Genome-wide Association Study of Seed Protein and Amino Acid Contents in Cultivated Lentils as Determined by Near-infrared Reflectance Spectroscopy

By

Jiayi Hang

A thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Food and Human Nutritional Sciences

University of Manitoba

Winnipeg, Manitoba, Canada

ABSTRACT

Lentil (*Lens culinaris* Medik.) is an important legume crop and is considered as a plant-based protein food to fight protein malnutrition and bring health benefits. The current study focused on the development and evaluation of near-infrared reflectance spectroscopy (NIRS) models to predict the protein and amino acid contents in lentils by two NIRS spectrometers: PerkinElmer DA7250 and FT 9700. In total, 361 lentil samples grown in Saskatchewan, Canada, were selected as a calibration set. NIRS calibration models developed by partial least squares (PLS) equation had a satisfactory performance for measuring protein and most amino acids (except for histidine, tyrosine, methionine and cysteine) in lentils with $R^2_C > 0.65$. The sample status, type of spectrometer, and amino acid/protein correlation could influence the NIRS models' predictive abilities. NIRS models from DA 7250 achieved similar accuracy for determination of crude protein and amino acids in whole and ground lentils. In the current study, the predictive ability of DA 7250 models and FT 9700 models was not significantly different for all compositions ($p<0.05$). For amino acids highly correlated to crude protein, NIRS generally predicted them with a higher accuracy. The protein and 18 amino acid contents of 1290 whole lentil samples predicted from DA 7250 models on a dry basis were used as the phenotypic data in the later genome-wide association study (GWAS). GWAS was conducted using phenotypic data from four environments in Saskatchewan, Canada and 266,164 single nucleotide polymorphism (SNP) markers for 324 lentil accessions to identify significantly associated markers. A total of 85 SNP markers were identified to have significant associations with protein and/or 18 amino acids. Only one identical SNP marker (SLCU.2RBY.CHR7_524204079) significantly associated with Val was identified in two environments, and other SNPs were identified only in one environment. These identified SNPs could be studied further to find potential genomic regions or candidate

genes. In summary, NIRS could be regarded as a highly promising method for rapid prediction of lentil seed protein and most amino acid contents, and GWAS had a great potential to dissect the genetic basis of these traits in cultivated lentils. Both NIRS technology and GWAS can facilitate breeding programs to enhance lentil seed protein content and quality.

ACKNOWLEDGMENTS

Foremost, I would like to express my deepest appreciation to my supervisor, Dr. James House, for invaluable guidance, novel insights, great support and full encouragement throughout this project. I am grateful and honored to work in his research team. During the two-year program, Dr. House provided me with patience and a great amount of assistance both in study and life. His profound knowledge and academic passion impressed me a lot, and his kindness and humor made these days precious and unforgettable. He could always make me feel excited about my project and help me to reinforce my determination to keep doing research in the future. Also, many thanks to my committee members: Dr. Rebecca Mollard and Dr. Mehmet Tulbek for giving me constructive advice and valuable feedback during meetings and reviewing my thesis.

I am deeply indebted to Jason Neufeld and Shusheng Zhao for teaching me lab skills, assisting me in sample preparation and data collection. Special thanks to Da Shi for his continuous assistance in sample analysis, insightful advice in data interpretation and encouragement in difficulty. I am also grateful for the assistance from administrative staff, my lab mates Zhongyang Wan, Junya Liu, Adam Franczyk, Shengnan Li, summer student Jennifer Nguyen, and all other graduate students that participated in my project.

I would like to express my sincere thanks to Derek Wright, Laura Jardine and Dr. Kirstin Bett (University of Saskatchewan) for project coordination, genotypic and environmental data collection, and other research and statistical assistance. I am thankful for the technical assistance that I received from Dr. David Honigs and Dr. Emily Moore (PerkinElmer company).

I gratefully acknowledge funding from Genome Canada, the University of Saskatchewan, and the University of Manitoba, which made this project a reality.

Finally, I want to express my appreciation to my family and friends for their love and support throughout my life.

DEDICATION

I would like to dedicate this thesis to my beloved parents, who trust me, love me and support me in every aspect of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AACC – American Association for Clinical Chemistry

Ala – Alanine

ANN – Artificial neural networks

ANOVA – Analysis of variance

AOAC – American Association of Analytical Chemists

AOTF – Acousto-optic tunable filter

Arg – Arginine

Asn – Asparagine

Asp – Aspartic acid

BIC – Bayesian information criteria

BLINK – Bayesian information and LD iteratively nested keyway

CDC – Crop Development Centre (University of Saskatchewan)

CGC – Canadian Grain Commission

Chr – Chromosome

CV – Coefficient of variation

Cys – Cysteine

DA – Diode array

DIAAS – Digestible indispensable amino acid score

FAO – Food and Agriculture Organization

FarmCPU – Fixed and random model circulating probability unification

FT – Fourier transform

FT-NIRS – Fourier transform near-infrared reflectance spectroscopy

G × E – Gene × environment interaction

GAPIT – Genome Association and Prediction Integrated Tool

Gln – Glutamine

Glu – Glutamic acid

Gly – Glycine

GWA – Genome-wide association

GWAS – Genome-wide association study

HCl – Hydrogen chloride

His – Histidine

IAA – Indispensable amino acid

Ile – Isoleucine

InGaAs – Indium–gallium–arsenide

ISO – International Organization for Standardization

K – Kinship

LCTF – Liquid crystal tunable filter

LD – Linkage disequilibrium

LED– Light emitting diodes

Leu – Leucine

Lys – Lysine

MAF – Minor allele frequency

Met – Methionine

MLM – Mixed linear model

MLMM – Multi-locus mixed model

MLR – Multiple linear regression

MSC – Multiplicative scatter correction

NIR – Near-infrared

NIRS – Near-infrared Reflectance Spectroscopy

NIST – National Institute of Standards and Technology

OPD – Optical path difference

PC – Principal component

PCA – Principal component analysis

PCR – Principal components regression

PDCAAS – Protein Digestibility Corrected Amino Acid Score

PER – Protein efficiency ratio

Phe – Phenylalanine

PLS – Partial least squares

Pro – Proline

Q – Population structure

Q-Q – Quantile-quantile

QTL – Quantitative trait locus

r – Correlation coefficient

R^2/ RSQ – Coefficient of determination

R^2_C – Coefficient of determination for calibration

R^2_{CV} – Coefficient of determination for cross-validation

R^2_P – Coefficient of determination for prediction

RMSEC – Root mean square error of calibration

RMSECV – Root mean square error of cross-validation

RMSEP – Root mean square error of prediction

RPD – Residual predictive deviation

SAA – Sulphur amino acids

SD – Standard deviation

SEC – Standard error of calibration

SECV – Standard error of cross-validation

SEP – Standard error of prediction

Ser – Serine

SNP – Single nucleotide polymorphism

SNV – Standard normal variate

SUPER – Settlement of MLM under progressively exclusive relationship

SVM – Support vector machines

TCD – Thermal conductivity detector

Thr – Threonine

Trp – Tryptophan

Tyr – Tyrosine

UNU – United Nations University

UPLC – Ultra performance liquid chromatography

USDA – United States Department of Agriculture

Val – Valine

WHO – World Health Organization

1-VR – 1 Minus the ratio of unexplained variance to total variance

CHAPTER 1: INTRODUCTION

There is a fast-growing market for plant-based protein all over the world because of its environmental, economic, social and health benefits (Health Canada, 2019). As a nutritious pulse, lentil plays an important role in the plant-based protein market (Khazaei et al., 2019). Similar to other legumes, lentil can increase soil fertility by fixing atmospheric nitrogen, thus reducing the need to provide nitrogen-based fertilizers and protecting the environment (Khazaei et al., 2019). Compared to animal protein, lentil protein can be more environmentally friendly and more affordable. Furthermore, the population growth and improvement of environmental awareness promote the development of lentil market (Health Canada, 2019). The relatively low content of sulphur amino acids and high content of lysine in lentils makes it an ideal complementary food to cereals, and their combination can generate a complete essential amino acid profile (Erskine, 2009; Moldovan et al., 2015). The consumption of lentil protein and cereal protein together can meet the amino acids requirement, and this is of great importance for vegans and vegetarians. Meanwhile, lentil protein plays a vital role in combating human protein malnutrition all over the world, especially in some low-income countries (Semba, 2016). In addition to being an important source of protein, lentil is a good source of fiber, many vitamins and minerals (Wright et al., 2021). The high nutritional value, low fat and calories of lentils are good for human health.

Due to these benefits of lentils, they have received more attention and have been cultivated in many countries. Canada is a leading country in lentil production and export in recent decades (Health Canada, 2019). FAO (2021) shows that Canada accounted for 38% of the global lentil production in 2019, followed by India (22%), Australia (9%) and Turkey (6%).

A growing trend in plant-based protein consumption along with the high production yield of lentils in Canada, has led to an increased focus on the study of the protein and amino acids of

lentils grown in Canada. The seed protein content and amino acid composition significantly affect the quality and uses of lentil. Determining the protein and amino acid concentrations of lentils can benefit lentil breeding program to improve the protein content and quality, can help food industries make correct nutrient content claims and create value-added products with new food formulations. However, the traditional reference methods for quantifying protein and amino acids are complicated, expensive, destructive, labor intensive, time consuming and dangerous. To overcome these disadvantages, the Near-infrared Reflectance Spectroscopy (NIRS) analytical method has attracted great attention (Baianu et al., 2004). NIRS can accurately predict the protein contents in several legumes including lentil (Quiñones et al., 2018; Moldovan et al., 2015; Revilla et al.,2019). Several studies have investigated the potential of NIRS to measure amino acid concentrations in legumes such as soybean (Fontaine et al., 2001; Kovalenko, Rippke & Hurburgh, 2006) and pea (Fontaine et al., 2001). It is important to study the potential of NIRS to predict the protein and amino acid contents in lentil.

What's more, after obtaining the data of protein and amino acid concentrations and genotypic data, a genome-wide association study (GWAS) can dissect the genetic basis of these traits and identify significantly associated molecular genetic markers (Zhang et al., 2018b). These markers could be studied further, verified and applied in lentil breeding programs (Pearson & Manolio, 2008; Tibbs Cortes, Zhang & Yu, 2021).

CHAPTER 2: LITERATURE REVIEW

2.1 Lentils

2.1.1 Background

Lentil (*Lens culinaris*) is an edible legume crop, which is classified as a pulse (Boye, 2015). Lentil has a long history of usage as a food in human diet, and Asian people first grew lentils around 7000BC (Boye, 2015). The lens shape of seeds gives this cultivar its name. Lentil is tolerant of different and unfertile soil types and thus adapted to grow in marginal lands (Erskine, 2009). Nowadays, lentil has become a staple in many countries (Khazaei et al., 2019). It is consumed in the form of whole seed, dehulled split grain and flour. According to the FAO, the global production of lentils was 5734.2 thousand tonnes in 2019 (FAO, 2021). In 2019, lentil production was 2166.9 thousand tonnes in Canada, which accounted for about 38% of the global production. India was the second-largest producer, accounting for about 22% of the world's total production. Australia (9%), Turkey (6%), Nepal (4%) and the USA (4%) were the other major producers of lentils (FAO, 2021).

Canada has led in lentil production and export in recent decades (Health Canada, 2019). Lentil is primarily grown in Saskatchewan, accounting for about 92% production in Canada, followed by Alberta and Manitoba (Wang, 2019). There are different varieties of lentils, with the red and green lentils being the predominant classes that are grown in Canada. Green lentils, which can be identified by grey green seed coats, are normally consumed as whole seeds mainly in Europe (Tahir et al., 2011). Red lentils with red cotyledons are usually available in the split form after removing the seed coat by decortication. Decorticated red lentils can be used in soups or stews (Tahir et al., 2011). Some varieties of lentils support fast cooking without the pre-soaking step. Compared with other pulses (>70min), the cooking time of lentils is much shorter

(23-26min) (Faris, Takruri & Issa, 2013). The convenient preparation and high nutritional values of lentils have promoted the rapid growth in lentil consumption (Khazaei et al., 2019).

2.1.2 Environmental, economic and social benefits

Cultivation of legumes has beneficial effects on environmental, economic and social aspects of the agri-food system. Firstly, crop rotations with legumes can maintain soil fertility and reduce the use of fertilizers because legumes can fix nitrogen in their roots and release part of the nitrogen into the soil (Khazaei et al. 2019). Animal-based protein usually has a significantly higher environmental cost, which also leads to higher emissions of greenhouse gases (Khazaei et al. 2019). Detzel et al. (2021) pointed out that livestock farming contributed to around 7% of the EU's greenhouse gas. They also concluded that lentil proteins had a great potential to make vegetable milk alternative with a significantly lower environmental footprint (Detzel et al., 2021). Also, proteins from legumes are more affordable than animal-based proteins. Lentils are suitable for human consumption and animal feeding (Khazaei et al. 2019). The sales of plant-based protein products increased by 7% in 2016-2017 (Health Canada, 2019). The rapidly increasing global trend of plant-based food provides an excellent chance for many food companies to create new food formulations with lentils (Detzel et al., 2021). These all promote economic growth. From a social aspect, a growing global population and a growing concern about health and environment from consumers contribute to the rapid growth of plant-based protein. More than 40% of Canadians are trying to consume more plant-based foods in their daily diet (Health Canada, 2019). Moreover, lentils can satisfy the needs of vegans and vegetarians, who rely on legumes as their major protein source. However, attention must be paid to the quality of dietary protein in lentils. A lack of essential amino acids in children's diets can

cause adverse effects on growth patterns (Semba, 2016). Children, especially from low-income countries, should consume sufficient high-quality protein to prevent protein malnutrition (Semba, 2016). Due to these reasons, scientists and industries are putting more effort into legumes such as lentils, with a greater focus on protein content and quality.

2.1.3 Chemical composition

Lentils contain around 26% crude protein on a dry basis, and lentils are usually regarded as a protein food that may replace animal-based proteins or soy proteins in creating new food formulations (Khazaei et al., 2019). Lentil proteins can provide a sufficient amount of some essential amino acids such as leucine, lysine, threonine and phenylalanine (Boye, 2015). The relatively high content of lysine makes it possible to generate a complete essential amino acid profile when consumed with cereal-based foods such as rice (Erskine, 2009). However, like many other leguminous seeds, lentils generally lack sulfur amino acids (methionine and cysteine) and tryptophan (Boye, 2015; Nosworthy et al., 2017). Meanwhile, lentil is an excellent source of fiber, many vitamins and minerals, and has low content of fat and is relatively low in calories (Wright et al., 2021). The most abundant micronutrient present in lentils is potassium, followed by phosphorus. The fiber content and relatively high amylose concentration in the starch fraction may lead to a low glycemic index because of reduced digestibility (Tahir et al., 2011). One study has also shown that lentils have relatively greater antioxidant activity compared with other legumes such as chickpeas and common beans (Grela et al.,2017). The high nutritional value, low calories and glycemic index of lentils can bring health benefits (Asif et al., 2013).

2.2 Proteins

2.2.1 Lentil protein content

Proteins are organic compounds which contain carbon, oxygen, hydrogen, nitrogen and in some cases sulfur. They are polymers made of amino acids linked together by peptide bonds (Watford & Wu, 2011). Protein is one of the macronutrients and it contains 4 kcal per gram. It also provides some essential amino acids that humans can't synthesize. Both protein quantity and protein quality are of great importance for the human diet.

More than 80% of the total proteins of lentils are located in the cotyledons as storage proteins (de Souza Cândido et al.,2011). Seed storage proteins have various functions such as enhancing seedling growth and antimicrobial activity (de Souza Cândido et al.,2011). The seed proteins can be classified into four groups according to their solubility in different solvents, including salt-soluble globulins, water-soluble albumins, acid-soluble glutelins and ethanol-soluble prolamins. The predominant lentil proteins are globulins, which account for greater than 40% of total proteins (Bhatty,1988; Boye et al.,2010). The globulins can be grouped into two classes according to different sedimentation coefficients, 7S vicilin-type and 11S legumin-type proteins (Khazaei et al.,2019). The 7S vicilin-type proteins lack disulfide bonds while 11S legumin-type proteins have disulfide bonds. The 7S/11S ratio was 2.78 in lentils, which was much higher than that in peas (Scippa et al., 2010). Compared to pea proteins, lentil proteins contain a higher concentration of 7S vicilin-type and have fewer disulfide bonds. The deficiency of disulfide bonds is due to the low availability of cysteine residues in the protein fractions. The second primary proteins in lentils are albumins (Khazaei et al.,2019). High variations were reported for albumins, ranging from 11% to 61% (Khazaei et al.,2019). The possible reasons

might be genotypic variations or different extraction methods. The difference in protein subunits can alter protein functional and nutritional properties.

Protein yield is an increasingly important target in lentil breeding programs, and it is calculated as protein fraction × grain yield. Protein yield values for lentils from previous studies were in the range of 0.2-0.6 t ha^{-1} (Lizarazo et al.,2015; Khatun et al.,2010; Subedi et al., 2021).

Many studies have illustrated a significant negative correlation between starch and protein concentration (Tahir et al., 2011; Wang & Daun, 2006). The selection of a cultivar with a desirable composition concentration depends on its end use.

Lentil is an important and good source of protein (Khazaei et al.,2019). The protein content and components may vary among different lentil seeds, which will influence the value, properties and applications of lentils (Jarpa-Parra, 2018).

2.2.2 Lentil protein quality

Protein quality of the product reflects: (1) the amino acid pattern of the protein relative to the requirements of the consumer (2) the extent to which the protein is digested, absorbed and made available to contribute to the amino acid needs of the consumer.

Protein quality can be influenced by the origins of proteins, amino acid composition and the presence of anti-nutritive factors (Leser, 2013; Nosworthy et al., 2018). Some anti-nutritive factors in lentils can change protein bioavailability, including trypsin inhibitors which inactivate digestive enzymes and tannins which bind to protein to reduce their digestibility (Nosworthy et al., 2018). Processing of lentils such as cooking and extrusion showed some evidence of increasing protein quality (Nosworthy et al., 2017; Nosworthy et al., 2018). Several methods can be used to measure protein quality and establish protein content claims. The protein efficiency

ratio (PER), based on the ratio of weight gain (g) to protein intake (g) in male rats for a 28-day growth period, is an approved method for assessing protein content claims in Canada (Health Canada., 1981). Protein digestibility corrected amino acid score (PDCAAS) can be measured through multiplying the limiting amino acid score by true protein digestibility (%), which is currently used in the United States (FAO/WHO, 1991). According to FAO/WHO (1991), the amino acid score of each indispensable amino acid (IAA) is calculated by dividing the content of IAA (mg/g protein) in the test sample by a reference pattern of the same amino acid. IAA with the lowest score determines the amino acid score of the sample. Furthermore, true digestibility can be calculated as:

$$\frac{\text{Nitrogen intake - (Fecal nitrogen - Metabolic nitrogen)}}{\text{Nitrogen intake}} \times 100$$

PDCAAS uses a single value of true protein digestibility instead of the specific digestibility of individual indispensable dietary amino acids, and this has been regarded as a drawback. Moreover, fecal digestibility is used to determine the true protein digestibility; however, FAO/WHO (2013) mentions ileal digestibility can be a better choice to correct for digestibility. To overcome these shortcomings, FAO/WHO (2013) recommended the use of the digestible indispensable amino acid score (DIAAS), which is calculated as: DIAAS% = 100× [(mg of digestible dietary indispensable amino acid in 1 g of the dietary protein)/(mg of the same dietary indispensable amino acid in 1 g of the reference protein)]. The digestibility in this method reflects the true ileal digestibility of each indispensable amino acid.

Sulfur amino acids and tryptophan are the most limiting amino acids that adversely influence the protein quality in pulses, including lentils (Nosworthy et al., 2018; Iqbal et al., 2006). Nosworthy et al. (2017) and Nosworthy et al. (2018) mentioned that the sulfur amino acids were the limiting amino acids in cooked lentils, leading to an amino acid score of cooked

lentils being between 0.57 and 0.68. Both Iqbal et al. (2006) and Wang and Daun (2006) reported that tryptophan was the most limiting amino acid of lentils with an amino acid score of 0.64. Cooked lentils had higher true protein digestibility (87.9% for green lentils and 90.6% for red lentils) than other pulses such as beans or peas (Nosworthy et al., 2017). Nosworthy et al. (2018) showed that the PDCAAS values for cooked lentils ranged from 47.1% to 63.0%. The PDCAAS value for cooked whole green lentils and split red lentils was 0.628 and 0.538, respectively (Nosworthy et al.,2017). The PDCAAS values of lentils were similar to other pulses (0.5-0.7) such as navy beans, chickpeas, kidney beans; however, the values were lower than milk or egg protein (1.0; full mark). The DIAAS values for lentils were between 0.50 and 0.58 (Nosworthy et al., 2018). The PER of processed lentils was from 0.98 to 1.41, which was higher than that of split green peas (0.86) but lower than that of chickpeas (2.32) (Nosworthy et al., 2018).

2.3 Amino Acids

Around 500 amino acids have been identified from nature; however, only 20 appear in the genetic code and play important roles in the body (Watford & Wu, 2011). Of these, nine amino acids are known as essential/indispensable amino acids (IAA) that people can't synthesize or synthesize adequately to meet body's needs. These include histidine (His), isoleucine (Ile), leucine (Leu), lysine (Lys), methionine (Met), phenylalanine (Phe), threonine (Thr), tryptophan (Trp) and valine (Val). Also, two amino acids can only be made from their essential precursors, tyrosine (Tyr) from phenylalanine (Phe) and cysteine (Cys) from methionine (Met) (Watford & Wu, 2011). Therefore, when calculating amino acid scores, Met and Cys are added to reflect total sulfur amino acids. The sum of Phe and Tyr is considered as the aromatic amino acid

content (FAO/WHO/UNU, 1985). The amino acid score is an important indicator of protein quality (Erskine, 2009).

Lentil has a generally good amino acid profile because it contains all essential amino acids. Most of them meet the requirements for humans by the World Health Organization except sulfur amino acids and tryptophan (Table 2.1).

Sulfur amino acids (Met + Cys) and tryptophan are the limiting amino acids in lentil proteins while the major amino acids are glutamic acid (Glu), aspartic acid (Asp), arginine (Arg), leucine (Leu) and lysine (Lys) (Bhatty et al.,1976; Erskine, 2009; Wang & Daun, 2006; Alghamdi et al., 2014; Nosworthy et al.,2017). The high level of lysine in lentils makes it an ideal complementary food to cereals, which have low lysine content and relatively good contents of Met and Cys (Erskine, 2009). A combination of these two proteins can achieve a balance of essential amino acids.

The protein and amino acid contents in lentils vary significantly and have a broad range (Kumar, Singh, Kanaujia & Gupta, 2016; Khazaei et al., 2019; Table 2.1). The phenotype can be influenced by many genotypes (G), environments (E) and gene × environment interaction (G × E) (Tibbs Cortes, Zhang & Yu, 2021; Subedi et al., 2021).

Table 2.1. Amino acid composition for whole lentil seeds from published literatures and from UDSA database, FAO/WHO/UNU (1985) and WHO (2007) reference requirement pattern of amino acid for child (total protein content = total nitrogen × 6.25)

Amino acid (mg/g protein)	Published range for lentil [a]	Values from USDA [b]	Pattern for pre-school child (2-5 years old) [c]	Pattern for child (1-2 years old) [d]
Histidine (His)	13-34	28	19	15
Isoleucine (Ile)	26-55	43	28	27
Leucine (Leu)	57-87	73	66	54
Lysine (Lys)	40-82	70	58	45
Threonine (Thr)	25-49	36	34	23
Tryptophan (Trp)	6-26	09	11	6.4
Valine (Val)	33-61	50	35	36
Tyrosine (Tyr)	11-36	27	63[e]	40 [e]
Phenylalanine (Phe)	36-58	49		
Methionine (Met)	7-13	09	25[f]	22 [f]
Cysteine (Cys)	7-15	13		
Alanine (Ala)	24-50	42		
Arginine (Arg)	39-111	77		
Aspartic acid (Asp)	93-159	111		
Glutamic acid (Glu)	128-185	155		
Glycine (Gly)	32-56	41		
Proline (Pro)	12-72	42		
Serine (Ser)	29-64	46		

[a] Values derived from the following references: Shekib et al. (1986); Combe et al. (1991); Kavas and Nehir (1992); Urbano et al. (1995); Carbonaro et al. (1997); Porres et al. (2002); Wang and Daun (2004, 2006); Lizarazo et al. (2015); Nosworthy et al. (2017); Nosworthy et al. (2018); Paucean et al. (2018).

[b] Lentils, raw. NDB Number:16069. USDA (2018).

[c] Values from FAO/WHO/UNU (1985).

[d] Values from WHO (2007).

[e] Combined recommended pattern for tyrosine plus phenylalanine (Aromatic amino acids).

[f] Combined recommended pattern for methionine plus cysteine (Sulfur amino acids).

2.4. Factors Influencing Protein and Amino Acid Content of Lentils

2.4.1 Environment

Environmental factors that can influence crop characteristics include uncontrolled weather effects and partially controlled agronomic effects such as planting density and fertilization (Wang & Daun, 2006). Among many environmental stresses in lentil production, drought and heat stress are considered the most important, which have been extensively studied recently in India (Sehgal et al.,2017; Sita et al.,2018). Seed filling is the ultimate stage of lentil growth, where carbohydrates, proteins and lipids are synthesized. This phase is the most vulnerable to stress because of the involvement of diverse enzymes and transporters (Sehgal et al., 2018).

Lentil cultivars are cool-season legume, which require low temperatures during vegetative growth and warm temperatures at maturity (Sehgal et al.,2017). Sehgal et al. (2017) reported that heat and drought stress inhibited lentil yield traits (seed number and seed weight). The drought stress caused worse impacts on seed filling than heat stress, and the combined stress had a significantly detrimental effect. Based on Sita et al. (2018), one set of cultivated lentils was maintained in a controlled environment (28/23◦C, as mean day and night temperature) and another set was exposed to heat stress (33/28◦C) during seed filling. The results showed that heat stress reduced protein content (26–41%) and inhibited the accumulation of storage proteins (22-42%). Most of the amino acids decreased significantly under heat stress; however, certain amino acids increased such as Pro, followed by Gly, Ala, Ile, Leu, and Lys. Choukri et al. (2020) observed that the heat and drought stress adversely affected the protein content of lentils. The combined heat-drought stress reduced crude protein by 57.2%, whereas heat stress alone reduced crude protein by 14.3%. These studies were conducted in India and Morocco. So far, limited

research was found regarding the influence of the environment on protein and amino acid composition of lentils grown in Canada.

Data on the composition of Canadian lentils is available from the Canadian Grain Commission. The 10-year mean of the protein content of western Canadian lentils is 26.7% on a dry basis, and the average content varies from year to year (Wang, 2020). The average protein contents of 674 lentils from southeast Saskatchewan, southwest Saskatchewan, northeast Saskatchewan and northwest Saskatchewan in 2019 were 27.0%, 27.2%, 26.8% and 26.6%, respectively (Wang, 2019). The protein difference between lentils from the eastern or western regions of Saskatchewan was ambiguous; however, there was a clear pattern that protein concentration is slightly higher in those lentils from the south than those grown in the north. Furthermore, the results from 2018 and 2020 cropping years showed the same pattern (Wang, 2019; Wang 2020).

2.4.2 Genotype

Wong et al. (2015) identified four gene pools based on their relatedness to *L. culinaris* (cultivated lentil), namely *L. culinaris*, *L. orientalis*, *L. tomentosus* (primary gene pool); *L. odemensis, L. lamottei,* (secondary gene pool); *L. ervoides* (tertiary gene pool); and *L. nigricans* (quaternary gene pools). Kumar et al. (2016) mentioned that the highest protein content was found in the *L. ervoides* accession, which could contain 32.7% protein. Kumar et al. (2016) showed that the average protein content in Mediterranean landraces, wild species and Indian breeding lines were 22.4%, 22.6% and 18.6%, respectively from 72 various accessions of lentils. These studies indicated that wild lentil species could become valuable genetic resources for identifying the potential genes governing protein contents.

Regarding cultivated lentils (*Lens culinaris*) grown in Canada, Boye et al. (2010) mentioned that the mean protein content of CDC Grandora green lentil was 23.03% on as is basis, which was lower than that of common blaze red lentil (25.88% on as is basis). According to Tahir et al. (2011), 22 lentil (*Lens culinaris* Medikus subsp. *culinaris*) genotypes were grown in 2005 in Saskatchewan, Canada. The protein content of lentil genotypes varied significantly ($P \leq 0.05$) and ranged from 23.8 to 29.3g per 100 g flour dry matter. The highest protein concentration was recorded for accession ILL 1704 (red cotyledon). Subedi et al. (2021) became the first to report the genetic stability for protein content and yield in lentils. Lentil samples with 34 genotypes were collected from five locations in Saskatchewan, Canada, during 2017 and 2018. They observed that the extra small red market class genotypes had 3% higher protein content than other marker classes. Genotypes IBC 1235, 3923-9, 3674-17, IBC 929R, and 4371-4 became the useful candidates in lentil breeding programs because they yielded higher seed protein content and protein yield in a relatively stable manner (Subedi et al.,2021). Based on the Canadian Grain Commission, the average protein content (% dry basis) of small lentils (CDC Imvincible, CDC Kermit and Eston) was 28.3%, which was higher than medium or large lentils in 2019 and 2020 (Wang, 2019; Wang, 2020). Wang & Daun (2006) also mentioned variety Eston (small-sized green lentil) had the highest mean protein content of 28.7% among other varieties (Crimson, Laird, Richlea). The potential reason is that small seeds have a proportionally larger part of embryo, which has higher protein content (Subedi et al., 2021).

Rozan, Kuo & Lambein (2001) determined amino acid compositions through acid hydrolysis of five species of lentils, including *L. culinaris, L. orientalis, L.ervoides, L. nigricans* and *L. odemensis.* Tyrosine was lowest (5.05mg/g) in *L. odemensis* and lysine was lowest in *L. culinaris* (4.54mg/g). It revealed that *L. orientalis* seeds had relatively higher AAs than

cultivated lentils. Regarding cultivated lentils (*Lens culinaris* Medik.), Alghamdi et al., (2014) concluded that FLIP2009-64L and FLIP2009-69L, among 35 lentil genotypes, could be used as a significant source for lentil genetic improvement in Saudi Arabia due to high yield, high content of total protein, and essential amino acids. In a study conducted in Turkey, lentil varieties Cagil and Altintoprak became the promising genotypes due to their high content of essential amino acids (Kahraman, 2016). According to Wang & Daun (2006), analysis of variance showed that variety had a significant effect on histidine and serine in lentils grown in Canada.

Understanding the influence of genotypes and identifying specific genes responsible for protein and amino acid content are of great importance for lentil breeding programs. Genome-wide association study (GWAS) can find the connection between each genotyped marker and specific phenotype, which has been widely used in recent years to investigate many traits in various plants (Tibbs Cortes, Zhang & Yu, 2021).

2.5. Genome-wide association study (GWAS)

2.5.1 Introduction to GWAS

Genome-wide association study (GWAS) has become a powerful and ubiquitous tool for dissecting the genetic basis of complex traits in plants (Zhang et al., 2018; Tibbs Cortes, Zhang and Yu, 2021). GWAS can assist gene cloning studies, accelerate crop breeding and enable genetic engineering (Tibbs Cortes, Zhang & Yu, 2021). The development of genomic technology and robust statistical methods, combined with a strong desire to identify genotype-phenotype association became the driving forces behind the growth of Genome-wide association (GWA) studies in different plant species (Zhu et al., 2008; Tibbs Cortes, Zhang & Yu, 2021).

In GWA studies, a genome-wide set of genetic variants are tested to identify candidate genes associated or genetic markers with traits of interest (Pearson & Manolio, 2008). Single-nucleotide polymorphism (SNP) is the substitution of a single nucleotide, and it is the most common form of genetic variation (Pearson & Manolio, 2008). As a molecular genetic marker, SNPs are well suited for GWA studies because of higher densities throughout the genome and a lower mutation rate (Zhu et al., 2008). Therefore, GWA studies are typically designed to identify SNPs associated with agriculturally important traits in plants (Pearson & Manolio, 2008).

GWAS depends on linkage disequilibrium (LD) between markers and causative genes. LD is the non-random association of alleles at different loci. Alleles located near each other on a chromosome have a higher frequency of association than expected by chance. Alleles of SNPs in high LD are very likely to be inherited together (Tibbs Cortes, Zhang and Yu, 2021). The SNPs near the causative genes can have a significant association with interested phenotype due to high LD. The genomic regions containing these SNPs and causative genes could be detected in GWAS.

