL/min), 27.1 min (100%, 0.62 mL/min), 27.9 min (100%, 0.63 mL/min), 28 min (100%, 0.25 mL/min,), and 33 min (100%, 0.25 mL/min). Column temperature was maintained at 40 ˚C. Online sample derivatization with *o*-

phthalaldehyde (OPA) and 9-fluorenylmethyl chloroformate (FMOC) was performed by the G1329A autosampler in 13 steps: draw 2.5 µL Agilent borate buffer, draw 0.5 µL sample, mix 3 µL air five times, wait 0.2 min, draw 0 µL water, draw 0.5 µL Agilent OPA (vial insert) using 2 mm offset, mix 3.5 µL air six times, draw 0 µL water, draw 0.4 µL Agilent FMOC (vial insert) using 2 mm offset, mix 3.9 µL air ten times, draw 32 µL injection diluent, mix 20 µL air eight times, and inject. A lab reference lentil sample was included in every digestion batch to monitor batch-to-batch variation and an amino acid standard mix was run on HPLC before analyzing each batch of samples. Calibration standards (9–900 pmol/µL) with internal standards norvaline and sarcosine (500 pmol/µL) were run, and linear calibration models were generated based on peak areas for calculating sample amino acid concentrations, which were converted into percent of lentil flour. Total amino acid concentration was calculated by summing all amino acid concentrations for each sample. The percent of total AA concentration was calculated for each amino acid. Consequently, the 17 amino acid concentrations resulted in 35 amino acid traits—17 amino acids, 17 amino acid percent of total AA (AA:TA), and total AA concentration.

In Vitro *Protein Digestibility Analysis*

Protein digestibility (PDg) was measured using the Megazyme Protein Digestibility Amino Acid Score assay kit with modified protocol for a 100 mg sample size (Megazyme 2019). The protocol was followed precisely except all masses and volumes were divided by 5. Due to expected underestimation of some amino acids caused by acid hydrolysis, reference amino acid values from the U.S. Department of Agriculture FoodData Central (2019) were used for the PDg calculations; these included: Pro (1.03%), Lys (1.72%), His (0.69%), and Arg (1.90%). The Megazyme Excel calculator was modified to change the approximate sample mass from 0.5 g to 0.1 g. In addition

to the controls included in the assay kit, a lab reference lentil sample was included in every batch to monitor batch-to-batch variation.

Summary Statistics and Correlations

Protein quality trait means, standard deviations, and ranges were calculated for each accession in JMP 14.0.0 (Tables 4.1 & S4.1 [Appendix B]). Histograms of trait distributions were fit with density curves for the normal distribution using estimates of the mean and standard deviation (Figure S4.1 [Appendix B]). Percent recommended dietary allowance estimates were calculated for the essential amino acids Cys, His, Iso, Leu, Lys, Met, Phe, Thr, and Val as well as for total AA concentration. Estimates were for a 72 kg adult consuming 100 mg of lentil (15% moisture content) per day given the following dietary requirements: 8–12 mg/kg His, 10 mg/kg Iso, 14 mg/kg Leu, 12 mg/kg Lys, 13 mg/kg Met + Cys, 14 mg/kg Phe + Tyr, 10 mg/kg Val, and 0.8 g/kg protein (National Research Council Subcommittee on the Tenth Edition of the Recommended Dietary Allowances, 1989). To estimate repeatability, a model was developed with trait concentration as the response variable and genotype as a random effect. Repeatability is the proportion of phenotypic variance attributable to genetic variance and provides an upper bound to broad-sense heritability (H^2). Repeatability equals H^2) when all differences between genotypes are assumed to be genetic (Kruijer *et al.*, 2014). Pearson's correlation coefficients (r) were calculated in JMP for protein quality traits using accession means across replicates (Table 4.2).

Genome-Wide Association Studies

A previously generated VCF file was used for genetic analyses (Johnson *et al.*, 2021). In brief, the TASSEL-GBS pipeline (Glaubitz *et al.*, 2014) was used to process raw genotyping-by-sequencing data (Amin, 2018) into SNP genotypes. The *Lens culinaris* CDC Redberry Genome Assembly v2.0 (Ramsay *et al.*, 2021) was used as a reference genome. SNPs identified in contigs

not incorporated into assembled chromosomes were removed from the analysis. Variants were filtered using VCFtools (Danecek *et al.*, 2011) to include only biallelic SNPs (--min-alleles 2 –max-alleles 2) with a 5% minimum minor allele frequency (--maf 0.05) and a maximum of 20% missing genotypes (--max-missing 0.2). Missing genotypes were then imputed using Beagle 5.4 (Browning *et al.*, 2018). Default parameters were used except for effective population size, which was set to ne = 100,000. Genotypes without AA data were removed, and chromosomes were renamed to integers (1–7) using BCFtools. The final VCF file contained 158 genotypes and 22,280 SNPs.

To mitigate batch effects from the amino acid and PDg analyses, Bayesian random effects (cf. BLUPs) were used instead of means in the genome-wide association study. Parameter estimates for the effect of genotype were calculated using the stanarm version 2.21.3 package in R (Goodrich *et al.*, 2022) by fitting the following model:

$$y = (1|\text{Genotype}) + (1|\text{Batch})$$

where y is the observed mean and *Genotype* and *Batch* are random effects.

Genome-wide associations were performed using the Genome Association and Prediction Integrated Tool (GAPIT) version 3 package in R (Wang & Zhang, 2021) using default settings. GAPIT's model selection with Bayesian Information Criterion feature determined that the kinship matrix sufficiently accounted for population structure and so principal components were not included in the analyses. However, a separate GAPIT analysis was performed to calculate principal component eigen values for later visualization with the admixture analysis (Figures 4.1c & 4.1d). The following models were employed for association analyses: Generalized Linear Model (GLM), Mixed Linear Model (MLM), Multiple Loci Mixed Model (MLMM), Compressed MLM (CMLM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Fixed and

random model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK). A Bonferroni threshold (0.05 / 22,280 = 2.24 x 10^{-6}) was used to determine significance. Manhattan plots (Figures 4.3 & S4.2 [Appendix B]) and QQ plots were drawn using the CMplot version 4.1.0 package in R (https://github.com/YinLiLin/CMplot).

LD blocks were determined using PLINK v1.07 (Purcell *et al.*, 2007) by calculating pair-wise correlations (r^2) of significant SNPs with adjacent SNPs within a 1 Mb window. LD blocks were determined to decay either at the first SNP with $r^2 < 0.4$ or at 100 kb past the final linked SNP, whichever was less. The 100 kb provision was to help account for the low SNP density of many regions of the genome which otherwise resulted in highly inflated LD block sizes. Genes within local LD with significant SNPs were identified using a custom python script (https://github.com/jlboat/features_from_snps) and were considered candidate genes.

Population Structure and Origin Analysis

Population structure was estimated using ADMIXTURE (Alexander & Lange, 2011). The optimal number of ancestral populations (K = 6) was determined by selecting the model with the lowest cross-validation error using five-fold cross-validation. The model generated a Q matrix containing ancestral coefficients for each genotype. Accessions were categorized into subpopulations based on their highest ancestry coefficient (> 0.5). An admixture plot (Figure 4.1b) was drawn using the R package gglot2 version 3.3.6 (Wickham, 2016).

The global distribution of accessions by ancestral subpopulation was then visualized (Figure 4.1a). ISO3 country codes were first assigned to accessions resulting in 52 unique countries of origin. Origin information was missing from two accessions resulting in the inclusion of 156 accessions in the figure. The ddply function in the R package plyr version 1.8.7 (Wickham, 2011)

was used to calculate the mean ancestral coefficients for each country of origin. These values were then multiplied by the number of accessions per country before being translated into a SpatialPolygonsDataFrame using the function joinCountryData2Map in rworldmap version 1.3-4 (South, 2011). The mapPies function in rworldmap was then used to draw a figure displaying ancestral population pie charts for each country of origin (Figure 4.1a). Pie charts depict average admixture composition of accessions from the same country of origin. Pie chart circumferences are proportional to the number of accessions sharing a country of origin. PCA scatter plots with accessions classified by ADMIXTURE subpopulation were drawn in ggplot2 using PCA variance components calculated in GAPIT (Figures 4.1c & 4.1d).

Analysis of variance was performed to determine if ancestral group had a significant effect on protein quality traits. Trait means across accession replicates, previously calculated in JMP, were combined with ADMIXTURE subpopulations and a model was developed in JMP with trait concentration as the response variable and subpopulation as a fixed effect. For models with a significant effect (Ala, Arg, Cys, and His), Fisher's protected LSD procedure was used to determine differences between subpopulations. Boxplots were drawn in JMP. All figures received final formatting using Adobe Illustrator 2019.

Code, Data, and Software Availability

Code and data can be found at https://github.com/njohns4/LentilProteinQualityGWAS. Linux shell and R loop scripts were used extensively to perform the multiplicative analyses and visualizations (36 traits x 7 models). Custom python scripts were used to generate an sqlite3 database from a GFF file (https://github.com/daler/gffutils) and to subsequently extract candidate genes from the database (https://github.com/jlboat/features_from_snps). The *Lens culinaris* CDC Redberry Genome Assembly v2.0 is available at https://knowpulse.usask.ca/genome-

assembly/Lcu.2RBY. The following software was used: BCFtools (Danecek *et al.*, 2021), Beagle 5.4 (Browning *et al.*, 2018), Megazyme Protein Digestibility Calculator (https://www.megazyme.com/documents/Data_Calculator/K-PDCAAS_CALC.xlsx), CMplot (https://github.com/YinLiLin/CMplot), GAPIT3 (Wang & Zhang, 2021), PLINK (Purcell *et al.*, 2007), TASSEL5 (Glaubitz *et al.*, 2014), and VCFtools (Danecek *et al.*, 2011).

Acknowledgements

This project was supported by the Good Food Institute and the USDA National Institute of Food and Agriculture including: the Pulse Health Initiative, [Hatch] project [1022664], and the Organic Agriculture Research and Extension Initiative (OREI) (award no. 2018-51300-28431/proposal no. 2018--02799).

Tables and Figures

Table 4.1: Mean concentration, concentration range, repeatability estimates , and %RDA for lentil protein quality traits

Trait	Mean (%) ± SD	Range (%)	%RDA[a]	Repeatability %[b]
Ala	1.24 ± 0.11	0.77–1.55	N/A	20.0
Arg	2.72 ± 0.33	1.53–3.62	N/A	19.5
Asp	3.87 ± 0.48	2.29–6.04	N/A	22.2
Cys	0.22 ± 0.03	0.15–0.29	20	18.8
Glu	4.45 ± 0.40	2.74–5.63	N/A	19.7
Gly	1.25 ± 0.11	0.81–1.56	N/A	19.3
His	0.61 ± 0.10	0.18–1.00	60–90	14.2
H-Pro	3.16 ± 0.95	1.5–6.01	N/A	11.6
Ile	1.24 ± 0.11	0.77–1.52	100	21.5
Leu	2.25 ± 0.20	1.37–2.8	100	21.4
Lys	1.33 ± 0.23	0.56–1.89	100	19.5
Met	0.21 ± 0.02	0.11–0.27	19	26.4
Phe	1.29 ± 0.23	0.39–1.72	100	3.8
Pro	2.49 ± 0.49	1.44–3.96	N/A	18.1
Ser	1.40 ± 0.13	0.85–1.76	N/A	20.7
Thr	1.05 ± 0.10	0.6–1.3	100	22.4
Val	0.21 ± 0.02	0.14–0.26	25	18.2
Total AA	28.99 ± 2.90	18.27–36.05	43	23.5
PDg	88 ± 1	0.86–0.91	N/A	34.9

[a] Percent recommended dietary allowance (%RDA) estimates were calculated for a 72 kg adult consuming a 100 g serving of lentil (15% moisture content) per day. [b] Repeatability is the proportion of phenotypic variance attributable to genetic variance and provides an upper bound to broad-sense heritability (H^2) (Kruijer *et al.,* 2014).