2.5.2 Approach

The GWA studies usually contain four parts: (1) sample selection; (2) genotyping and phenotyping; (3) identifying the associations between the SNPs and interested traits by different statistical models, (4) and then the potential significant SNPs can be further investigated and validated, and the casual genes might be identified (Pearson & Manolio, 2008; Tibbs Cortes, Zhang & Yu, 2021). In the first step “sample selection”, a diversity panel of plants with interested traits can be assembled for GWAS (Pearson & Manolio, 2008). Secondly, the genotype data generally indicate genome-wide SNP, which can be identified by high-throughput

genotyping and sequencing technologies with lower cost (Tibbs Cortes, Zhang & Yu, 2021; Zhu et al., 2008). Various phenotypic traits of plants have been investigated via GWAS, including flowering time (Neupane, 2019), stress tolerance (Chen et al., 2021; Dissanayake et al., 2021), composition concentration (Karaca et al.,2019; Leamy et al., 2017), and seed weight (Karikari et al., 2020). Thirdly, multiple statistical methods are now available to identify SNPs significantly associated with phenotypic trait variation. Different methods have been developed to reduce the false positives, increase computing speed and enhance statistical power (Zhang, 2020). The mixed linear model (MLM) is a widely accepted and used method for GWAS, and it controls both population structure (Q) and kinship (K) to reduce the false positives (Yu et al., 2006). Population structure, as a fixed effect, can be controlled through STRUCTURE software (Pritchard et al., 2000) or principal component analysis (Price et al., 2006; Zhao et al., 2007). Kinship matrix is about the relationships among individuals, which is a random effect (Yu et al., 2006). To improve the statistical power and decrease the false discovery rate of single-locus methods, Segura et al. (2012) first proposed the multi-locus mixed model (MLMM). It incorporates multiple markers as covariates via a forward-backward stepwise approach (Segura et al., 2012). Other multi-locus models are built based on MLMM, including fixed and random model circulating probability unification (FarmCPU) (Liu et al., 2016) and Bayesian information and LD iteratively nested keyway (BLINK) (Huang et al., 2019). FarmCPU uses a fixed effect model and a random effect model iteratively, and FarmCPU can remove the confounding between markers and kinship completely. In the random effect model, a maximum likelihood method is used to select associated markers (Liu et al., 2016). However, the random effect model has a relatively higher computing cost. A fixed effect model by using Bayesian information criteria (BIC) is used in Blink to replace the random effect model in FarmCPU.

Unlike FarmCPU, Blink eliminates the assumption that quantitative trait nucleotides (QTNs) are evenly distributed throughout the genome (Huang et al., 2019). Consequently, compared to FarmCPU, Blink has a higher statistical power and a significantly higher computational efficiency (Zhang, 2020; Huang et al., 2019). More methods can be used to conduct GWAS such as settlement of MLM under progressively exclusive relationship (SUPER) method (Wang et al., 2014b) and Bayesian methods (Tibbs Cortes, Zhang & Yu, 2021). Lastly, the potential candidates identified by GWAS needs to be validated by comparing with previous literature, performing biological and statistical methods (Tibbs Cortes, Zhang & Yu, 2021).

2.5.3 Advantages and limitations of GWAS

GWAS has many advantages, as it (i) examines the entire genome in many unrelated individuals, (ii) has high resolution, (iii) has large allele numbers, (iv) is a hypothesis free (non-hypothesis driven) method, (v) can detect common variants of modest effect on the phenotype (Alqudah et al., 2020; Tibbs Cortes, Zhang & Yu, 2021). However, the method still has some limitations. It has the potential to generate false-positive results due to population structure and kinship, especially for some naïve GWAS analysis models (Tibbs Cortes, Zhang & Yu, 2021). Also, it can miss rare variants with low minor allele frequency and alleles with low effects on the phenotype (Alqudah et al., 2020). Moreover, it can be challenging to determine the causative variant. Sample selection and phenotyping errors could cause biases (Pearson & Manolio, 2008; Tibbs Cortes, Zhang & Yu, 2021).

2.5.4 Application of GWAS

GWAS was first used to detect the associations between the entire human genome and common human diseases (myocardial infarction) in 2002 (Ozaki et al., 2002). Aranzana et al. (2005) were the first to use GWAS in plants by searching for associations with flowering time and pathogen resistance of model plant *Arabidopsis thaliana*. In recent years, a lot of GWA studies have been carried out in many major crop species, including cereal crops like wheat, rice and maize, legumes like soybeans, peas, chickpeas, and also the model plant species *Arabidopsis* and model legume *Medicago truncatula* (Tibbs Cortes, Zhang & Yu, 2021; Chen et al., 2021; Alseekh et al., 2021). GWA studies have been used to study various agronomic, physiological, metabolic and fitness traits including flowering time, height, metabolites, yield, biotic resistance, stress tolerance, etc. (Tibbs Cortes, Zhang & Yu, 2021; Alseekh et al. 2021). As important nutritional composition traits, seed protein and amino acid contents have been focused and studied in GWAS to identify associated genomic regions. Lee et al. (2019) conducted a GWAS in soybean and reported that three, one, three, one and four genomic regions were associated with seed protein, cysteine, methionine, lysine and threonine content, respectively. Several GWA studies were conducted to uncover the genetic basis for protein or amino acid contents in legumes, including soybeans (https://www.soybase.org/), common beans (Katuuramu et al., 2018), chickpea (Karaca et al. 2019). Determining the protein and amino acid contents in food crops is an essential step for GWA studies of these traits.

2.6. Traditional Methods for Protein and Amino Acid Determination

2.6.1 Reference methods for protein determination

The Kjeldahl method and Dumas combustion method are the two most prevalent methods to measure protein content in crops (Nielsen, 2010). The crude protein content is measured indirectly by determining the nitrogen (N) content; therefore, the results include nitrogen from nonprotein components (Nielsen, 2010). A conversion factor is used to convert percent N to percent crude protein. Because most proteins contain 16% N and most nitrogen is derived from protein, the conversion factor is usually set as 6.25 (Nielsen, 2010). However, 6.25 is not always an accurate conversion factor for legumes, which leads to overestimating the protein content due to nonprotein nitrogen and different amino acid profiles (Krul, 2019). Sosulski & Holt (1980) pointed out that the average conversion factor for 11 different legumes was 5.6, and specifically 5.72 for lentils. However, changing the conversion factor from 6.25 to other factors could have negative impacts such as food relabeling cost and devaluation on legume industries (Krul, 2019). Therefore, the conversion factor 6.25 is still used in most applications to calculate the crude protein content.

The traditional Kjeldahl method is based on sample digestion with a strong acid in the presence of catalysts to convert organic nitrogen to ammonium sulfate (Nielsen, 2010). The digest is neutralized and distilled into a boric acid solution. The nitrogen content is determined by the titration of ammonium borate with standardized acid (Nielsen, 2010). The basic Kjeldahl procedure, semi automation, automation and micro Kjeldahl methods have been established by AOAC in methods 955.04, 976.06, 976.05, and 960.52, respectively (*Official methods of analysis of AOAC International*, n.d.). The Kjeldahl method is inexpensive, accurate and relatively straightforward. Many researchers have used the Kjeldahl or modified Kjeldahl

methods to determine lentil protein content (Karaköy et al.,2012; Kumar et al.,2016). This official method can measure all types of foods, and it can apply to macro- or micro-scale. However, this method measures organic nitrogen content plus ammonia, not just protein nitrogen. It is time-consuming (>2h/sample) and involves the use of hazardous chemical reagents (Nielsen, 2010).

The Dumas combustion method is based on the combustion of the entire sample and selective detection of released nitrogen, which has been established by AOAC in method 992.23 (Tahir et al., 2011; *Official methods of analysis of AOAC International*, n.d.). N_2 content is measured using gas chromatography with a thermal conductivity detector (TCD). Compared with the Kjeldahl method, the combustion method is safer, faster (3min/sample), and it can analyze more samples since autosamplers can handle around 150 samples. However, this method has a high initial cost due to expensive equipment. Meanwhile, this complex instrumentation needs regular maintenance and dedicated technicians. It measures total nitrogen, including the inorganic part (Nielsen, 2010). Thompson et al. (2002) reported that the Dumas method usually provides a slightly higher result (about 1.4%) than the Kjeldahl method.

Another option to determine protein content directly is to measure the content of individual amino acids and calculate the sum of these contents (Krul, 2019). Understanding the amino acid profile is necessary to analyze protein quality.

2.6.2 Reference methods for amino acid analysis

The traditional wet chemical methods of amino acid analysis are based on the release of amino acids through acid or alkaline hydrolysis followed by separation and detection through chromatographic techniques (Nielsen, 2010). Differences in the stability of specific amino acids

to the hydrolysis conditions requires separate hydrolysis methods to be employed to determine the full amino acid profile: 1) regular acid hydrolysis (all amino acids without sulfur amino acid and tryptophan); 2) pre-oxidation step followed by acid hydrolysis (sulfur amino acids: methionine and cysteine); and 3) alkaline hydrolysis (tryptophan) (*Official methods of analysis of AOAC International*, n.d.). In the regular acid hydrolysis, a protein sample is hydrolyzed under an oxygen-free environment in constant boiling 6N HCl at 110^{o}C for 24h to release amino acids. Microwave-assisted acid hydrolysis has been reported to speed up the hydrolysis, which could become a rapid alternative to conventional hydrolysis (Kabaha et al., 2011; Themelis et al., 2019). Asparagine and glutamine lose amide N during acid hydrolysis to aspartic acid and glutamic acid, respectively (Rutherfurd & Gilani, 2009). Hydrolysates are neutralized and derivatized before separation on chromatography. Sulfur groups in the sulfur amino acids are unstable during regular acid hydrolysis. During this hydrolysis, the oxidation of sulphur amino acids is variable. A pre-oxidation step can convert all sulphur amino acids to their stable oxidized forms. In this method, performic acid is added prior to hydrolysis to oxidize cysteine and methionine to cysteic acid and methionine sulfone, respectively (*Official methods of analysis of AOAC International*, n.d.). Sodium metabisulfite is added to stop oxidation after 16 hours. The remaining steps are the same as for the regular amino acid analysis procedure. Tryptophan is completely destroyed by acid hydrolysis; therefore, the tryptophan content of the samples is analyzed after alkaline hydrolysis with barium hydroxide for 20 h at 110^{o}C in an autoclave. The hydrolysate is neutralized and separated by chromatographic methods (Fontaine et al., 2001). Various detectors such as ultraviolet and fluorescence can be used to measure the content of amino acids. The regular acid hydrolysis, pre-oxidation step followed by acid hydrolysis and alkaline hydrolysis methods agree with AOAC official methods 982.30, AOAC official methods

985.28 and ISO 13904:2005(E), respectively (*Official methods of analysis of AOAC International*, n.d.; ISO, 2005).

These wet chemical methods are very complicated, labor intensive, expensive and time consuming, and these drawbacks indicate that these methods are impractical to measure a larger number of samples from breeding programs (Baianu et al., 2004). Additionally, the high temperatures and dangerous chemical reagents may lead to unsafe conditions (González-Martín et al., 2006). Furthermore, these destructive and invasive methods make it impossible for samples to be used for further analysis. This can be a major problem when dealing with limited volumes of seeds from a breeding program. The Near-infrared (NIR) spectroscopy method can overcome many disadvantages of traditional methods to measure grain composition (Table 2.3). As a new, quick and non-destructive technology, NIR spectroscopy is worthy of in-depth study.

2.7. Near-infrared Reflectance Spectroscopy (NIRS)

2.7.1 Introduction

The foundation of Near-infrared (NIR) spectroscopy is that different functional groups and bonds from a sample absorb energy from electromagnetic radiation at wavelengths in the NIR region (780-2500nm) (Osborne, 2006). The NIR spectra occur when chemical bonds of a compound bend and stretch at matching frequencies, and NIR energy transfers from the radiation to the molecule (Nielsen, 2010). Bending and stretching are two major types of fundamental vibrations of covalent bonds (García-Sánchez et al., 2017). Nielsen (2010) mentions that the NIR-induced transitions of vibrating molecules include fundamental transitions (transition from energy level 0 to 1), overtones (transition from energy level 0 to >1) and combinations (transition of NIR energy to more than one fundamental absorptions). NIR spectra of legumes contain

many absorption bands mainly due to overtones and combinations of different chemical bonds such as N-H, C-H and O-H from proteins, carbohydrates and water (Blanco & Villarroya, 2002; *Official methods of analysis of AOAC International*, n.d). Therefore, the spectrum is a graph of absorbance versus wavelength (Osborne, 2006). Wavelength can be calculated as: wavelength = velocity of light/frequency (Osborne, 2006). The broad overlapping absorption bands (mainly overtones and combinations) of NIR spectra contain sufficient information of chemical compositions of samples; therefore, NIR spectroscopic method can be used to determine constituents quantitatively after establishing NIR models (Nielson, 2010).

NIR spectroscopy can make several measurements based on different characteristics of samples, usually reflectance mode for solids, transmittance for liquids, and transflectance for emulsions and turbid liquids (Blanco& Villarroya, 2002). Baianu et al. (2004) mentioned that reflectance is usually used to calculate the absorbance in practice because it is easier to measure, and absorbance = log (1/ reflectance). Reflectance = intensity of radiation reflected from the sample at a given wavelength (I) / intensity of radiation reflected from the reference at the same wavelength (I_0). Using near-infrared reflectance spectroscopy combined with linear chemometric algorithms is the basis for the prediction of crude protein and amino acid contents in solid samples like grains.

The main advantage of NIR spectrometry is that this method requires little sample preparation (Osborne, 2006). Moreover, the analysis of samples is rapid (30s to 2min/sample), applicable to on-line analysis, non-destructive, non-invasive, safe and simple, which experimenters can conduct with minimal training (Blanco & Villarroya, 2002; Nielsen, 2010). Non-destructive testing has significant applying value to deal with a limited volume of seeds from breeding programs. Furthermore, it can measure several constituents simultaneously

(Osborne, 2006). Compared to traditional wet chemistry methods, the cost per sample of the NIRS test can be significantly lower (Williams, Manley & Antoniszyn, 2019). Moreover, Khazaei et al. (2019) mentioned that the NIR method is a green technique without using chemical reagents. NIR also has relatively good accuracy and higher precision due to unnecessary sample treatment (Blanco & Villarroya, 2002).

However, this method still has some limitations. The initial cost is high because of expensive instruments (Nielson, 2010). Moreover, physical properties such as particle size and shape can obscure the chemical information from the spectra (Blanco & Villarroya, 2002; Osborne, 2006). Also, NIR spectroscopy needs extensive calibration against a reference method (Osborne, 2006). The instruments should be calibrated properly, which indicates that the content of interested compositions (proteins, amino acids, etc.) from representative samples for the calibration should be equivalent to or broader than that of unknown samples (Blanco & Villarroya, 2002). Lastly, as an indirect method, NIR spectroscopy depends on the precision of the reference methods (Osborne, 2006; Nielson, 2010).

2.7.2 Instrumentation for NIR

A NIR spectrometer usually consists of a radiation source (e.g. halogen lamp), a wavelength selection device, a sample presentation device (reflectance, transmittance and transflectance), a detector and a computer to show and process spectral data (Porep, Kammerer & Carle, 2015). There are several technologies to select the measurement wavelength, which leads to various NIR instrumentation. These technologies are listed in chronological order, including filters, scanning gratings, light emitting diodes (LED) with filters, diode arrays, Fourier transform near-infrared (FT-NIR), Acousto-optic tunable filter (AOTF), Liquid crystal tunable

filter (LCTF) and hyper-spectral imaging focal plane array (Stark & Luchter, 2005; Ozaki, McClure & Christy, 2006).

NIR instrumentation can be classified into two main groups with respect to wavelength selection, including discrete wavelength and whole spectrum (Blanco& Villarroya, 2002; Stark & Luchter, 2005). Filter instruments, as discrete-wavelength spectrophotometers, are the oldest, simplest and cheapest NIR instruments (Osborne, 2006). Filter instruments depend on narrow-band interference filters to represent the absorptions of protein, moisture, etc. (Osborne, 2006).

On the other hand, several technologies are available to develop whole-spectrum NIR instruments, including scanning grating monochromators, fixed grating detector diode array technology and FT-NIR interferometer technology (Stark & Luchter, 2005; Andersen, Wedelsback & Hansen, 2013).

Scanning monochromators based on diffraction grating are pre-dispersive instruments (Agelet & Hurburgh Jr, 2010). Monochromators are the most versatile instruments which can be used to measure the full NIR spectrum in various situations (Osborne, 2006). The grating functions as the dispersing element to disperse the radiation by wavelength (Nielsen, 2010).

Another instrument that contains a grating is the diode array spectrometer, which is a post-dispersive instrument (Agelet & Hurburgh Jr, 2010). The light over the entire NIR region irradiates samples and the reflected light is then directed onto a fixed grating which helps the multichannel diode array detectors to detect dispersed wavelengths (Blanco& Villarroya, 2002; Osborne, 2006; Nielsen, 2010). Therefore, the diode array (DA) spectrometers are good at dealing with a high sample throughput in that they can measure samples very fast, and they are qualified for on-line measurements (Osborne, 2006). Compared to the moving grating

spectrometers, a stationary grating with diode array detectors has greater wavelength accuracy because of simultaneous detection of reflectance at different wavelengths (Baianu et al. 2004).

Unlike scanning monochromators and diode array spectrometers, the radiation of Fourier transform (FT) instruments is not dispersed by a grating (Ozaki, McClure & Christy, 2006). Fourier transform near-infrared reflectance spectroscopy (FT-NIRS) is mainly comprised of a NIR light source, Michelson interferometer, sample compartment, detector amplifier and computer (Khan et al.,2018). The basic principle of FT-NIRS is the generation of a sample's interferogram and its subsequent conversion into a typical NIR spectrum through Fourier transformation, a mathematical treatment (Nielsen, 2010). The Michelson interferometer is the core and important part of the FT instrument, consisting of a beam splitter, a fixed mirror and a movable mirror (Khan et al.,2018). The NIR beam is split into two beams by beam splitter (transmitted beam to the fixed mirror while reflected beam to the moving mirror) and then recombined at the beam splitter by reflecting the radiation back with mirrors (Andersen, Wedelsback & Hansen, 2013). The pathlength of one beam can be changed by moving a mirror to a different distance, which results in two beams undergoing different constructive or destructive interference (Nielson, 2010). The resulting signal pattern is known as an interferogram, which is a pattern of radiation intensity obtained as a function of optical path difference (OPD) (Nielsen, 2010). This interferogram giving intensity vs. OPD can be converted into a spectrum showing absorbance vs. frequency through Fourier transformations, which can be done quickly by a computer (Nielsen, 2010). This machine has one detector, which can be either a photon detector or a thermal detector (Khan et al.,2018). According to Nielson (2010), compared to dispersive spectrometers, FT-NIRS instruments are more rapid, precise, sensitive with a higher signal-to-noise ratio. Moreover, they can measure a wider or the entire range of

wavelengths (Stark & Luchter, 2005). FT-NIRS instrument can have a high and changeable spectral resolution (Andersen et al., 2013; Baianu et al.,2019). The diversity of NIR spectrometers is increasing to satisfy the need for measuring different samples with higher speed, flexibility and reproducibility (Stark & Luchter, 2005).

2.7.3 Calibration of NIRS

The typically broad, extensively overlapping bands of NIR spectra make it necessary to use chemometrics to extract useful information from the convoluted spectra (Osborne, 2006; Porep, Kammerer& Carle, 2015; García-Sánchez et al. 2017). During the development of calibration models, absorption bands are assigned to specific functional groups (Williams, Manley& Antoniszyn, 2019).

There are seven steps to construct a multivariate model, including (1) selecting samples as a calibration set, (2) determining the protein and amino acid contents by using the reference methods, (3) getting the NIR spectra of all samples, (4) pre-treating spectra, (5) constructing the model by using multivariate methods, (6) validating the model and (7) predicting unknown samples (Blanco & Villarroya, 2002).

The selected samples should be appropriately normally distributed and cover a broad range in protein and amino acid contents of population to ensure representativeness (García-Sánchez et al. 2017).

2.7.3.1 Preprocessing of NIR spectra data

The absorbance calculated from reflectance can be influenced by the light scattering effect (Baianu et al., 2004). Physical characteristics of the material such as sample heterogeneities, particle size and shape can cause the scattering effect (Huang, Romero-Torres & Moshgbar, 2010). The scattering effect can be categorized into three groups: additive effect (baseline shift), multiplicative effect and wavelength-dependent baseline variation (Huang, Romero-Torres & Moshgbar, 2010). Before developing calibration equations, appropriate spectra preprocessing should be applied to correct noises and improve the signal-to-noise ratio. The preprocessing aims at minimizing unrelated variation (Agelet & Hurburgh Jr, 2010).

Several pre-treatments or combinations can be used for NIR spectra. The most popular methods include derivatives, de-trending, standard normal variate (SNV) and multiplicative scatter correction (MSC) (Blanco & Villarroya, 2002; Huang, Romero-Torres & Moshgbar, 2010). The purpose of derivatives is to enhance signal through resolving overlapping peaks and to remove constant baseline drift and baseline slope (Agelet & Hurburgh Jr, 2010). De-trending, also known as baseline corrections, can remove additive effect by subtracting a polynomial fit of baseline from each spectrum (Huang, Romero-Torres & Moshgbar, 2010). The SNV and MSC are two commonly used methods to minimize both additive and multiplicative effects, and these two methods can generate similar results (Agelet & Hurburgh Jr, 2010). The SNV method is preferred because of its simplicity and effectiveness (Huang, Romero-Torres & Moshgbar, 2010). Moreover, SNV corrects each spectrum individually by subtracting the mean and dividing by the standard deviation for that spectrum. The MSC method is more complicated and requires a reference spectrum (Agelet & Hurburgh Jr, 2010).

2.7.3.2 Regression techniques for NIR

According to García-Sánchez et al. (2017), different multivariate analytical methods can be used to build calibration models and to obtain a correlation between the spectral data and the reference concentration, including linear methods such as partial least squares (PLS) regression, principal components regression (PCR) and multiple linear regression (MLR), and nonlinear methods such as artificial neural networks (ANN) and support vector machines (SVM).

The most used linear chemometric algorithm for calibration is partial least squares (PLS) regression. Unlike MLR, PLS and PCR can deal with correlated wavelengths. Compared to PCR, PLS has a faster algorithm and provides more precise models (Agelet & Hurburgh Jr, 2010).

Kovalenko, Rippke & Hurburgh (2006) used PLS and ANN to build calibration equations for predicting amino acid composition in whole soybeans, and they concluded that the model performance of PLS was significantly better than that of ANN.

2.7.3.3 Outlier detection

The first step is to plot the spectra, and the abnormal spectra can be detected by visual check (Agelet & Hurburgh Jr, 2010). The samples of those spectra should be rescanned to get the right spectra. Moreover, after building the first attempt of the calibration model, samples with a high leverage and a high residual value can become a potential outlier. Leverage or Hotelling's T^2 belong to influence measures. The high values show that the samples have a strong influence on the model. Samples with high model residuals indicate they are badly described by the model. The exclusion of outliers can improve the performance of calibration models; however, the outliers should be removed carefully without significantly reducing the representativeness of models (Agelet & Hurburgh Jr, 2010).

2.7.3.4 Validation of NIR models

Validation can evaluate the accuracy of the calibration model, determine the ability to predict the contents of new samples, and avoid overfitting (Nicolai et al., 2007). Two commonly used validation methods are external validation and cross validation.

In external validation, several samples in the calibration set are used to build the calibration model, and this model is used to predict the interested compositions of samples from a separate test set. This independent validation can give the best estimate of prediction error (Nicolai et al., 2007; Williams, Antoniszyn & Manley, 2019). However, withholding data will decrease sample variation and adversely influence the model performance.

In cross validation, the samples are the same as for calibration models. Usually, a group of samples have been left out and predicted by the calibration model built from the remaining samples. Another group of samples replace it until all samples have been predicted and used in the calibration models (Agelet & Hurburgh Jr, 2010). Cross validation has been broadly accepted and widely used in NIRS studies. It becomes a powerful method to determine the PLS factors without overfitting (Williams, Antoniszyn & Manley, 2019). However, compared to external validation, it can't provide the best estimate of future prediction error (Williams, Antoniszyn & Manley, 2019).

2.7.3.5 Statistical terms for model evaluation

Several statistical terms are useful to evaluate the efficiency of NIRS models, including Coefficient of Determination for Calibration (R^2_C), Coefficient of Determination for Cross-Validation / 1 minus the ratio of unexplained variance to total variance (R^2_{CV} / 1-VR), Coefficient of Determination for Prediction (R^2_P), Root Mean Square Error of Calibration

(RMSEC), Root Mean Square Error of Cross-Validation (RMSECV), Root Mean Square Error of Prediction (RMSEP), Standard Error of Calibration (SEC), Standard Error of Cross-Validation (SECV), Standard Error of Prediction (SEP), Residual predictive deviation (RPD), slope and bias (Nicolai et al., 2007; Agelet & Hurburgh Jr, 2010; García-Sánchez et al. 2017; Shi & Yu, 2017; Quiñones et al., 2018). Table 2.2 shows the explanation and units of several statistical terms.

According to Williams, Manley & Antoniszyn (2019), when the NIR model has a high R^2, a low SEP, a low bias and a slope close to 1, this model is highly efficient and predicts the results accurately.

Table 2.2. Common Statistical terms when reporting the efficiency of an NIRS model given from Agelet & Hurburgh Jr (2010) and Williams Manley & Antoniszyn (2019)

Statistics	Explanation	Units
R^2/ RSQ	Explained variance	Unitless
RMSE (/-C/-CV/-P)	Calibration (validation) error	Same as reference values
SE (/-C/-CV/-P)	The standard deviation of residuals[a]	Same as reference values
RPD	SD/SEP	Unitless
slope	Slope of regression line	Unitless
bias	Average values of residuals	Same as reference values

[a] Residuals: differences between NIRS predicted and reference values.

2.7.4 Application

NIRS is applicable to many food categories such as cereals, dairy products, meat, vegetables, etc. (Osborne, 2006; Quiñones et al., 2018). Canada is a pioneer in utilizing near-infrared spectroscopy (NIRS) to measure protein content. In 1975, the Canadian Grain Commission (CGC) used NIRS to replace the Kjeldahl system in the Canadian wheat protein segregation program (Williams, Manley & Antoniszyn, 2019). The replacement of the Kjeldahl system resulted in significantly reduced cost, reduced chemical waste and a larger scale for testing protein in an environmentally friendly way (Blanco & Villarroya, 2002; Osborne, 2006). The AOAC has accepted the NIRS method 997.06 for the analysis of crude protein in wheat (*Official methods of analysis of AOAC International*, n.d.). AACC International has approved the NIR technique for measuring protein in cereal crops and soybeans (AACC International, 2010). Near-infrared reflectance spectroscopy (NIRS) has become a robust and routine method to determine protein contents in legumes (Quiñones et al., 2018). Wang et al.(2014a) developed NIRS model to quantify the protein content in faba bean with coefficient of determination (R^2) values equal to 0.97 and 0.88 for milling power and intact seeds, respectively. The results indicated that NIRS could be a reliable method for predicting protein content accurately. Plans et al. (2013) pointed out that FT-NIR had the best predictive performance for the determination of protein contents in common beans (R^2=0.97; RPD=3.65). NIRS can predict crude protein content of lentils with great accuracy (Moldovan et al., 2015; Revilla et al.,2019). In terms of soybeans, Ferreira, Pallone & Poppi (2013) developed a near-infrared calibration model for the protein content based on 100 ground soybean samples. The model presented high R^2 (0.81) which indicated that FT-NIRS had the high predictability for soybean seed protein content. Zhu et al. (2018) reported soybean powder with particle sizes of 0.3mm was most suitable to construct

NIRS protein model with R^2_{CV} = 0.953. The protein content of single soybean seed could be accurately predicted by four different NIR instruments and R^2 of all calibration models were higher than 0.92 (Esteve Agelet et al., 2012). Moreover, there are several studies about the potential of using NIRS to predict amino acid compositions.

Rubenthaler & Bruinsma (1978) reported the first successful determination of the amino acid lysine through Near-infrared reflectance spectroscopy (NIRS), and the results showed that calibrated model could predict lysine content from wheat precisely (r=0.98). More research was focused on the prediction of amino acids in crops through NIRS in the following years. In 1997, Pazdernik, Killam& Orf concluded that NIRS could be used as a gross screening method for most amino acids in soybeans, and NIRS was more accurate for analysis of ground seed than whole seed samples. Zhang et al. (2011) reported that the NIRS calibration equations had good performance for most amino acids (except for Cys, Met, Tyr and Trp) in brown rice. Those calibration models showed high coefficients of determination (0.837-0.947) and low standard errors (Zhang et al., 2011). Li et al. (2011) reported that NIRS coupled with the modified partial least squares permitted the accurate and fast analysis of Asp, Thr, Ser, Gly, Ile, Leu, Lys and Pro in stevia powders. According to Fontaine, Hoerr, Schirmer & Fontaine (2001), NIRS calibrations were developed to predict amino acid contents of some protein-rich feedstuffs successfully and rapidly, including soy, rapeseed meal, sunflower meal, peas, fishmeal, meat meal products and poultry meal. Kovalenko, Rippke & Hurburgh (2006) developed NIRS models for the estimation of amino acid composition of soybeans. For the spectrometer DA 7200 (Perten Instruments Inc., Springfield, IL), PLS regression produced models with the highest RPD values. Except for Cys, Glu, Met, Ser and Trp, other amino acids had RPD values higher than 2, demonstrating that NIRS could be a reliable and rapid method for determining several amino acids in soybeans

(Kovalenko, Rippke & Hurburgh, 2006). Carbas et al. (2020) showed the potential use of NIRS to determine the several amino acids in common beans. For legumes, values of methionine, cystine and tryptophan might be difficult to measure by NIRS due to low concentration, which requires more studies to develop the accurate calibration models (Fontaine, Hoerr, Schirmer & Fontaine, 2001).

Overall, using NIRS to determine protein and amino acid contents is more time-efficient, cost-effective, safer and simpler. The test is non-destructive and non-invasive, which makes it possible for further tests of samples. This technology can help farm managers understand more about the crops, making it possible to segregate or blend crops based on their protein content (Osborne, 2006). The amino acid profiles of lentils can reflect protein quality. It is an ideal and high throughout technology for rapid screening in quality detection systems and lentil breeding programs (Zhang et al.,2011; Khazaei et al., 2019).

Table 2.3. The advantages and disadvantages of reference methods and near-infrared reflectance spectroscopy to measure protein and amino acid contents

Methods	Advantages	Disadvantages
Kjeldahl Method (Traditional method to measure proteins)	➢ Applicable to all types of foods ➢ Relatively straightforward ➢ Relatively simple equipment ➢ Precise and accurate ➢ Inexpensive (relative to automated Dumas) ➢ Applicable to macro- or micro-scale ➢ Can be automated ➢ Preferred method for high-fat samples ➢ AOAC, AACC reference methods	➢ Measure all organic nitrogen content instead of true protein content ➢ Various conversion factors ➢ Use of hazardous chemical reagents ➢ time consuming
Dumas combustion method (Traditional method to measure proteins)	➢ Quick (<5 min/sample) ➢ Easy to use ➢ No corrosive or toxic chemical agents ➢ Many samples can be analyzed since that autosamplers can handle more than 100 samples ➢ A lower detection limit, compared to Kjeldahl method ➢ AOAC reference methods	➢ Measure all nitrogen content instead of true protein ➢ Various conversion factors ➢ High initial cost ➢ Complex instrumentation ➢ Small sample size makes it difficult to obtain representative sample ➢ Unsuitable for high-fat samples
Traditional method to measure amino acids	➢ Relatively accurate ➢ AOAC reference methods	➢ Time consuming (≥ 3 days) ➢ Labor intensive ➢ Use of hazardous chemical reagents ➢ High cost per sample
Near-infrared Reflectance Spectroscopy	➢ Little or no sample preparation ➢ Very safe ➢ Easy to install and use ➢ Non-destructive; non-invasive ➢ Quick (30s to 2min/sample) ➢ Simultaneous analysis of several parameters ➢ Green technique; environmentally friendly ➢ Low cost per sample ➢ Comparable in accuracy to traditional methods ➢ Great precision ➢ Versatile and flexible ➢ Small-size and durable equipment ➢ Networking (use the same calibration)	➢ Extensive calibration ➢ Relatively expensive equipment

CHAPTER 3: HYPOTHESES & OBJECTIVES

3.1 Hypotheses

NIRS models from DA 7250 have the same predictive ability as models from FT 9700 for the measurement of protein and amino acid contents in lentils.

Several single nucleotide polymorphisms (SNP) markers significantly associated with protein and amino acid of lentil seeds can be identified through genome-wide association study (GWAS).