Table 4.2: Correlations between protein quality traits

	Serine Family			Pyruvate Family			Aspartate Family					Glutamate Family				Other			
Trait	**Cys**	**Gly**	**Ser**	**Ala**	**Leu**	**Val**	**Asp**	**Ile**	**Lys**	**Met**	**Thr**	**Arg**	**Glu**	**Pro**	**H-Pro**	**His**	**Phe**	**TA**	**PDg**
Cys	1.00*	0.36*	0.48*	0.49*	0.51*	0.41*	0.38*	0.49*	0.26*	0.68*	0.56*	0.44*	0.44*	0.54*	0.07	-0.03	0.20*	0.47*	0.12
Gly	0.36*	1.00*	0.94*	0.95*	0.93*	0.94*	0.87*	0.93*	0.62*	0.77*	0.93*	0.86*	0.94*	0.42*	0.16*	0.22*	0.46*	0.85*	0.20*
Ser	0.48*	0.94*	1.00*	0.93*	0.94*	0.89*	0.84*	0.93*	0.58*	0.82*	0.95*	0.82*	0.95*	0.47*	0.17*	0.14	0.43*	0.85*	0.15*
Ala	0.49*	0.95*	0.93*	1.00*	0.97*	0.96*	0.89*	0.97*	0.64*	0.82*	0.96*	0.89*	0.94*	0.50*	0.15*	0.21*	0.51*	0.88*	0.20*
Leu	0.51*	0.93*	0.94*	0.97*	1.00*	0.95*	0.87*	0.99*	0.64*	0.81*	0.96*	0.88*	0.94*	0.52*	0.12	0.19*	0.55*	0.87*	0.16*
Val	0.41*	0.94*	0.89*	0.96*	0.95*	1.00*	0.89*	0.96*	0.64*	0.74*	0.91*	0.88*	0.94*	0.46*	0.16*	0.25*	0.48*	0.86*	0.22*
Asp	0.38*	0.87*	0.84*	0.89*	0.87*	0.89*	1.00*	0.86*	0.60*	0.74*	0.84*	0.87*	0.84*	0.44*	0.2*	0.21*	0.46*	0.85*	0.40*
Ile	0.49*	0.93*	0.93*	0.97*	0.99*	0.96*	0.86*	1.00*	0.67*	0.78*	0.96*	0.87*	0.93*	0.49*	0.11	0.20*	0.53*	0.86*	0.15*
Lys	0.26*	0.62*	0.58*	0.64*	0.64*	0.64*	0.60*	0.67*	1.00*	0.44*	0.64*	0.65*	0.57*	0.20*	-0.09	0.37*	0.49*	0.56*	0.07
Met	0.68*	0.77*	0.82*	0.82*	0.81*	0.74*	0.74*	0.78*	0.44*	1.00*	0.86*	0.71*	0.77*	0.49*	0.15*	0.05	0.33*	0.73*	0.19*
Thr	0.56*	0.93*	0.95*	0.96*	0.96*	0.91*	0.84*	0.96*	0.64*	0.86*	1.00*	0.85*	0.92*	0.52*	0.15*	0.02*	0.49*	0.86*	0.15*
Arg	0.44*	0.86*	0.82*	0.89*	0.88*	0.88*	0.87*	0.87*	0.65*	0.71*	0.85*	1.00*	0.85*	0.50*	0.12	0.30*	0.50*	0.84*	0.39*
Glu	0.44*	0.94*	0.95*	0.94*	0.94*	0.94*	0.84*	0.93*	0.57*	0.77*	0.92*	0.85*	1.00*	0.43*	0.11	0.16*	0.42*	0.83*	0.20*
Pro	0.54*	0.42*	0.47*	0.50*	0.52*	0.46*	0.44*	0.49*	0.2*	0.49*	0.52*	0.50*	0.43*	1.00*	0.66*	0.33*	0.38*	0.77*	0.19*
H-Pro	0.07	0.16*	0.17*	0.15*	0.12	0.16*	0.20*	0.11	-0.09	0.15*	0.15*	0.12	0.11	0.66*	1.00*	0.32*	0.00	0.55*	0.16*
His	-0.03	0.22*	0.14	0.21*	0.19*	0.25*	0.21*	0.20*	0.37*	0.05	0.20*	0.30*	0.16*	0.33*	0.32*	1.00*	0.40*	0.40*	0.06
Phe	0.20*	0.46*	0.43*	0.51*	0.55*	0.48*	0.46*	0.53*	0.49*	0.33*	0.49*	0.50*	0.42*	0.38*	0.00	0.40*	1.00*	0.52*	0.00
TA	0.47*	0.85*	0.85*	0.88*	0.87*	0.86	0.85*	0.86*	0.56*	0.73*	0.86*	0.84*	0.83*	0.77*	0.55*	0.40*	0.52*	1.00*	0.28*
PDg	0.12	0.20*	0.15*	0.20*	0.16*	0.22	0.40*	0.15*	0.07	0.19*	0.15*	0.39*	0.20*	0.19*	0.16*	0.06	0.00	0.28*	1.00*

Correlation coefficients greater than 0.70 were considered strongly correlated and are in shaded boxes. * Significant at $p < 0.05$. TA = total amino acid concentration. PDg = *in vitro* protein digestibility

Table 4.3: Protein quality traits with significantly associated SNPs and candidate genes.

Traits	SNPs[a]	Models	*p*-value[b]	maf[c]	Genes[d]
Alanine	CHR1_111531458	FarmCPU	3E-07	0.06	Lcu.2RBY.1g016210, Lcu.2RBY.1g016220, Lcu.2RBY.1g016230, Lcu.2RBY.1g016240
	CHR2_55767900	FarmCPU	1E-06	0.06	
	CHR3_424796277	BLINK, GLM, MLM, MLMM	3E-09	0.04	Lcu.2RBY.3g073770, Lcu.2RBY.3g073780, Lcu.2RBY.3g073790, Lcu.2RBY.3g073800
Arginine	CHR6_293947596	SUPER	2E-06	0.39	Lcu.2RBY.6g041200, Lcu.2RBY.6g041210, Lcu.2RBY.6g041220, Lcu.2RBY.6g041230, Lcu.2RBY.6g041240, Lcu.2RBY.6g041250, Lcu.2RBY.6g041260, Lcu.2RBY.6g041270, Lcu.2RBY.6g041280, Lcu.2RBY.6g041290, Lcu.2RBY.6g041300, Lcu.2RBY.6g041310, Lcu.2RBY.6g041320, Lcu.2RBY.6g041330, Lcu.2RBY.6g041340, Lcu.2RBY.6g041350, Lcu.2RBY.6g041360, Lcu.2RBY.6g041370
	CHR6_293947618	SUPER	2E-06	0.39	Lcu.2RBY.6g041200, Lcu.2RBY.6g041210, Lcu.2RBY.6g041220, Lcu.2RBY.6g041230, Lcu.2RBY.6g041240, Lcu.2RBY.6g041250, Lcu.2RBY.6g041260, Lcu.2RBY.6g041270, Lcu.2RBY.6g041280, Lcu.2RBY.6g041290, Lcu.2RBY.6g041300, Lcu.2RBY.6g041310, Lcu.2RBY.6g041320, Lcu.2RBY.6g041330, Lcu.2RBY.6g041340, Lcu.2RBY.6g041350, Lcu.2RBY.6g041360, Lcu.2RBY.6g041370
Aspartate	CHR3_115494955	MLMM	2E-06	0.02	Lcu.2RBY.3g018060, Lcu.2RBY.3g018070, Lcu.2RBY.3g018080, Lcu.2RBY.3g018090, Lcu.2RBY.3g018100, Lcu.2RBY.3g018110, Lcu.2RBY.3g018120, Lcu.2RBY.3g018130, Lcu.2RBY.3g018140, Lcu.2RBY.3g018150, Lcu.2RBY.3g018160, Lcu.2RBY.3g018170, Lcu.2RBY.3g018180, Lcu.2RBY.3g018190, Lcu.2RBY.3g018200

	CHR3_115582822	MLMM	2E-06	0.02	Lcu.2RBY.3g018060, Lcu.2RBY.3g018070, Lcu.2RBY.3g018080, Lcu.2RBY.3g018090, Lcu.2RBY.3g018100, Lcu.2RBY.3g018110, Lcu.2RBY.3g018120, Lcu.2RBY.3g018130, Lcu.2RBY.3g018140, Lcu.2RBY.3g018150, Lcu.2RBY.3g018160, Lcu.2RBY.3g018170, Lcu.2RBY.3g018180, Lcu.2RBY.3g018190, Lcu.2RBY.3g018200
Aspartate:TotalAA	CHR3_115494955	CMLM, GLM, MLM, MLMM	6E-08	0.02	Lcu.2RBY.3g018060, Lcu.2RBY.3g018070, Lcu.2RBY.3g018080, Lcu.2RBY.3g018090, Lcu.2RBY.3g018100, Lcu.2RBY.3g018110, Lcu.2RBY.3g018120, Lcu.2RBY.3g018130, Lcu.2RBY.3g018140, Lcu.2RBY.3g018150, Lcu.2RBY.3g018160, Lcu.2RBY.3g018170, Lcu.2RBY.3g018180, Lcu.2RBY.3g018190, Lcu.2RBY.3g018200
	CHR3_115582822	CMLM, GLM, MLM	7E-07	0.02	Lcu.2RBY.3g018060, Lcu.2RBY.3g018070, Lcu.2RBY.3g018080, Lcu.2RBY.3g018090, Lcu.2RBY.3g018100, Lcu.2RBY.3g018110, Lcu.2RBY.3g018120, Lcu.2RBY.3g018130, Lcu.2RBY.3g018140, Lcu.2RBY.3g018150, Lcu.2RBY.3g018160, Lcu.2RBY.3g018170, Lcu.2RBY.3g018180, Lcu.2RBY.3g018190, Lcu.2RBY.3g018200
	CHR5_271369658	FarmCPU	1E-07	0.04	
	CHR5_445247913	FarmCPU	2E-08	0.19	Lcu.2RBY.5g065860, Lcu.2RBY.5g065870, Lcu.2RBY.5g065880, Lcu.2RBY.5g065890, Lcu.2RBY.5g065900, Lcu.2RBY.5g065910
	CHR6_335566577	FarmCPU	7E-08	0.19	Lcu.2RBY.6g049180, Lcu.2RBY.6g049190
	CHR7_185325635	FarmCPU	2E-06	0.05	
	CHR7_185744579	FarmCPU	6E-07	0.16	Lcu.2RBY.7g029130, Lcu.2RBY.7g029140, Lcu.2RBY.7g029150, Lcu.2RBY.7g029160, Lcu.2RBY.7g029170, Lcu.2RBY.7g029180, Lcu.2RBY.7g029190, Lcu.2RBY.7g029200
Cystine	CHR4_306390322	SUPER	1E-06	0.31	Lcu.2RBY.4g043680
	CHR4_317014401	SUPER	1E-06	0.25	Lcu.2RBY.4g045850, Lcu.2RBY.4g045860
	CHR4_317014408	SUPER	1E-06	0.25	Lcu.2RBY.4g045850, Lcu.2RBY.4g045860
Digestibility	CHR1_142769633	BLINK	5E-11	0.01	

	CHR1_330575676	CMLM, GLM, MLM	6E-07	0.02	Lcu.2RBY.1g040750, Lcu.2RBY.1g040760, Lcu.2RBY.1g040770, Lcu.2RBY.1g040780, Lcu.2RBY.1g040790
	CHR2_14447425	CMLM, GLM, MLM	2E-08	0.05	Lcu.2RBY.2g006630, Lcu.2RBY.2g006640
	CHR2_484921441	CMLM, GLM, MLM	9E-07	0.04	Lcu.2RBY.2g074730, Lcu.2RBY.2g074740, Lcu.2RBY.2g074750
	CHR3_115494955	BLINK, FarmCPU, MLMM	2E-12	0.02	Lcu.2RBY.3g018060, Lcu.2RBY.3g018070, Lcu.2RBY.3g018080, Lcu.2RBY.3g018090, Lcu.2RBY.3g018100, Lcu.2RBY.3g018110, Lcu.2RBY.3g018120, Lcu.2RBY.3g018130, Lcu.2RBY.3g018140, Lcu.2RBY.3g018150, Lcu.2RBY.3g018160, Lcu.2RBY.3g018170, Lcu.2RBY.3g018180, Lcu.2RBY.3g018190, Lcu.2RBY.3g018200
	CHR3_288258714	FarmCPU	1E-07	0.03	Lcu.2RBY.3g044660, Lcu.2RBY.3g044670
	CHR5_155570229	BLINK, MLMM	3E-13	0.01	Lcu.2RBY.5g028750, Lcu.2RBY.5g028760
	CHR6_289995023	FarmCPU	1E-08	0.24	Lcu.2RBY.6g040510, Lcu.2RBY.6g040520
	CHR7_244870870	CMLM, GLM, MLM	6E-10	0.00	
	CHR7_497443978	MLMM	3E-10	0.00	Lcu.2RBY.7g065930, Lcu.2RBY.7g065940, Lcu.2RBY.7g065950, Lcu.2RBY.7g065960
Glutamate:TotalAA	CHR1_137107598	SUPER	9E-07	0.47	Lcu.2RBY.1g018720, Lcu.2RBY.1g018730, Lcu.2RBY.1g018740
Glycine:TotalAA	CHR2_319072281	SUPER	1E-06	0.06	Lcu.2RBY.2g050080, Lcu.2RBY.2g050090
	CHR5_107992651	BLINK, FarmCPU, SUPER	3E-07	0.09	Lcu.2RBY.5g022850, Lcu.2RBY.5g022860, Lcu.2RBY.5g022870, Lcu.2RBY.5g022880
Histidine:TotalAA	CHR1_519949144	FarmCPU	1E-08	0.39	Lcu.2RBY.1g072230, Lcu.2RBY.1g072240
	CHR2_16164950	FarmCPU	5E-07	0.04	Lcu.2RBY.2g007300, Lcu.2RBY.2g007310
	CHR6_301590681	BLINK, GLM, MLMM	8E-09	0.04	Lcu.2RBY.6g042780, Lcu.2RBY.6g042790
Isoleucine	CHR3_424796277	BLINK, FarmCPU, MLMM	2E-08	0.04	Lcu.2RBY.3g073770, Lcu.2RBY.3g073780, Lcu.2RBY.3g073790, Lcu.2RBY.3g073800
Leucine	CHR2_601711657	FarmCPU	2E-08	0.23	Lcu.2RBY.2g093390, Lcu.2RBY.2g093400, Lcu.2RBY.2g093410, Lcu.2RBY.2g093420, Lcu.2RBY.2g093430, Lcu.2RBY.2g093440, Lcu.2RBY.2g093450