3.2 Objectives

(i) To develop and evaluate the NIRS models for the prediction of protein and amino acid composition in whole and ground lentils for two types of NIR spectrometers: DA 7250 and FT 9700 (PerkinElmer Health Sciences Canada Inc., Winnipeg, MB, Canada)

(ii) To analyze the effects of sample status, spectrometers and amino acid correlations to protein on the performance of NIR models

(iii) To evaluate variation of protein and amino acid compositions in cultivated lentils from Canada

(iv) To detect SNPs significantly associated with protein and 18 amino acids content by using GWAS

CHAPTER 4: PREDICTION OF PROTEIN AND AMINO ACID CONTENTS IN WHOLE AND GROUND LENTILS USING NEAR-INFRARED REFLECTANCE SPECTROSCOPY

4.1 Abstract

Lentil is an important source of plant-based protein, and the protein and amino acid contents have a significant influence on its nutritional quality and value. The conventional reference methods for protein and amino acid determination are very complicated, time consuming, labor intensive, expensive and destructive. Therefore, near-infrared reflectance spectroscopy (NIRS), as a quick, simple and non-destructive analytical method, has high practical value to measure these compositions. This study developed NIRS calibration models by partial least squares (PLS) regression to predict the protein and 18 amino acid contents of lentil seeds. The effects of sample status (whole and ground), type of spectrometer (PerkinElmer DA 7250 and FT 9700), and amino acid/protein correlation on model performance were analyzed and evaluated. The DA 7250 equations of protein and 14 amino acids, except histidine, tyrosine, methionine and cysteine, showed a relatively high coefficient of determination for calibration (R^2_C = 0.652 - 0.918) for both whole and ground lentils. FT 9700 models had R^2_C from 0.665 to 0.927 for most compositions (except tyrosine, methionine and cysteine) in ground lentils. Generally, most compositions could be predicted with similar accuracy in whole and ground status. DA 7250 models had a slightly better predictive ability with higher R^2_{CV} and RPD values than FT 9700 models for all compositions except histidine. However, the difference between the predicted data from two spectrometers was not significant ($p>0.05$) for each composition. NIRS models performed better for certain amino acids when they were highly correlated to protein. The NIRS calibration models were clearly superior to crude protein regressions in predicting the amino acid

contents. Overall, NIRS combined with PLS regression had a significant potential for rapid and simultaneous prediction of protein and most amino acid contents in lentils with satisfactory accuracy, and these models were usable for research purposes or sample screening.

KEYWORDS: *Near-Infrared Reflectance Spectroscopy, lentils, protein, amino acids, calibration, partial least squares regression*

4.2 Introduction

Lentil (*Lens culinaris*) is a nutritious pulse which has health benefits, supports fast cooking compared to other main pulses and can be consumed in whole, dehulled and split form (Khazaei et al., 2019). As a leguminous plant, lentil can fix nitrogen from the atmosphere in their root nodules through symbiotic rhizobia. Similar to other legumes, lentil is a good and important source of protein, containing around 26% crude protein on a dry basis (Khazaei et al., 2019). Previous studies reported that the protein contents of lentils can vary from 10.5% - 36.4% (Hawtin, Rachie & Green, 1977; Kumar et al., 2016; Khazaei et al., 2019). Lentil protein contains all essential amino acids, and it provides a sufficient amount of some essential amino acids such as leucine, lysine, threonine and phenylalanine (Boye, 2015). However, like many other leguminous seeds, lentils generally lack sulfur amino acids (methionine and cysteine) and tryptophan (Boye, 2015; Nosworthy et al., 2017). The combination of cereal and lentils can provide a complete essential amino acid profile, enhancing the protein quality and nutritional value (Erskine, 2009; Moldovan et al., 2015).

Lentil is essential in the plant-based protein market. According to Health Canada (2019), the plant-based protein market is booming globally due to the environmental, economic, social and health benefits of plant proteins. Several factors have led to increasing consumer demand for lentils, including (i) protection of environment and animal welfare, (ii) low growing cost and low purchase price, (iii) increasing demand for protein due to a growing global population, (v) great choice for vegans and vegetarians as a protein source, (vi) health benefits because of a high content of protein, fiber, many vitamins and minerals (Health Canada, 2019; Tahir et al., 2011).

Due to these benefits of lentils, lentils have received more attention and have been cultivated in many countries. Canada leads the global market in lentil production and export,

accounting for around 40% of total global production in 2019 (FAO, 2021). Lentils are primarily grown in Saskatchewan, with this province accounting for approximately 87% of Canada's lentil production, followed by Alberta and Manitoba (Wang, 2020).

The new trend in plant-based protein consumption combined with the high production yield of Canadian lentils suggest a closer look at the protein and amino acid contents of lentils grown in Canada. Understanding the protein content and amino acid profile of lentils can help the food industry to make value-added products with correct nutritional claims and can benefit lentil breeding programs. However, the conventional reference methods, also known as wet chemistry methods, for protein and amino acid determination are very complicated, labor intensive, expensive and time consuming (Baianu et al., 2004). These drawbacks indicate that these methods are unsuitable for measuring large numbers of samples from breeding programs. In addition, the high temperatures and dangerous chemical reagents used in the reference methods may lead to unsafe conditions (González-Martín et al., 2006). Furthermore, these destructive and invasive methods make it impossible for samples to be used for further analysis, which can be highly problematic when dealing with small mass samples.

The Near-infrared Reflectance Spectroscopy (NIRS) analytical method can overcome many disadvantages of conventional reference methods for the measurement of grain composition, as it (i) requires little sample preparation, (ii) is rapid (30s to 2min/sample), non-destructive, non-invasive, safe and simple, (iii) can measure several constituents simultaneously, (iv) is a green technique without using chemical reagents, (v) has a low cost per sample, and (vi) is capable of analyzing a large number of samples with a limited volume per sample from breeding programs (Blanco & Villarroya, 2002; García-Sánchez et al. 2017; Khazaei et al.,2019; Nielsen, 2010). This method is based on different functional groups and bonds from a sample

absorbing energy from electromagnetic radiation in the NIR region 780-2500nm (Osborne, 2006). Chemometrics can extract useful information from the broad overlapping absorption bands corresponding to overtones and combinations of chemical bonds (Nielson, 2010).

Canada is a pioneer in utilizing near-infrared (NIR) spectroscopy to measure protein content. In 1975, the Canadian Grain Commission (CGC) used NIR spectroscopy to replace the Kjeldahl system in the Canadian wheat protein segregation program (Williams, Manley & Antoniszyn, 2019). NIR spectroscopy has been applied by AOAC in method 997.06 for analysis of crude protein in wheat and AACC International for measuring protein in cereal crops and soybeans (*Official methods of analysis of AOAC International*, n.d.; AACC International, 2010). In the case of lentils, CGC has determined the crude protein content by NIR spectroscopy calibrated against the Dumas combustion reference method since 1999 (Daun, 1999). Some studies show that NIRS can predict the crude protein content of lentils with great accuracy (Quiñones et al., 2018; Revilla et al.,2019). Moreover, since 1978, many researchers have investigated the potential of NIRS to predict amino acid contents in different crops such as wheat (Rubenthaler & Bruinsma,1978; Fontaine, Schirmer & Hörr, 2002), soybean (Fontaine et al., 2001; Kovalenko, Rippke & Hurburgh, 2006; Baianu et al., 2004), pea (Fontaine et al., 2001), milled rice (Wu, Shi & Zhang, 2002), peanut (Wang et al., 2013; Yu et al., 2020), brown rice (Zhang et al., 2011), and stevia leaf powder (Li et al., 2011). These studies show that the predictive ability of NIRS calibration models can be influenced by grain type, sample status (whole, dehulled, ground, etc.), sample size, specific compound (amino acid, protein, etc.), type of NIR spectrometers and different regression methods (Kovalenko, Rippke & Hurburgh, 2006; Wang et al., 2013; Zhang et al., 2011). According to Kovalenko, Rippke & Hurburgh (2006), the

NIRS model performance of each amino acid was also dependent on the correlation between respective amino acid and crude protein content.

As far as we know, no works were done to explore the feasibility of NIRS for quantitatively determining the 18 amino acids in whole and ground lentils. Also, the effects of NIR spectrometers and correlation of amino acid concentration with protein on the predictive ability of NIRS models have not been sufficiently studied. Therefore, the objectives of present study were to i) develop and evaluate the NIRS models for the prediction of protein and amino acid composition in whole and ground lentils for two types of NIR spectrometers and ii) analyze the effects of sample status, spectrometers and amino acid correlations to protein on the performance of NIR models.

4.3. Methods and Materials

4.3.1 Samples and sample preparation

A total of 1290 lentil samples were collected and sent by the Crop Development Centre (CDC) at the University of Saskatchewan. These samples consisted of 324 diverse lentil genotypes and were grown in 2 locations (Rosthern, Canada and Sutherland, Canada) for two years (2016 and 2017). The names, origins, and sources of 324 lentil genotypes are available fromhttps://github.com/derekmichaelwright/AGILE_LDP_Phenology/blob/master/Supplemental_Table_01.csv (Wright et al., 2021). Six samples were missing due to protection by a licensing agreement, including Entry 109 from 16Sut, Entry 7, 319 from 17Ros, Entry 7, 239, 319 from 17Sut.

These samples from a lentil diversity panel showed high variation in adaptation and traits. The protein contents of all lentils were estimated from the existing factory NIR calibrations, and

around 90 samples were randomly selected from each quartile of crude protein % of lentils. Therefore, 361 representative lentil samples were selected as a calibration set. These samples were selected and ground into powder by an Ultra Centrifugal Mill ZM 200 (Retsch, Haan, Germany) with a 0.75 mm sieve for further NIRS scanning and wet chemistry analysis. The particle size distribution of these powders was measured using Mastersizer 2000 Version 5.22 (Malvern Instruments, UK). The median particle size (d_{50}) was around 475µm.

4.3.2 Spectra collection

All samples as received were scanned by an at-line NIRS analyzer, DA 7250 (PerkinElmer Health Sciences Canada Inc., Winnipeg, MB, Canada). The DA 7250 belongs to the family of diode array spectrometers, and it can analyze several components in samples within 6 seconds. The wavelength range is from 950nm to 1650nm with an interval of 5nm. Every sample (around 50g) was scanned twice in the rotating sample tray and the average spectrum was used for analysis. The spectra of whole lentil samples were collected by DA 7250. The ground lentil samples were scanned in the same DA 7250 system to collect the spectra of ground lentil samples. Moreover, ground samples were scanned on a FT 9700 (PerkinElmer Health Sciences Canada Inc., Winnipeg, MB, Canada) between 14304 cm-^{1}(699nm) and 3856 cm^{-1}(2593nm) with an interval of 8 cm^{-1}. The FT 9700 is a Fourier transform near-infrared reflectance spectroscopy (FT-NIRS) instrument, which has a broader spectral range that covers the entire NIR region. Therefore, the spectra of ground lentil samples were collected by both DA 7250 and FT 9700. Specifications of these two NIR instruments are listed in Table 4.1.

Table 4.1. Specifications of the NIR Spectrometers

Characteristic	DA 7250	FT 9700
Technology	Fixed grating detector diode array Thermoelectrically cooled 256-element InGaAs[a] detector	FT-NIR Interferometer, InGaAs [a] detector
Radiation source	Halogen lamp	Halogen lamp
Mode	Reflectance	Reflectance
Wavelength range	950nm-1650nm	14304 cm^{-1}(699nm) – 3856 cm^{-1}(2593nm)
Wavelength spacing	5nm (default and recommended setting)	8 cm^{-1} (default and recommended setting)
Number of data points	141	1307
Number of scans	~20 spectra/second	32
Analysis time	6 seconds	Less than 25 seconds

[a] indium–gallium–arsenide

4.3.3 Compositional analysis by reference methods (conventional wet chemistry methods)

Protein contents were determined by the Dumas Combustion method according to AOAC method 990.03 (*Official methods of analysis of AOAC International*, n.d.). Nitrogen Analysis was conducted by Central Testing Labs (Winnipeg, MB, Canada). The conversion factor was set as 6.25 to convert percent nitrogen to percent crude protein (N × 6.25).

Dry matter or moisture was determined by weighing approximately 0.5g of the sample in a tared metal dish and putting them into the drying oven at 104^{o}C overnight. Samples were cooled in a desiccator and weighed again. Samples were measured twice, and the average value was used for calculation. Dry matter % = 1- moisture% = (Weight of dried samples and metal dish – weight of metal dish)/ Weight of samples * 100%

The conventional wet chemistry methods determined the contents of the following 18 amino acids: histidine (His), serine (Ser), arginine (Arg), glycine (Gly), aspartic acid (Asp), glutamic acid (Glu), threonine (Thr), alanine (Ala), proline (Pro), cysteine (Cys), lysine (Lys), tyrosine (Tyr), methionine (Met), valine (Val), isoleucine (Ile), leucine (Leu), phenylalanine (Phe) and tryptophan (Trp). Separate hydrolysis methods were required to determine the complete amino acid profile because amino acids have different stabilities to the hydrolysis conditions.

All amino acids, except Met, Cys and Trp, were analyzed by the regular acid hydrolysis. Asparagine (Asn) and Glutamine (Gln) were deaminated during acid hydrolysis to Asp and Glu, respectively (Rutherfurd & Gilani, 2009). Approximately 50mg of a lentil sample was weighed in 11ml Pyrex glass tubes. One external standard, SRM 3234 – Soy Flour (a reference material from the National Institute of Standards and Technology), was included per batch. For the analysis, 4mL of 0.1% phenol 1.25 mM norvaline 6N HCl was added, with norvaline functioning as an internal standard. Two drops of 2-octanol were added to each sample to mitigate bumping

during hydrolysis. After purging these tubes with nitrogen gas to prevent oxidation, tubes were placed in a conventional air oven and hydrolyzed at 110^{o}C for 24 hours. After cooling down to room temperature, the pH of the samples was adjusted to pH 5.5-6 by the addition of approximately 4 ml 25% NaOH. The deionized water was added to bring the volume to 50ml in a volumetric flask. Approximately 1ml of samples were transferred into cryovial via a syringe with a 13mm 0.22µm Nylon luer lock filter. The hydrolysates were stored at -20^{o}C. The commercially available AccQ•Tag^{TM} Ultra Derivatization Kit was used as a pre-column derivatization method and analysis technique (Waters Corporation, 2014). Sample analysis was conducted using a Shimadzu Nexera ultra-high performance liquid chromatography (UHPLC) system (Kyoto, Japan) equipped with a Waters AccQ C18 column (100mm×2.1mm, 1.7µm). The running eluents included Buffer A, prepared by diluting AccQ-Tag Ultra Eluent A 20 times, and Buffer B, 2% formic acid in acetonitrile. The column oven temperature was set as 51^{o}C. The derivatives were determined by UV detection at 260nm and the run time was 17 minutes. The Lab Solutions software (Shimadzu, Kyoto, Japan) was used to process data from the UHPLC. The free amino acid molecular weights were used to calculate amino acid contents.

Two sulfur amino acids (i.e., Met and Cys) were determined through performic acid oxidation followed by acid hydrolysis. Approximately 50mg of a lentil sample was weighed in a glass tube with a glass stopper. Performic acid was made up of 9 parts 0.1% phenolic 88% formic acid with 1 part 30% hydrogen peroxide. 2ml of freshly prepared and cold performic acid was added for sample oxidation in a fridge overnight. The cysteine and methionine were oxidized to cysteic acid and methionine sulfone, respectively, and 0.35g of sodium metabisulfite was added to stop oxidation. The sample was hydrolyzed with 2mL of 2.5mM norvaline in concentrated HCl and placed in the heating block at 110^{o}C for 18 hours. The following

procedures to prepare hydrolysates and derivatization were the same as regular acid hydrolysis. The column oven temperature was set at 40°C and 60°C for analysis of cysteine and methionine, respectively. Detection was by fluorescence with excitation at 266 nm and emission at 473 nm. The run time was 30 minutes for each sulfur amino acid.

The Trp content was analyzed by alkaline hydrolysis. To approximately 50mg of sample, 14 mL deionized water and 8.4 g barium hydroxide octahydrate were added in a polypropylene Erlenmeyer flask, which was loosely capped and autoclaved for 20 h at 110 °C. After removing samples from the autoclave, 30ml deionized water, 5ml of 0.5M orthophosphoric acid, and 8.3 ml of 6N HCl were added to flasks. The pH of the hydrolysate was adjusted in the range of 2.95 to 3.2. The 20ml methanol and deionized water were added to bring the volume to 100ml. The sample was filtered through a 0.22μmnylon filter and injected into a Phenomenex Luna C18 column (250mm×4.6mm, 3μm) with a flow rate of 1ml/min. The buffer, as well as the mobile phase, consisted of 0.3% glacial acetic acid and 0.05% 1,1,1-trichloro-2-methyl-2-propanol in water, brought to a pH of 5 with ethanolamine. The running time of a sample by reversed phase UPLC was 34 minutes. Specific fluorescence detection was applied using an excitation wavelength of 280 nm and an emission wavelength of 356 nm.

Regular and oxidized amino acid analytical methods are according to AOAC official methods 982.30 and 985.28, respectively (*Official methods of analysis of AOAC International*, n.d.). The method for tryptophan analysis followed the method positioned by the International Organizations for Standardization 13904:2005 (ISO, 2005).

4.3.4 Development and evaluation of calibration models

The spectral data and data from reference methods of 361 selected lentil samples were imported in the Unscrambler® X version 10.3 software (CAMO Software, Oslo, Norway), which was used to perform spectral data preprocessing, build calibrations, cross-validation models and make predictions. As recommended by the instrument supplier (PerkinElmer), spectra were first treated with de-trending and standard normal variate (SNV) to minimize scattering effects and reduce particle size noise. Partial least squares (PLS) regression, as a linear chemometric algorithm, was used to build calibration and cross-validation models. The samples for cross-validation were the same for calibration development. The number of segments in cross validation was set to 20, which indicated that the samples were randomly divided into 20 groups. In all, 19 groups were used to build the calibration model, which predicted the results of the remaining group, and this procedure was repeated 20 times to ensure that all the groups were predicted. The maximum number of factors was chosen when standard error of a twenty-fold cross-validation achieved the lowest value to prevent overfitting. The outliers were identified as chemical outliers when a point was outside ±3SDs of calibration y-residuals, which indicated that these samples didn't fit the calibration model.

Several statistical terms were used to evaluate the performance of NIRS models, including the Coefficient of Determination for Calibration (R^2_C), Coefficient of Determination for Cross-Validation (R^2_{CV} / 1-VR), Standard Error of Calibration (SEC), Standard Error of Cross-Validation (SECV) and Residual predictive deviation (RPD). The coefficient of determination represents explained variance. Standard error indicates the standard deviation of residuals. Residual predictive deviation, calculated as SD/SECV, is a non-dimensional statistic that can be used for model evaluation (Agelet & Hurburgh Jr, 2010).

4.4. Results and Discussion

4.4.1 Spectral characteristics

Figure 4.1 and Figure 4.2 provide the raw NIRS spectra (950–1650nm) from the DA 7250 spectrometer of 361 whole and ground lentil samples, respectively. Figure 4.3 depicts the raw NIRS spectra from the FT 9700 spectrometer of 361 ground lentil samples over the spectral range of 14304 - 3856 cm^{-1} (699nm - 2593nm). The absorptions in the NIR region consist of overtone and combination bands of C-H, O-H, N-H, C=O, S-H, etc. (Williams, Manley & Antoniszyn, 2019; *Official methods of analysis of AOAC International*, n.d.).

Using the same spectrometer, the spectra of different lentil samples were very similar to one another. Therefore, no apparent outliers came from plotting the spectra. The spread in the y-axis or scatter in spectral data was larger when the spectral data was above 1450nm/6896.6cm^{-1}. The stronger absorber and scattering effects at the higher wavelengths were the primary cause. Similar results related to increased scatter in the longer wavelength region was reported in wheat samples (Williams, Manley & Antoniszyn, 2019).

The spectra from Figure 4.1 showed a considerable variation on the y-axis. The greater variation in particle size and shape of lentil samples could cause a larger scattering effect (Williams, Manley & Antoniszyn, 2019). On the other hand, Figure 4.2 shows smaller scatter in spectral data, possibly because ground samples had reduced particle size with the median particle size d_{50} = 475µm. Williams, Manley & Antoniszyn (2019) reported a similar reduced scattering effect in wheat flour compared with whole wheat kernels. Some negative values of absorbance indicated that the reflections of samples were higher than the reflection of the reference at certain wavelengths, and the potential reason was the finely powdered material became whiter with stronger reflection. The negative values did not influence the subsequent chemometrics. Figure

4.1 and Figure 4.2 illustrated two similar peaks: the peaks near 1205nm were related to C-H second overtone and the absorption bands around 1480nm were related to N-H stretching first overtone of protein (Williams, Manley & Antoniszyn, 2019; Saha et al.,2017).

The FT-NIRS spectra showed more peaks in regions with longer wavelengths (Figure 4.3). The principal absorption bands of protein could be observed between 4808 cm^{-1} (2080nm) and 4544 cm^{-1} (2201nm). These absorption bands could be related to N-H bending second overtone, N-H combination, C-H stretching and C=O bending of protein (Shi & Yu, 2017; Saha et al., 2017). These absorption bands around 2100nm could also be related to C-O, O-H stretching combination of carbohydrates (Nielson, 2010; Williams, Manley & Antoniszyn, 2019). The typically broad, extensively overlapping bands due to overtones and combination make it necessary to use chemometrics to extract useful information from the convoluted NIRS spectra.

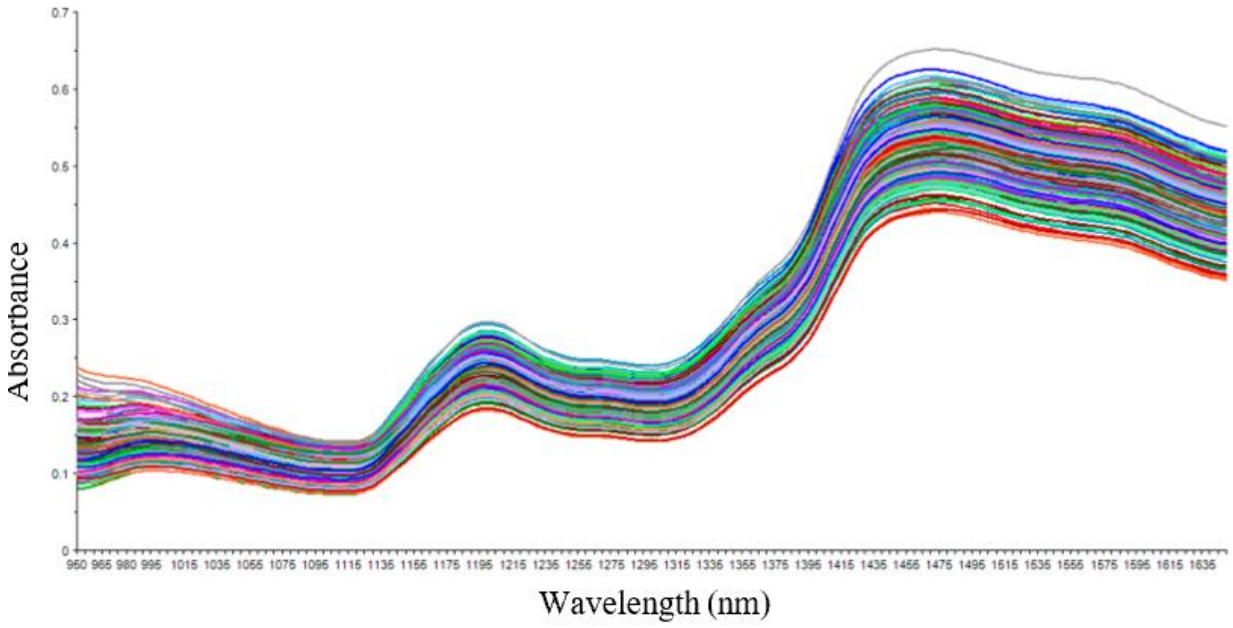


Figure 4.1. Raw infrared spectra of whole lentil samples from DA 7250 (950 -1650nm)

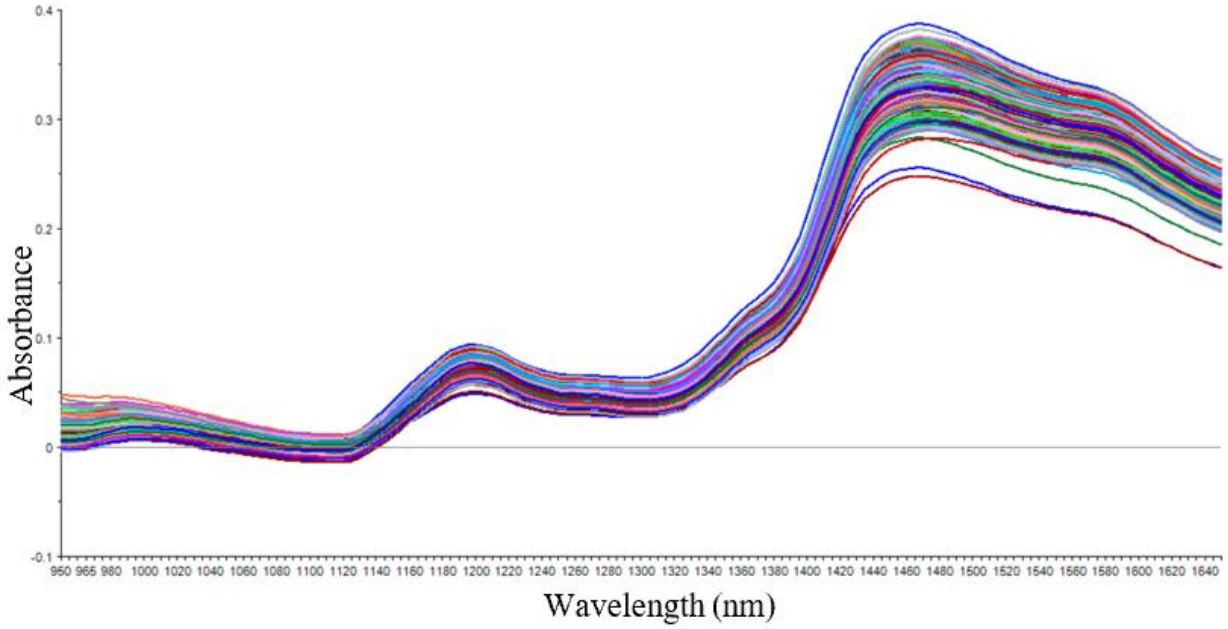


Figure 4.2. Raw infrared spectra of ground lentil samples from DA 7250 (950 -1650nm)

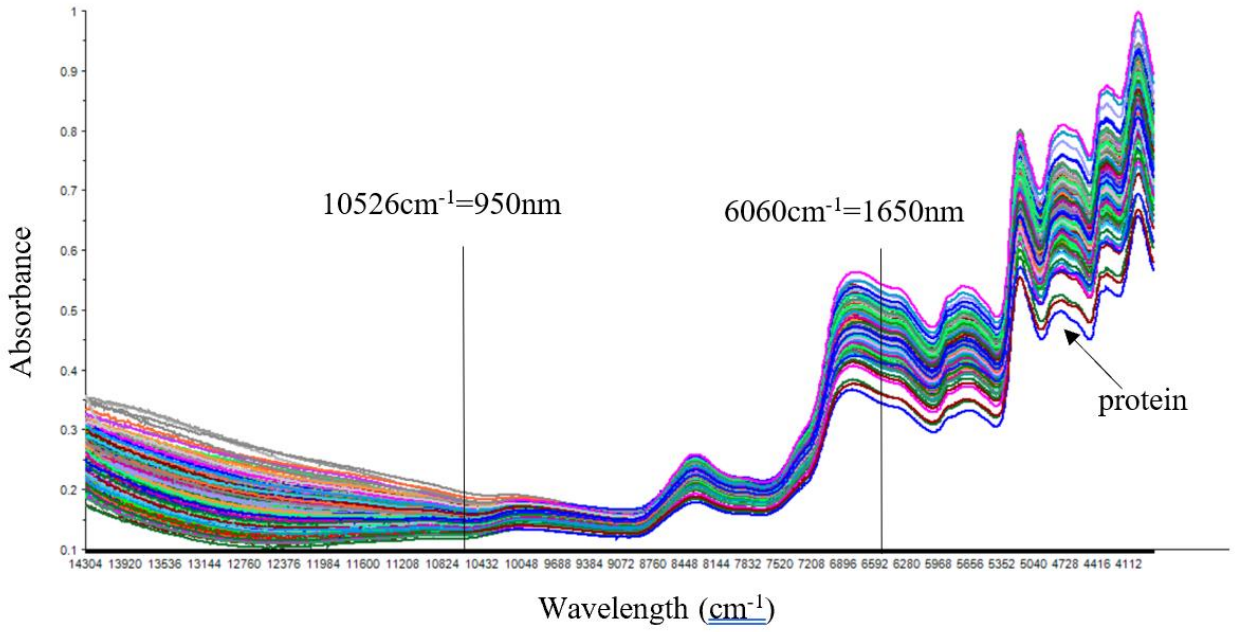


Figure 4.3. Raw infrared spectra of ground lentil samples from FT 9700 (14304 cm^{-1}(699nm) and 3856 cm^{-1}(2593nm))

4.4.2 Sample statistics

The statistics of protein and amino acid concentrations on an as is basis and on a dry basis by reference methods are given in Table 4.2 and Table 4.3, respectively. These results showed a broad range in protein and amino acid contents to ensure representativeness, and it was essential to enhance the performance of NIRS models. Table 4.2 shows the protein content for selected 361 lentils grown in 2016 and 2017 ranged from 21.6% to 32.4% on a wet basis. The mean crude protein content was 26.8%, which was a slightly higher than that (24.6%) from the database of USDA (2018) and Health Canada (2018). Iqbal et al. (2006) reported a similar protein content (26.1%) in lentils. The average value of all amino acids except Cys, Lys, Met and Trp from this study was slightly higher than the corresponding value from the USDA (2018). The two most abundant amino acids in lentils were Glu and Asp. One potential reason was that the Glu content was the sum of Glu and Gln and the Asp value was the combination of Asp and Asn. Data from this study and official database (USDA, 2018) both indicated that lentils contained low amounts (0.2%-0.3%) of Trp, Met and Cys on an as-is basis. These observations were consistent with those reported by Iqbal et al. (2006) and Nosworthy et al. (2018).

Table 4.3 shows all compositions on a dry basis, and these values are more commonly used in literature. The average protein content of lentils was 29.6% on a dry matter basis. Khazaei et al. (2019) mentioned that the protein content of whole dry lentil seeds ranged from 10.5% - 36.4% from previous studies, which covered the range of protein contents (24.3%-35.0%) in lentils from this study. Canadian Grain Commission measured the protein content of western Canadian lentils by near-infrared measurement calibrated against the Dumas combustion method (Wang, 2016). The mean protein content was 27.1% and 25.6% on a dry basis in 2016 and 2017, respectively (Wang, 2016; Wang 2017). The mean protein content in this study was

higher than the results from the Canadian Grain Commission, possibly because of measuring different lentil types.

Moreover, the slopes of linear regression of all amino acids to crude protein were all positive (Table 4.2, Table 4.3), and these slopes were all significantly different from zero ($p<0.001$). Among these amino acids, Leu, Gly, Glu, Ile, Phe, Ala, Asp, Val and Pro were strongly correlated with protein with RSQ > 0.5 on both as is basis and dry basis.

Table 4.2. Descriptive statistics of protein and amino acid contents (% as is basis) of 361 lentil samples measured by reference methods [a]

						Linear regression of amino acids to crude protein		
	Mean	SD	CV	Min	Max	slope	intercept	RSQ
Protein	26.814	1.752	6.535	21.625	32.375			
His	0.886	0.166	18.690	0.439	1.435	0.055	-0.576	0.332
Ser	1.280	0.097	7.563	0.843	1.531	0.041	0.186	0.546
Arg	1.933	0.172	8.922	1.500	2.948	0.061	0.290	0.387
Gly	1.027	0.066	6.471	0.854	1.264	0.032	0.159	0.729
Asp	3.029	0.249	8.211	2.101	3.842	0.113	-0.002	0.634
Glu	4.185	0.298	7.127	3.373	5.033	0.145	0.296	0.726
Thr	0.978	0.075	7.692	0.789	1.184	0.007	0.783	0.029
Ala	1.056	0.079	7.495	0.867	1.389	0.036	0.091	0.634
Pro	1.093	0.066	6.062	0.930	1.331	0.030	0.294	0.620
Cys	0.236	0.026	10.946	0.167	0.334	0.003	0.148	0.050
Lys	1.713	0.120	7.008	1.367	2.088	0.052	0.315	0.579
Tyr	0.771	0.055	7.198	0.642	0.947	0.020	0.246	0.382
Met	0.236	0.024	10.180	0.153	0.302	0.006	0.084	0.171
Val	1.263	0.077	6.080	1.057	1.504	0.034	0.345	0.610
Ile	1.110	0.067	6.058	0.918	1.305	0.031	0.268	0.670
Leu	1.891	0.120	6.349	1.544	2.267	0.060	0.284	0.765
Phe	1.289	0.074	5.751	1.062	1.524	0.034	0.380	0.643
Trp	0.204	0.014	6.955	0.165	0.248	0.006	0.054	0.479
SAA	0.472	0.042	8.960	0.356	0.616	0.009	0.232	0.138

[a] Abbreviation: SD, standard deviation; CV, coefficient of variation; Min, minimum; Max, maximum; SAA, sulphur amino acids (Cys + Met)

Table 4.3. Descriptive statistics of protein and amino acid contents (% dry basis) of 361 lentil samples measured by reference methods [a]

						Linear regression of amino acids to crude protein		
	Mean	SD	CV	Min	Max	slope	intercept	RSQ
Protein	29.615	1.690	5.707	24.326	34.997			
His	0.978	0.177	18.076	0.496	1.553	0.053	-0.596	0.258
Ser	1.414	0.099	7.012	0.960	1.699	0.040	0.231	0.464
Arg	2.136	0.197	9.224	1.642	3.294	0.075	-0.082	0.413
Gly	1.135	0.066	5.805	0.952	1.367	0.032	0.197	0.660
Asp	3.345	0.251	7.500	2.392	4.201	0.111	0.067	0.556
Glu	4.622	0.295	6.386	3.860	5.484	0.142	0.427	0.658
Thr	1.082	0.093	8.642	0.888	1.331	0.012	0.737	0.044
Ala	1.166	0.080	6.831	0.954	1.501	0.035	0.121	0.561
Pro	1.207	0.065	5.379	1.060	1.463	0.028	0.385	0.523
Cys	0.261	0.029	11.096	0.183	0.368	0.003	0.162	0.038
Lys	1.891	0.118	6.228	1.550	2.302	0.048	0.466	0.477
Tyr	0.851	0.061	7.177	0.707	1.024	0.021	0.231	0.336
Met	0.261	0.025	9.628	0.166	0.331	0.005	0.123	0.098
Val	1.395	0.082	5.865	1.157	1.676	0.036	0.341	0.541
Ile	1.227	0.069	5.630	1.028	1.445	0.032	0.295	0.594
Leu	2.088	0.118	5.665	1.773	2.483	0.059	0.353	0.701
Phe	1.424	0.078	5.483	1.219	1.699	0.035	0.397	0.564
Trp	0.226	0.015	6.655	0.179	0.271	0.006	0.058	0.408
SAA	0.522	0.046	8.751	0.391	0.666	0.008	0.286	0.087

[a] Abbreviation: SD, standard deviation; CV, coefficient of variation; Min, minimum; Max, maximum; SAA, sulphur amino acids (Cys + Met)

4.4.3 Evaluating the performance of models

After removing outliers for each composition, basic statistics and performance of the DA 7250 NIRS models for protein and amino acid compositions on a dry basis and on as is basis in whole lentil samples are summarized in Table 4.4 and Table 4.5, respectively. For ground lentil samples on a dry basis, the calibration statistics from the DA 7250 and FT 9700 NIRS models are shown in Table 4.6 and Table 4.7, respectively.