	CHR3_424796277	BLINK, CMLM, GLM, MLM, MLMM	3E-09	0.04	Lcu.2RBY.3g073770, Lcu.2RBY.3g073780, Lcu.2RBY.3g073790, Lcu.2RBY.3g073800
	CHR5_202476372	FarmCPU	8E-07	0.07	Lcu.2RBY.5g033320
Lysine	CHR2_292740642	BLINK	1E-06	0.14	Lcu.2RBY.2g046140
	CHR3_152229376	BLINK	9E-08	0.24	Lcu.2RBY.3g022880, Lcu.2RBY.3g022890, Lcu.2RBY.3g022900, Lcu.2RBY.3g022910, Lcu.2RBY.3g022920, Lcu.2RBY.3g022930, Lcu.2RBY.3g022940, Lcu.2RBY.3g022950
	CHR7_7686751	BLINK	9E-07	0.32	
Methionine	CHR1_141754068	SUPER	2E-07	0.34	Lcu.2RBY.1g019160
	CHR3_424796277	BLINK, MLMM	1E-07	0.04	Lcu.2RBY.3g073770, Lcu.2RBY.3g073780, Lcu.2RBY.3g073790, Lcu.2RBY.3g073800
	CHR4_209096920	SUPER	1E-06	0.22	Lcu.2RBY.4g029800, Lcu.2RBY.4g029810, Lcu.2RBY.4g029820, Lcu.2RBY.4g029830, Lcu.2RBY.4g029840, Lcu.2RBY.4g029850
	CHR4_209096949	SUPER	2E-06	0.27	Lcu.2RBY.4g029800, Lcu.2RBY.4g029810, Lcu.2RBY.4g029820, Lcu.2RBY.4g029830, Lcu.2RBY.4g029840, Lcu.2RBY.4g029850
	CHR5_11042934	SUPER	1E-06	0.21	Lcu.2RBY.5g006620, Lcu.2RBY.5g006630, Lcu.2RBY.5g006640, Lcu.2RBY.5g006650
	CHR5_167207846	SUPER	2E-06	0.23	Lcu.2RBY.5g030100, Lcu.2RBY.5g030110
	CHR6_326624017	SUPER	1E-06	0.41	Lcu.2RBY.6g047800
Methionine:TotalAA	CHR1_518846076	FarmCPU	2E-11	0.09	Lcu.2RBY.1g071860, Lcu.2RBY.1g071870, Lcu.2RBY.1g071880, Lcu.2RBY.1g071890, Lcu.2RBY.1g071900, Lcu.2RBY.1g071910
	CHR1_96093588	FarmCPU	6E-08	0.06	Lcu.2RBY.1g014500
	CHR2_6756842	FarmCPU	9E-09	0.03	
	CHR5_214999927	BLINK, GLM	9E-10	0.20	
	CHR6_42597339	FarmCPU	2E-06	0.06	Lcu.2RBY.6g006710
	CHR7_431083997	FarmCPU	7E-07	0.07	
Phenylalanine	CHR4_413695971	SUPER	2E-06	0.38	Lcu.2RBY.4g065210, Lcu.2RBY.4g065220, Lcu.2RBY.4g065230, Lcu.2RBY.4g065240, Lcu.2RBY.4g065250, Lcu.2RBY.4g065260, Lcu.2RBY.4g065270, Lcu.2RBY.4g065280,

					Lcu.2RBY.4g065290, Lcu.2RBY.4g065300, Lcu.2RBY.4g065310
	CHR4_99139105	SUPER	8E-07	0.25	Lcu.2RBY.4g017470, Lcu.2RBY.4g017480, Lcu.2RBY.4g017490
Threonine	CHR3_424796277	BLINK, FarmCPU, MLMM	2E-08	0.04	Lcu.2RBY.3g073770, Lcu.2RBY.3g073780, Lcu.2RBY.3g073790, Lcu.2RBY.3g073800
Valine	CHR3_424796277	BLINK, MLMM	4E-10	0.04	Lcu.2RBY.3g073770, Lcu.2RBY.3g073780, Lcu.2RBY.3g073790, Lcu.2RBY.3g073800
	CHR4_385425795	BLINK	8E-08	0.30	Lcu.2RBY.4g059370, Lcu.2RBY.4g059380, Lcu.2RBY.4g059390, Lcu.2RBY.4g059400, Lcu.2RBY.4g059410, Lcu.2RBY.4g059420, Lcu.2RBY.4g059430, Lcu.2RBY.4g059440, Lcu.2RBY.4g059450

[a] SNPs exceeding a significance threshold of 0.05/22,280 (Bonferroni correction) in association with a trait. [b] The smallest *p*-value associating the trait by any model. [c] maf = minor allele frequency [d] Genes within the linkage disequilibrium block of the associated SNP

Table 4.4: Subset of linkage disequilibrium blocks associated with protein quality traits[a]

LD Block ID	Size (kb)	Associated Traits	Genes	Gene Description[b]
Chr2_14447415–14547425	100	Digestibility	Lcu.2RBY.2g006630	Integrin-linked kinase family protein
			Lcu.2RBY.2g006640	Uncharacterized protein
Chr2_292740630–292740648	< 1	Lysine	Lcu.2RBY.2g046140	Replication factor-A carboxy-terminal domain protein
Chr3_115394955–116212912	818	Aspartate, Aspartate:TotalAA, Digestibility	Lcu.2RBY.3g018060	Uncharacterized protein
			Lcu.2RBY.3g018070	Glutathione S-transferase
			Lcu.2RBY.3g018080	Glutathione S-transferase
			Lcu.2RBY.3g018090	Glutathione S-transferase
			Lcu.2RBY.3g018100	Glutathione S-transferase; amino-terminal domain protein
			Lcu.2RBY.3g018110	Uncharacterized protein
			Lcu.2RBY.3g018120	Uncharacterized protein
			Lcu.2RBY.3g018130	Eukaryotic aspartyl protease family protein
			Lcu.2RBY.3g018140	60S ribosomal protein L18a
			Lcu.2RBY.3g018150	3-hydroxyisobutyryl-CoA hydrolase-like protein
			Lcu.2RBY.3g018160	Polyprotein
			Lcu.2RBY.3g018170	Subtilisin-like serine protease
			Lcu.2RBY.3g018180	Ulp1 protease family, carboxy-terminal domain protein
			Lcu.2RBY.3g018190	Uncharacterized protein
			Lcu.2RBY.3g018200	Lipid transfer protein
Chr3_151509045–152255260	746	Lysine	Lcu.2RBY.3g022880	Uncharacterized protein
			Lcu.2RBY.3g022890	DUF295 family protein
			Lcu.2RBY.3g022900	NB-ARC domain disease resistance protein
			Lcu.2RBY.3g022910	IPP transferase
			Lcu.2RBY.3g022920	Ankyrin repeat plant-like protein
			Lcu.2RBY.3g022930	Uncharacterized protein
			Lcu.2RBY.3g022940	Beta-(1,2)-xylosyltransferase
			Lcu.2RBY.3g022950	Ulp1 protease family, carboxy-terminal domain protein

Chr3_424696277–424813245	117	Alanine, Isoleucine, Leucine, Methionine, Threonine, Valine	Lcu.2RBY.3g073770	Gibberellin 2-beta-dioxygenase
			Lcu.2RBY.3g073780	Gibberellin 2-beta-dioxygenase
			Lcu.2RBY.3g073790	Stem 28 kDa glycoprotein
			Lcu.2RBY.3g073800	Plant receptor-like kinase
Chr4_385392509–385525795	133	Valine	Lcu.2RBY.4g059370	Global transcription factor group protein
			Lcu.2RBY.4g059380	ORF1
			Lcu.2RBY.4g059390	Uncharacterized protein
			Lcu.2RBY.4g059400	Uncharacterized protein
			Lcu.2RBY.4g059410	Uncharacterized protein
			Lcu.2RBY.4g059420	Ulp1 protease family, carboxy-terminal domain protein
			Lcu.2RBY.4g059430	Putative AC transposase
			Lcu.2RBY.4g059440	Heat shock 70 kDa protein, mitochondrial (Precursor)
			Lcu.2RBY.4g059450	Polynucleotidyl transferase, Ribonuclease H fold
Chr5_107892651–107992664	100	Glycine:TotalAA	Lcu.2RBY.5g022850	Clustered mitochondria protein homolog
			Lcu.2RBY.5g022860	Clustered mitochondria protein homolog
			Lcu.2RBY.5g022870	Uncharacterized protein
			Lcu.2RBY.5g022880	RNA-directed DNA polymerase (Reverse transcriptase) Chromo Zinc finger, CCHC-type Peptidase aspartic, active site Polynucleotidyl transferase, Ribonuclease H fold
Chr6_301590674–301590682	< 1	Histidine:TotalAA	Lcu.2RBY.6g042780	Alpha-mannosidase
			Lcu.2RBY.6g042790	RNA-directed DNA polymerase (Reverse transcriptase) Chromo Zinc finger, CCHC-type Peptidase aspartic, active site Polynucleotidyl transferase, Ribonuclease H fold

[a] Subset contains LD blocks associated with multiple traits and/or with SNP minor allele frequencies exceeding 0.05. Blocks solely identified by the models SUPER and FarmCPU were excluded. [b] Descriptions were taken from a GFF file. See Supplemental Data: GWAS_Exhaustive (https://github.com/njohns4/LentilProteinQualityGWAS) for source information.

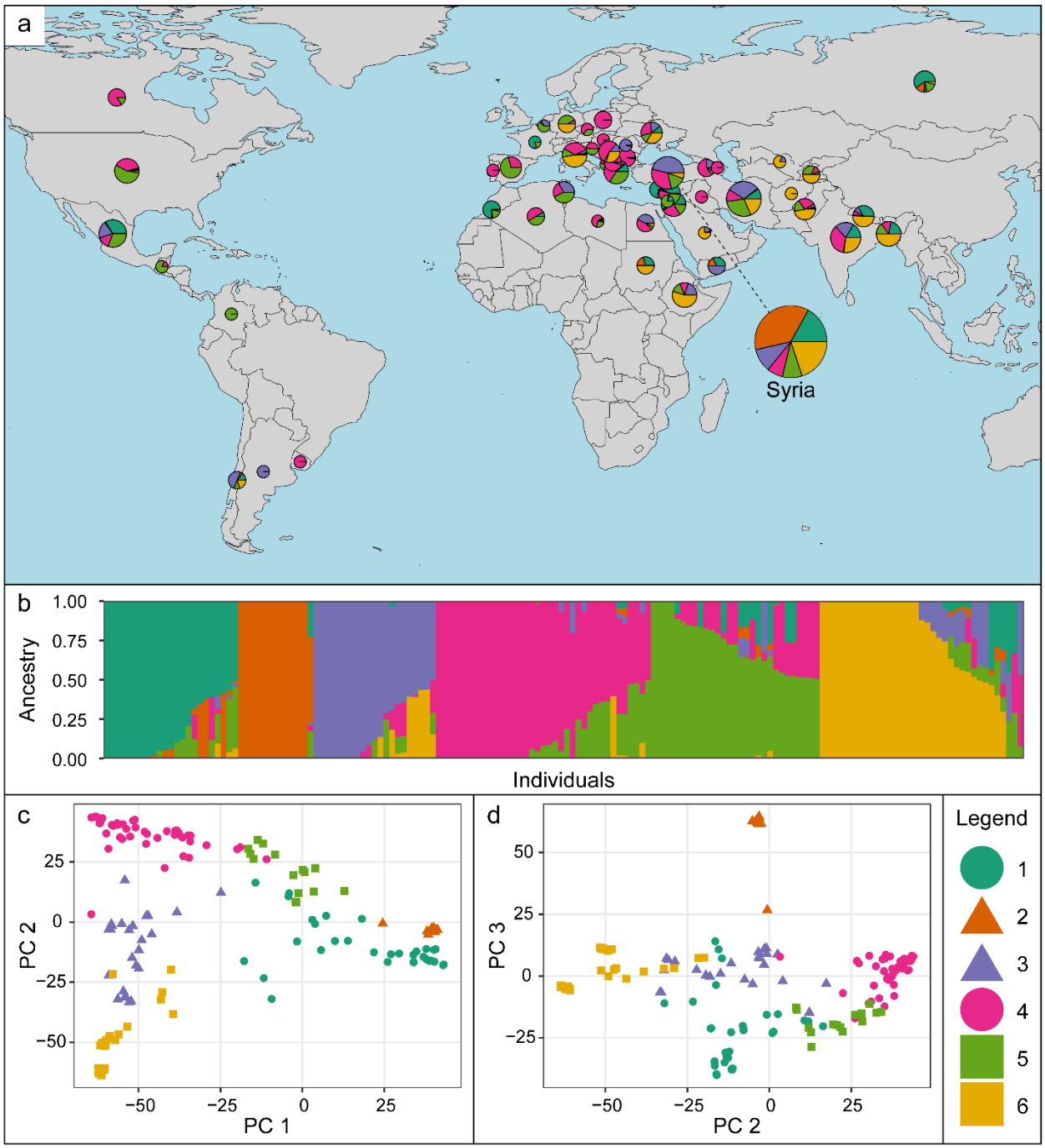

Figure 4.1: Lentil population origin and population structure analysis. A) Pie charts depict average admixture composition of accessions from the same country of origin. Pie chart circumferences are proportional to the number of accessions sharing each country of origin. The colors depict the average ancestral subpopulation composition of each location as determined by ADMIXTURE analysis where k = 6 (B). C and D depict the first three principal components with points representing accessions that have been colored corresponding to their ADMIXTURE ancestral subpopulation classification.