Several statistical terms can be used for the interpretation of NIRS results and evaluation of model efficiency. Models with a higher coefficient of determination (R^2), higher RPD and lower standard error values were deemed to be more accurate and efficient (Williams & Norris, 2001; Wang et al., 2013; Lohr et al., 2016). According to standard equations, R^2 and RPD are unitless while the unit of standard error is same as reference values (Agelet & Hurburgh Jr, 2010). RPD and R^2 provided similar statistical information, and RPD could be calculated from R^2 ($RPD = 1/\sqrt{1 - R^2}$) when SD is unavailable (Kovalenko, Rippke & Hurburgh, 2006; Fernández-Novales et al.,2019; Williams, Manley & Antoniszyn, 2019).

Table 4.4 shows the crude protein model had the best performance, with the highest calibration and cross-validation determination coefficients (R^2_c=0.90; R^2_{cv}=0.88) and the highest RPD (2.84). The protein content was much higher than other amino acids; therefore, the standard errors were higher than other models. Moreover, the calibration and cross-validation equations for crude protein, Gly, Ala, Val, Ile, Leu and Phe had high R^2_c (0.82-0.90), R^2_{cv} (0.78-0.88) and RPD (2.16-2.84) values. These results suggested that those models were suitable for analytical purpose, which could predict variables with good accuracy (Williams & Norris, 2001; Shi and Yu, 2017). The R^2_c (0.67-0.81), R^2_{cv} (0.61-0.77) and RPD (1.62-2.06) of Ser, Arg, Asp, Glu, Thr, Pro, Lys and Trp indicated that these models were usable for sample screening and provided

approximate quantitative prediction (Williams & Norris, 2001; Shi and Yu, 2017; Smyth et al., 2008). For His, Cys, Tyr and Met, the relatively low R^2_c (0.41-0.61), R^2_{cv} (0.35-0.53) and RPD (1.23-1.46) indicated that NIRS models for these compositions were only acceptable for very rough screening to distinguish between high and low values with careful use (Williams & Norris, 2001; García-Sánchez et al., 2017). The RPD values of 18 amino acids in lentils ranged from 1.23 to 2.42. A similar range of RPDs (1.09-3.24) for 18 amino acids in whole soybeans was reported by Kovalenko, Rippke & Hurburgh (2006). Fernández-Novales et al. (2019) said that the RPDs for analyzed amino acids in grape berries were in the range of 1.21-1.64, which were slightly lower than results from current study.

More attention was given to two sulfur amino acids: Cys and Met. The model performance of these two amino acids were inferior to NIRS models of other compositions. The models for Cys (R^2_c=0.41; R^2_{cv}=0.35) could explain less variation than that for Met (R^2_c=0.56; R^2_{cv}= 0.46). Similar results were reported by Fontaine et al. (2001) in peas (R^2_c=0.43 for Cys and R^2_c=0.68 for Met) and Kovalenko, Rippke & Hurburgh (2006) in soybeans (RPD = 1.25 for Cys and RPD = 1.51 for Met). Compared with the models of individual Cys or Met, the sulphur amino acids (Cys+Met) calibration model could perform slightly better with higher determination coefficients (R^2_c=0.57; R^2_{cv}=0.47). Fontaine et al. (2001) also reported the Cys+Met models had higher RPD values than the individual amino acid regression model in soybean meal and sunflower meal. Nevertheless, it was obvious that regression models of sulphur amino acids were inferior to other models. The potential reasons were as follows: (i) lentils had low content of sulphur amino acids, and their values were more susceptible to noises or interferences; (ii) the peroxidation step might cause additional preparation error.

To improve predictive abilities of models of Cys, Met, Tyr and His, further studies can be done to include more lentil samples from different environments in the calibration set to cover a wider range of these compositions. Lentil samples, especially outlier samples, could be analyzed by reference methods again to obtain an accurate lab result and enhance the integrity of the reference analysis. Moreover, reference wet chemistry methods could be modified and optimized for lentils, possibly through changing hydrolysis time, UPLC condition, etc.

Models from Table 4.4 had stable numbers of PLS factors, ranging from 11 to 16 for protein and all amino acids except Cys (factor = 8). Williams, Manley & Antoniszyn (2019) indicated that successful PLS models usually had factors higher than 6-8 to ensure that these factors can account for most of the variance in the calibration system. However, models with a factor higher than 15 should be used more cautiously to prevent overfitting (Williams, Manley & Antoniszyn, 2019). The number of factors in the Thr model was 16, and only this one exceeded 15. The number of factors could depend on external variation in crop varieties, temperature, moisture, solar radiation, etc. Since there was high variation in the calibration set, it was acceptable for models to have relatively high factors.

In Table 4.5, the models were developed based on the wet chemistry data on as is basis, and the results were similar to those in Table 4.4. The determination coefficients and RPD values of most compositions (except Arg, Thr, Cys, Tyr and Trp) on as is basis were slightly higher than those on a dry basis. The difference illustrated that NIRS models were slightly more efficient in quantifying concentrations of most compositions on a wet basis as compared to on a dry basis. The reason might be that the moisture content was measured after grinding, and the grinding step could cause errors in the calculation of dry matter due to moisture loss. The best performance was also observed for the prediction of protein (R^2_C=0.92, SEC=0.49, SECV=0.55, and

RPD=3.11). Revilla et al. (2019) reported that R^2_c, SEC, SECV, and RPD of protein in whole lentils on a wet basis were 0.96, 0.36, 0.33 and 5.4, respectively. Their study showed better predictability for the determination of protein due to higher R^2 and lower errors. The possible reason was that samples in this study were from two different cropping years while their samples were harvested the same year. Since the protein seems to be sensitive to environmental stress, the protein of lentils from different years may have different compositions or structures (Revilla et al., 2019). This could influence the model performance. Furthermore, 9.5% (4 out of 42) samples were removed as outliers in their studies. In this study, only 1.1% (4 out of 361) samples were removed to build the calibration model for protein. In terms of amino acids, the NIR models for all amino acids except His, Cys, Tyr and Met had satisfactory performance with $R^2_c > 0.5$ and RPD > 1.5.

The calibration statistics for ground lentils are shown in Table 4.6. The R^2_C (0.83-0.91), R^2_{CV} (0.78-0.90) and RPD (2.15-3.10) values for crude protein, Gly, Ala, Val, Leu and Phe were relatively high, which indicated that NIRS models could predict the contents of these components with satisfactory accuracy. These models were suitable for research use in most situations. Moreover, NIRS models for Ser, Arg, Asp, Glu, Thr, Pro, Lys, Ile, Trp were able for sample screening because R^2_C values for these eight amino acids were >0.65, R^2_{CV} values and RPD were all higher than 0.62 and 1.64, respectively. It was worth noting that R^2_C (0.53-0.57) values of all remaining amino acids were higher than 0.50. However, the R^2_{CV} values (0.43-0.53) and RPD (1.33-1.45) values of His, Cys, Tyr and Met were still lower than other models. These models with lowest fit may only be suitable for very rough screening with careful use (Williams & Norris, 2001; Shi and Yu, 2017). These model with $R^2c > 0.4$ and RPD >1.2 could be still useful to distinguish between high and low values (García-Sánchez et al., 2017). Wu, Shi &

Zhang (2002) also mentioned that NIRS equations of Cys, Met and His showed low R^2 in milled rice mainly because of relatively low concentrations in rice samples, and these amino acids could not be predicted with confidence.

The calibration and cross validation statistics for the FT 9700 NIR models are shown in Table 4.7. The best performance was observed for the prediction of crude protein content (R^2_C = 0.93; R^2_{CV} = 0.89 and SECV = 0.56). Ferreira et al. (2014) mentioned the R^2 and SECV of FT-NIR models for protein in ground soybeans were 0.88 and 1.76, respectively, which were a little inferior to my results. One possible reason was that this study included 361 samples in the calibration set while their study only had 30 samples. Regarding amino acids, NIRS could predict the content of Leu with relatively higher accuracy (R^2_C = 0.88 and RPD = 2.16). The R^2_C (0.67-0.86) and RPD (1.64-2.05) values for Ser, Arg, Gly, Asp, Glu, Thr, Ala, Val, Ile, Phe and Trp indicated these NIRS models were acceptable for sample screening. The R^2_C value of Pro and Lys were both higher than 0.5 together with RPD ≥ 1.5, indicating these models were suitable for very rough screening (Williams & Norris, 2001; Shi and Yu, 2017). The relatively low R^2_C (0.46-0.69), R^2_{CV} (0.34-0.54) and RPD (1.22-1.46) values of His, Tyr and Met models reflected that these models may be used to distinguish between high and low values with caution. Cys model had poor correlation (R^2_C = 0.38, RPD = 1.18), thus FT 9700 couldn't produce suitable analytical models for this amino acid (Shi and Yu, 2017).

Table 4.4. Calibration and cross-validation results for the DA 7250 NIR models (950 –1650 nm) of protein and amino acids content (% dry basis) in whole lentils [a]

								Calibration		Cross-validation		
Compound	Mean	SD	CV	Min	Max	N	Factor	SEC	R^2_C	SECV	R^2_{CV}	RPD
Protein	29.642	1.659	5.598	25.692	34.997	357	14	0.521	0.901	0.585	0.877	2.837
His	0.973	0.171	17.561	0.496	1.474	357	14	0.116	0.539	0.131	0.42	1.304
Ser	1.421	0.088	6.193	1.180	1.699	351	12	0.044	0.75	0.048	0.709	1.834
Arg	2.130	0.184	8.618	1.642	2.642	358	12	0.086	0.782	0.092	0.749	1.995
Gly	1.135	0.065	5.753	0.952	1.367	357	15	0.024	0.863	0.027	0.83	2.418
Asp	3.349	0.244	7.289	2.671	4.201	357	14	0.113	0.784	0.126	0.733	1.937
Glu	4.621	0.295	6.387	3.860	5.484	358	13	0.13	0.806	0.143	0.768	2.064
Thr	1.082	0.093	8.613	0.894	1.331	358	16	0.042	0.8	0.05	0.718	1.864
Ala	1.167	0.079	6.770	0.966	1.501	360	13	0.033	0.825	0.036	0.793	2.194
Pro	1.206	0.064	5.284	1.060	1.463	358	11	0.037	0.672	0.039	0.623	1.634
Cys	0.261	0.028	10.824	0.183	0.344	359	8	0.022	0.406	0.023	0.351	1.226
Lys	1.892	0.117	6.173	1.550	2.302	360	13	0.065	0.686	0.072	0.618	1.622
Tyr	0.851	0.061	7.180	0.707	1.024	360	13	0.038	0.61	0.042	0.528	1.455
Met	0.261	0.024	9.282	0.203	0.329	356	14	0.016	0.555	0.018	0.46	1.344
Val	1.395	0.082	5.869	1.157	1.676	359	14	0.035	0.82	0.038	0.782	2.155
Ile	1.226	0.068	5.566	1.028	1.445	359	12	0.028	0.83	0.03	0.806	2.275
Leu	2.088	0.118	5.672	1.773	2.483	360	14	0.045	0.856	0.05	0.824	2.369
Phe	1.424	0.077	5.428	1.219	1.699	360	13	0.031	0.834	0.034	0.803	2.273
Trp	0.226	0.015	6.539	0.186	0.271	358	13	0.008	0.682	0.009	0.61	1.639
SAA	0.521	0.045	8.606	0.391	0.666	358	13	0.03	0.565	0.033	0.472	1.360

[a] Abbreviation: SD, standard deviation; CV, coefficient of variation; Min, minimum; Max, maximum; N, number of samples used for calibration and cross-validation; SEC, standard error of calibration; R^2_c, coefficient of determination for calibration; SECV, standard error of cross-validation; R^2_{cv}, coefficient of determination for cross-validation; RPD, residual predictive deviation (SD/SECV); SAA, sulphur amino acids (Cys + Met).

Table 4.5. Calibration and cross-validation results for the DA 7250 NIR models (950 –1650 nm) of protein and amino acids content (% as is basis) in whole lentils [a]

								Calibration		Cross-validation		
Compound	Mean	SD	CV	Min	Max	N	Factor	SEC	R^2_C	SECV	R^2_{CV}	RPD
Protein	26.840	1.721	6.411	22.500	32.375	357	15	0.493	0.918	0.553	0.897	3.111
His	0.882	0.161	18.227	0.439	1.347	358	14	0.107	0.559	0.12	0.446	1.340
Ser	1.286	0.088	6.820	0.989	1.531	352	13	0.041	0.785	0.045	0.741	1.950
Arg	1.927	0.161	8.349	1.500	2.338	358	14	0.073	0.796	0.081	0.751	1.987
Gly	1.027	0.066	6.421	0.854	1.264	359	14	0.024	0.872	0.026	0.84	2.537
Asp	3.033	0.243	7.999	2.374	3.842	357	13	0.106	0.81	0.118	0.765	2.056
Glu	4.184	0.298	7.128	3.373	5.033	359	13	0.126	0.821	0.138	0.786	2.161
Thr	0.978	0.075	7.673	0.789	1.184	358	16	0.036	0.77	0.043	0.679	1.745
Ala	1.056	0.079	7.442	0.883	1.389	360	15	0.03	0.85	0.034	0.811	2.312
Pro	1.092	0.066	6.011	0.930	1.331	359	10	0.036	0.704	0.038	0.67	1.728
Cys	0.236	0.025	10.658	0.167	0.318	359	8	0.02	0.366	0.021	0.315	1.197
Lys	1.713	0.119	6.949	1.367	2.088	360	13	0.059	0.752	0.065	0.704	1.832
Tyr	0.770	0.055	7.171	0.642	0.947	360	10	0.036	0.574	0.038	0.52	1.454
Met	0.236	0.023	9.894	0.185	0.302	356	14	0.015	0.589	0.017	0.479	1.374
Val	1.263	0.077	6.080	1.057	1.504	359	12	0.031	0.832	0.034	0.806	2.259
Ile	1.110	0.067	5.996	0.918	1.305	359	12	0.027	0.841	0.028	0.818	2.377
Leu	1.890	0.120	6.329	1.544	2.267	358	13	0.043	0.874	0.047	0.849	2.545
Phe	1.289	0.074	5.720	1.062	1.524	360	13	0.029	0.84	0.032	0.808	2.303
Trp	0.204	0.014	6.887	0.167	0.248	360	13	0.008	0.696	0.009	0.624	1.564
SAA	0.472	0.042	8.887	0.356	0.616	359	12	0.028	0.541	0.031	0.47	1.352

[a] Abbreviation: SD, standard deviation; CV, coefficient of variation; Min, minimum; Max, maximum; N, number of samples used for calibration and cross-validation; SEC, standard error of calibration; R^2_c, coefficient of determination for calibration; SECV, standard error of cross-validation; R^2_{cv}, coefficient of determination for cross-validation; RPD, residual predictive deviation (SD/SECV); SAA, sulphur amino acids (Cys + Met).

Table 4.6. Calibration and cross-validation results for the DA 7250 NIR models (950 –1650 nm) of protein and amino acids content (% dry basis) in ground lentils [a]

								Calibration		Cross-validation		
Compound	Mean	SD	CV	Min	Max	N	Factor	SEC	R^2_C	SECV	R^2_{CV}	RPD
Protein	29.640	1.653	5.576	25.692	34.997	355	11	0.503	0.907	0.533	0.897	3.101
His	0.975	0.173	17.786	0.496	1.478	359	11	0.119	0.527	0.127	0.463	1.366
Ser	1.421	0.088	6.197	1.180	1.699	350	11	0.042	0.773	0.045	0.741	1.957
Arg	2.128	0.182	8.568	1.642	2.642	356	10	0.078	0.816	0.082	0.798	2.224
Gly	1.135	0.066	5.810	0.952	1.367	360	14	0.026	0.848	0.03	0.8	2.198
Asp	3.351	0.241	7.181	2.671	4.201	354	11	0.108	0.798	0.117	0.766	2.057
Glu	4.621	0.295	6.379	3.860	5.484	357	10	0.131	0.801	0.139	0.779	2.121
Thr	1.080	0.091	8.460	0.894	1.331	357	13	0.043	0.78	0.048	0.726	1.904
Ala	1.166	0.079	6.773	0.966	1.501	359	16	0.03	0.854	0.036	0.791	2.194
Pro	1.205	0.062	5.178	1.060	1.439	356	7	0.037	0.652	0.038	0.626	1.642
Cys	0.261	0.028	10.740	0.183	0.344	357	14	0.019	0.54	0.021	0.426	1.333
Lys	1.891	0.116	6.151	1.579	2.302	357	11	0.064	0.701	0.069	0.653	1.686
Tyr	0.851	0.061	7.153	0.707	1.024	359	9	0.04	0.572	0.042	0.529	1.449
Met	0.261	0.024	9.353	0.203	0.329	359	13	0.016	0.558	0.018	0.44	1.354
Val	1.395	0.081	5.832	1.157	1.676	357	12	0.033	0.834	0.036	0.804	2.260
Ile	1.226	0.068	5.579	1.028	1.445	359	9	0.031	0.8	0.032	0.783	2.137
Leu	2.088	0.118	5.672	1.773	2.483	360	13	0.045	0.858	0.05	0.82	2.369
Phe	1.423	0.077	5.431	1.219	1.699	359	14	0.031	0.839	0.036	0.782	2.147
Trp	0.226	0.015	6.655	0.179	0.271	361	16	0.008	0.726	0.009	0.63	1.669
SAA	0.521	0.045	8.691	0.391	0.666	360	13	0.029	0.594	0.032	0.506	1.416

[a] Abbreviation: SD, standard deviation; CV, coefficient of variation; Min, minimum; Max, maximum; N, number of samples used for calibration and cross-validation; SEC, standard error of calibration; R^2_c, coefficient of determination for calibration; SECV, standard error of cross-validation; R^2_{cv}, coefficient of determination for cross-validation; RPD, residual predictive deviation (SD/SECV); SAA, sulphur amino acids (Cys + Met).

Table 4.7. Calibration and cross-validation results for the FT 9700 NIR models (14304 cm^{-1} – 3856 cm^{-1}) of protein and amino acids content (% dry basis) in ground lentils [a]

								Calibration		Cross-validation		
Compound	Mean	SD	CV	Min	Max	N	Factor	SEC	R^2_C	SECV	R^2_{CV}	RPD
Protein	29.623	1.669	5.633	25.692	34.997	356	11	0.45	0.927	0.555	0.89	3.007
His	0.978	0.176	17.959	0.496	1.553	359	11	0.099	0.685	0.12	0.536	1.463
Ser	1.421	0.088	6.193	1.180	1.699	351	12	0.036	0.837	0.049	0.694	1.796
Arg	2.130	0.184	8.618	1.642	2.642	358	9	0.086	0.78	0.097	0.722	1.892
Gly	1.134	0.066	5.780	0.952	1.367	359	10	0.026	0.839	0.032	0.769	2.049
Asp	3.349	0.241	7.190	2.692	4.201	356	10	0.112	0.785	0.138	0.674	1.745
Glu	4.622	0.292	6.323	3.873	5.484	358	10	0.117	0.839	0.147	0.749	1.988
Thr	1.082	0.093	8.598	0.894	1.331	360	9	0.05	0.713	0.055	0.655	1.692
Ala	1.167	0.079	6.748	0.966	1.501	359	10	0.034	0.818	0.041	0.723	1.921
Pro	1.206	0.064	5.284	1.060	1.463	358	10	0.033	0.736	0.04	0.604	1.593
Cys	0.261	0.028	10.824	0.183	0.344	359	8	0.022	0.378	0.024	0.264	1.175
Lys	1.892	0.116	6.141	1.550	2.302	359	9	0.065	0.684	0.074	0.597	1.570
Tyr	0.852	0.061	7.121	0.707	1.024	359	7	0.042	0.527	0.044	0.481	1.379
Met	0.261	0.024	9.353	0.203	0.329	359	9	0.018	0.463	0.02	0.337	1.219
Val	1.395	0.081	5.829	1.157	1.676	357	9	0.036	0.802	0.041	0.751	1.983
Ile	1.227	0.069	5.630	1.028	1.445	361	11	0.027	0.844	0.036	0.724	1.918
Leu	2.088	0.119	5.678	1.773	2.483	359	11	0.041	0.883	0.055	0.788	2.155
Phe	1.424	0.077	5.428	1.219	1.699	360	12	0.029	0.862	0.04	0.734	1.932
Trp	0.226	0.015	6.528	0.186	0.271	359	9	0.008	0.665	0.009	0.592	1.637
SAA	0.522	0.045	8.631	0.391	0.666	359	9	0.032	0.481	0.036	0.358	1.251

[a] Abbreviation: SD, standard deviation; CV, coefficient of variation; Min, minimum; Max, maximum; N, number of samples used for calibration and cross-validation; SEC, standard error of calibration; R^2_c, coefficient of determination for calibration; SECV, standard error of cross-validation; R^2_{cv}, coefficient of determination for cross-validation; RPD, residual predictive deviation (SD/SECV); SAA, sulphur amino acids (Cys + Met).

4.4.4 Potential influencing factors on predicative ability of NIR models

By comparing results from Table 4.4 and from Table 4.6, the influence of sample status (whole or ground) on model performance could be analyzed. Generally, the number of factors for ground samples was lower than that for whole samples, except Ala, Cys, Phe and Trp. Considering R^2_C and R^2_{CV} values, the grinding procedure could improve the predictive ability of models for crude protein, Ser, Arg, Asp, Cys, Lys, Val and Trp. The R^2_C, R^2_{CV} and SECV difference between ground samples and whole samples for protein were 0.006, 0.02 and -0.052, respectively. A slight improvement of model performance was observed for the prediction of protein after grinding. The absolute values of the difference between R^2_{CV} for ground samples and R^2_{CV} for whole samples of every composition were lower than 0.05. The absolute values of SECV difference were lower than 0.01 for all amino acids. Moreover, the values of R^2_{CV} difference were in the range of -0.03 – 0.075 for all compositions. Therefore, NIRS calibration equations developed from ground samples showed similar performance and accuracy to those with whole samples. The possible explanation was that the pre-treatments (De-trending and SNV) could minimize the influence of particle size and background noise. Revilla et al. (2019) drew a similar conclusion that NIRS could predict the protein content for both whole and ground lentils. Their results showed that NIRS models had a slightly better predictability for crude protein in ground lentils than in whole lentils, but generally these two models achieved similar and high accuracy (Revilla et al., 2019). This is highly important for lentil breeders to estimate the protein and amino acid contents in whole lentils without grinding by using NIR models, and these intact lentils are suitable for further uses.

These two spectrometers (DA 7250 and FT 9700) displayed different predictability for the determination of protein and amino acids in ground lentils (Table 4.6 and Table 4.7). Apart

from His, DA 7250 could predict all other compositions with higher R^2_{CV}, lower SECV and higher RPD values than FT 9700. Therefore, models from DA 7250 might have greater suitability for predicting the contents of most analyzed compositions (except His) than the FT 9700 models. It was worth noting that the FT 9700 models were developed with lower numbers of factors (between seven to twelve), which could be regarded as a positive outcome. Carbas et al. (2020) mentioned a low number of PLS factors may indicate a more robust and representative model. To test the agreement of these two spectrometers, the protein and amino acid contents of 361 samples were predicted by models developed from DA 7250 and FT 9700, respectively. The samples used for calibration were also used for prediction. Table 4.8 summarizes statistics from Bland-Altman analysis and P value from paired t-test to analyze the agreement between FT 9700 and DA 7250 spectrometers. Bland and Altman analysis was used to assess the agreement between the two instruments (Bland & Altman, 1986). The largest bias/mean difference was 0.007% for protein. The bias was lower than 0.003% for all amino acids. All the differences of individual compositions were normally distributed through the Shapiro-Wilk test. Bland and Altman (1986) recommended that 95% of the data points should lie within ± 2SD of the mean difference. In this case, around 361*5% ≈ 18 could be outside the limits, and the results (13-21) from Table 4.8 conformed to this recommendation. The 95% limits of agreement (-0.75% ~ 0.763%) for protein are small; therefore, these two spectrometers have an acceptable degree of agreement for predicting protein content. The discrepancy for His could be up to 0.15% while the reference His content was around 0.98%, which indicated that FT 9700 and DA 7250 had a relatively poor agreement to measure His. The paired t test showed the predicted data of two spectrometers didn't have a significant difference ($p>0.05$) for every composition (Table 4.8). In

general, these two machines had a relatively good agreement for measuring protein and most amino acids.

The performance of NIRS models might be influenced by correlations between amino acids and crude protein. Figure 4.4 illustrates a positive relationship between R^2 from crude protein regression and R^2_C from NIR models. The slope was 0.46, and it was significantly different from 0 ($p<0.05$). When the amino acid was highly correlated to crude protein, the NIRS could predict this amino acid with higher accuracy. NIRS models for amino acids strongly correlated with protein (RSQ > 0.5) could be more effective ($R^2_C > 0.67$). A similar observation was found in soybean (Kovalenko, Rippke & Hurburgh, 2006). The R^2_C from NIR models was always higher than the corresponding R^2 from crude protein regression. Compared to the crude protein regressions, the NIRS calibrations were clearly better to predict the amino acid contents in lentils. The results showed that more information could be extracted from NIR spectra than only the protein bands (Fontaine et al., 2001). Wang et al. (2013) also mentioned that in amino acid analysis of peanuts, NIRS could explain more of the variance than the crude protein regression.

Table 4.8. Statistics from Bland-Altman analysis and P value from paired t-test to analyze the agreement between FT 9700 and DA 7250 spectrometers [a]

Compound	Bias (%) [b]	SD of bias	95% Limits of Agreement		N	P value
			Upper limits (%)	Lower limits (%)		
Protein	0.007	0.386	0.763	-0.750	17	0.742
His	0.002	0.077	0.153	-0.149	21	0.628
Ser	0.001	0.029	0.058	-0.056	21	0.664
Arg	0.003	0.054	0.109	-0.103	20	0.332
Gly	0.001	0.020	0.039	-0.038	15	0.554
Asp	-0.002	0.074	0.144	-0.148	17	0.563
Glu	0.000	0.089	0.174	-0.175	12	0.966
Thr	0.002	0.031	0.062	-0.059	17	0.257
Ala	0.001	0.025	0.050	-0.049	15	0.491
Pro	0.001	0.020	0.041	-0.039	15	0.427
Cys	0.000	0.013	0.025	-0.025	20	0.982
Lys	0.001	0.032	0.064	-0.061	17	0.407
Tyr	0.002	0.019	0.040	-0.036	14	0.080
Met	0.000	0.011	0.021	-0.021	19	0.934
Val	0.000	0.024	0.047	-0.047	16	0.775
Ile	0.001	0.022	0.043	-0.041	20	0.492
Leu	0.001	0.033	0.065	-0.063	13	0.622
Phe	0.000	0.024	0.047	-0.046	10	0.709
Trp	0.000	0.006	0.011	-0.011	18	0.994
SAA	0.000	0.020	0.039	-0.038	17	0.705

[a] Abbreviation: SD, standard deviation of difference; N, number of samples outside the limits; P, P value for paired student t-test; SAA, sulphur amino acids (Cys + Met).
[b] Bias, mean difference.

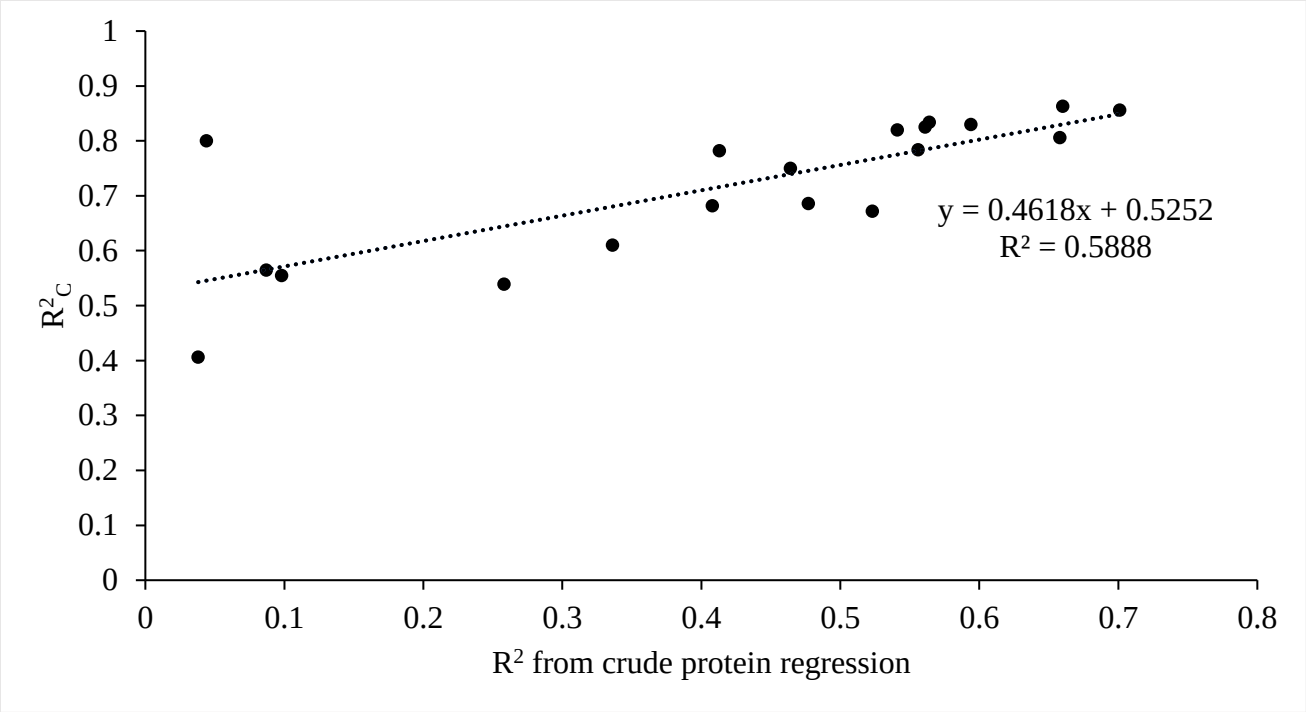


Figure 4.4. Coefficient of determination (RSQ from Table 4.3) from regression of amino acids to crude protein versus coefficient of determination (R^2_C from Table 4.4) of DA 7250 NIR calibration models

4.5 Conclusion

This study evaluated the performance of NIRS combined with PLS regression for the prediction of crude protein and amino acid contents in whole and ground lentils. The DA 7250 spectrometer could predict contents of protein and most amino acids with satisfactory accuracy in whole and ground lentil samples. However, the predictive ability of DA 7250 models for His, Tyr, Met and Cys was relatively inferior due to low values of R^2_C, R^2_{CV} and RPD. Except for His, Tyr, Met and Cys, FT 9700 NIRS had satisfactory predictive accuracy for protein and other amino acids in ground lentils. Continued refinement of the calibration equations is needed to enhance the model performance for these four amino acids.

The effects of sample status, type of spectrometer, and amino acid/protein correlation on predictive ability were analyzed. NIRS had slightly better predictability for crude protein and most amino acids in ground lentils than in whole lentils. Still, these compositions could be predicted for both whole and ground samples with similar accuracy. The DA 7250 models had a slightly higher R^2_{CV} and RPD values than FT 9700 models for all compositions except His. The predicted data of two spectrometers didn't have a significant difference ($p>0.05$) for every composition. NIRS could have a better performance for certain amino acids in lentils when they were highly correlated to protein. The NIRS calibration models were clearly better to predict the amino acid contents than crude protein regressions.

Overall, the NIRS was a highly promising method for rapid prediction of protein and most amino acid contents in lentils. It is an ideal and high throughout technology to measure theses compositions in lentil breeding programs, which can improve selections of lentils with high protein content and quality.

CHAPTER 5: GENOME-WIDE ASSOCIATION STUDY OF SEED PROTEIN AND AMINO ACID CONTENTS IN CULTIVATED LENTILS

5.1 Abstract

Lentil (*Lens culinaris* Medik.) is an important food legume crop and is regarded as an affordable source of plant-based proteins. The protein and amino acid contents of lentil seeds are important nutritional composition traits. The objectives of this study were to evaluate variation in seed protein and 18 amino acid compositions and to detect single nucleotide polymorphism (SNP) markers significantly associated with these traits. The present Genome-wide association study (GWAS) incorporated 324 accessions in a lentil diversity panel with 266,164 SNP markers in four environments (Rosthern, Saskatchewan, Canada 2016 and 2017, Sutherland, Saskatchewan, Canada 2016 and 2017) to identify significantly associated markers for seed protein and amino acids. A total of 85 SNP markers associated with seed protein or amino acid contents were identified in lentils. Only one identical SNP marker located on chromosome 7 was associated with one trait (Val) in two environments, and other markers were identified only in one site-year. 17 SNP markers associated with two or more traits simultaneously could be further studied and used in marker-assisted breeding programs to improve several amino acid contents at the same time. This study showed that GWAS was a promising method to identify significantly associated SNP markers, and these identified markers had a good potential to be used in marker-assisted programs to improve the protein content and quality of lentil seeds.

KEYWORDS: *Genome-wide association study (GWAS), lentils, single nucleotide polymorphism (SNP) markers, protein, amino acids*

5.2 Introduction

Lentil (*Lens culinaris* Medik.) is a diploid (2n=2x=14), annual self-pollinating pulse crop with a haploid genome size of approximately 4 Gb (Arumuganathan & Earle, 1991). Lentil is a good protein source (around 26% protein) for human consumption. Meanwhile, this affordable source plays a vital role in combating human protein malnutrition all over the world, especially in some developing countries (Kumar et al., 2015). While lentil seed protein contains all essential amino acids, the sulfur amino acids and tryptophan tend to be the most limiting amino acids in lentils (Boye, 2015; Nosworthy et al., 2017). The seed amino acid composition has a significant influence on the quality and uses of lentil protein (Nosworthy & House, 2017; Nosworthy et al., 2017). Canada is the largest producer and exporter of lentils in recent years (FAO, 2021). Enhancing the protein content and quality of lentil seeds is of great importance for Canadian lentil breeders, food manufactures and consumers. The increasing demands for good-quality plant-based protein prompt further investigation into the genetic architecture of lentil seed protein and amino acid contents.

Genome-wide association studies (GWASs) are powerful tools to dissect the genetic architecture of quantitative traits including protein and amino acid composition. As a hypothesis-free, genome-wide approach, GWAS has received increasing interest in mapping traits and identifying markers associated with complex traits (Yuan et al., 2021). Compared to traditional quantitative trait locus (QTL) analysis, this method eliminates the need to develop time-consuming crosses, has a higher resolution and larger allele numbers (Tibbs Cortes, Zhang & Yu, 2021; Korte & Farlow, 2013; Alqudah et al., 2020). However, it can be challenging to detect rare alleles and alleles with low effect (Alqudah et al., 2020). GWAS can be regarded as complementary to QTL analysis because they overcome each other's limitations (Korte &

Farlow, 2013; Tibbs Cortes et al., 2021). Results from GWASs could be used to generate markers that could facilitate the marker-assisted selection in breeding programs focusing on protein and /or amino acid contents (Khazaei et al., 2017).

Several GWA studies have been conducted to study the genetic association with protein and amino acids in various crops, such as soybeans (Lee et al., 2019; Zhang et al., 2018a; Yuan et al., 2021), chickpea (Karaca et al., 2019) and wheat (Nigro et al., 2019).

Regarding lentils, a limited number of studies have investigated the genetic basis of some important traits of lentils, such as iron and zinc concentrations (Khazaei et al., 2017), seed dimensions (Khazaei et al., 2018), salt tolerance (Dissanayake, Cogan, Smith & Kaur, 2021), and anthracnose race 1 resistance (Gela et al., 2021).

To the best of our knowledge, currently there are no studies investigating the genetic control of variation in protein and amino acid contents in cultivated lentil seeds by GWAS. Therefore, the objective of the current study was to (i) evaluate protein and amino acid compositions in cultivated lentils from Canada, and (ii) detect SNPs significantly associated with protein and 18 amino acids content by using GWAS.

5.3 Materials and methods

5.3.1 Plant Materials

A lentil diversity panel of 324 *Lens culinaris* cultivars was used in this study. Details about plant material can be found in Chapter 4. The growing environmental data were measured by the University of Saskatchewan and were summarized in Table 5.1 (https://github.com/derekmichaelwright/AGILE_LDP_Phenology/blob/master/Supplemental_Table_02.csv; Wright et al., 2021).

5.3.2 Determination of protein and amino acid contents

The lentil diversity panel has been phenotyped for 19 traits, including protein and 18 amino acids. The protein and amino acid contents in whole lentil seed on a dry basis were estimated by a DA 7250 Near-infrared (NIR) reflectance spectrometer calibrated against reference methods (AOAC method 990.03; AOAC method 982.30; AOAC method 985.28; ISO 13904:2005(E)). The calibration methods and model performance are mentioned in Chapter 4.

5.3.3 Statistical analysis of phenotypic data

Statistical analysis was performed using GraphPad Prism, version 9 for Windows (GraphPad Software, CA). Correlation analyses were carried out between every composition. Statistical comparison was conducted using one-way analysis of variance (ANOVA) and Tukey's multiple comparisons test to analyze whether the protein and 18 essential amino acid contents were significantly different among different environments (site-year) or genotypes. The p-value <0.05 was considered significant.

5.3.4 Genotyping

Genotyping of the lentil diversity panel was done by the University of Saskatchewan using an exome capture array. Markers have been filtered with several criteria as follows: (1) only biallelic markers; (2) a minor allele frequency (MAF) is high than or equal to 5%; (3) less than or equal to 20% of missing genotypes (i.e. No undefined genotypes); (4) less than or equal to 20% heterozygosity (Wright, 2021). The remaining 266,164 SNPs, with high density across the lentil genome were used for GWAS (Figure 5.1).

5.3.5 Genome-wide association study

Genome-wide association analysis was performed by using the Bayesian information and LD iteratively nested keyway (BLINK) model (Huang et al., 2019). Genome Association and Prediction Integrated Tool (GAPIT) (Lipka et al. 2012) in R was used for the BLINK analysis. As mentioned by Zhang (2020), the BLINK model has the highest statistical power compared to other models in the GAPIT R package. Additionally, BLINK has a high computing speed and less false positives (Zhang, 2020). Principal component analysis (PCA) (Patterson et al., 2006) was conducted to determine the optimal principal component (PC). The turning point was chosen, and the four principal component (PC) model was identified as the best fit for all traits (Figure 5.2). The first four PCs can explain most of the total variances, and PCA was set as 4 to correct for the population structure. GWAS was conducted separately for each site-year (16Ros, 16Sut, 17Ros, 17Sut) for 19 traits.

The default threshold [-log10(p-value)] for significant association was set at 6.73 which is equal to -log10(0.05 / 266,164) [i.e. (P = 0.05 / no. of markers)] based on the Bonferroni

correction method (Holm, 1979). The SNP markers with [-log10(p-value)] $\geqslant$6.73 were regarded significantly associated with traits.

Manhattan plots and Quantile-quantile (Q-Q) plots were built using the R package "qqman" (Turner 2014). Manhattan plots are scatter plots that can display the significance of SNP markers across the genome (Zhang, 2020). The X-axis is the genomic position of every SNP, and the number represents the number of the chromosome. The Y-axis is -log10(p-value) of every SNP. SNP with a stronger association with the trait will have a higher Y value. Q-Q plots can evaluate the performance of the models and how well the model was corrected for population structure and kinship (Zhang, 2020). The null hypothesis (H0) is that no association between the SNP and traits of interest. The expected -log10(P) values (X-axis) are plotted against the observed -log10(P) values (Y-axis) under the null hypothesis H0 (Zhang, 2020). Most of the points at the beginning of the line should fit the diagonal line (red line) because most SNPs have no association with the trait. SNPs are expected to deviate at the end of the line (Zhang, 2020).

Table 5.1. Average data of day length, precipitation, relative humidity and temperature in four different growing environments: Rosthern, Canada 2016 and 2017, Sutherland, Canada 2016 and 2017

Location	Year	Short Name	Day Length (hours)	Precipitation (mm)	Relative Humidity (%)	Temperature (^{o}C)
Rosthern	2016	16Ros	16.2	15.8	73.7	17.2
Sutherland	2016	16Sut	15.9	3.2	73.8	16.7
Rosthern	2017	17Ros	16.4	8.3	68.7	17.5
Sutherland	2017	17Sut	16.1	1.7	66.7	15.7

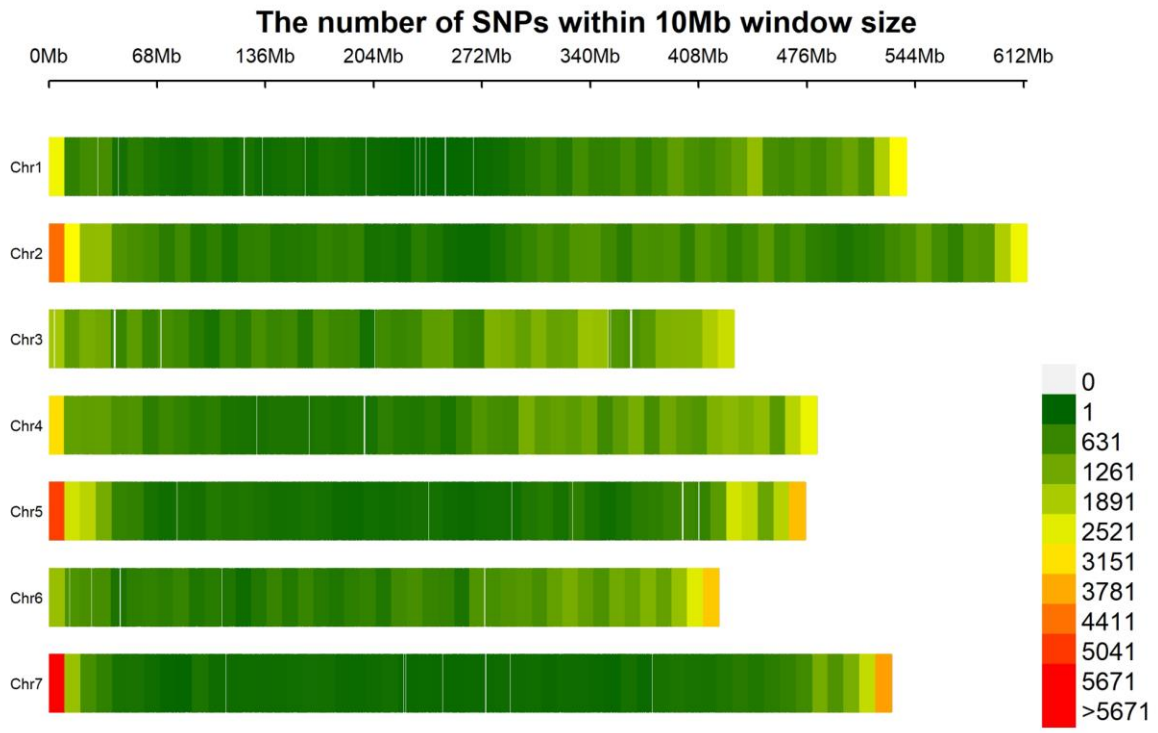


Figure 5.1. SNP density plot chromosome wise representing number of SNPs within 10Mb window size. Different colors depict marker densities, with the dark green indicating the lowest density while the red color indicating the highest density

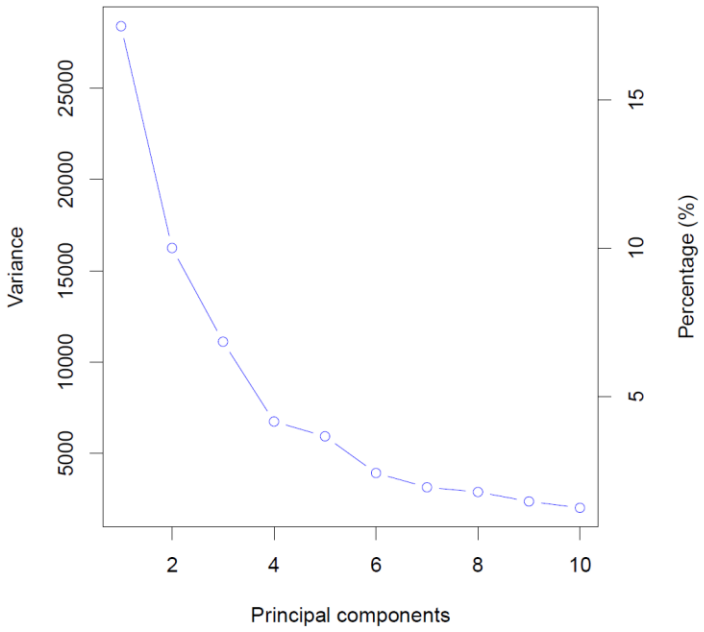


Figure 5.2. Principal component analysis results to evaluate population structure in the lentil diversity panel. The turning point was chosen as the optimal number of principal components (PC=4)

5.4 Results and Discussion

5.4.1 Phenotypic variations and correlations

Table 5.2 summarizes the mean, standard deviation (SD), coefficient of variation (CV), minimum and maximum of protein and nine essential amino acids (His, Thr, Lys, Met, Val, Ile, Leu, Phe, Trp), and Table 5.3 shows these statistical data of nine conditionally essential or nonessential amino acids (Ser, Arg, Gly, Asp, Glu, Ala, Pro, Cys, Tyr). The concentration of protein and 18 amino acids on a seed weight dry basis showed wide and continuous distribution (Table 5.2; Table 5.3). The distributions of all compositions were near normal, with $W>0.97$ from the Shapiro-Wilk test, which indicated that these compositions were complex quantitative traits. These compositions were controlled by multiple genes with different levels of effects, and they were often affected by the environment and G × E interaction (Kumar et al., 2015; Karikari et al., 2020).

The protein content on a dry basis ranged from 24.6% to 35.2%, and the overall mean content was 29.53%. The highest and lowest average protein content on a dry basis was 30.22% and 28.67% in seeds from 16Sut and 17Ros, respectively. Among these 18 amino acids, Glu had the highest average concentration (4.5%-4.7%), followed by Asp (3.2%-3.5%) in four environments. Arg and Leu were also abundant in lentil seed protein, and the average content for both was approximately 2.1% (Table 5.2; Table 5.3). In the 2016 cropping year, the average Leu content was higher than the average Arg content in lentil seeds. The Leu became the third most common amino acid; however, the opposite situation was observed in the 2017 cropping year that the Arg was the third most abundant amino acid. Trp, Met and Cys ranging from 0.2%-0.3%, were the three amino acids with lowest concentration in lentils. Regarding essential amino acids, Leu had the highest concentration (2.1%), followed by Lys (1.9%) and Phe (1.4%). The average

concentration of Trp and Met was 0.23% and 0.26%, respectively, which were much lower than other essential amino acids. The results agreed with previous studies (Nosworthy et al.,2017; Nosworthy et al.,2018). Protein and most amino acids had relatively small CV values (approximately 5%); however, His showed the largest CV, which was from 9.7% to 13.2%. Zhang et al. (2018a) also observed large variations for His concentration in soybeans. Moreover, it was obvious that the protein and amino acid contents varied with different environments.

Table 5.4 shows that most of the correlation coefficients (r) among protein and 18 amino acids were significant ($p<0.0001$) except for the correlation coefficient between Met and His ($p>0.05$). Positive coefficients were observed among all traits except for His and Thr, His and Cys, His and Met (Table 5.4). Several amino acids were highly correlated with protein with correlation coefficients greater than 0.9, including Ser (r=0.93), Gly(r=0.91), Glu(r=0.95), Pro(r=0.94) and Leu (r=0.95). Strong correlations with $r>0.95$ could be found between Gly and Ser (r=0.95), Glu and Ser (r=0.96), Ile and Ser (r=0.97), Leu and Ser (r=0.98), Leu and Gly (r=0.97), Leu and Glu (r=0.99), Leu and Lys (r=0.95), Ile and Val (r=0.98), Phe and Val (r=0.99), Leu and Ile (r=0.97), Phe and Ile (r=0.98). However, some amino acids like His, Thr and Cys, were poorly related with other compositions with most correlation coefficients lower than 0.5 (Table 5.4).

Ten samples were removed from the one-way ANOVA due to missing samples, including Entry 7, 109, 239 and 319. Therefore, the ANOVA included 320 different lentil genotypes in four environments. A significant environment effect was noted for Phe and Trp at the 0.05 probability level, and for Protein, His, Thr, Lys, Met, Ile, Leu at the 0.0001 probability level (Table 5.5). However, the Val content was not significantly different among the four environments (Table 5.5). For the protein content, Tukey's multiple comparisons tests revealed a

significant pairwise difference between the cropping year 2016 and the year 2017 ($p<0.0001$). The average protein content of lentil seeds from 2016 was significantly higher than that from 2017. Wang (2017) also mentioned that the mean protein content of 2017 western Canadian lentils was lower than the mean for 2016. Moreover, a significant difference in protein content was observed at the 0.001 probability level between 17Ros (28.7%) and 17Sut (29.1%). Nevertheless, the seed protein content between 16Ros (30.1%) and 16Sut (30.2%) was non-significant.

A significant genotype effect was observed for all traits ($p<0.0001$; Table 5.5). Regarding protein, Entry 154 had the highest average protein content (32.95%), followed by Entry 157 (32.87%), 181 (32.84%), 268 (32.63%) and 47 (32.59%). The protein contents among these five accessions were not significantly different.

Table 5.2. Summary descriptive statistics of lentil seed protein and nine essential amino acid contents (% dry basis) in four environments: Rosthern, Canada 2016 and 2017 (Ros16, Ros17), Sutherland, Canada 2016 and 2017 (Sut16, Sut17) [a]

		Protein	His	Thr	Lys	Met	Val	Ile	Leu	Phe	Trp
16Ros	Mean	30.11	0.98	1.04	1.95	0.27	1.40	1.24	2.13	1.43	0.23
	SD	1.54	0.10	0.06	0.09	0.02	0.08	0.07	0.11	0.07	0.01
	CV	5.12	10.05	5.83	4.49	5.60	5.63	5.33	5.20	5.19	5.81
	Min	25.64	0.63	0.85	1.69	0.23	1.18	1.06	1.81	1.22	0.20
	Max	35.14	1.26	1.21	2.22	0.32	1.65	1.43	2.46	1.65	0.27
16Sut	Mean	30.22	1.05	1.03	1.92	0.25	1.39	1.23	2.11	1.42	0.23
	SD	1.54	0.10	0.06	0.08	0.01	0.08	0.06	0.11	0.07	0.01
	CV	5.11	9.74	5.62	4.39	5.78	5.42	5.00	5.11	4.98	6.00
	Min	26.00	0.84	0.88	1.67	0.22	1.18	1.07	1.81	1.22	0.19
	Max	35.18	1.51	1.19	2.17	0.30	1.63	1.42	2.43	1.63	0.26
17Ros	Mean	28.67	0.84	1.18	1.86	0.26	1.39	1.22	2.05	1.43	0.22
	SD	1.26	0.10	0.05	0.08	0.01	0.07	0.05	0.09	0.06	0.01
	CV	4.38	12.15	4.65	4.24	5.47	4.79	4.49	4.56	4.43	4.78
	Min	24.73	0.57	1.05	1.64	0.22	1.20	1.08	1.81	1.27	0.19
	Max	32.54	1.25	1.31	2.09	0.29	1.56	1.36	2.28	1.58	0.25
17Sut	Mean	29.10	0.95	1.11	1.83	0.25	1.40	1.22	2.05	1.42	0.23
	SD	1.46	0.13	0.06	0.08	0.01	0.07	0.06	0.10	0.07	0.01
	CV	5.02	13.19	5.53	4.52	5.52	5.10	4.90	4.97	4.80	4.86
	Min	24.60	0.65	0.93	1.61	0.22	1.20	1.05	1.76	1.23	0.19
	Max	35.10	1.34	1.26	2.10	0.29	1.66	1.43	2.44	1.67	0.27
Overall	Mean	29.53	0.96	1.09	1.89	0.26	1.40	1.23	2.08	1.42	0.23
	SD	1.60	0.13	0.09	0.10	0.02	0.07	0.06	0.11	0.07	0.01
	CV	5.40	13.95	7.82	5.06	6.47	5.25	4.99	5.26	4.87	5.43
	Min	24.60	0.57	0.85	1.61	0.22	1.18	1.05	1.76	1.22	0.19
	Max	35.18	1.51	1.31	2.22	0.32	1.66	1.43	2.46	1.67	0.27

[a] Abbreviation: SD, standard deviation; CV, coefficient of variation; Min, minimum; Max, maximum

Table 5.3. Summary descriptive statistics of nine conditionally essential or nonessential amino acid contents (% dry basis) of lentil seeds in four environments: Rosthern, Canada 2016 and 2017 (Ros16, Ros17), Sutherland, Canada 2016 and 2017 (Sut16, Sut17) [a]

		Ser	Arg	Gly	Asp	Glu	Ala	Pro	Cys	Tyr
16Ros	Mean	1.45	2.09	1.16	3.45	4.71	1.20	1.23	0.26	0.85
	SD	0.08	0.16	0.06	0.19	0.25	0.08	0.05	0.02	0.05
	CV	5.23	7.58	5.55	5.61	5.32	6.77	4.37	6.36	5.45
	Min	1.24	1.62	0.98	2.90	3.95	1.01	1.09	0.22	0.72
	Max	1.68	2.56	1.35	4.11	5.45	1.50	1.39	0.32	1.01
16Sut	Mean	1.44	2.08	1.15	3.39	4.70	1.17	1.22	0.25	0.84
	SD	0.07	0.16	0.06	0.21	0.26	0.07	0.05	0.02	0.05
	CV	5.22	7.84	5.37	6.16	5.48	6.01	4.22	6.27	5.68
	Min	1.22	1.62	1.00	2.85	3.94	0.98	1.09	0.21	0.72
	Max	1.67	2.66	1.31	4.11	5.47	1.36	1.35	0.29	0.98
17Ros	Mean	1.40	2.17	1.12	3.29	4.54	1.14	1.18	0.27	0.86
	SD	0.07	0.14	0.05	0.18	0.23	0.06	0.04	0.02	0.04
	CV	4.79	6.56	4.62	5.48	5.05	4.86	3.44	5.69	4.82
	Min	1.20	1.77	0.99	2.82	3.91	0.99	1.08	0.23	0.75
	Max	1.57	2.54	1.26	3.73	5.14	1.30	1.28	0.31	0.97
17Sut	Mean	1.39	2.18	1.11	3.22	4.49	1.15	1.19	0.26	0.85
	SD	0.07	0.16	0.05	0.20	0.25	0.06	0.05	0.01	0.04
	CV	5.16	7.44	4.90	6.11	5.67	4.86	3.81	5.67	5.20
	Min	1.18	1.70	0.97	2.69	3.73	1.01	1.06	0.22	0.71
	Max	1.68	2.88	1.32	4.08	5.45	1.36	1.36	0.32	1.02
Overall	Mean	1.42	2.13	1.13	3.34	4.61	1.17	1.20	0.26	0.85
	SD	0.08	0.16	0.06	0.22	0.27	0.07	0.05	0.02	0.05
	CV	5.33	7.63	5.37	6.44	5.77	6.05	4.31	6.55	5.37
	Min	1.18	1.62	0.97	2.69	3.73	0.98	1.06	0.21	0.71
	Max	1.68	2.88	1.35	4.11	5.47	1.50	1.39	0.32	1.02

[a] Abbreviation: SD, standard deviation; CV, coefficient of variation; Min, minimum; Max, maximum

Table 5.4 Correlation coefficients among protein and 18 amino acids (% dry basis) of 1290 lentil samples

	His	Ser	Arg	Gly	Asp	Glu	Thr	Ala	Pro	Cys	Lys	Tyr	Met	Val	Ile	Leu	Phe	Trp
CP	0.67*	0.93*	0.75*	0.91*	0.90*	0.95*	0.18*	0.85*	0.94*	0.27*	0.89*	0.74*	0.50*	0.86*	0.89*	0.95*	0.86*	0.86*
His		0.42*	0.34*	0.40*	0.40*	0.46*	-0.32*	0.34*	0.55*	-0.24*	0.37*	0.18*	-0.05ns	0.30*	0.34*	0.44*	0.30*	0.37*
Ser			0.79*	0.95*	0.93*	0.96*	0.35*	0.89*	0.89*	0.47*	0.94*	0.86*	0.60*	0.95*	0.97*	0.98*	0.94*	0.89*
Arg				0.74*	0.69*	0.76*	0.66*	0.68*	0.72*	0.55*	0.61*	0.88*	0.34*	0.89*	0.84*	0.78*	0.91*	0.80*
Gly					0.92*	0.94*	0.38*	0.92*	0.90*	0.46*	0.92*	0.85*	0.66*	0.92*	0.94*	0.97*	0.91*	0.93*
Asp						0.94*	0.30*	0.87*	0.84*	0.47*	0.94*	0.78*	0.72*	0.85*	0.88*	0.94*	0.84*	0.86*
Glu							0.30*	0.86*	0.92*	0.33*	0.95*	0.81*	0.60*	0.91*	0.95*	0.99*	0.92*	0.85*
Thr								0.33*	0.19*	0.69*	0.23*	0.70*	0.34*	0.55*	0.46*	0.33*	0.55*	0.49*
Ala									0.89*	0.49*	0.89*	0.78*	0.69*	0.88*	0.88*	0.91*	0.86*	0.89*
Pro										0.26*	0.87*	0.73*	0.53*	0.84*	0.88*	0.94*	0.86*	0.83*
Cys											0.35*	0.68*	0.55*	0.56*	0.49*	0.39*	0.53*	0.60*
Lys												0.75*	0.70*	0.86*	0.89*	0.95*	0.85*	0.81*
Tyr													0.60*	0.92*	0.89*	0.83*	0.92*	0.87*
Met														0.56*	0.56*	0.62*	0.54*	0.60*
Val															0.98*	0.94*	0.99*	0.91*
Ile																0.97*	0.98*	0.88*
Leu																	0.94*	0.89*
Phe																		0.88*

*: $p<0.0001$; ns: $p>0.05$

Table 5.5. One-way analysis of variance (ANOVA) of protein and nine essential amino acid contents in seeds of 320 lentil accessions grown in Rosthern and Sutherland, Saskatchewan, Canada in 2016–2017

Source of Variation	df	Mean Squares									
		Protein	His	Thr	Lys	Met	Val	Ile	Leu	Phe	Trp
Environment	3	184.7**	2.668**	1.628**	0.943**	0.031**	0.012^{ns}	0.030**	0.549**	0.015*	0.001*
Residual	1276	2.105	0.012	0.003	0.007	0.000	0.001	0.004	0.011	0.005	0.000
Genotype	319	5.710**	0.028**	0.009**	0.020**	0.001**	0.016**	0.011**	0.031**	0.014**	0.000**
Residual	960	1.478	0.014	0.007	0.005	0.000	0.002	0.001	0.006	0.002	0.000

**: $p<0.0001$; *: $p<0.05$; ns: $p>0.05$

5.4.2 GWAS for protein and 18 amino acids / Marker-Trait Associations

Association analyses were implemented for protein and 18 amino acid contents with BLINK models by each environment (16Ros, 16Sut, 17Ros and 17Sut). In most Q-Q plots, the majority of blue points fitted the red line (diagonal line) at the beginning and showed deviations from this line at the tail (Appendix G). The red line indicates the situation under null hypothesis that no SNP is significantly associated with the trait. Most of the observed associations (blue points) should conform to the red line because most SNPs have no association with the trait. Only a few SNPs should be identified to be above the red line. Q-Q plots in current study indicated that most BLINK models effectively accounted for population structure and kinship and controlled false-positive associations (Lee et al., 2019). The significant SNPs in the upper-right corner had a high potential to associate with traits.

A total of 85 SNP markers were identified to be significantly associated with seed protein and amino acids, with 11, 19, 11, 8, 10, 11 and 15 SNPs located on chromosomes 1, 2, 3, 4, 5, 6 and 7, respectively.

For protein and essential amino acids, the current study has identified 3, 9, 7, 1, 1, 9, 5, 12, 6 and16 SNPs significantly associated with protein, His, Thr, Lys, Met, Val, Ile, Leu, Phe and Trp contents in lentil samples, respectively (Table 5.6). Moreover, 8, 6, 9, 5, 8, 14, 8, 6, and 4 SNP markers were significant for Ser, Arg, Gly, Asp, Glu, Ala, Pro, Cys and Tyr, respectively (Table 5.7). Most of these significant markers varied with the change of environment. Lee et al. (2019) also mentioned that the significance of the association between SNPs and amino acids in soybeans often depended on environments. In this study of lentils, the majority of the markers were significant only in one site-year because most SNP markers were environmentally specific. Only one identical SNP marker (SLCU.2RBY.CHR7_524204079) significantly associated with

Val was identified in two environments (16Ros and 16Sut). This SNP was located on chromosome 7, with -log10(p-value) higher than 16.0 and 7.0 in 16Ros and 16Sut, respectively. More interestingly, this marker was also significant for Trp in 16Ros with -log10(p-value) higher 10.3. This SNP marker almost passed the threshold for Trp in 17Ros with -log10(p-value) equaled to 6.5 (Figure 5.3). The default threshold was set at 6.73 based on the Bonferroni correction. Many studies mentioned that this threshold could be too conservative for GWAS because it assumes that each marker is independent (Pearson & Manolio, 2008; Tibbs Cortes, Zhang & Yu, 2021). The relatively good stability of this SNP across the environments indicated that it has a good potential value in marker-assisted programs to improve Val and Trp contents of lentils simultaneously.

According to Table 5.8, 17 significant SNP markers were associated with two or more traits simultaneously. One SNP (SLCU.2RBY.CHR5_467478441) on Chr 5 was associated with 11 amino acids (Thr, Lys, Val, Ile, Leu, Phe, Ser, Gly, Asp, Glu, Ala) in 17Ros. SLCU.2RBY.CHR7_524176613 on Chr 7 was associated with nine compositions, including Protein, Ile, Leu, Phe, Ser, Gly, Glu, Pro and Tyr in one environment (16Sut). SLCU.2RBY.CHR2_582173746 on Chr 2 and SLCU.2RBY.CHR5_467611866 on Chr 5 showed significant association with seven and five traits, respectively. It was interesting to notice that four SNP markers were significantly associated with different traits in different environments. For instance, SLCU.2RBY.CHR2_128835034 on Chr 2 showed the significant associations with Tyr in 16Ros and Val, Trp and Gly in 17Ros. Other remaining nine SNP markers were associated with two or three or four traits simultaneously under the same environment. Two markers (SLCU.2RBY.CHR7_524176613 and SLCU.2RBY.CHR7_524204079) at chromosome 7 were very close. Through personal

communication with Derek M. Wright, the distance between these two markers was shorter than linkage disequilibrium decay, and these two markers could be considered as the same quantitative trait locus (QTL). Moreover, two markers (SLCU.2RBY.CHR5_467478441 and SLCU.2RBY.CHR5_467611866) at chromosome 5 could also be clustered in the same QTL. These identified SNP markers could be studied further and used to find potential candidate genes and genomic regions for controlling protein and amino acid concentrations. SNP markers associated with multiple amino acids have a good potential to improve several amino acid concentrations and protein quality in lentil seeds at the same time, which are worthy of further study.

Table 5.8 shows that under the same environment, five identical significant SNP makers could be found between Val and Gly, Leu and Glu, Phe and Leu. For example, for Val and Gly, SLCU.2RBY.CHR1_46680284 in 16Ros, SLCU.2RBY.CHR2_128835034 in 17Ros, SLCU.2RBY.CHR3_270889818 in 16Sut, SLCU.2RBY.CHR5_467478441 in 17Ros and SLCU.2RBY.CHR7_524204079 in 16Ros were identified as significant markers. Moreover, four identical significant SNP makers could be found between Leu and Ser, Phe and Glu, and Leu and Pro. The significant correlation coefficients of these pairs were also greater than 0.9 (Table 5.4). Two compositions with strong correlations ($r>0.95$) were found to have one or more identical significantly associated SNPs in different environments (Table 5.4; Table 5.8). For His and Cys, no similar significant SNP markers could be detected with other traits under the same environment. This was in agreement with the correlation analysis since these two amino acids were in weak correlations with other compositions. Regarding Thr, this amino acid generally was in poor correlation with other traits; however, there were 10 traits (Ser, Gly, Asp, Glu, Ala, Lys, Val, Ile, Leu, Phe) showing the same SNPs with Thr. The results from table 5.8 were consistent

with the findings of correlation analysis for most amino acids except Thr. The amino acids that were highly correlated were more likely to have identical significant SNP markers under the same environment; however, the SNP markers could vary with various environments.