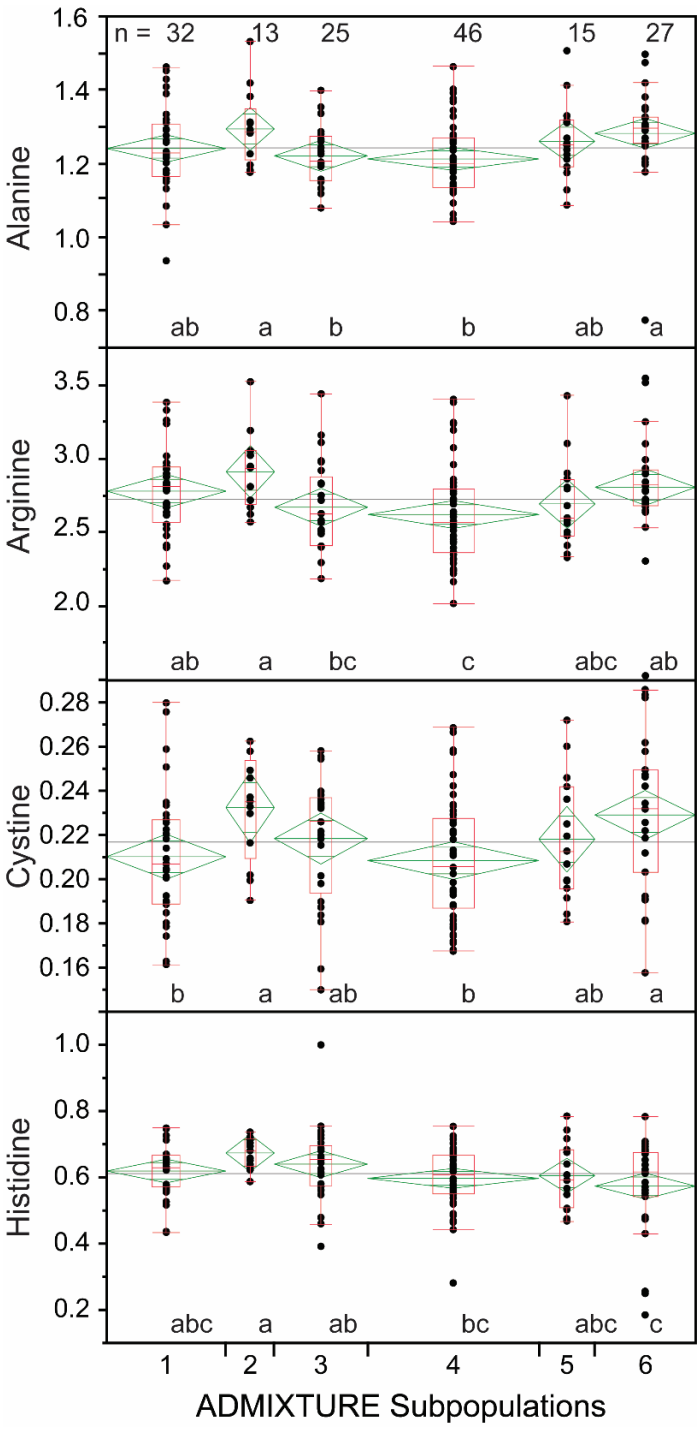

Figure 4.2: Boxplots depicting one-way analysis of variance of amino acid concentrations by ADMIXTURE ancestral subpopulation classifications. Boxplots connected by different letters have significantly different means ($p < 0.05$) as determined by Fisher's protected LSD. Green diamonds indicate the 95% confidence interval of the mean. The diamond width is proportional to the number of samples belonging to the subpopulation classification ($n_1 = 32$, $n_2 = 13$, $n_3 = 25$, $n_4 = 46$, $n_5 = 15$, $n_6 = 27$).

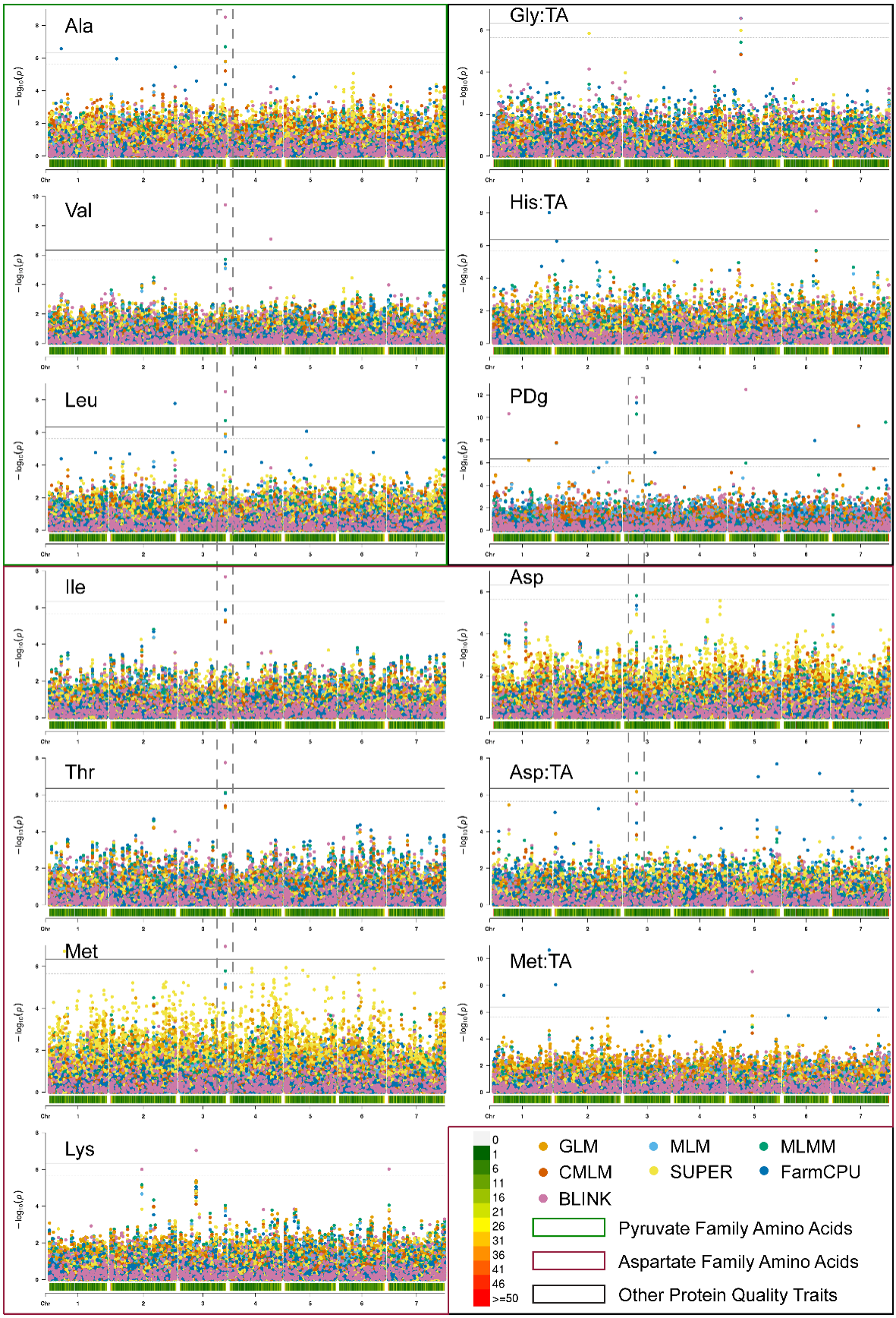
Ala
Gly:TA
Val
His:TA
Leu
PDg
Ile
Asp
Thr
Asp:TA
Met
Met:TA
Lys
0
1
6
11
16
21
26
31
36
41
46
>=50
GLM
MLM
MLMM
CMLM
SUPER
FarmCPU
BLINK
Pyruvate Family Amino Acids
Aspartate Family Amino Acids
Other Protein Quality Traits

Figure 4.3: Manhattan plots of traits with at least one SNP significantly associated with the trait by multiple models. Different color points represent different GWAS models. Significance thresholds are indicated by dotted and solid grey horizontal lines and correspond to –log(0.05/22,280) and –log(0.01/22,280), respectively (Bonferroni correction). Colored outlines represent pyruvate family amino acids, aspartate family amino acids, and other protein quality traits (Gly, His:TA, PDg). SNP density plots are located above chromosome numbers on a red to green scale of 1 to 50 SNPs per 1 Mb. Grey dashed line boxes indicate significant loci shared across multiple traits.

References

Abdelhalim, T. S., Kamal, N. M., & Hassan, A. B. (2019). Nutritional potential of wild sorghum: Grain quality of Sudanese wild sorghum genotypes (Sorghum bicolor L. Moench). *Food Science and Nutrition*, *7*(4), 1529–1539. https://doi.org/10.1002/fsn3.1002

Agilent Application Note. (2010). Separation of two sulfurated amino acids with other seventeen amino acids by HPLC with pre-column derivatization. *Agilent Technologies*.

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-246

Alghamdi, S. S., Khan, A. M., Ammar, M. H., El-Harty, E. H., Migdadi, H. M., El-Khalik, S. M. A., … Al-Faifi, S. A. (2014). Phenological, nutritional and molecular diversity assessment among 35 introduced lentil (lens culinaris medik.) genotypes grown in saudi arabia. *International Journal of Molecular Sciences*, *15*(1), 277–295. https://doi.org/10.3390/ijms15010277

Amin, M. N. (2018). *Molecular Analysis of Abiotic Stress in Lentil (Lens culinaris Medik.)*. Washington State University.

Bailey-Serres, J., Fukao, T., Ronald, P., Ismail, A., Heuer, S., & Mackill, D. (2010). Submergence tolerant rice: SUB1's journey from landrace to modern cultivar. *Rice*, *3*(2–3), 138–147. https://doi.org/10.1007/s12284-010-9048-5

Barbosa, P., Melnyk, S., Bennuri, S. C., Delhey, L., Reis, A., Moura, G. R., … Carvalho, E. (2021). Redox Imbalance and Methylation Disturbances in Early Childhood Obesity. *Oxidative Medicine and Cellular Longevity*, *2021*. https://doi.org/10.1155/2021/2207125

Baudoin, J. P., & Maquet, A. (1999). Improvement of protein and amino acid contents in seeds

of food legumes. A case study in Phaseolus. *Biotechnology, Agronomy, Society and Environment*, *3*(4), 220–224.

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, *103*(3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Chen, W., Wang, W., Peng, M., Gong, L., Gao, Y., Wan, J., … Luo, J. (2016). Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms12767

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., … Ng, E. H. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theoretical and Applied Genetics*, *132*(3), 627–645. https://doi.org/10.1007/s00122-019-03317-0

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/gigascience/giab008

Desai, K., Tank, C., Gami, R., & Patel, A. (2015). Genetic variability in indigenous collection of chickpea (Cicer arietinum L.) genotypes for seed yield and quality traits. *Journal of Progressive Agriculture*, *9*(January 2015), 59–62.