In Chapter 4, relatively low values of the R^2_C for His, Met and Cys calibration models indicated that the predictive ability of NIRS models was relatively inferior for these amino acids. Current calibrations didn't produce suitable models to predict the amino acids of these compositions with high accuracy. The relatively unreliable phenotypic data could be the possible reason that these amino acids didn't have good relationships with protein and other amino acids (Table 5.4).

Table 5.6 Significant SNPs associated with protein and nine essential amino acids (His, Thr, Lys, Met, Val, Ile, Leu, Phe, Trp) in multiple environments from BLINK model analyses [a]

Trait	SNP marker	Chr	Position(Mb)	P value	maf	Site-year
Protein	SLCU.2RBY.CHR5 467611866	5	467611866	3.16E-11	0.45	16Sut
	SLCU.2RBY.CHR7_524176613	7	524176613	3.35E-09	0.46	16Sut
	SLCU.2RBY.CHR5_169306351	5	169306351	3.74E-09	0.46	16Sut
His	SLCU.2RBY.CHR3_169946051	3	169946051	1.86E-11	0.15	16Ros
	SLCU.2RBY.CHR4_393259943	4	393259943	1.17E-09	0.39	16Ros
	SLCU.2RBY.CHR2_6708220	2	6708220	1.58E-09	0.41	16Ros
	SLCU.2RBY.CHR2_452083	2	452083	3.23E-08	0.14	16Ros
	SLCU.2RBY.CHR7_420073193	7	420073193	8.61E-08	0.10	16Ros
	SLCU.2RBY.CHR4_12555114	4	12555114	7.97E-12	0.14	16Sut
	SLCU.2RBY.CHR2_391487900	2	391487900	1.04E-08	0.10	16Sut
	SLCU.2RBY.CHR1_401855397	1	401855397	4.6E-16	0.14	17Sut
	SLCU.2RBY.CHR6_353602378	6	353602378	1.49E-07	0.44	17Sut
Thr	SLCU.2RBY.CHR3_428502810	3	428502810	1.16E-08	0.23	16Ros
	SLCU.2RBY.CHR3_270889818	3	270889818	8.9E-12	0.29	16Sut
	SLCU.2RBY.CHR4_406204422	4	406204422	4.99E-11	0.18	16Sut
	SLCU.2RBY.CHR7_520668776	7	520668776	1.46E-08	0.44	16Sut
	SLCU.2RBY.CHR5_467478441	5	467478441	4.95E-12	0.48	17Ros
	SLCU.2RBY.CHR3_77153912	3	77153912	5.74E-09	0.11	17Sut
	SLCU.2RBY.CHR1_427926834	1	427926834	1.02E-08	0.27	17Sut
Lys	SLCU.2RBY.CHR5_467478441	5	467478441	1.95E-11	0.48	17Ros
Met	SLCU.2RBY.CHR5_200627011	5	200627011	1.22E-09	0.47	16Ros
Val	SLCU.2RBY.CHR7_524204079	7	524204079	9.8E-17	0.35	16Ros
				8.6E-08	0.35	16Sut
	SLCU.2RBY.CHR2_571776341	2	571776341	2.41E-12	0.16	16Ros
	SLCU.2RBY.CHR2_18927314	2	18927314	7.01E-11	0.42	16Ros
	SLCU.2RBY.CHR1_46680284	1	46680284	8.8E-11	0.17	16Ros
	SLCU.2RBY.CHR5_466841061	5	466841061	1.28E-07	0.36	16Ros
	SLCU.2RBY.CHR3_270889818	3	270889818	1.62E-08	0.29	16Sut
	SLCU.2RBY.CHR5_467611866	5	467611866	3.89E-08	0.45	16Sut
	SLCU.2RBY.CHR2_128835034	2	128835034	6.78E-10	0.06	17Ros
	SLCU.2RBY.CHR5_467478441	5	467478441	1.68E-09	0.48	17Ros
Ile	SLCU.2RBY.CHR7_474941369	7	474941369	4.53E-18	0.39	16Sut
	SLCU.2RBY.CHR7_524176613	7	524176613	5.97E-12	0.46	16Sut
	SLCU.2RBY.CHR1_406743581	1	406743581	8.8E-10	0.28	16Sut
	SLCU.2RBY.CHR4_374653055	4	374653055	2.89E-08	0.19	16Sut
	SLCU.2RBY.CHR5_467478441	5	467478441	9.37E-10	0.48	17Ros
Leu	SLCU.2RBY.CHR7_474327484	7	474327484	6.59E-18	0.41	16Sut
	SLCU.2RBY.CHR7_524176613	7	524176613	4.71E-13	0.46	16Sut
	SLCU.2RBY.CHR5_15051870	5	15051870	2.57E-12	0.10	16Sut

	SLCU.2RBY.CHR5_467611866	5	467611866	1.06E-09	0.45	16Sut
	SLCU.2RBY.CHR7_2031129	7	2031129	1.43E-09	0.42	16Sut
	SLCU.2RBY.CHR3_270889818	3	270889818	9.41E-08	0.29	16Sut
	SLCU.2RBY.CHR1_217285859	1	217285859	1.68E-13	0.36	17Ros
	SLCU.2RBY.CHR5_467478441	5	467478441	3.98E-12	0.48	17Ros
	SLCU.2RBY.CHR2_396979253	2	396979253	1.14E-11	0.12	17Ros
	SLCU.2RBY.CHR7_478047589	7	478047589	2.03E-10	0.49	17Ros
	SLCU.2RBY.CHR2_582173746	2	582173746	3.03E-10	0.32	17Ros
	SLCU.2RBY.CHR4_479372389	4	479372389	4.51E-08	0.22	17Ros
Phe	SLCU.2RBY.CHR7_524176613	7	524176613	5.67E-10	0.46	16Sut
	SLCU.2RBY.CHR1_217285859	1	217285859	1.84E-11	0.36	17Ros
	SLCU.2RBY.CHR2_396979253	2	396979253	3.19E-11	0.12	17Ros
	SLCU.2RBY.CHR5_467478441	5	467478441	6.79E-11	0.48	17Ros
	SLCU.2RBY.CHR2_603259068	2	603259068	9.75E-09	0.28	17Ros
	SLCU.2RBY.CHR2_582173746	2	582173746	2.33E-08	0.32	17Ros
Trp	SLCU.2RBY.CHR1_46680284	1	46680284	1.39E-13	0.17	16Ros
	SLCU.2RBY.CHR7_524204079	7	524204079	4.78E-11	0.35	16Ros
	SLCU.2RBY.CHR5_437222344	5	437222344	5.33E-10	0.08	16Ros
	SLCU.2RBY.CHR7_6284456	7	6284456	7.67E-09	0.22	16Ros
	SLCU.2RBY.CHR7_505141647	7	505141647	1.52E-07	0.24	16Ros
	SLCU.2RBY.CHR6_180664667	6	180664667	6.86E-10	0.13	16Sut
	SLCU.2RBY.CHR6_392062289	6	392062289	2.43E-09	0.33	16Sut
	SLCU.2RBY.CHR2_604968961	2	604968961	1.24E-08	0.49	16Sut
	SLCU.2RBY.CHR4_428734293	4	428734293	1.75E-11	0.13	17Ros
	SLCU.2RBY.CHR2_4306617	2	4306617	6.71E-09	0.21	17Ros
	SLCU.2RBY.CHR2_128835034	2	128835034	1.03E-08	0.06	17Ros
	SLCU.2RBY.CHR2_582173746	2	582173746	2.39E-08	0.32	17Ros
	SLCU.2RBY.CHR6_396645507	6	396645507	7.15E-08	0.10	17Ros
	SLCU.2RBY.CHR2_12554117	2	12554117	1.69E-07	0.18	17Ros
	SLCU.2RBY.CHR5_388203666	5	388203666	1.7E-07	0.26	17Ros
	SLCU.2RBY.CHR7_439628225	7	439628225	3.24E-10	0.19	17Sut

[a] Abbreviation: chr, chromosome; maf, minor allele frequency

Table 5.7 Significant SNPs associated with nine conditionally essential or nonessential amino acids (Ser, Arg, Gly, Asp, Glu, Ala, Pro, Cys, Tyr) in multiple environments from BLINK model analyses [a]

Trait	SNP marker	Chr	Position(Mb)	P value	maf	Site-year
Ser	SLCU.2RBY.CHR7_476576495	7	476576495	7.29E-13	0.41	16Sut
	SLCU.2RBY.CHR5_467611866	5	467611866	1.4E-09	0.45	16Sut
	SLCU.2RBY.CHR7_524176613	7	524176613	3.62E-09	0.46	16Sut
	SLCU.2RBY.CHR3_30641397	3	30641397	7.48E-09	0.17	16Sut
	SLCU.2RBY.CHR5_467478441	5	467478441	3.32E-11	0.48	17Ros
	SLCU.2RBY.CHR3_65942068	3	65942068	4.57E-10	0.36	17Ros
	SLCU.2RBY.CHR1_217285859	1	217285859	7.5E-10	0.36	17Ros
	SLCU.2RBY.CHR7_474941369	7	474941369	2.56E-08	0.39	17Ros
Arg	SLCU.2RBY.CHR5_467611866	5	467611866	1.2E-10	0.45	16Sut
	SLCU.2RBY.CHR1_438115679	1	438115679	5.17E-08	0.25	16Sut
	SLCU.2RBY.CHR7_364263709	7	364263709	9.26E-08	0.34	16Sut
	SLCU.2RBY.CHR2_582173746	2	582173746	3.29E-17	0.32	17Ros
	SLCU.2RBY.CHR4_479372389	4	479372389	4.92E-08	0.22	17Ros
	SLCU.2RBY.CHR6_414535406	6	414535406	2.41E-09	0.18	17Sut
Gly	SLCU.2RBY.CHR1_46680284	1	46680284	1.57E-11	0.17	16Ros
	SLCU.2RBY.CHR1_516348614	1	516348614	4.06E-10	0.48	16Ros
	SLCU.2RBY.CHR7_524204079	7	524204079	3.12E-09	0.35	16Ros
	SLCU.2RBY.CHR7_524176613	7	524176613	2.34E-13	0.46	16Sut
	SLCU.2RBY.CHR3_270889818	3	270889818	1.79E-10	0.29	16Sut
	SLCU.2RBY.CHR2_604066634	2	604066634	3.91E-10	0.09	16Sut
	SLCU.2RBY.CHR6_315678917	6	315678917	1.44E-08	0.28	16Sut
	SLCU.2RBY.CHR2_128835034	2	128835034	1.04E-08	0.06	17Ros
	SLCU.2RBY.CHR5_467478441	5	467478441	1.53E-08	0.48	17Ros
Asp	SLCU.2RBY.CHR7_2040151	7	2040151	5.72E-09	0.34	16Sut
	SLCU.2RBY.CHR6_392062289	6	392062289	3.41E-08	0.33	16Sut
	SLCU.2RBY.CHR5_467478441	5	467478441	6.16E-14	0.48	17Ros
	SLCU.2RBY.CHR1_109406086	1	109406086	9.6E-13	0.42	17Ros
	SLCU.2RBY.CHR7_479246092	7	479246092	4.81E-09	0.37	17Ros
Glu	SLCU.2RBY.CHR7_524176613	7	524176613	2.38E-10	0.46	16Sut
	SLCU.2RBY.CHR4_374653055	4	374653055	6.04E-10	0.19	16Sut
	SLCU.2RBY.CHR5_467478441	5	467478441	1.96E-11	0.48	17Ros
	SLCU.2RBY.CHR2_582173746	2	582173746	2.67E-11	0.32	17Ros
	SLCU.2RBY.CHR2_396979253	2	396979253	3.35E-10	0.12	17Ros
	SLCU.2RBY.CHR7_478047589	7	478047589	7.28E-10	0.49	17Ros
	SLCU.2RBY.CHR1_184695941	1	184695941	2.16E-08	0.41	17Ros
	SLCU.2RBY.CHR2_1973630	2	1973630	2.84E-08	0.25	17Ros
Ala	SLCU.2RBY.CHR3_235311412	3	235311412	7.33E-08	0.33	16Ros
	SLCU.2RBY.CHR6_28765615	6	28765615	1.36E-07	0.15	16Sut
	SLCU.2RBY.CHR5_467478441	5	467478441	5.62E-13	0.48	17Ros

	SLCU.2RBY.CHR2_396979253	2	396979253	5.91E-10	0.12	17Ros
	SLCU.2RBY.CHR1_293275629	1	293275629	1.84E-09	0.39	17Ros
	SLCU.2RBY.CHR2_580506970	2	580506970	9.26E-09	0.28	17Ros
	SLCU.2RBY.CHR2_603259068	2	603259068	1.99E-08	0.28	17Ros
	SLCU.2RBY.CHR1_503487	1	503487	4.41E-08	0.06	17Ros
	SLCU.2RBY.CHR2_2217769	2	2217769	4.68E-08	0.38	17Ros
	SLCU.2RBY.CHR3_130281462	3	130281462	5.06E-08	0.16	17Ros
	SLCU.2RBY.CHR3_52343909	3	52343909	1.09E-07	0.20	17Ros
	SLCU.2RBY.CHR6_94019559	6	94019559	1.41E-09	0.07	17Sut
	SLCU.2RBY.CHR4_175118045	4	175118045	5.99E-09	0.48	17Sut
	SLCU.2RBY.CHR6_411536500	6	411536500	2.65E-08	0.16	17Sut
Pro	SLCU.2RBY.CHR5_15051870	5	15051870	6.32E-11	0.10	16Sut
	SLCU.2RBY.CHR7_524176613	7	524176613	3.57E-08	0.46	16Sut
	SLCU.2RBY.CHR2_603259068	2	603259068	6.63E-08	0.28	16Sut
	SLCU.2RBY.CHR5_160215588	5	160215588	8.59E-08	0.09	16Sut
	SLCU.2RBY.CHR4_479372389	4	479372389	4.38E-10	0.22	17Ros
	SLCU.2RBY.CHR2_582173746	2	582173746	7.83E-10	0.32	17Ros
	SLCU.2RBY.CHR2_396131130	2	396131130	5.51E-09	0.11	17Ros
	SLCU.2RBY.CHR6_419415324	6	419415324	4.51E-09	0.09	17Sut
Cys	SLCU.2RBY.CHR3_6491845	3	6491845	7.38E-09	0.32	16Sut
	SLCU.2RBY.CHR3_269662193	3	269662193	3.53E-08	0.49	16Sut
	SLCU.2RBY.CHR2_520274159	2	520274159	6.25E-08	0.25	16Sut
	SLCU.2RBY.CHR2_536983132	2	536983132	6.28E-08	0.16	16Sut
	SLCU.2RBY.CHR4_1857563	4	1857563	1.68E-07	0.23	17Ros
	SLCU.2RBY.CHR6_96597650	6	96597650	1.74E-08	0.38	17Sut
Tyr	SLCU.2RBY.CHR2_128835034	2	128835034	2.53E-14	0.06	16Ros
	SLCU.2RBY.CHR7_524176613	7	524176613	3.37E-08	0.46	16Sut
	SLCU.2RBY.CHR5_467492181	5	467492181	1.63E-13	0.44	17Ros
	SLCU.2RBY.CHR2_582173746	2	582173746	3.29E-08	0.32	17Ros

[a] Abbreviation: chr, chromosome; maf, minor allele frequency

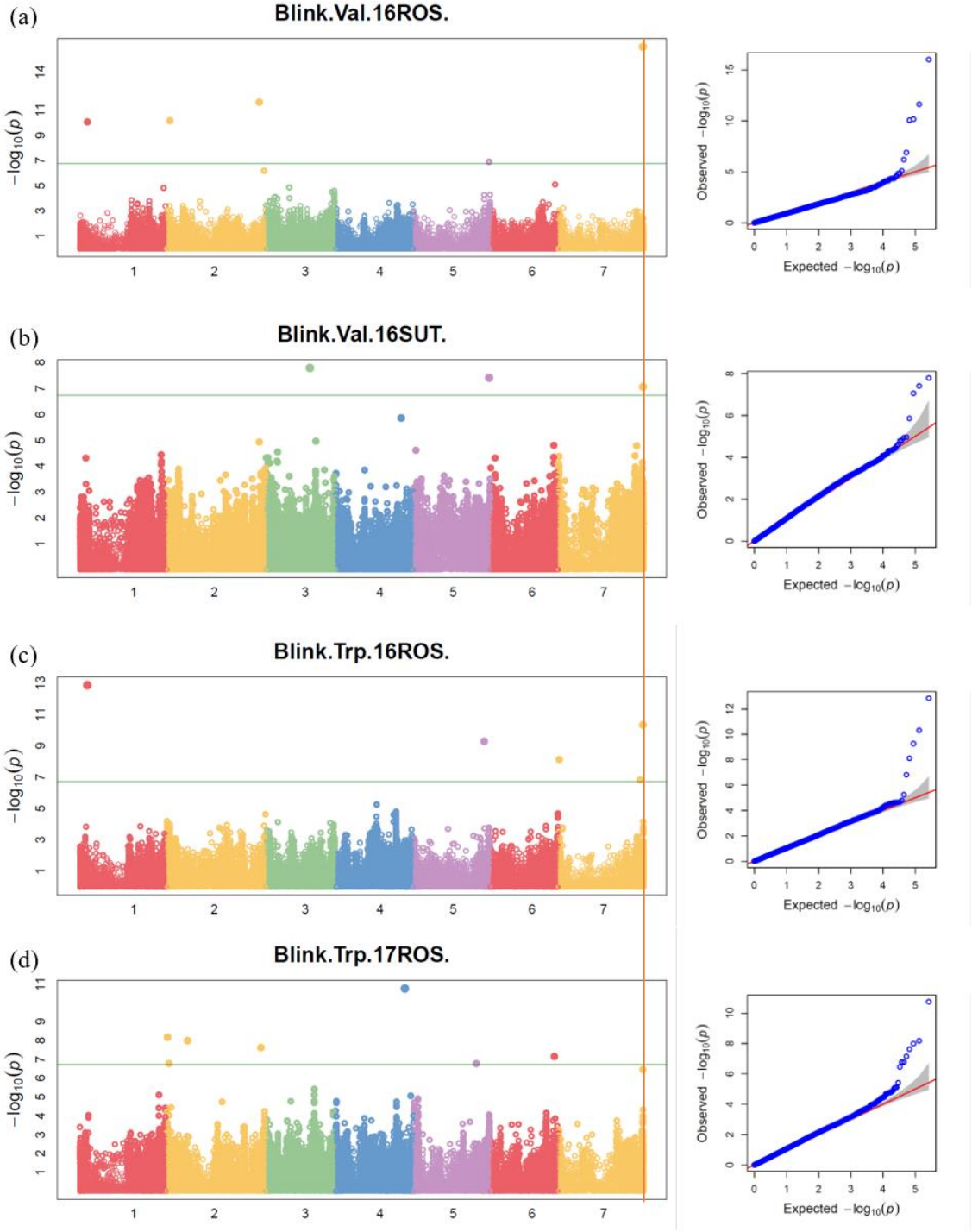


Figure 5.3. Manhattan plots (left) and quantile-quantile plots (right) for GWAS of the 324 lentil accessions for (a) Val content in 16Ros, (b) Val content in 16Sut, (c) Trp content in 16Ros and (d) Trp content in 17Ros. The green horizontal lines in the Manhattan plots represent the default significant threshold (-log10(p) = 6.73). The orange vertical line indicates the significant marker (SLCU.2RBY.CHR7_524204079).

Table 5.8 17 SNP markers associated with two or more traits in lentils simultaneously

SNP marker	Chr	Site-year	Trait associated
SLCU.2RBY.CHR1_46680284	1	16Ros	Val,Trp,Gly
SLCU.2RBY.CHR1_217285859	1	17Ros	Leu,Phe,Ser
SLCU.2RBY.CHR2_128835034	2	16Ros	Tyr
		17Ros	Val,Trp,Gly
SLCU.2RBY.CHR2_396979253	2	17Ros	Leu,Phe,Glu,Ala
SLCU.2RBY.CHR2_582173746	2	17Ros	Leu,Phe,Trp,Arg,Glu,Pro,Tyr
SLCU.2RBY.CHR2_603259068	2	16Sut	Pro
		17Ros	Phe,Ala
SLCU.2RBY.CHR3_270889818	3	16Sut	Thr,Val,Leu,Gly
SLCU.2RBY.CHR4_374653055	4	16Sut	Ile,Glu
SLCU.2RBY.CHR4_479372389	4	17Ros	Leu,Arg,Pro
SLCU.2RBY.CHR5_15051870	5	16Sut	Leu,Pro
SLCU.2RBY.CHR5_467478441	5	17Ros	Thr,Lys,Val,Ile,Leu,Phe,Ser,Gly,Asp,Glu,Ala
SLCU.2RBY.CHR5_467611866	5	16Sut	Protein,Val,Leu,Ser,Arg
SLCU.2RBY.CHR6_392062289	6	16Sut	Trp,Asp
SLCU.2RBY.CHR7_474941369	7	16Sut	Ile
		17Ros	Ser
SLCU.2RBY.CHR7_478047589	7	17Ros	Leu, Glu
SLCU.2RBY.CHR7_524176613	7	16Sut	Protein,Ile,Leu,Phe,Ser,Gly,Glu,Pro,Tyr
SLCU.2RBY.CHR7_524204079	7	16Ros	Val,Trp,Gly
		16Sut	Val

5.5 Conclusion

This study was the first to employ GWAS to identify SNP markers significantly associated with protein and 18 amino acids in cultivated lentils. These 19 traits showed a wide and continuous distribution. A total of 85 SNP markers were identified to be significantly associated with seed protein and amino acids. Most SNP markers were detected in only one of the four environments indicating that they may be environmentally dependent. Only one identical SNP marker (SLCU.2RBY.CHR7_524204079) on chromosome 7 significantly associated with Val was identified in two environments. Some amino acids were highly correlated and had the identical significantly associated SNPs in the same environment. These amino acids were likely to be improved simultaneously in the marker-assisted breeding program. In general, GWAS was a promising method to dissect the genetic basis of protein and amino acid contents in cultivated lentils. These identified SNPs could be studied further and facilitate selection in breeding programs to enhance seed protein content and quality.

CHAPTER 6. GENERAL DISCUSSION

Lentil (*Lens culinaris*) is a cool season pulse, which has high nutritional value due to high protein and fiber content, high digestibility, low fat content, and low calories (Subedi et al., 2021). Canada is the leading producer and exporter of lentils because lentils are suitable to grow in boreal climates. More importantly, extensive research and effort in Canada contribute to this success (Lizarazo et al., 2015). Lentil seeds are a good source of plant-based protein, which can combat protein malnutrition, replace soy protein or animal protein in food industries to generate new value-added products. Therefore, quantification of protein and amino acid contents in lentils is of great importance in breeding programs, food industries, and scientific research.

The traditional analytical methods for crude protein and amino acids have some drawbacks, which can be overcome by the Near-infrared Reflectance Spectroscopy (NIRS). Previous studies have confirmed that NIRS can be a robust method to determine the protein and most amino acid contents in several crops (Fontaine, Schirmer & Hörr, 2002; Fontaine et al., 2001; Kovalenko, Rippke & Hurburgh, 2006; Baianu et al., 2004; Wang et al., 2013; Yu et al., 2020; Zhang et al., 2011). However, little is known about the performance of NIRS models to determine these compositions in lentils. The current study showed that protein and most amino acids could be predicted by NIRS with satisfactory accuracy. Continued refinement of the calibration equations is needed to enhance the predictive value of these models for sulphur amino acids, His and Tyr.

NIRS models were grouped based on statistical parameters as follows: (a) $R^2_C \geq 0.82$ and RPD ≥ 2.14, models with good accuracy and usable for analytical purpose in most situations; (b) $R^2_C \geq 0.65$ and RPD ≥ 1.60, models suitable for sample screening; (c) $R^2_C \geq 0.50$ and RPD ≥ 1.50, models suitable for very rough to rough screening; (d) $R^2_C \geq 0.40$ and RPD ≥ 1.20, models can

distinguish between high and low values with careful use; (d) $R^2_C < 0.40$ or RPD < 1.20, insufficient and poor correlation models. This guideline was based on Williams & Norris (2001) with minor modifications.

The R^2 values highly depend on the range of the analyte of interest, and this statistical term may not reflect model performance completely (Esteve Agelet et al., 2012). Standard error statistics (SEC, SECV, SEP) should be emphasized in evaluating model performance. Several statistical terms should be considered together to analyze efficiency of NIRS models.

The NIRS analytical method has several strengths. Using NIRS for predicting protein and amino acid contents in lentil seeds can result in a significantly reduced cost, reduced chemical waste and a larger scale for testing samples in an environmentally friendly way. As a non-destructive and non-invasive method, NIRS offers an attractive method to measure protein and amino acid contents in seeds with limited volume such as those derived from breeding programs. It can assist the breeding program in selecting lentils with higher protein content and quality. However, NIRS is a secondary analytical method, and the NIR model performance depends on the accuracy of analytical methods. Additionally, this method needs extensive calibration. The compound concentration of predicted samples should fall within the concentration range of calibration samples.

The established NIRS models were used to predict the protein and amino acid contents in 1290 lentil seeds, and these data became the phenotypic data in the following genome-wide association study (GWAS). This study used SNP arrays as a genotyping method. This genotyping method is popular, cost-effective and accurate; however, it may not cover rare variants. Another genotyping method is whole-genome sequencing. Although it is more expensive and relatively less accurate, it is the gold standard in GWAS because it covers all

variants and overcomes potential limits such as identification of missed signals and ultra-rare mutations (Alseekh et al., 2021). Several GWASs have used the phenotypic data determined by NIRS, which indicated that the NIRS could be a suitable quantification method for GWAS (Lee et al., 2019; Zhang et al, 2021; Zhang et al., 2018a; Zhang et al., 2018b). However, NIRS models with a relatively lower R^2 could generate less accurate results, which adversely affected the GWAS and generated inconsistent results (Pearson & Manolio, 2008). Therefore, refined and frequently updated NIRS models are a must for GWA studies, especially for Cys, Met, His and Tyr calibration models.

BLINK model was applied in the current GWAS to identify potential SNP markers. This method is very fast and has high statistical power; however, it was first introduced in 2019 (Huang et al., 2019), and this relatively new method has had fewer citations compared to other methods (Zhang, 2020). This study could be improved by using multiple methods, and the markers detected by several methods could be regarded as credible markers which would be in the primary position for further studies. These significantly associated SNPs could be used to identify potential genes in the following research.

CHAPTER 7. FUTURE DIRECTIONS

Future research could focus on the following directions:

- Several feasible improvement methods can be applied in future studies to enhance the NIRS model performance, including (i) enhancing the accuracy of reference methods by doing replicates (Fontaine et al., 2001) or/and by optimizing the methods for specific crop in question (ii) updating sample calibration pool by adding new samples to increase model representativeness (Agelet & Hurburgh Jr, 2010), (iii) investigating the influence of different spectral preprocessing methods and wavelength ranges on the model performance to find the best methods and ranges for each composition (Shi and Yu, 2017), and (iv) constructing models using different linear and nonlinear regression methods and finding the most suitable method.
- In addition to cross-validation, external validation is another way to evaluate the quality and estimate the future prediction error of the calibration model. External validation is widely used in recent NIRS studies. New lentil samples from different cropping years or different locations can create a test set for external validation in future works.
- Seed yield is a valuable agronomic trait, which is worth examining in further works. Seed yield can be used to calculate protein yield. After determining the protein yield, the relationship between the protein concentration and protein yield can be further studied. Moreover, the effect of genotype, environment, and G × E interaction on protein yield of lentil can be investigated. This can help us understand more about lentil protein, which can also benefit the breeding program.
- Planting seeds in two or three replications in each site-year can enhance the accuracy of the experiments and reduce some random errors. In addition, this design can help

scientists to determine whether there is a significant G × E interaction effect on seed compositions.

- Planting seeds in different environments (i.e. locations such as tropics and/or years) could evaluate the performance of genotypes across diverse environments to identify significant and broadly adapted markers (Alqudah et al.,2020).
- These GWA studies only used the BLINK model to detect SNP molecular markers. Karikari et al. (2020) mentioned that using multiple models is more effective to detect markers. Zhang et al. (2018a) also reported the complementarity of MLM and MLMM models. Applying other statistical models in GWAS can generate more convincing and complete results. The peaks detected by the BLINK model could be compared with results from other models for statistical validation.
- These identified SNPs can be studied, verified, and confirmed in further research. Furthermore, these identified SNPs could be used to deduce potential genomic regions and candidate genes associated with these traits.
- The causal genes underlying detected peaks could be further validated through functional studies, including overexpression, CRISPR-Cas9 mediated knockout and other gene editing methods (Tibbs Cortes, Zhang & Yu, 2021).

REFERENCES

AACC International (2010). Approved methods of analysis, 11th edn (On-line). The American Association of Cereal Chemists, St. Paul, MN

Agelet, L. E., & Hurburgh Jr, C. R. (2010). A tutorial on near-infrared spectroscopy and its calibration. *Critical Reviews in Analytical Chemistry*, *40*(4), 246-260.

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.

Alqudah, A. M., Sallam, A., Baenziger, P. S., & Börner, A. (2020). GWAS: Fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley–A review. *Journal of advanced research*, *22*, 119-135.

Alseekh, S., Kostova, D., Bulut, M., & Fernie, A. R. (2021). Genome-wide association studies: assessing trait characteristics in model and crop plants. *Cellular and Molecular Life Sciences*, 1-12.

Andersen, H. V., Wedelsback, H., & Hansen, P. W. (2013). NIR spectrometer technology comparison. *White Paper from FOSS*, 1-14.

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., ... & Nordborg, M. (2005). Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS genetics*, *1*(5), e60.

Arumuganathan, K., & Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant molecular biology reporter*, 9(3), 208-218.

Asif, M., Rooney, L., Ali, R., & Riaz, M. (2013). Application and Opportunities of Pulses in Food System: A Review. *Critical Reviews in Food Science and Nutrition*, *53*(11), 1168–1179. https://doi.org/10.1080/10408398.2011.574804

Baianu, I. C., You, T., Costescu, D. M., Lozano, P. R., Prisecaru, V., & Nelson, R. L. (2004). High-resolution nuclear magnetic resonance and near-infrared determination of soybean oil, protein, and amino acid residues in soybean seeds. *Oil extraction and analysis: Critical issues and ComparatiVe studies*, 193-340.

Bhatty R.S., Slinkard A.E., & Sosulski F.W. (1976). Chemical composition and protein characteristics of lentils. *Canadian Journal of Plant Science*, *56*(4), 787–794.

Bhatty, R. (1988). In vitro hydrolysis of pea, faba bean and lentil meals and isolated protein fractions by pepsin and trypsin. *Canadian Institute of Food Science and Technology Journal : Journal de l'Institut Canadien de Science et Technologie Alimentaire*, *21*(1), 66–71. https://doi.org/10.1016/S0315-5463(88)70719-1

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, *327*(8476), 307-310.

Blanco, M., & Villarroya, I. N. I. R. (2002). NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry*, *21*(4), 240-250.

Boye, J. (2015). Lentil. *The Canadian Encyclopedia.* Retrieved from https://www.thecanadianencyclopedia.ca/en/article/lentil.

Boye, J., Aksay, S., Roufik, S., Ribéreau, S., Mondor, M., Farnworth, E., & Rajamohamed, S. (2010). Comparison of the functional properties of pea, chickpea and lentil protein concentrates processed using ultrafiltration and isoelectric precipitation techniques. *Food Research International*, *43*, 537–546.

Carbas, B., Machado, N., Oppolzer, D., Ferreira, L., Brites, C., Rosa, E. A., & Barros, A. I. (2020). Comparison of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for the determination of nutritional and antinutritional parameters in common beans. *Food chemistry*, *306*, 125509.

Carbonaro, M., Cappelloni, M., Nicoli, S., Lucarini, M., & Carnovale, E. (1997). Solubility−Digestibility Relationship of Legume Proteins. *Journal of Agricultural and Food Chemistry*, *45*(9), 3387–3394. https://doi.org/10.1021/jf970070y

Chen, Z., Vu, J. L., Vu, B. L., Buitink, J., Leprince, O., & Verdier, J. (2021). Genome-wide association studies of seed performance traits in response to heat stress in Medicago truncatula uncover MIEL1 as a regulator of seed germination plasticity. *Frontiers in plant science*, *12*.

Choukri, H., Hejjaoui, K., & El-Baouchi, A. (2020). Heat and drought stress impact on phenology, grain yield, and nutritional quality of lentil (Lens culinaris Medikus). *Frontiers in Nutrition*, *7*.

Combe, E., Achi, T., & Pion, R. (1991). Comparative digestive and metabolic utilization of beans, lentils and chick peas in the rat. *Reproduction, nutrition, development*, *31*(6), 631-646.Daun, J. K. (1999). Quality of western Canadian lentils. Retrieved from https://publications.gc.ca/collections/collection_2012/ccg-cgc/A92-16-1999-eng.pdf

Daun, J.K. (1999). Quality of western Canadian lentils (1999). Retrieved from https://publications.gc.ca/collections/collection_2012/ccg-cgc/A92-16-1999-eng.pdf

de Souza Cândido, E., Pinto, M. F. S., Pelegrini, P. B., Lima, T. B., Silva, O. N., Pogue, R., ... & Franco, O. L. (2011). Plant storage proteins with antimicrobial activity: novel insights into plant defense mechanisms. *The FASEB Journal*, *25*(10), 3290-3305.

Detzel, A., Krüger, M., Busch, M., Blanco-Gutiérrez, I., Varela, C., Manners, R., ... & Zannini, E. (2021). Life cycle assessment of animal-based foods and plant-based protein-rich alternatives: an environmental perspective. *Journal of the Science of Food and Agriculture*.Directive, C. (2000). Establishing community methods for the determination of vitamin A, vitamin E and tryptophan, annex part C. Determination of Tryptophan. *Official J European Communities L*, *174*, 45-50.

Dissanayake, R., Cogan, N. O., Smith, K. F., & Kaur, S. (2021). Application of Genomics to Understand Salt Tolerance in Lentil. *Genes*, *12*(3), 332.

Erskine, W. (2009). *The lentil botany, production and uses.* Cambridge, Mass: CABI.

Esteve Agelet, L., Armstrong, P. R., Romagosa Clariana, I., & Hurburgh, C. R. (2012). Measurement of single soybean seed attributes by near-infrared technologies. A comparative study. *Journal of agricultural and food chemistry*, *60*(34), 8314-8322.

FAO (2021). Production quantities of Lentils by country 2019. Retrieved from http://www.fao.org/faostat/en/#data/QC/visualize

FAO/WHO/UNU, E. C. (1985). Energy and protein requirements. *World Health Organ Tech Rep Ser*, *724*, 1-206. Retrieved from http://www.fao.org/3/aa040e/aa040e00.htm

FAO/WHO (1991). Protein quality evaluation. Report of the Joint FAO/WHO Expert Consultation. Food and Nutrition Paper No. 51. Rome: Food and Agriculture Organizations and the World Health Organization.

FAO/WHO (2013). Dietary protein quality evaluation in human nutrition Report of an FAO Expert Consultation. Food and Nutrition Paper No. 92. Rome: Food and Agriculture Organizations and the World Health Organization.

Faris, M., Takruri, H., & Issa, A. (2013). Role of lentils (Lens culinaris L.) in human health and nutrition: a review. *Mediterranean Journal of Nutrition and Metabolism*, *6*(1), 3–16. https://doi.org/10.1007/s12349-012-0109-8

Fernández-Novales, J., Garde-Cerdán, T., Tardáguila, J., Gutiérrez-Gamboa, G., Pérez-Álvarez, E. P., & Diago, M. P. (2019). Assessment of amino acids and total soluble solids in intact grape berries using contactless Vis and NIR spectroscopy during ripening. *Talanta*, *199*, 244-253.

Ferreira, D. S., Galão, O. F., Pallone, J. A. L., & Poppi, R. J. (2014). Comparison and application of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for determination of quality parameters in soybean samples. *Food Control*, *35*(1), 227-232.

Ferreira, D. S., Pallone, J. A. L., & Poppi, R. J. (2013). Fourier transform near-infrared spectroscopy (FT-NIRS) application to estimate Brazilian soybean [Glycine max (L.) Merril] composition. *Food Research International*, *51*(1), 53-58.

Fontaine, J., Hoerr, J., Schirmer, B., & Fontaine, J. (2001). Near-Infrared Reflectance Spectroscopy Enables the Fast and Accurate Prediction of the Essential Amino Acid Contents in Soy, Rapeseed Meal, Sunflower Meal, Peas, Fishmeal, Meat Meal Products, and Poultry Meal. *Journal of Agricultural and Food Chemistry*, *49*(1), 57–66. https://doi.org/10.1021/jf000946s

Fontaine, J., Schirmer, B., & Hörr, J. (2002). Near-infrared reflectance spectroscopy (NIRS) enables the fast and accurate prediction of essential amino acid contents. 2. Results for wheat, barley, corn, triticale, wheat bran/middlings, rice bran, and sorghum. *Journal of Agricultural and Food Chemistry*, *50*(14), 3902-3911.

García-Sánchez, F., Galvez-Sola, L., Martínez-Nicolás, J. J., Muelas-Domingo, R., & Nieves, M. (2017). Using near-infrared spectroscopy in agricultural systems. *Developments in near-infrared spectroscopy*, *1*, 97-127.

Gela, T. S., Ramsay, L., Haile, T., Vandenberg, A., & Bett, K. (2021). Identification of anthracnose (Colletotrichum lentis) race 1 resistance loci in lentil by integrating linkage mapping and a genome-wide association study. *bioRxiv*.

González-Martín, I., Álvarez-García, N., & González-Cabrera, J. (2006). Near-infrared spectroscopy (NIRS) with a fibre-optic probe for the prediction of the amino acid composition in animal feeds. *Talanta*, *69*(3), 706–710. https://doi.org/10.1016/j.talanta.2005.11.015

Grela, E., Kiczorowska, B., Samolińska, W., Matras, J., Kiczorowski, P., Rybiński, W., & Hanczakowska, E. (2017). Chemical composition of leguminous seeds: part I—content of basic nutrients, amino acids, phytochemical compounds, and antioxidant activity. *European Food Research and Technology*, *243*(8), 1385–1395. https://doi.org/10.1007/s00217-017-2849-7

Hawtin, G. C., Rachie, K. O., & Green, J. M. (1977). Breeding strategy for the nutritional improvement of pulses. In *Nutritional standards and methods of evaluation for food legume breeders*. IDRC, Ottawa, ON, CA.

Health Canada. (1981). Determination of protein rating FO-1. http://www.hc-sc.gc.ca/fnan/alt_formats/hpfb-dgpsa/pdf/res-rech/fo-1-eng.pdf Accessed 20.01.17.

Health Canada. (2018). Lentils, raw/. Retrieved from https://food-nutrition.canada.ca/cnf-fce/report-rapport.do

Health Canada. (2019). Plant-based protein market: global and Canadian market analysis. Retrieved from https://nrc.canada.ca/en/research-development/research-collaboration/programs/plant-based-protein-market-global-canadian-market-analysis

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Huang, J., Romero-Torres, S., & Moshgbar, M. (2010). Practical Considerations in Data Pre-treatment for NIR and Raman Spectroscopy, American Pharmaceutical Review. *Dostopno na: http://www. americanpharmaceuticalreview. com/Featured-Articles/116330-Practical-Considerations-in-Data-Pre-treatment-for-NIR-and-Raman-Spectroscopy/.[Dostop: 10-Sep-2019]*.

Huang, M., Liu, X., Zhou, Y., Summers, R. M., & Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *GigaScience*, *8*(2), giy154.

Iqbal, A., Khalil, I., Ateeq, N., & Sayyar Khan, M. (2006). Nutritional quality of important food legumes. *Food Chemistry*, *97*(2), 331–335. https://doi.org/10.1016/j.foodchem.2005.05.011

ISO. (2005). ISO 13904:2005 Animal feeding stuffs — Determination of tryptophan content. Retrieved from https://www.iso.org/standard/37259.html

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 & Technology*, *53*(4), 892-903.

Kabaha, K., Taralp, A., Cakmak, I., & Ozturk, L. (2011). Accelerated Hydrolysis Method To Estimate the Amino Acid Content of Wheat (Triticum durum Desf.) Flour Using

Microwave Irradiation. *Journal of Agricultural and Food Chemistry*, *59*(7), 2958–2965. https://doi.org/10.1021/jf103678c

Kahraman, A. (2016). Nutritional components and amino acids in lentil varieties. *Selcuk Journal of Agriculture and Food Sciences*, *30*(1), 34-38.

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.

Karaköy, T., Erdem, H., Baloch, F., Toklu, F., Eker, S., Kilian, B., & Özkan, H. (2012). Diversity of macro- and micronutrients in the seeds of lentil landraces. *TheScientificWorldJournal*, *2012*, 710412.

Karikari, B., Wang, Z., Zhou, Y., Yan, W., Feng, J., & Zhao, T. (2020). Identification of quantitative trait nucleotides and candidate genes for soybean seed weight by multiple models of genome-wide association study. *BMC plant biology*, *20*(1), 1-14.

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), 1-18.

Kavas, A., & Nehir, S. (1992). Changes in nutritive value of lentils and mung beans during germination. *Chemie, Mikrobiologie, Technologie der Lebensmittel*, *14*(1-2), 3-9.

Khan, S. A., Khan, S. B., Khan, L. U., Farooq, A., Akhtar, K., & Asiri, A. M. (2018). Fourier Transform Infrared Spectroscopy: Fundamentals and Application in Functional Groups and Nanomaterials Characterization. *In Handbook of Materials Characterization (pp. 317-344)*. Springer, Cham.

Khatun, A., Bhuiyan, M. A. H., & Dey, T. K. (2010). Nitrogen uptake and protein yield in lentil as influenced by seed collection from different parts of plants. Bangladesh Journal of Agricultural Research, 35(3), 515-523.

Khazaei, H., Fedoruk, M., Caron, C. T., Vandenberg, A., & Bett, K. E. (2018). Single nucleotide polymorphism markers associated with seed quality characteristics of cultivated lentil. *The plant genome*, *11*(1), 170051.

Khazaei, H., Podder, R., Caron, C. T., Kundu, S. S., Diapari, M., Vandenberg, A., & Bett, K. E. (2017). Marker–trait association analysis of iron and zinc concentration in lentil (Lens culinaris Medik.) seeds. *The plant genome*, *10*(2), plantgenome2017-02.

Khazaei, H., Subedi, M., Nickerson, M., Martínez-Villaluenga, C., Frias, J., Vandenberg, A., & Khazaei, H. (2019). Seed Protein of Lentils: Current Status, Progress, and Food Applications. *Foods (Basel, Switzerland)*, *8*(9). https://doi.org/10.3390/foods8090391

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, *9*(1), 1-9.

Kovalenko, I. V., Rippke, G. R., & Hurburgh, C. R. (2006). Determination of amino acid composition of soybeans (Glycine max) by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry,* 54(10), 3485-3491.

Krul, E. S. (2019). Calculation of nitrogen-to-protein conversion factors: A review with a focus on soy protein. *Journal of the American Oil Chemists' Society*, *96*(4), 339-364.

Kumar, J., Singh, J., Kanaujia, R., & Gupta, S. (2016). Protein content in wild and cultivated taxa of lentil (Lens culinaris ssp. culinaris Medikus). *Indian J. Genet. Plant Breed*, *76*, 631-634.

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*, 78.

Leamy, L. J., Zhang, H., Li, C., Chen, C. Y., & Song, B. H. (2017). A genome-wide association study of seed composition traits in wild soybean (Glycine soja). BMC genomics, 18(1), 1-15.

Lee, S., Van, K., Sung, M., Nelson, R., LaMantia, J., McHale, L. K., & Mian, M. R. (2019). Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. *Theoretical and Applied Genetics*, *132*(6), 1639-1659.

Leser, S. (2013). The 2013 FAO report on dietary protein quality evaluation in human nutrition: Recommendations and implications. *Nutrition Bulletin*, *38*(4), 421-428.

Li, G., Wang, R., Quampah, A., Rong, Z., Shi, C., Wu, J., & Li, G. (2011). Calibration and prediction of amino acids in stevia leaf powder using near-infrared reflectance spectroscopy. *Journal of Agricultural and Food Chemistry, 59*(24), 13065–13071. https://doi.org/10.1021/jf2035912

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

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

Lizarazo, C., Lampi, A., Liu, J., Sontag-Strohm, T., Piironen, V., & Stoddard, F. (2015). Nutritive quality and protein production from grain legumes in a boreal climate. *Journal*

of the Science of Food and Agriculture, *95*(10), 2053–2064. https://doi.org/10.1002/jsfa.6920

Lohr, D., Tillmann, P., Zerche, S., Druege, U., Rath, T., & Meinken, E. (2016). Non-destructive measurement of nitrogen status of leafy ornamental cuttings by near infrared reflectance spectroscopy (NIRS) for assessment of rooting capacity. *Biosystems Engineering*, *148*, 157-167.

Moldovan, O., Păucean, A., Vlaic, R., BORȘ, M. D., & Muste, S. (2015). Preliminary assessment of the nutritional quality of two types of lentils (Lens culinaris) by near-infrared reflectance spectroscopy technology (Nirs). *Bulletin UASVM Food Science and Technology*, *72*, 1.

Neupane, S. (2019). *Flowering Time Response of Diverse Lentil (Lens Culinaris Medik.) Germplasm Grown in Multiple Environments* (Doctoral dissertation, University of Saskatchewan).

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest biology and technology*, *46*(2), 99-118.

Nielsen, S. (2010). *Food Analysis* (4th ed.). https://doi.org/10.1007/978-1-4419-1478-1

Nigro, D., Gadaleta, A., Mangini, G., Colasuonno, P., Marcotuli, I., Giancaspro, A., ... & Blanco, A. (2019). Candidate genes and genome-wide association study of grain protein content and protein deviation in durum wheat. *Planta*, *249*(4), 1157-1175.

Nosworthy, M., & House, J. (2017). Factors Influencing the Quality of Dietary Proteins: Implications for Pulses. *Cereal Chemistry*, *94*(1), 49–57. https://doi.org/10.1094/CCHEM-04-16-0104-FI

Nosworthy, M. G., Medina, G., Franczyk, A. J., Neufeld, J., Appah, P., Utioh, A., ... & House, J. D. (2018). Effect of processing on the in vitro and in vivo protein quality of red and green lentils (Lens culinaris). *Food chemistry, 240*, 588-593.

Nosworthy, M., Neufeld, J., Frohlich, P., Young, G., Malcolmson, L., & House, J. (2017). Determination of the protein quality of cooked Canadian pulses. *Food Science & Nutrition*, *5*(4), 896–903. https://doi.org/10.1002/fsn3.473

Official methods of analysis of AOAC International (CD-ROM). (n.d.). Gaithersburg, Md: AOAC International. Retrieved from http://www.eoma.aoac.org/

Osborne, B. G. (2006). Near-infrared spectroscopy in food analysis. *Encyclopedia of analytical chemistry: applications, theory and instrumentation*.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., ... & Tanaka, T. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, *32*(4), 650-654.

Ozaki, Y., McClure, W. F., & Christy, A. A. (Eds.). (2006). *Near-infrared spectroscopy in food science and technology*. John Wiley & Sons.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, *2*(12), e190.

Paucean, A., Moldovan, O., Mureșan, V., Socaci, S., Dulf, F., Alexa, E., … Muste, S. (2018). Folic acid, minerals, amino-acids, fatty acids and volatile compounds of green and red lentils. Folic acid content optimization in wheat-lentils composite flours. *Chemistry Central Journal*, *12*(1), 1–9. https://doi.org/10.1186/s13065-018-0456-8

Pazdernik, D. L., Killam, A. S., & Orf, J. H. (1997). Analysis of amino and fatty acid composition in soybean seed, using near-infrared reflectance spectroscopy. *Agronomy Journal*, *89*(4), 679-685.

Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association study. *Jama*, *299*(11), 1335-1344.

Plans, M., Simó, J., Casañas, F., Sabaté, J., & Rodriguez-Saona, L. (2013). Characterization of common beans (Phaseolus vulgaris L.) by infrared spectroscopy: Comparison of MIR, FT-NIR and dispersive NIR using portable and benchtop instruments. *Food Research International*, *54*(2), 1643–1651. https://doi.org/10.1016/j.foodres.2013.09.003

Porep, J. U., Kammerer, D. R., & Carle, R. (2015). On-line application of near-infrared (NIR) spectroscopy in food production. *Trends in Food Science & Technology*, *46*(2), 211-230.

Porres, J. M., Urbano, G., Fernández-Fígares, I., Prieto, C., Perez, L., & Aguilera, J. F. (2002). Digestive utilisation of protein and amino acids from raw and heated lentils by growing rats. *Journal of the Science of Food and Agriculture*, *82*(14), 1740-1747.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904-909.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Qin, J., Shi, A., Song, Q., Li, S., Wang, F., Cao, Y., ... & Zhang, M. (2019). Genome wide association study and genomic selection of amino acid concentrations in soybean seeds. *Frontiers in plant science*, *10*, 1445.

Quiñones, M. D. C. S., Martínez, L. A. O., Herrera, S. M. G., Quiñones, O. M. R., Laredo, R. F. G., & y Bioquímica, Q. (2018). Near-Infrared Spectroscopy (NIRS) applied to legume analysis: A Review. *Spectroscopy*, *8*(4).

Revilla, I., Lastras, C., González-Martín, M., Vivar-Quintana, A., Morales-Corts, R., Gómez-Sánchez, M., & Pérez-Sánchez, R. (2019). Predicting the physicochemical properties and geographical ORIGIN of lentils using near-infrared spectroscopy. *Journal of Food Composition and Analysis*, *77*, 84–90. https://doi.org/10.1016/j.jfca.2019.01.012

Rozan, P., Kuo, Y., & Lambein, F. (2001). Amino acids in seeds and seedlings of the genus Lens. *Phytochemistry*, *58*(2), 281–289. https://doi.org/10.1016/S0031-9422(01)00200-X

Rubenthaler, G. L., & Bruinsma, B. L. (1978). Lysine Estimation in Cereals by Near-Infrared Reflectance 1. *Crop Science*, *18*(6), 1039-1042.

Rutherfurd, S. M., & Gilani, G. S. (2009). Amino acid analysis. *Current protocols in protein science*, *58*(1), 11-9.

Saha, U., Endale, D., Tillman, P. G., Johnson, W. C., Gaskin, J., Sonon, L., ... & Yang, Y. (2017). Analysis of various quality attributes of sunflower and soybean plants by near-infrared reflectance spectroscopy: Development and validation calibration models. *American Journal of Analytical Chemistry*, *8*(7), 462-492.

Scippa, G., Rocco, M., Ialicicco, M., Trupiano, D., Viscosi, V., Di Michele, M., … Scaloni, A. (2010). The proteome of lentil (Lens culinaris Medik.) seeds: Discriminating between landraces. *ELECTROPHORESIS*, *31*(3), 497–506. https://doi.org/10.1002/elps.200900459

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, *44*(7), 825-830.

Sehgal, A., Sita, K., Kumar, J., Kumar, S., Singh, S., Siddique, K., & Nayyar, H. (2017). Effects of Drought, Heat and Their Interaction on the Growth, Yield and Photosynthetic Function of Lentil (Medikus) Genotypes Varying in Heat and Drought Sensitivity. *Frontiers in Plant Science*, *8*, 1776. https://doi.org/10.3389/fpls.2017.01776

Sehgal, A., Sita, K., Siddique, K., Kumar, R., Bhogireddy, S., Varshney, R., … Sehgal, A. (2018). Drought or/and Heat-Stress Effects on Seed Filling in Food Crops: Impacts on Functional Biochemistry, Seed Yields, and Nutritional Quality. *Frontiers in Plant Science*, *9*, 1705–1705. https://doi.org/10.3389/fpls.2018.01705

Semba, R. (2016). The Rise and Fall of Protein Malnutrition in Global Health. *Annals of Nutrition and Metabolism*, *69*(2), 79–88. https://doi.org/10.1159/000449175

Shekib, L. A., 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.

Shi, H., & Yu, P. (2017). Comparison of grating-based near-infrared (NIR) and Fourier transform mid-infrared (ATR-FT/MIR) spectroscopy based on spectral preprocessing and wavelength selection for the determination of crude protein and moisture content in wheat. *Food Control*, *82*, 57-65.

Sita, K., Sehgal, A., Bhandari, K., Kumar, J., Kumar, S., Singh, S., … Nayyar, H. (2018). Impact of heat stress during seed filling on seed quality and seed yield in lentil (*Lens culinaris*

Medikus) genotypes. *Journal of the Science of Food and Agriculture*, *98*(13), 5134–5141. https://doi.org/10.1002/jsfa.9054

Smyth, H., Cozzolino, D., Cynkar, W., Dambergs, R., Sefton, M., & Gishen, M. (2008). Near-infrared spectroscopy as a rapid tool to measure volatile aroma compounds in Riesling wine: possibilities and limits. *Analytical and Bioanalytical Chemistry*, *390*(7), 1911–1916. https://doi.org/10.1007/s00216-008-1940-0

Sosulski, F. W., & Holt, N. W. (1980). Amino acid composition and nitrogen-to-protein factors for grain legumes. *Canadian Journal of Plant Science*, *60*(4), 1327-1331.

Stark, E., & Luchter, K. (2005). NIR instrumentation technology. *NIR news*, *16*(7), 13-16.

Subedi, M., Khazaei, H., Arganosa, G., Etukudo, E., & Vandenberg, A. (2021). Genetic stability and genotype× environment interaction analysis for seed protein content and protein yield of lentil. *Crop Science*, *61*(1), 342-356.

Tahir, M., Lindeboom, N., Baga, M., Vandenberg, A., & Chibbar, R. (2011). Composition and correlation between major seed constituents in selected lentil (Lens culinaris. Medik) genotypes. *Canadian Journal of Plant Science*, *91*(5), 825–835. https://doi.org/10.4141/cjps2011-010

Themelis, T., Gotti, R., Orlandini, S., & Gatti, R. (2019). Quantitative amino acids profile of monofloral bee pollens by microwave hydrolysis and fluorimetric high performance liquid chromatography. *Journal of pharmaceutical and biomedical analysis*, *173*, 144-153.

Thompson, M., Owen, L., Wilkinson, K., Wood, R., & Damant, A. (2002). A comparison of the Kjeldahl and Dumas methods for the determination of protein in foods, using data from a proficiency testing scheme. *Analyst*, *127*(12), 1666-1668.

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome, 14*(1), e20077.

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *Biorxiv*, 005165.

Urbano, G., Lopez-Jurado, M., Hernandez, J., Fernandez, M., Moreu, M. C., Frias, J., ... & Vidal-Valverde, C. (1995). Nutritional assessment of raw, heated, and germinated lentils. *Journal of Agricultural and Food Chemistry*, *43*(7), 1871-1877.

USDA. (2018). Lentils, raw. Retrieved from https://fdc.nal.usda.gov/fdc-app.html#/food-details/172420/nutrients

Wang, L., Wang, Q., Liu, H., Liu, L., & Du, Y. (2013). Determining the contents of protein and amino acids in peanuts using near-infrared reflectance spectroscopy. *Journal of the Science of Food and Agriculture*, *93*(1), 118–124. https://doi.org/10.1002/jsfa.5738

Wang, J., Liu, H., & Ren, G. (2014a). Near-infrared spectroscopy (NIRS) evaluation and regional analysis of Chinese faba bean (Vicia faba L.). *The Crop Journal*, *2*(1), 28–37. https://doi.org/10.1016/j.cj.2013.10.001

Wang, N., & Daun, J. K. (2004). The chemical composition and nutritive value of Canadian pulses. *Canadian Grain Commission Report*, 19-29.

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*, 493–502.

Wang, N. (2016). Quality of western Canadian lentils. Retrieved from https://www.grainscanada.gc.ca/en/grain-research/export-quality/pulses/lentils/2016/lentils-quality-report-16.pdf

Wang, N. (2017). Quality of western Canadian lentils. Retrieved from https://www.grainscanada.gc.ca/en/grain-research/export-quality/pulses/lentils/2017/lentils-quality-report-17.pdf

Wang, N. (2019). Quality of western Canadian lentils. Retrieved from https://grainscanada.gc.ca/en/grain-research/export-quality/pulses/lentils/2019/pdf/report2019.pdf

Wang, N. (2020). Quality of western Canadian lentils. Retrieved from https://www.grainscanada.gc.ca/en/grain-research/export-quality/pulses/lentils/2020/pdf/quality-report-lentil-2020.pdf

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., & Zhang, Z. (2014b). A SUPER powerful method for genome wide association study. *PloS One*, *9*(9), e107684–e107684. https://doi.org/10.1371/journal.pone.0107684

Waters Corporation. (2014). AccQ•Tag Ultra Derivatization Kit. Retrieved from https://www.waters.com/webassets/cms/support/docs/715001331.pdf

Watford, M., & Wu, G. (2011). Protein. *Advances in Nutrition (Bethesda, Md.)*, *2*(1), 62–63. https://doi.org/10.3945/an.110.000091

WHO, J. (2007). Protein and amino acid requirements in human nutrition. *World Health Organization technical report series*, (935), 1.

Williams, P., Manley, M., & Antoniszyn, J. (2019). *Near-infrared technology: getting the best out of light*. African Sun Media.

Williams, P., & Norris, K. (2001). *Near-infrared technology: in the agricultural and food industries* (2nd ed.). American Association of Cereal Chemists.

Wong, M., Gujaria-Verma, N., Ramsay, L., Yuan, H., Caron, C., Diapari, M., … Wong, M. (2015). Classification and characterization of species within the genus lens using genotyping-by-sequencing (GBS). *PloS One*, *10*(3), e0122025–e0122025. https://doi.org/10.1371/journal.pone.0122025

Wright, D. (2021). GWAS Vignette. Retrieved November 4, 2021 from https://derekmichaelwright.github.io/htmls/academic/gwas_tutorial.html#data

Wright, D. M., Neupane, S., Heidecker, T., Haile, T. A., Chan, C., Coyne, C. J., ... & Bett, K. E. (2021). Understanding photothermal interactions will help expand production range and increase genetic diversity of lentil (Lens culinaris Medik.). *Plants, People, Planet*, *3*(2), 171-181.

Wu, J. G., Shi, C., & Zhang, X. (2002). Estimating the amino acid composition in milled rice by near-infrared reflectance spectroscopy. *Field Crops Research*, *75*(1), 1-7.

Yu, H., Liu, H., Erasmus, S., Zhao, S., Wang, Q., & van Ruth, S. (2020). Rapid high-throughput determination of major components and amino acids in a single peanut kernel based on portable near-infrared spectroscopy combined with chemometrics. *Industrial Crops and Products*, *158*. https://doi.org/10.1016/j.indcrop.2020.112956

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., ... & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, *38*(2), 203-208.

Yuan, W., Wu, Z., Zhang, Y. E., Yang, R., Wang, H., Kan, G., & Yu, D. (2021). Genome-wide association studies for sulfur-containing amino acids in soybean seeds. *Euphytica*, *217*(8), 1-14.

Zhang, B., Rong, Z., Shi, Y., Wu, J., & Shi, C. (2011). Prediction of the amino acid composition in brown rice using different sample status by near-infrared reflectance spectroscopy. *Food Chemistry, 127*(1), 275–281. https://doi.org/10.1016/j.foodchem.2010.12.110

Zhang, J., Wang, X., Lu, Y., Bhusal, S. J., Song, Q., Cregan, P. B., ... & Jiang, G. L. (2018a). Genome-wide scan for seed composition provides insights into soybean quality improvement and the impacts of domestication and breeding. *Molecular plant*, *11*(3), 460-472.

Zhang, K., Liu, S., Li, W., Liu, S., Li, X., Fang, Y., ... & Ning, H. (2018b). Identification of QTNs controlling seed protein content in soybean using multi-locus genome-wide association studies. *Frontiers in plant science*, *9*, 1690.

Zhang, Z. (2020). User Manual for GAPIT. Genomic Association and Prediction Integrated Tool(Version 3). Retrieved from https://zzlab.net/GAPIT/gapit_help_document.pdf

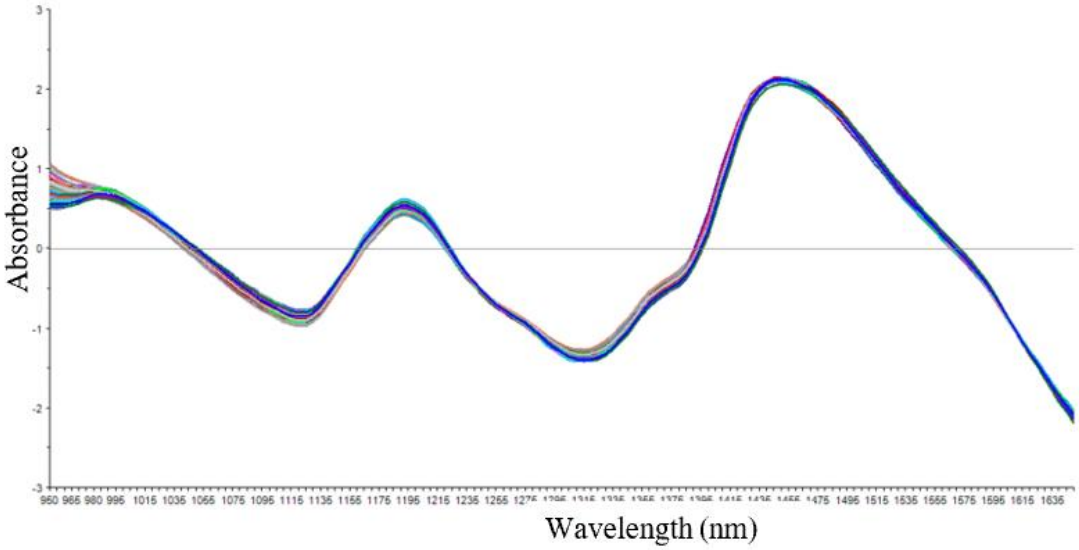
Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., ... & Nordborg, M. (2007). An Arabidopsis example of association mapping in structured samples. *PLoS genetics*, *3*(1), e4.

Zhu, C., Gore, M., Buckler, E. S., & Yu, J. (2008). Status and prospects of association mapping in plants. *The plant genome*, *1*(1).

Zhu, Z., Chen, S., Wu, X., Xing, C., & Yuan, J. (2018). Determination of soybean routine quality parameters using near-infrared spectroscopy. *Food science & nutrition*, *6*(4), 1109-1118.