Dissanayake, R., Cogan, N. O. I., Smith, K. F., & Kaur, S. (2021). Application of genomics to understand salt tolerance in lentil. *Genes*, *12*(3), 1–18.

https://doi.org/10.3390/genes12030332

Foyer, C. H., Lam, H. M., Nguyen, H. T., Siddique, K. H. M., Varshney, R. K., Colmer, T. D., … Considine, M. J. (2016). Neglecting legumes has compromised human health and sustainable food production. *Nature Plants*, *2*(8), 1–10. https://doi.org/10.1038/NPLANTS.2016.112

García-Lorenzo, M., Sjödin, A., Jansson, S., & Funk, C. (2006). Protease gene families in Populus and Arabidopsis. *BMC Plant Biology*, *6*, 1–24. https://doi.org/10.1186/1471-2229-6-30

Gautam, N. K., Bhardwaj, R., Yadav, S., Suneja, P., Tripathi, K., & Ram, B. (2018). Identification of lentil (Lens culinaris Medik.) germplasm rich in protein and amino acids for utilization in crop improvement. *Indian Journal of Genetics and Plant Breeding*, *78*(04 SE-Research Article), 470–477. https://doi.org/10.31742/IJGPB.78.4.9

Gehrke, C. W., Wall Sr, L. L., Absheer, J. S., Kaiser, F. E., & Zumwalt, R. W. (1985). Sample Preparation for Chromatography of Amino Acids: Acid Hydrolysis of Proteins. *Journal of Association of Official Analytical Chemists*, *68*(5), 811–821. https://doi.org/10.1093/jaoac/68.5.811

Ghumman, A., Singh, N., Kaur, A., & Rana, J. C. (2019). Diversity in protein secondary structure, molecular weight, mineral and amino acid composition of lentil and horse gram germplasm. *Journal of Food Science and Technology*, *56*(3), 1601–1612. https://doi.org/10.1007/s13197-019-03676-y

Gilbert, J. A., Bendsen, N. T., Tremblay, A., & Astrup, A. (2011). Effect of proteins from different sources on body composition. *Nutrition, Metabolism and Cardiovascular Diseases*, *21*(SUPPL. 2), B16–B31. https://doi.org/10.1016/j.numecd.2010.12.008

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0090346

Gong, Q., Yang, Z., Chen, E., Sun, G., He, S., Butt, H. I., … Li, F. (2018). A Phi-Class Glutathione S-Transferase Gene for Verticillium Wilt Resistance in Gossypium arboreum Identified in a Genome-Wide Association Study. *Plant and Cell Physiology*, *59*(2), 275–289. https://doi.org/10.1093/pcp/pcx180

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2022). rstanarm: Bayesian applied regression modeling via Stan. Retrieved from https://mc-stan.org/rstanarm/

Gullner, G., Komives, T., Király, L., & Schröder, P. (2018). Glutathione S-transferase enzymes in plant-pathogen interactions. *Frontiers in Plant Science*, *871*(December), 1–19. https://doi.org/10.3389/fpls.2018.01836

Hang, J. (2021). *Genome-wide association study of seed protein and amino acid contents in cultivated lentils as determined by Near-infrared Reflectance Spectroscopy*. University of Manitoba.

Hang, J., Shi, D., Neufeld, J., Bett, K. E., & House, J. D. (2022). Prediction of protein and amino acid contents in whole and ground lentils using near-infrared reflectance spectroscopy. *LWT*, *165*(June), 113669. https://doi.org/10.1016/j.lwt.2022.113669

Jalgaonkar, S., Gajbhiye, S., Sayyed, M., Tripathi, R., Khatri, N., Parmar, U., & Shankar, A. (2022). S-adenosyl methionine improves motor co-ordination with reduced oxidative stress, dopaminergic neuronal loss, and DNA methylation in the brain striatum of 6-hydroxydopamine-induced neurodegeneration in rats. *Anatomical Record*, (September 2021), 1–11. https://doi.org/10.1002/ar.24948

Jiang, G. L., & Katuuramu, D. N. (2021). Comparison of seed fatty and amino acids in edamame dried using two oven-drying methods and mature soybeans. *Journal of the Science of Food and Agriculture*, *101*(4), 1515–1522. https://doi.org/10.1002/jsfa.10766

Johnson, N., Boatwright, J. L., Bridges, W., Thavarajah, P., Kumar, S., Shipe, E., & Thavarajah, D. (2021). Genome-wide association mapping of lentil (Lens culinaris Medikus) prebiotic carbohydrates toward improved human health and crop stress tolerance. *Scientific Reports*, 1–12. https://doi.org/10.1038/s41598-021-93475-3

Johnson, N., Johnson, C. R., Thavarajah, P., Kumar, S., & Thavarajah, D. (2020). The roles and potential of lentil prebiotic carbohydrates in human and plant health. *Plants, People, Planet*, *00*, 1–10. https://doi.org/10.1002/ppp3.10103

Karaca, N., Ates, D., Nemli, S., Ozkuru, E., Yilmaz, H., Yagmur, B., … Tanyolac, M. B. (2019). Genome-wide association studies of protein, lutein, vitamin C, and fructose concentration in wild and cultivated chickpea seeds. *Crop Science*, *59*(6), 2652–2666. https://doi.org/10.2135/cropsci2018.12.0738

Katuuramu, D. N., Hart, J. P., Porch, T. G., Grusak, M. A., Glahn, R. P., & Cichy, K. A. (2018). Genome-wide association analysis of nutritional composition-related traits and iron bioavailability in cooked dry beans (Phaseolus vulgaris L.). *Molecular Breeding*, *38*(4). https://doi.org/10.1007/s11032-018-0798-x

Khazaei, H., Caron, C. T., Fedoruk, M., Diapari, M., Vandenberg, A., Coyne, C. J., … Bett, K. E. (2016). Genetic diversity of cultivated lentil (Lens culinaris Medik.) and its relation to the world's agro-ecological zones. *Frontiers in Plant Science*, *7*(JULY2016), 1–7. https://doi.org/10.3389/fpls.2016.01093

Kruijer, W., Boer, M. P., Malosetti, M., Flood, P. J., Engel, B., Kooke, R., … Van Eeuwijk, F.

A. (2014). Marker-based estimation of heritability in immortal populations. *Genetics*, *199*(2), 379–398. https://doi.org/10.1534/genetics.114.167916

Kumar, H., Singh, A., Dikshit, H. K., Mishra, G. P., Aski, M., Meena, M. C., & Kumar, S. (2019). Genetic dissection of grain iron and zinc concentrations in lentil (Lens culinaris Medik.). *Journal of Genetics*, *98*(3), 1–14. https://doi.org/10.1007/s12041-019-1112-3

Kumar, J., Gupta, D. Sen, Kumar, S., Gupta, S., & Singh, N. P. (2016). Current Knowledge on Genetic Biofortification in Lentil. *Journal of Agricultural and Food Chemistry*, *64*, 6383–6396. https://doi.org/10.1021/acs.jafc.6b02171

Kumar, J., Sen Gupta, D., Baum, M., Varshney, R. K., & Kumar, S. (2021). Genomics-assisted lentil breeding: Current status and future strategies. *Legume Science*, *3*(3), 1–20. https://doi.org/10.1002/leg3.71

Liber, M., Duarte, I., Maia, A. T., & Oliveira, H. R. (2021). The History of Lentil (Lens culinaris subsp. culinaris) Domestication and Spread as Revealed by Genotyping-by-Sequencing of Wild and Landrace Accessions. *Frontiers in Plant Science*, *12*(March), 1–18. https://doi.org/10.3389/fpls.2021.628439

Long, W. (2015). Automated Amino Acid Analysis Using an Agilent Poroshell HPH-C18 Column. *Agilent Technologies*.

Ma, Y., Marzougui, A., Coyne, C. J., Sankaran, S., Main, D., Porter, L. D., … McGee, R. J. (2020). Dissecting the Genetic Architecture of Aphanomyces Root Rot Resistance in Lentil by QTL Mapping and Genome-Wide Association Study. *International Journal of Molecular Sciences*, *21*(6). https://doi.org/10.3390/ijms21062129

Madurapperumage, A., Johnson, N., Tang, L., & Thavarajah, D. (2022). Fourier-transform infrared spectroscopy (FTIR) as a high-throughput phenotyping tool for quantifying

protein quality in pulse crops, (April), 1–10. https://doi.org/10.1002/ppj2.20047

Manneberg, M., Lahm, H. W., & Fountoulakis, M. (1995). Quantification of cysteine residues following oxidation to cysteic acid in the presence of sodium azide. *Analytical Biochemistry*, *231*(2), 349–353. https://doi.org/10.1006/abio.1995.9988

Martín-Cabrejas, M. A., Aguilera, Y., Pedrosa, M. M., Cuadrado, C., Hernández, T., Díaz, S., & Esteban, R. M. (2009). The impact of dehydration process on antinutrients and protein digestibility of some legume flours. *Food Chemistry*, *114*(3), 1063–1068. https://doi.org/10.1016/j.foodchem.2008.10.070

Monsoor, M. A., & Yusuf, H. K. M. (2002). In vitro protein digestibility of lathyrus pea (Lathyrus sativus), lentil (Lens culinaris), and chickpea (Cicer arietinum). *International Journal of Food Science and Technology*, *37*(1), 97–99. https://doi.org/10.1046/j.1365-2621.2002.00539.x

Muthusamy, V., Hossain, F., Thirunavukkarasu, N., Choudhary, M., Saha, S., Bhat, J. S., … Gupta, H. S. (2014). Development of β-carotene rich maize hybrids through marker-assisted introgression of β-carotene hydroxylase allele. *PLoS ONE*, *9*(12), 1–22. https://doi.org/10.1371/journal.pone.0113583

Nair, B. M., Oste, R., Asp, N. G., & Dahlqvist, A. (1976). Enzymic hydrolysis of food protein for amino acid analysis. I. Solubilization of the protein. *Journal of Agricultural and Food Chemistry*, *24*(2), 386–389. https://doi.org/10.1021/jf60204a043

National Research Council Subcommittee on the Tenth Edition of the Recommended Dietary Allowances. (1989). Recommended dietary allowances, tenth edition. In *Recommended Dietary Allowances* (10th ed.). National Academies Press. https://doi.org/10.1016/s0002-8223(21)22412-7

Papandreou, C., Becerra-Tomás, N., Bulló, M., Martínez-González, M. Á., Corella, D., Estruch, R., … Salas-Salvadó, J. (2019). Legume consumption and risk of all-cause, cardiovascular, and cancer mortality in the PREDIMED study. *Clinical Nutrition*, *38*(1), 348–356. https://doi.org/10.1016/j.clnu.2017.12.019

Patil, B., Lad, D., Bhagat, A., & Nawale, S. (2020). Assessment of biochemical parameters and genetic variability in chickpea (Cicerar arietinum L.) genotypes. *International Journal of Chemical Studies*, *8*(3), 1305–1308. https://doi.org/10.22271/chemi.2020.v8.i3r.9378

Pavan, S., Bardaro, N., Fanelli, V., Marcotrigiano, A. R., Mangini, G., Taranto, F., … Ricciardi, L. (2019). Genotyping by Sequencing of Cultivated Lentil (Lens culinaris Medik.) Highlights Population Structure in the Mediterranean Gene Pool Associated With Geographic Patterns and Phenotypic Variables. *Frontiers in Genetics*, *10*(September), 1–9. https://doi.org/10.3389/fgene.2019.00872

Pfeiffer, B. (2017). *The Improvement of Grain Sorghum Productivity, Black Pericarp Color, and Protein Digestibility*. Texas A&M University. Retrieved from https://hdl.handle.net/1969.1/173127

Pratap, A., & Kumar, J. (2011). *Biology and Breeding of Food Legumes*. CABI. Retrieved from https://books.google.com/books?id=3Z-BW3V9nL8C

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rajendran, K., Coyne, C. J., Zheng, P., Saha, G., Main, D., Amin, N., … Kumar, S. (2021). Genetic diversity and GWAS of agronomic traits using an ICARDA lentil (Lens culinaris

Medik.) Reference plus collection. *Plant Genetic Resources: Characterisation and Utilisation*, *19*, 279–288. https://doi.org/10.1017/S147926212100006X

Ramsay, L., Koh, C. S., Kagale, S., Gao, D., Kaur, S., Haile, T., … Bett, K. E. (2021). Genomic rearrangements have consequences for introgression breeding as revealed by genome assemblies of wild and cultivated lentil species. *BioRxiv*, 2021.07.23.453237. Retrieved from https://www.biorxiv.org/content/10.1101/2021.07.23.453237v1%0Ahttps://www.biorxiv.org/content/10.1101/2021.07.23.453237v1.abstract

Salaria, S., Boatwright, J. L., Thavarajah, P., Kumar, S., & Thavarajah, D. (2022). Protein Biofortification in Lentils (Lens culinaris Medik.) Toward Human Health. *Frontiers in Plant Science*, *13*(April). https://doi.org/10.3389/fpls.2022.869713

Semba, R. D., Ramsing, R., Rahman, N., Kraemer, K., & Bloem, M. W. (2021). Legumes as a sustainable source of protein in human diets. *Global Food Security*, *28*(June 2020), 100520. https://doi.org/10.1016/j.gfs.2021.100520

Shekib, L. A. H., Zoueil, M. E., Youssef, M. M., & Mohamed, M. S. (1986). Amino acid composition and In vitro digestibility of lentil and rice proteins and their mixture (Koshary). *Food Chemistry*, *20*(1), 61–67. https://doi.org/https://doi.org/10.1016/0308-8146(86)90167-6

Singh, U., & Jambunathan, R. (1981). Studies on Desi and Kabull Chickpea (Cicer arietinum L.) Cultivars: Levels of Protease Inhibitors, Levels of Polyphenolic Compounds and in vitro Protein Digestibility. *Journal of Food Science*, *46*(5), 1364–1367. https://doi.org/https://doi.org/10.1111/j.1365-2621.1981.tb04176.x

South, A. (2011). rworldmap: A New R package for Mapping Global Data. *The R Journal*.

Retrieved from http://journal.r-project.org/archive/2011-1/RJournal_2011-1_South.pdf

Tabangin, M. E., Woo, J. G., & Martin, L. J. (2009). The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proceedings*, *3*(S7), 5–8. https://doi.org/10.1186/1753-6561-3-s7-s41

Talukdar, D. (2016). Exogenous thiourea modulates antioxidant defence and glyoxalase systems in lentil genotypes under arsenic stress. *Journal of Plant Stress Physiology*, *2*, 9. https://doi.org/10.19071/jpsp.2016.v2.3041

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *Plant Genome*, *14*(1), 1–17. https://doi.org/10.1002/tpg2.20077

U.S. Department of Agriculture. (2019). Lentils, raw. Retrieved from https://fdc.nal.usda.gov/fdc-app.html#/food-details/172420/nutrients

Wang, J., & Zhang, Z. (2021). GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics, Proteomics and Bioinformatics*, *19*(4), 629–640. https://doi.org/10.1016/j.gpb.2021.08.005

Wang, N., & Daun, J. K. (2006). Effects of variety and crude protein content on nutrients and anti-nutrients in lentils (Lens culinaris). *Food Chemistry*, *95*(3), 493–502. https://doi.org/10.1016/j.foodchem.2005.02.001

Wang, Y., Zhang, L., & Zhu, S. (2014). 1-Methylcyclopropene (1-MCP)-induced protein expression associated with changes in Tsai Tai (Brassica chinensis) leaves during low temperature storage. *Postharvest Biology and Technology*, *87*, 120–125. https://doi.org/10.1016/j.postharvbio.2013.08.016

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*. Retrieved from https://www.jstatsoft.org/v40/i01/

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

APPENDICES

Appendix A

Chapter 3 Supplemental Materials

Table S3.1: The spectral ranges associated with the chemometric models.

Model Name	Spectral range cm^{-1}
Chickpea Total Protein	1718.30–1487.21 | 3682.61–3006.98
Dry Pea Total Protein	1718.30–1487.21 | 3682.61–3006.98
Lentil Total Protein	1718.30–1487.21 | 3682.61–3006.98
Total Lentil SAA	721.24–867.07 | 1231.88–1469.96 | 1904.20–2241.99 | 2825.78–2994.91
Lentil Methionine	674.65–808.37 | 1182.03–1484.41| 1975.49–2158.59 | 2658.52–2991.19

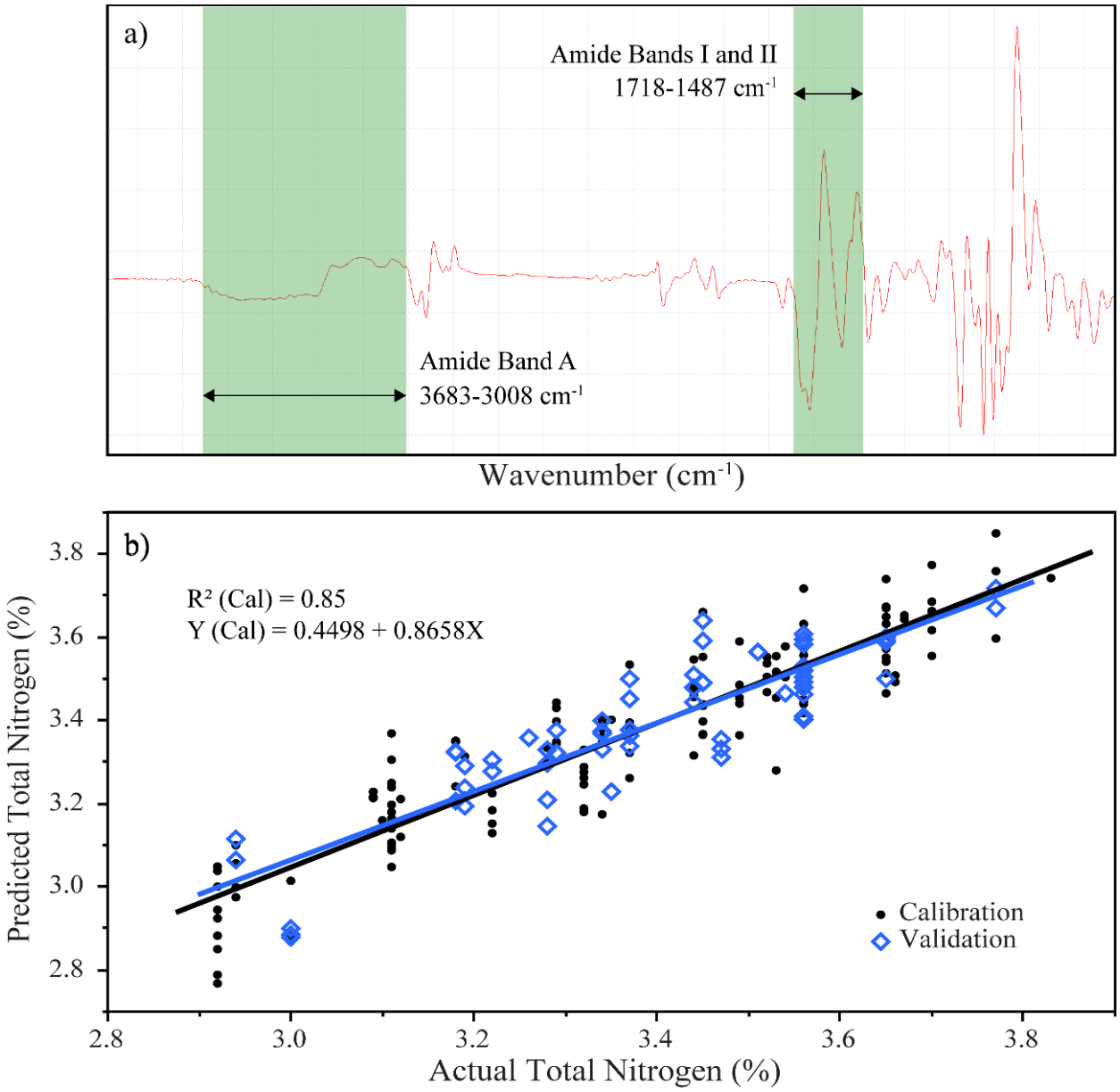

Figure S3.1: (a) Average dry pea MIR 1st-derivative absorbance spectrum. Regions in green were selected for the total nitrogen model in dry pea. (b) Scatter plot of actual vs. predicted total nitrogen (%) of calibration and validation data with lines of best fit.

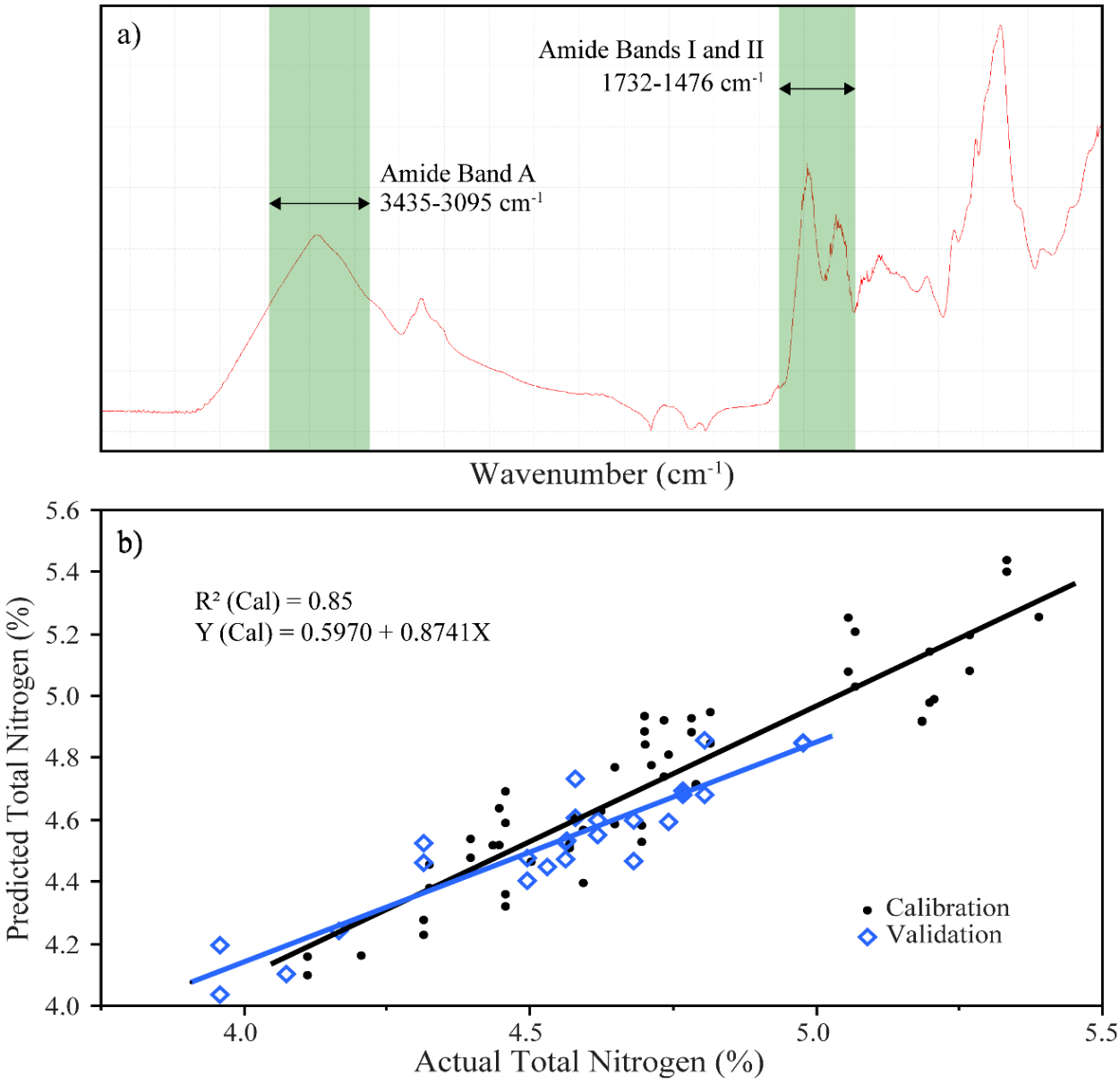

Figure S3.2: (a) Average lentil MIR absorbance spectrum. Regions in green were selected for the total nitrogen model in lentil. (b) Scatter plot of actual vs. predicted total nitrogen (%) of calibration and validation data with lines of best fit.

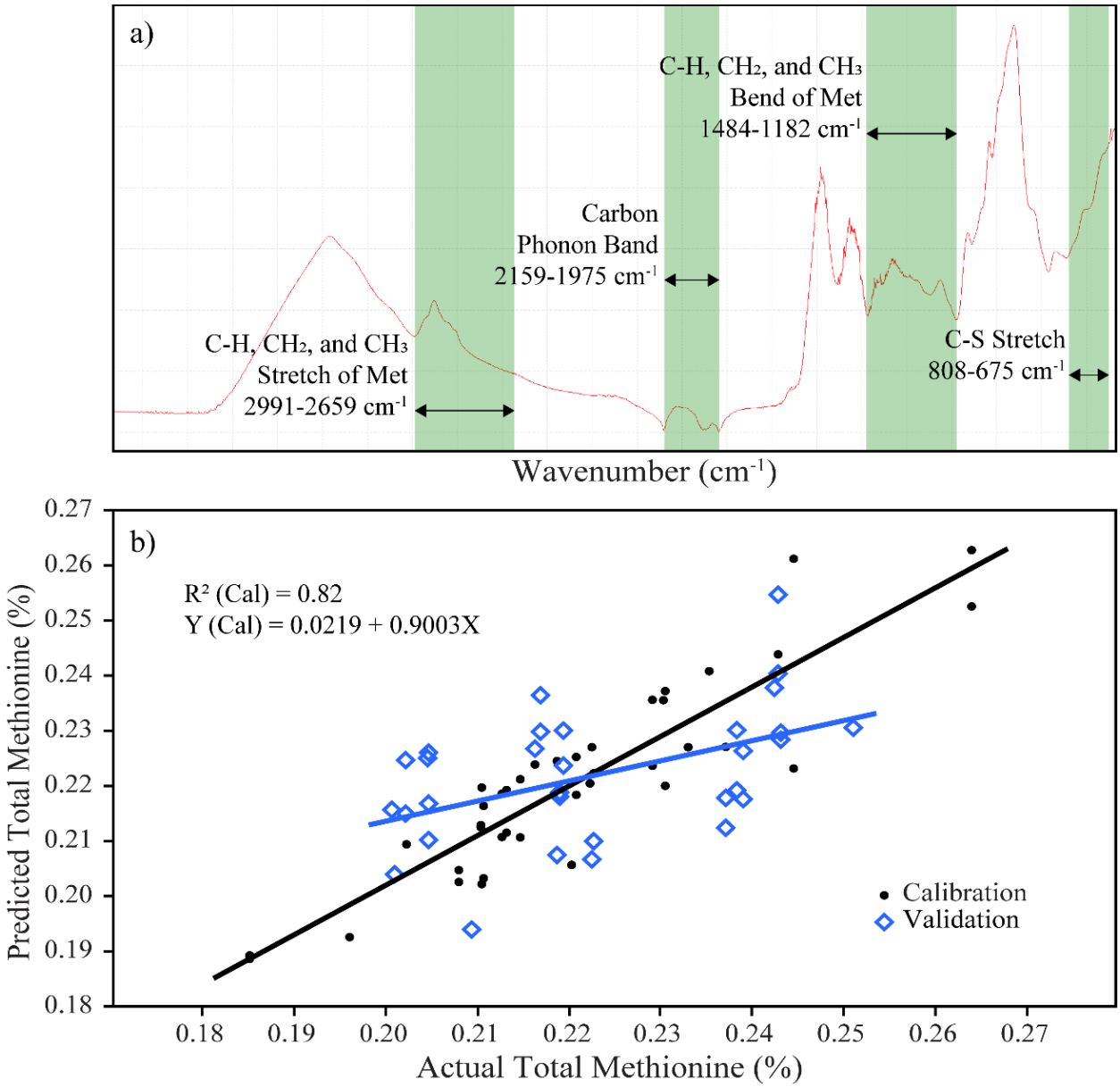

Figure S3.3: (a) Average lentil MIR absorbance spectrum. Regions in green were selected for the total methionine (Met) model in lentil. (b) Scatter plot of actual vs. predicted Met (%) values of calibration and validation data with lines of best fit.

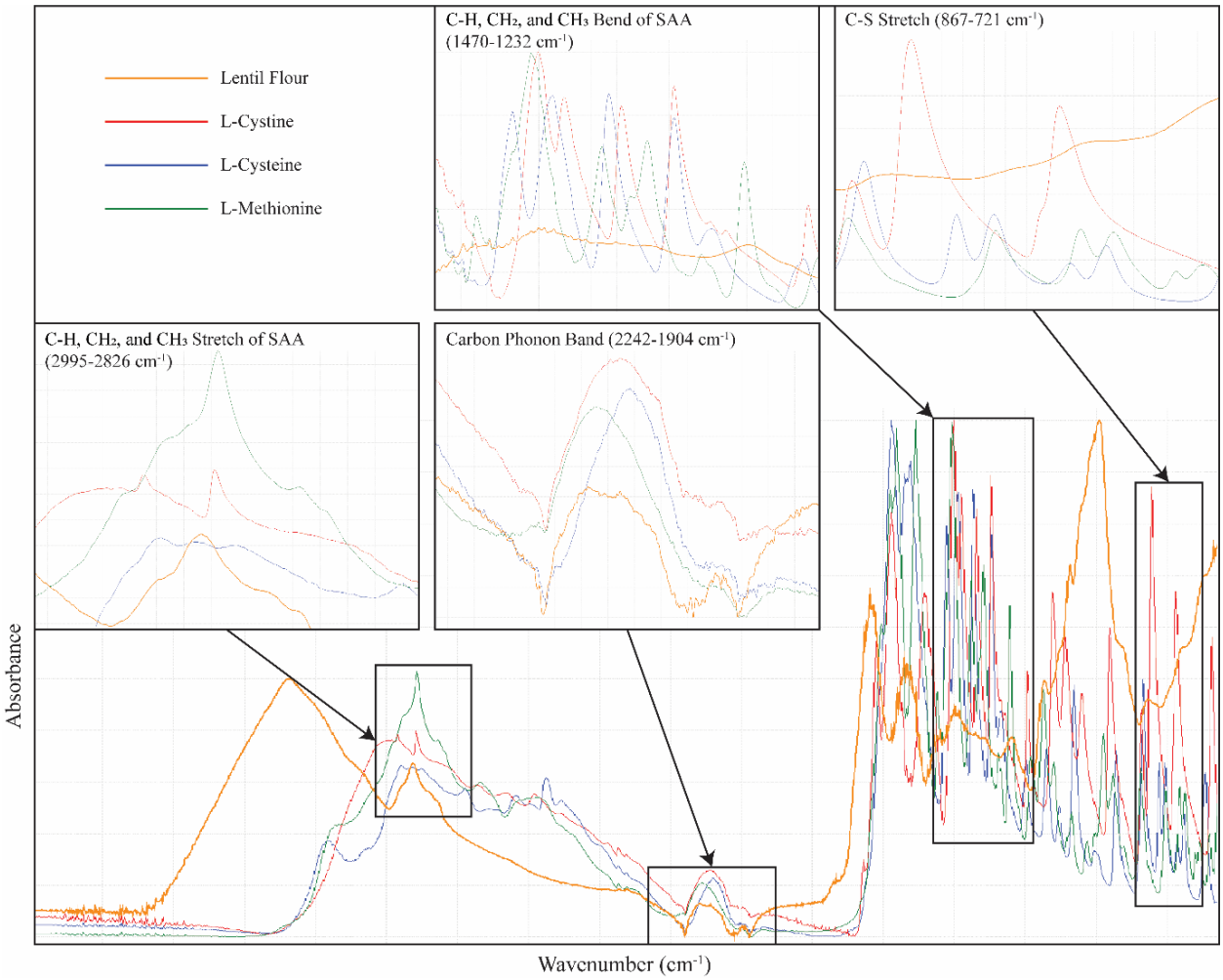

Figure S3.4: Identification of associated regions in lentil flour with powdered L-Cystine, L-Cysteine, and L-Methionine standards. All spectra normalized from 0 to 1.

Appendix B

Chapter 4 Supplemental Materials

Table S4.1: Mean, range, and repeatability of percent ratios of amino acids to total amino acid concentration

Trait	Mean (%) ± SD	Range (%)	Repeatability %
Ala:TA	4.31 ± 0.2	3.71d–4.9	13.1
Asp:TA	13.34 ± 0.77	10.91–16.9	13.1
Arg:TA	9.38 ± 0.61	7.38–12.02	15.6
Cys:TA	0.76 ± 0.10	0.55–1.03	15.5
Glu:TA	15.42 ± 0.85	13.01–17.85	17.5
Gly:TA	4.32 ± 0.22	3.7–4.92	14.2
His:TA	2.12 ± 0.33	0.64–3.2	6.5
H-Pro:TA	10.78 ± 2.78	5.33–19.91	9.6
Ile:TA	4.29 ± 0.22	3.68–4.82	18.3
Leu:TA	7.81 ± 0.37	6.69–8.77	19.7
Lys:TA	4.59 ± 0.66	2.72–6.14	9.9
Met:TA	0.71 ± 0.06	0.58–0.94	27.0
Phe:TA	4.45 ± 0.68	1.8–5.47	0.0
Pro:TA	8.50 ± 1.12	6.08–11.73	10.2
Ser:TA	4.86 ± 0.26	4.18–5.64	17.6
Thr:TA	3.64 ± 0.18	3.17–4.25	15.2
Val:TA	0.74 ± 0.04	0.65–0.83	12.4

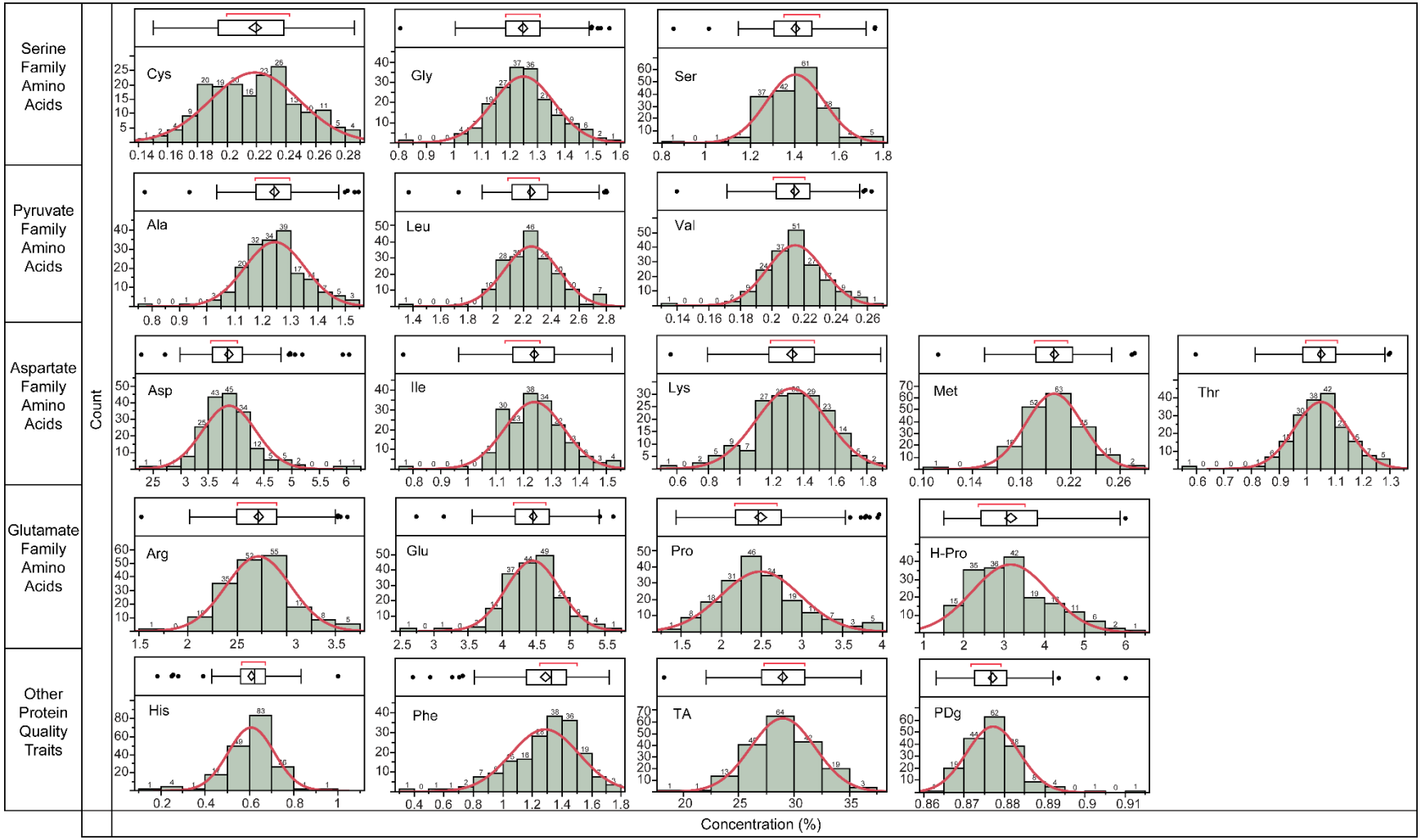

Figure S4.1: Histograms of trait distributions fit with density curves for the normal distribution using estimates of the mean and standard deviation.

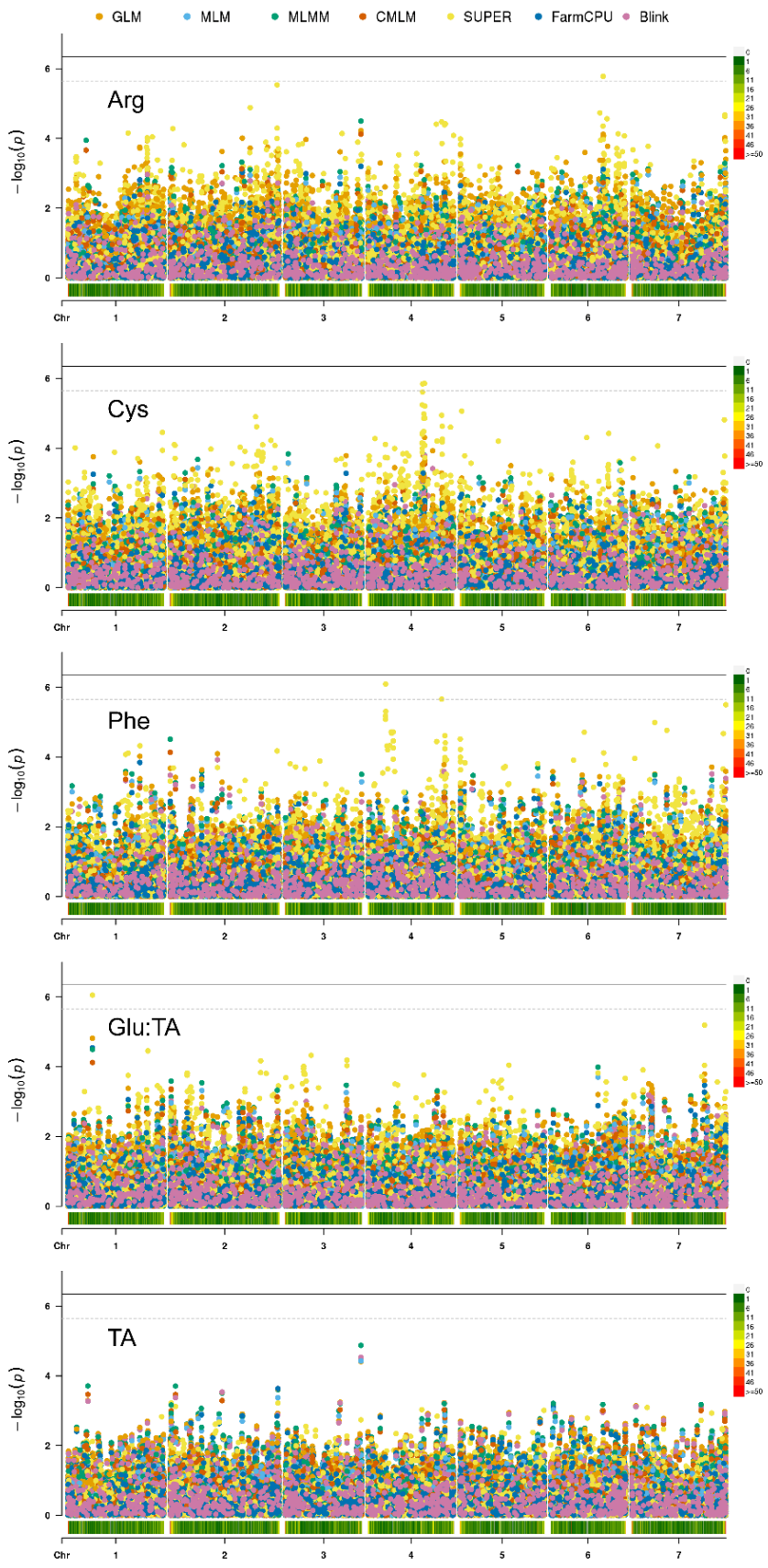
GLM
MLM
MLMM
CMLM
SUPER
FarmCPU
Blink
Arg
Cys
Phe
Glu:TA
TA
$-\log_{10}(p)$
Chr

Figure S4.2: Manhattan plots of traits with at least one SNP significantly associated with the trait by any model and TA which does not have significantly associated SNPs but was included for comparison with ratio traits. Different color points represent different GWAS models. Significance thresholds are indicated by dotted and solid grey horizontal lines and correspond to –log(0.05/22,280) and –log(0.01/22,280), respectively (Bonferroni correction). SNP density plots are located above chromosome numbers on a red to green scale of 1 to 50 SNPs per 1 Mb.

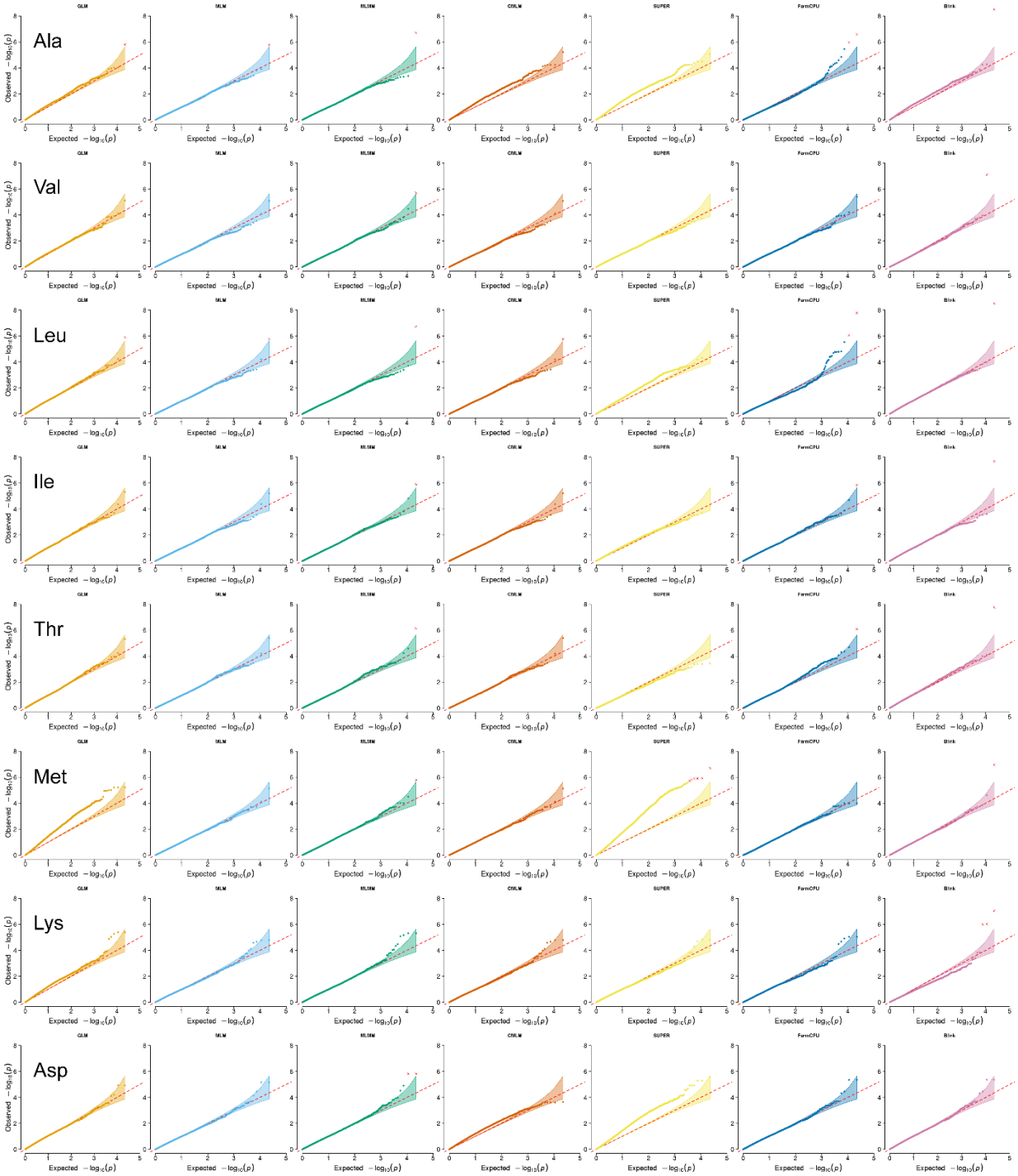

Figure S4.3 QQ plots of alanine, valine, leucine, isoleucine, threonine, methionine, lysine, and aspartate fitting the following genome-wide association models from GAPIT: GLM, MLM, MLMM, CMLM, SUPER, FarmCPU, and Blink.

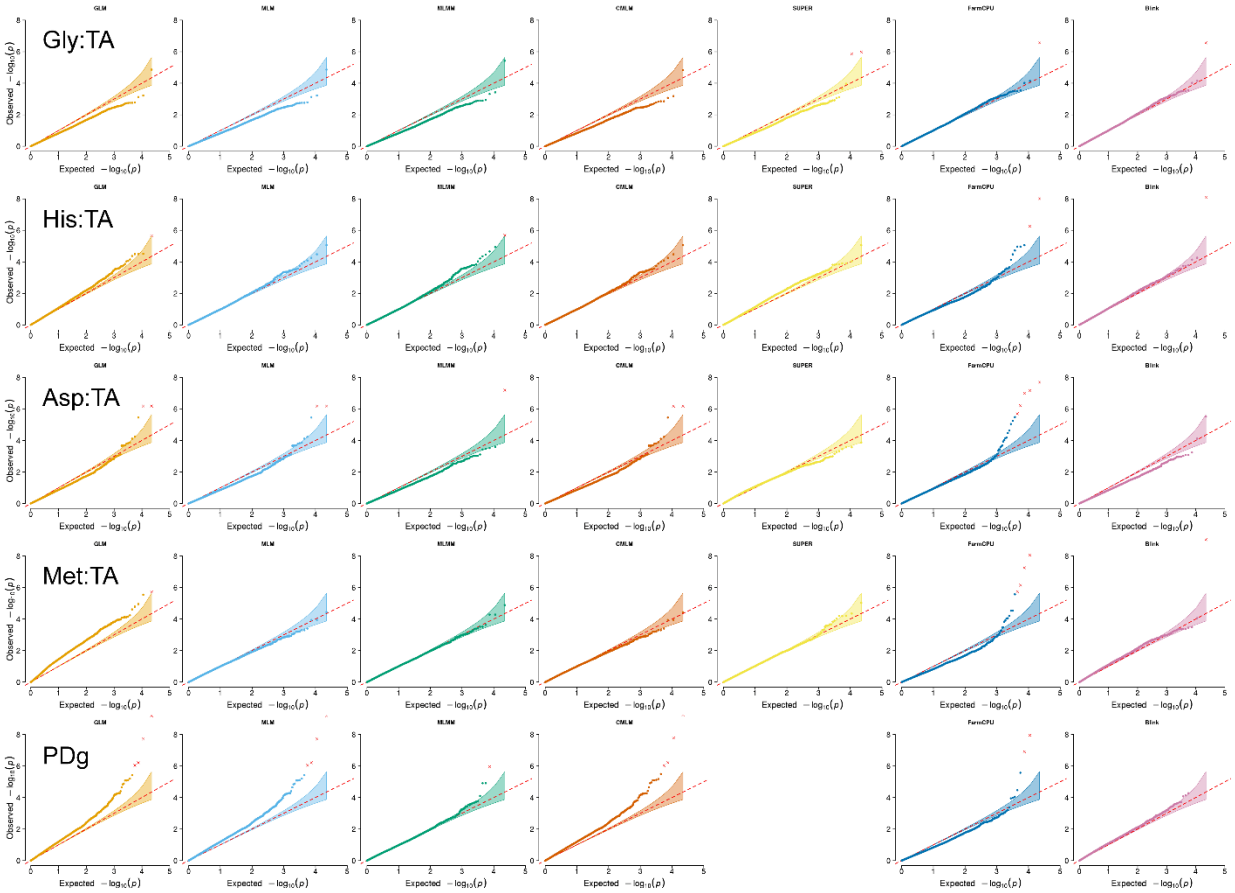

Figure S4.4: QQ plots of the ratio of glycine, histidine, aspartate, and methionine to total amino acid concentration and digestibility fitting the following genome-wide association models from GAPIT: GLM, MLM, MLMM, CMLM, SUPER, FarmCPU, and Blink

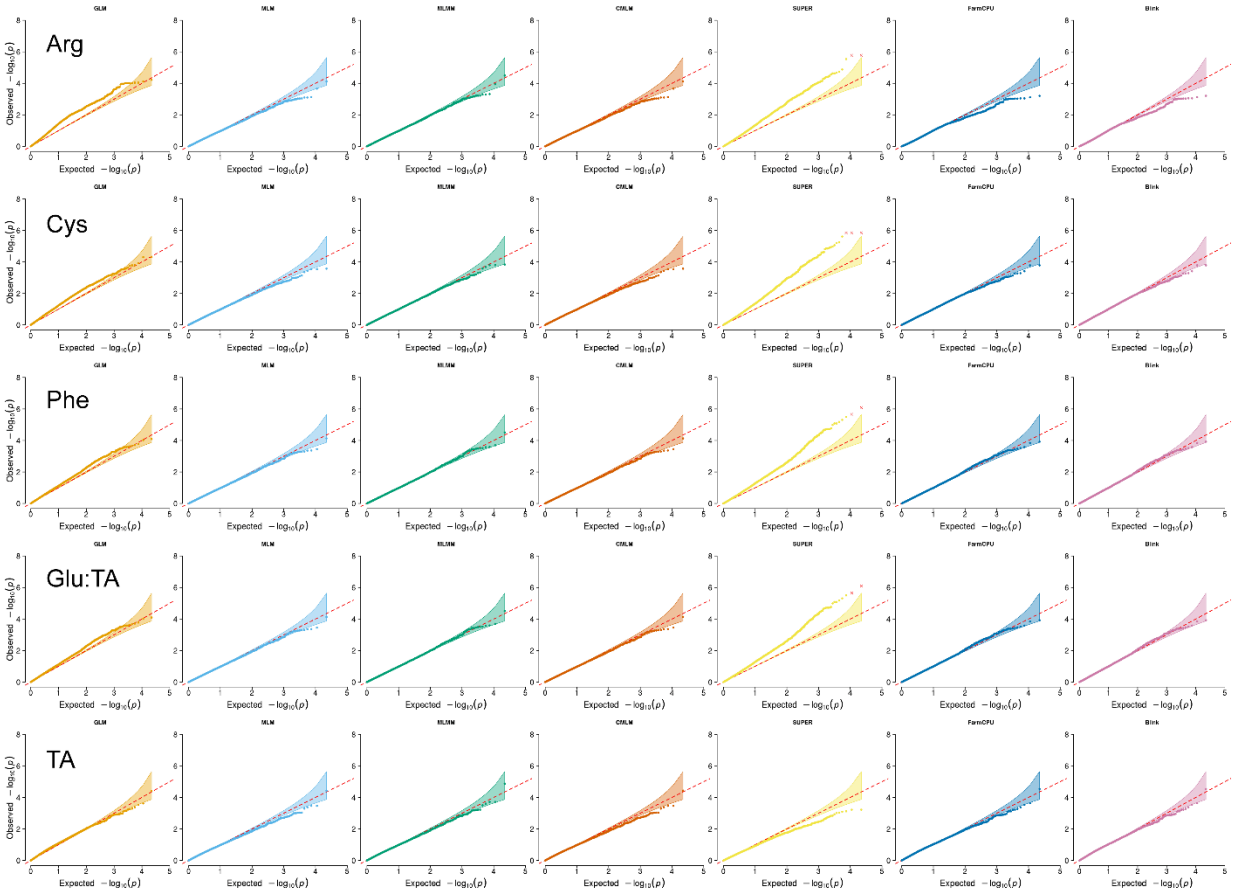

Figure S4.5: QQ plots of arginine, cystine, phenylalanine, the ratio of glutamate to total amino acid concentration, and total amino acid concentration fitting the following genome-wide association models from GAPIT: GLM, MLM, MLMM, CMLM, SUPER, FarmCPU, and Blink