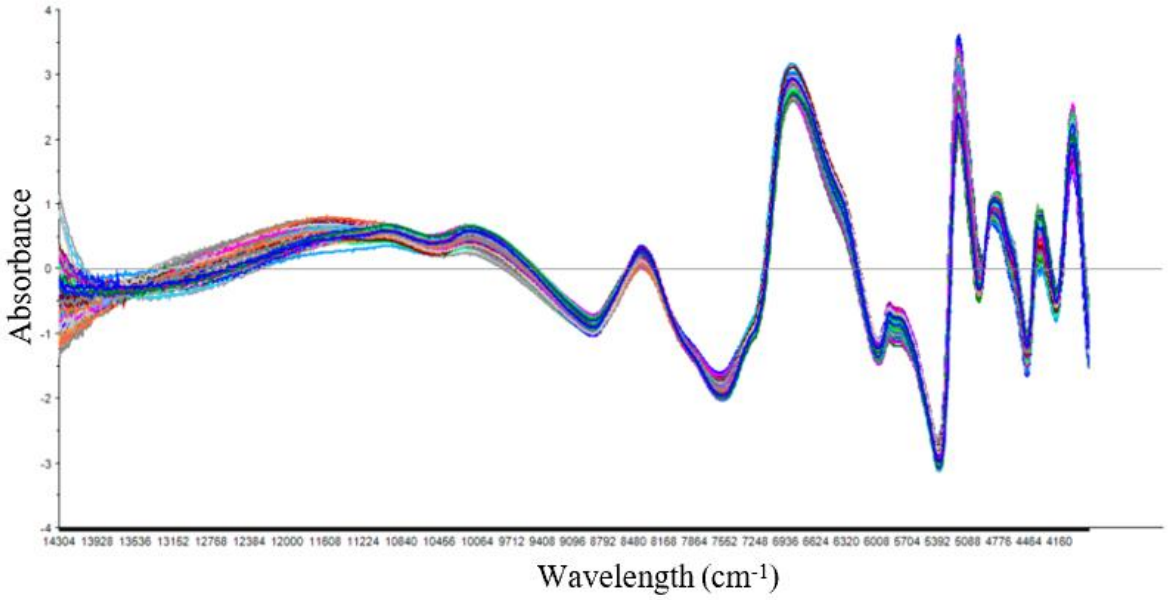
APPENDICES

Appendix A



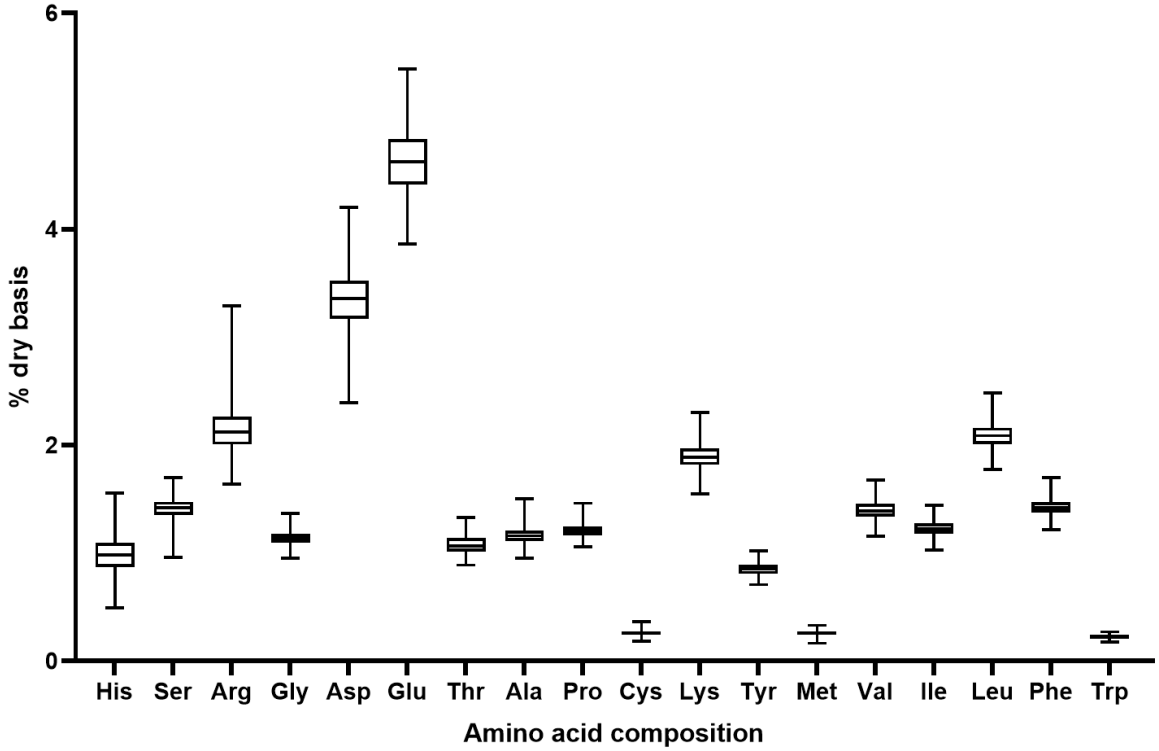
De-trending and SNV pretreated spectra of whole lentil samples from DA 7250 (950 -1650nm)

Appendix B



De-trending and SNV pretreated spectra of ground lentil samples from FT 9700 (14304 cm^{-1}(699nm) and 3856 cm^{-1}(2593nm))

Appendix C



Phenotypic distribution of 18 amino acids (% dry basis) in 361 lentil samples

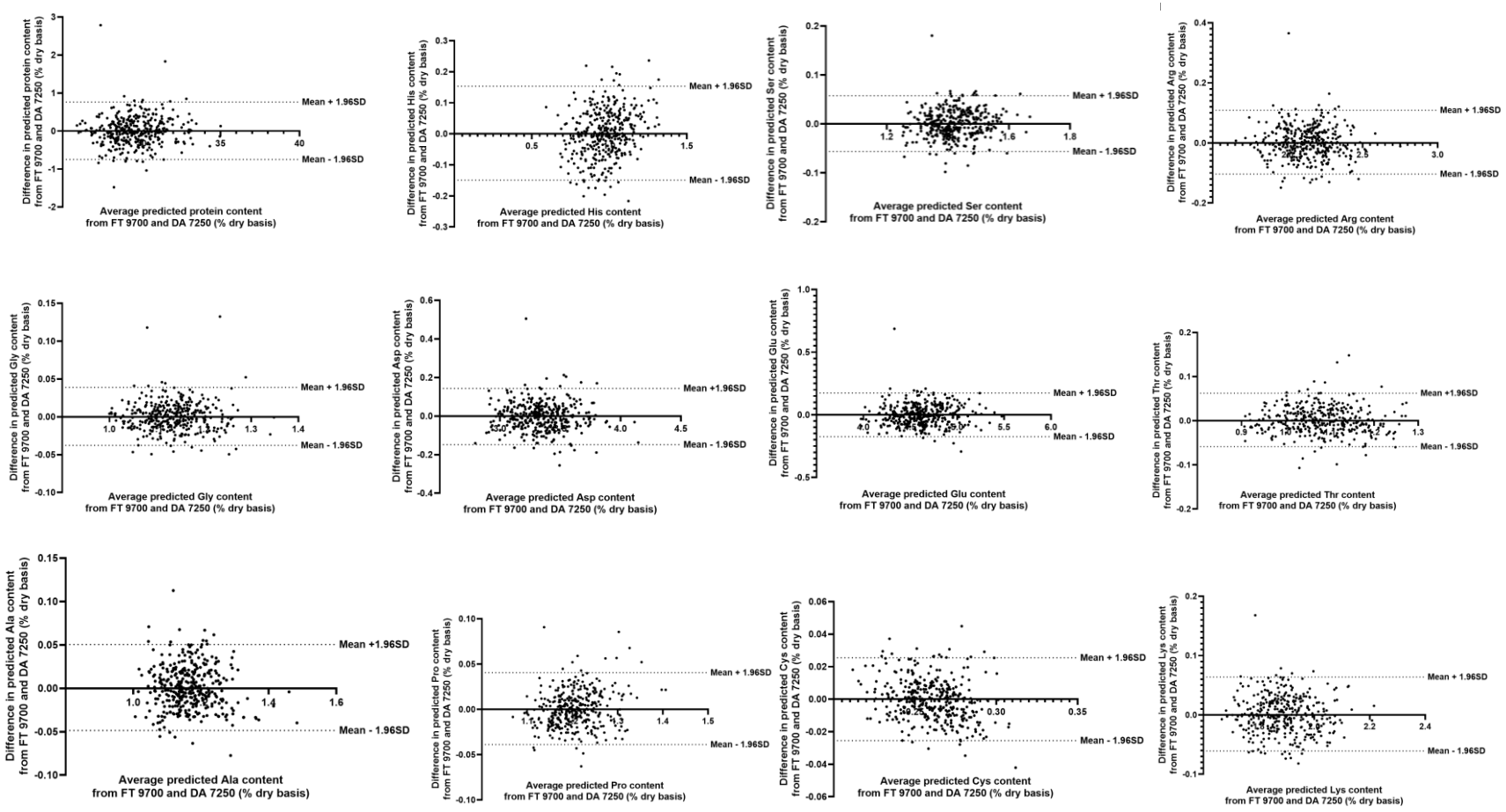
Appendix D

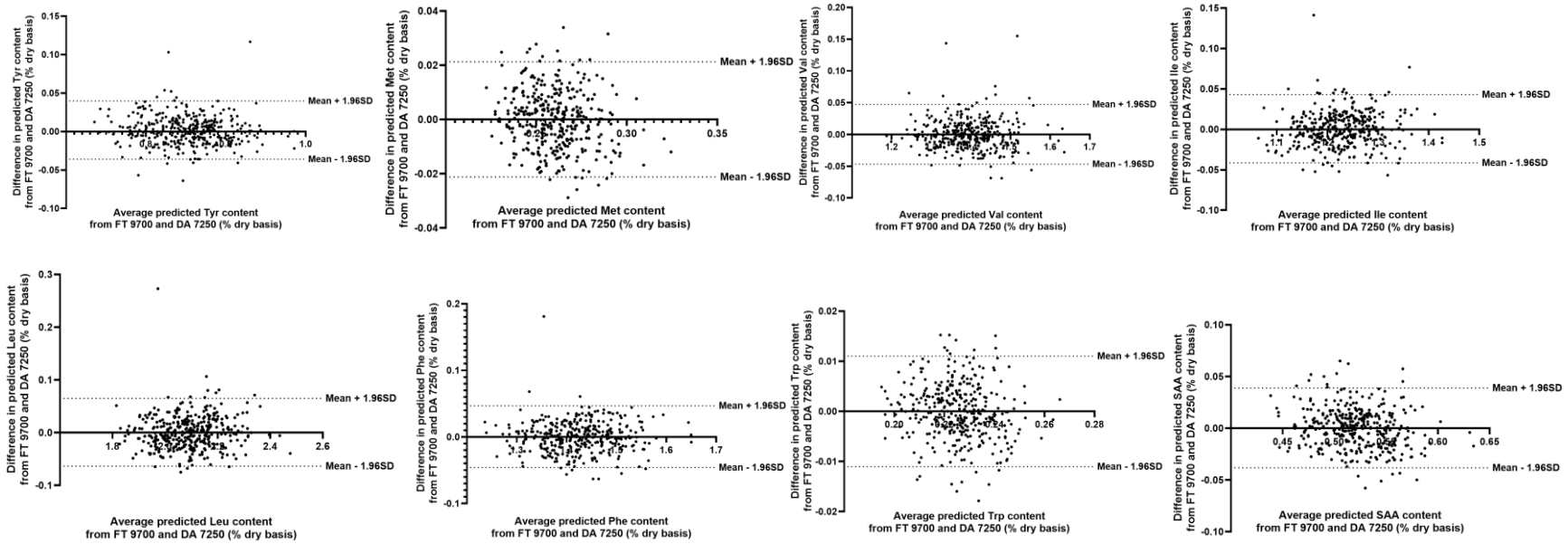
Guidelines for the interpretation of r and R^2 proposed from Williams & Norrris (2001) [a]

r	R^2	Interpretation
0-0.50	0-0.25	Unusable models
0.51-0.70	0.26-0.49	Poor correlation models
0.71-0.80	0.50-0.64	Acceptable for very rough to rough screening
0.81-0.90	0.66-0.81	Acceptable for sample screening
0.91-0.95	0.83-0.90	Usable with caution for most applications, including research
0.96-0.98	0.92-0.96	Usable in most applications, including quality assurance
0.99+	0.98+	Excellent, usable in any application

[a] No R^2 values of 0.65, 0.82, 0.91 and 0.97 due to rounding off.

Appendix E


Difference in predicted protein content from FT 9700 and DA 7250 (% dry basis)
Mean + 1.96SD
Mean - 1.96SD
Average predicted protein content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted His content from FT 9700 and DA 7250 (% dry basis)
Mean + 1.96SD
Mean - 1.96SD
Average predicted His content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Ser content from FT 9700 and DA 7250 (% dry basis)
Mean + 1.96SD
Mean - 1.96SD
Average predicted Ser content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Arg content from FT 9700 and DA 7250 (% dry basis)
Mean + 1.96SD
Mean - 1.96SD
Average predicted Arg content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Gly content from FT 9700 and DA 7250 (% dry basis)
Mean + 1.96SD
Mean - 1.96SD
Average predicted Gly content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Asp content from FT 9700 and DA 7250 (% dry basis)
Mean +1.96SD
Mean - 1.96SD
Average predicted Asp content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Glu content from FT 9700 and DA 7250 (% dry basis)
Mean + 1.96SD
Mean - 1.96SD
Average predicted Glu content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Thr content from FT 9700 and DA 7250 (% dry basis)
Mean +1.96SD
Mean - 1.96SD
Average predicted Thr content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Ala content from FT 9700 and DA 7250 (% dry basis)
Mean +1.96SD
Mean - 1.96SD
Average predicted Ala content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Pro content from FT 9700 and DA 7250 (% dry basis)
Mean + 1.96SD
Mean - 1.96SD
Average predicted Pro content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Cys content from FT 9700 and DA 7250 (% dry basis)
Mean + 1.96SD
Mean - 1.96SD
Average predicted Cys content from FT 9700 and DA 7250 (% dry basis)
Difference in predicted Lys content from FT 9700 and DA 7250 (% dry basis)
Mean + 1.96SD
Mean - 1.96SD
Average predicted Lys content from FT 9700 and DA 7250 (% dry basis)



Bland-Altman for protein, 18 amino acids and SAA predicted from FT 9700 and DA 7250 spectrometers.

Appendix F

R-scripts used for GWAS by using Blink model.

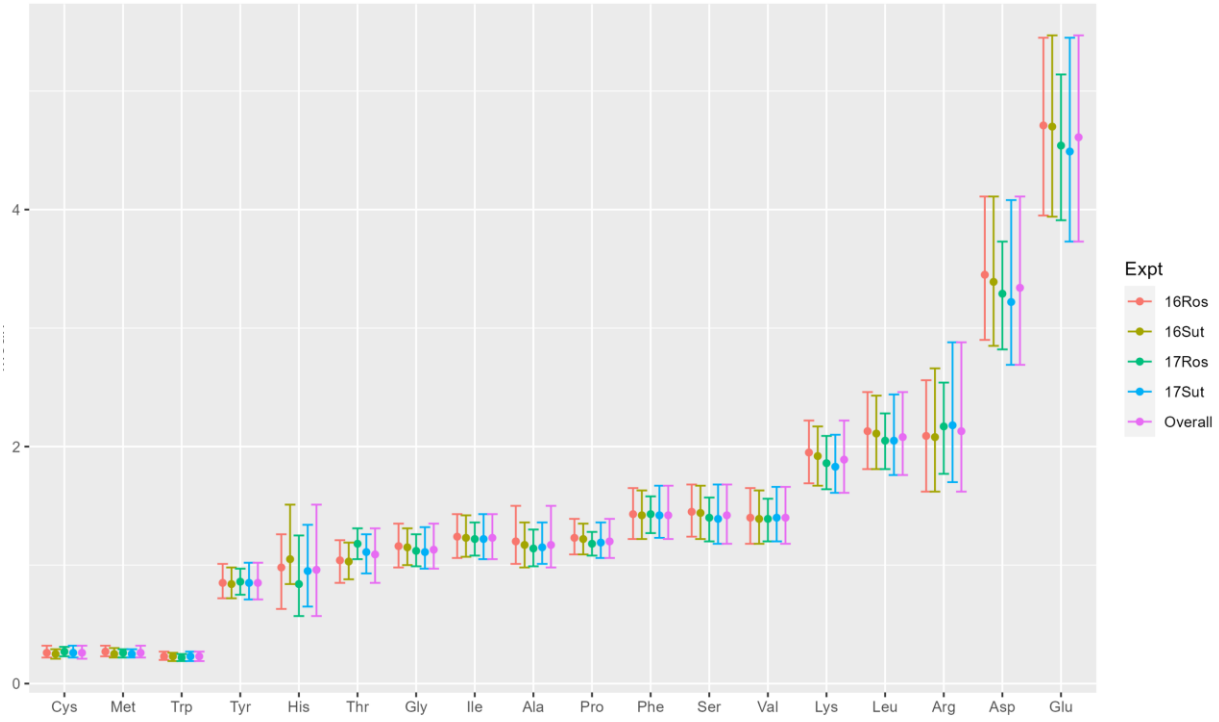
```
source("http://zzlab.net/GAPIT/GAPIT.library.R")
source("http://zzlab.net/GAPIT/gapit_functions.txt")
setwd("~/R")
myY <- read.csv("myY.csv", header = TRUE)
myG <- read.csv("myG.hmp.csv",header = F)
myGAPIT <- GAPIT(
  Y=myY,
  G=myG,
  PCA.total=4,
  model="Blink"
)
```

R-scripts used for generating SNP density plot.

```
install.packages("CMplot")
library("CMplot")
install.packages("tidyverse")
library("CMplot")
library("tidyverse")
setwd("~/R")
myG <- read.csv("myG.hmp.csv")
data(lens)
lens <- myG %>% select(rs, chrom, pos)
```

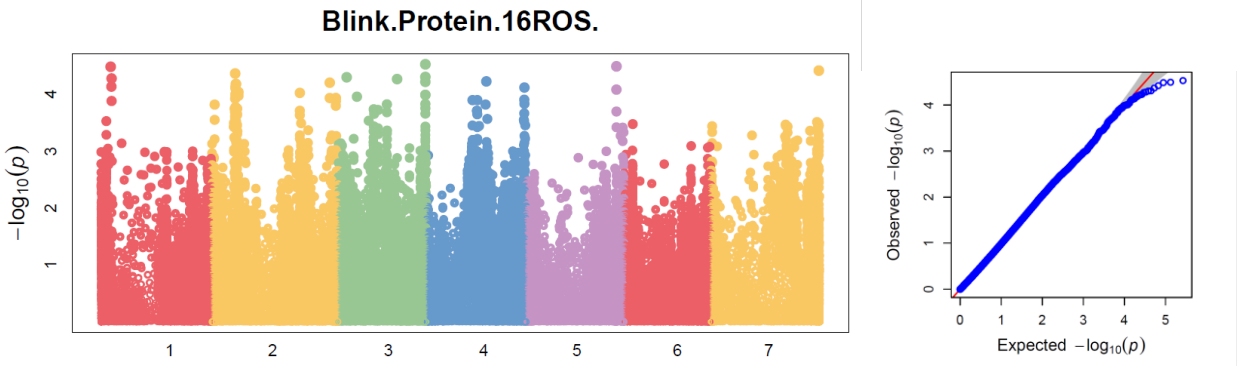
```
CMplot(lens, plot.type="d", bin.size=1e7, chr.den.col=c("darkgreen", "yellow", "red"),
    file="jpg", file.output=T, verbose=F, width=9, height=6)
```

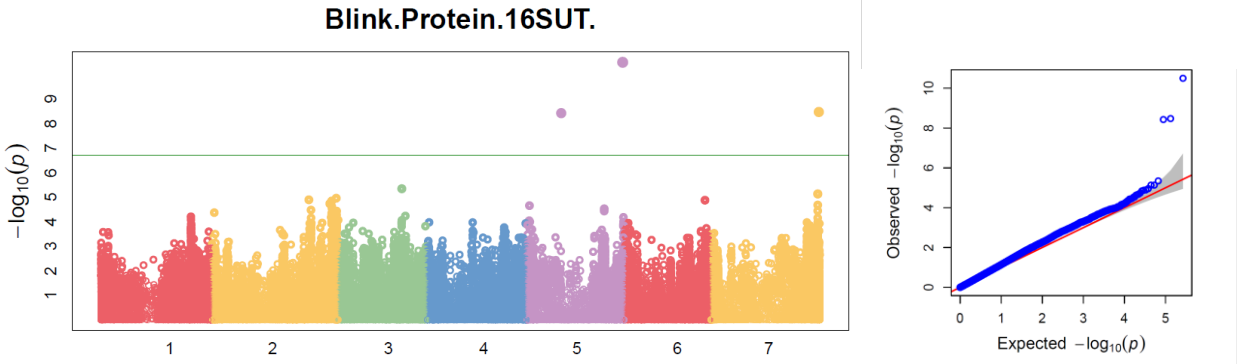
Appendix G

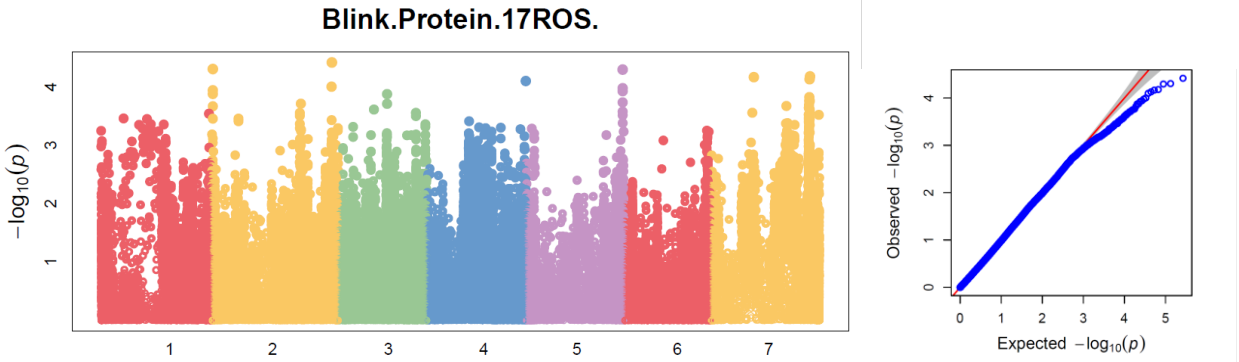


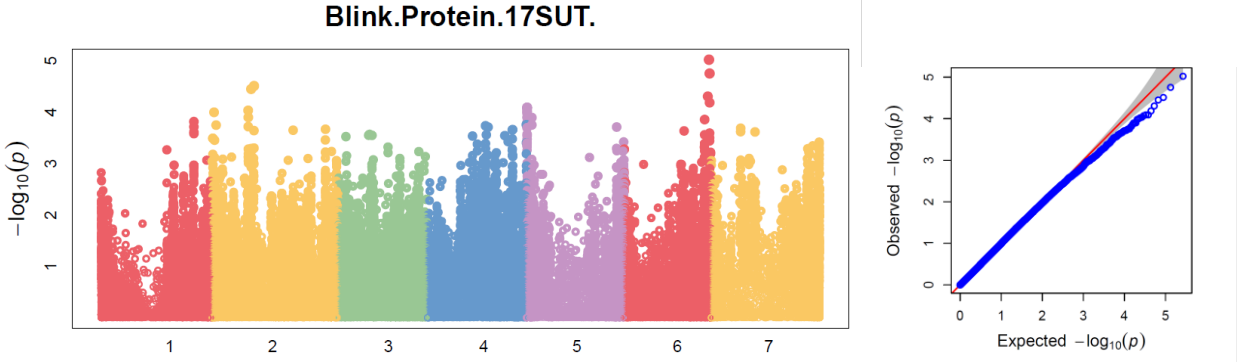
Phenotypic distribution of 18 amino acids (% dry basis) contents for the lentil diversity panel in four individual and overall environments

Appendix H

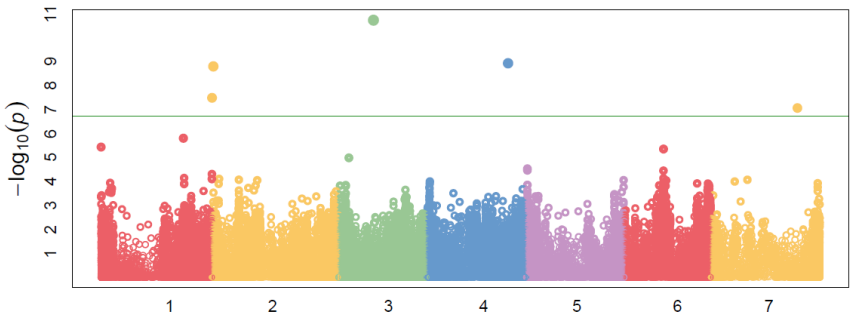

Blink.Protein.16ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)

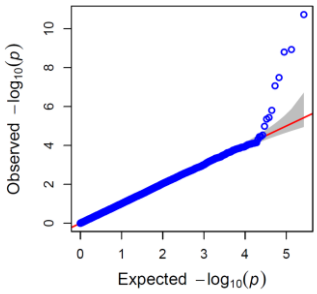

Blink.Protein.16SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)


Blink.Protein.17ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)

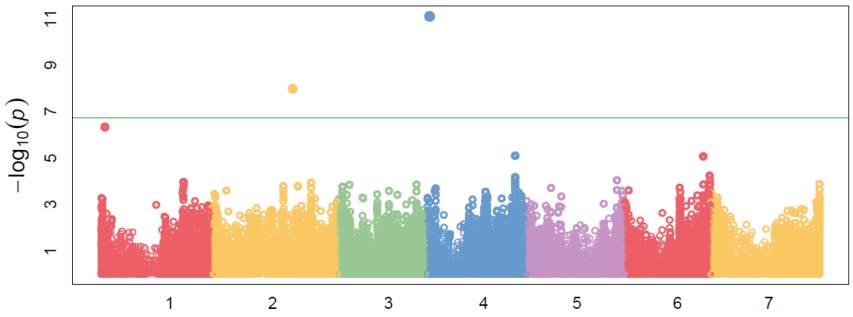

Blink.Protein.17SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)

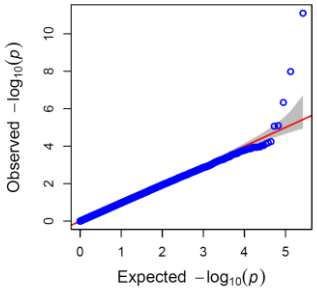
Blink.His.16ROS.



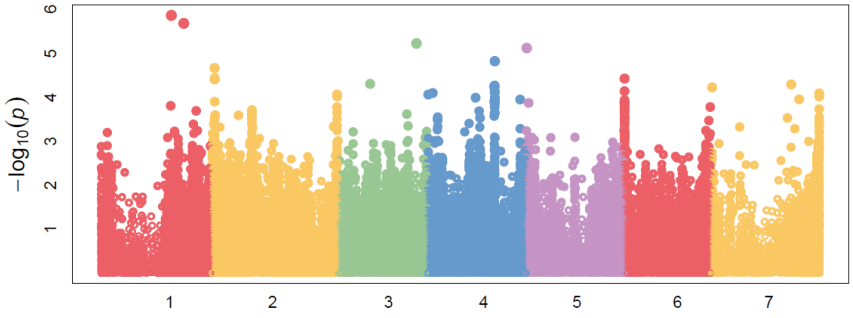


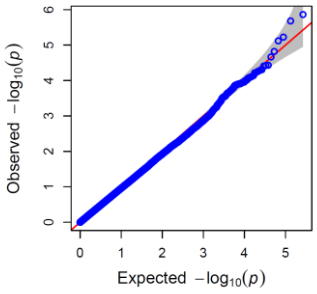
Blink.His.16SUT.



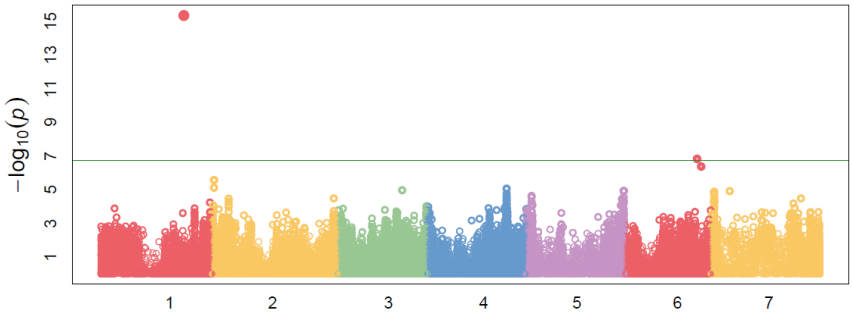


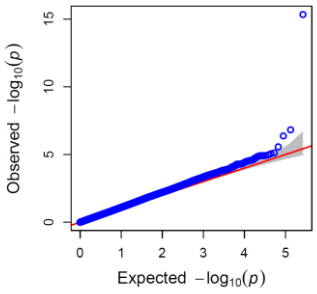
Blink.His.17ROS.



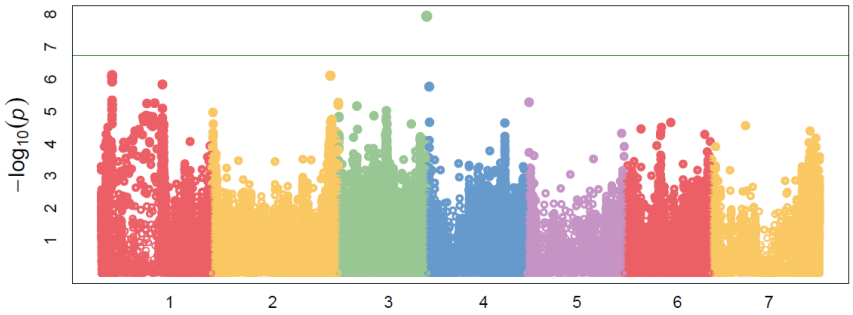


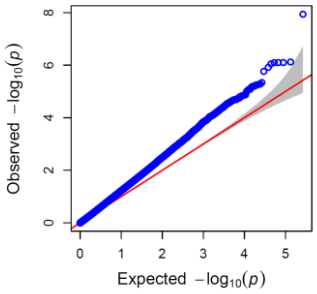
Blink.His.17SUT.



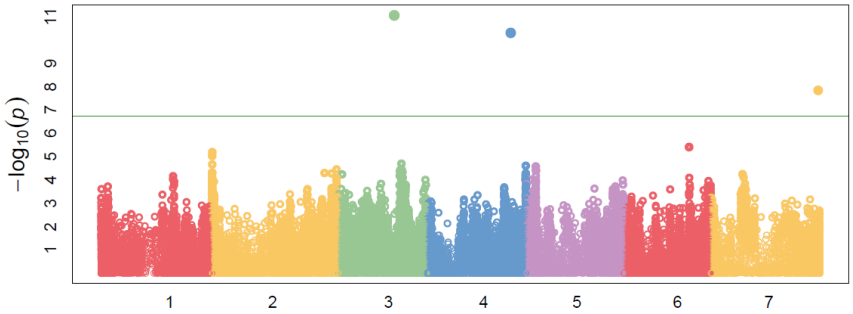


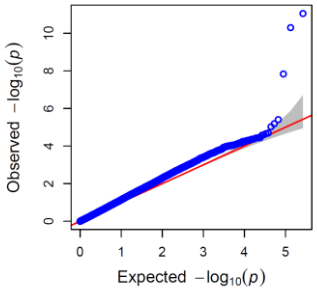
Blink.Thr.16ROS.



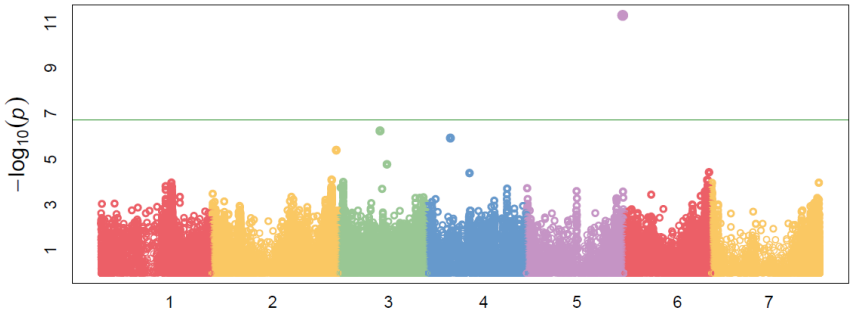


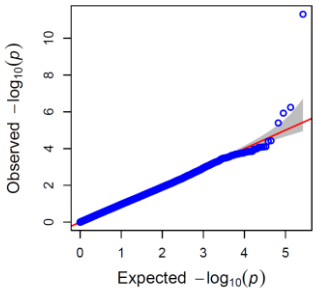
Blink.Thr.16SUT.



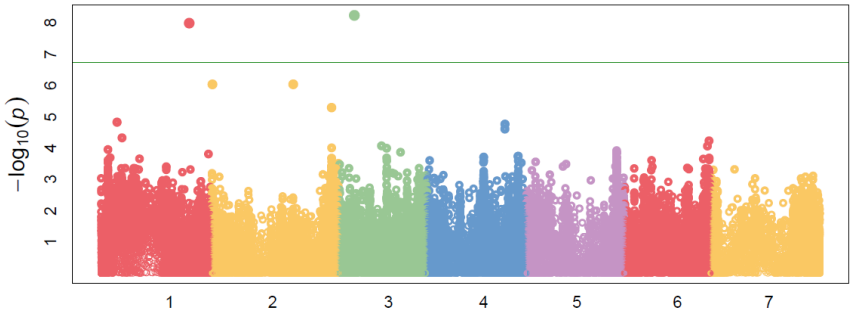


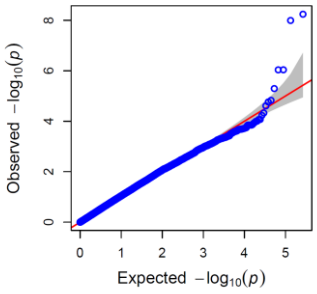
Blink.Thr.17ROS.



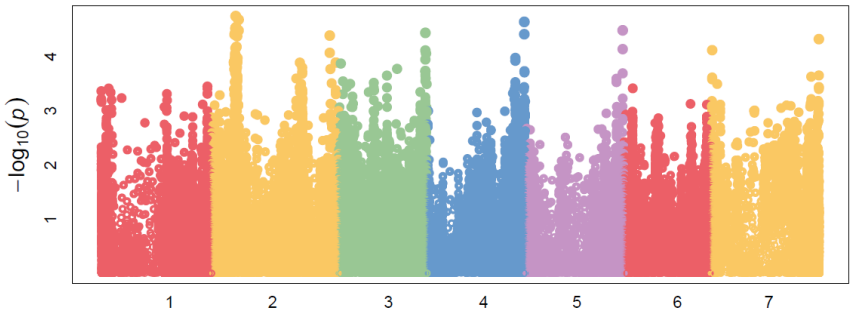


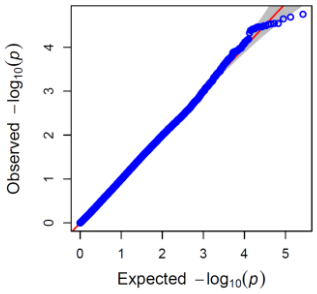
Blink.Thr.17SUT.



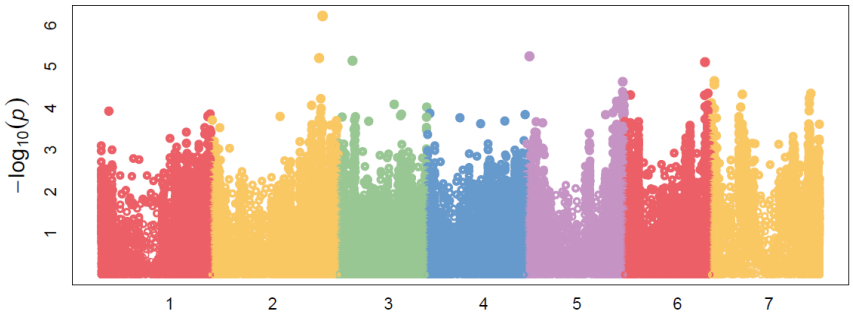


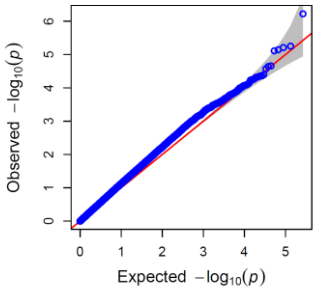
Blink.Lys.16ROS.



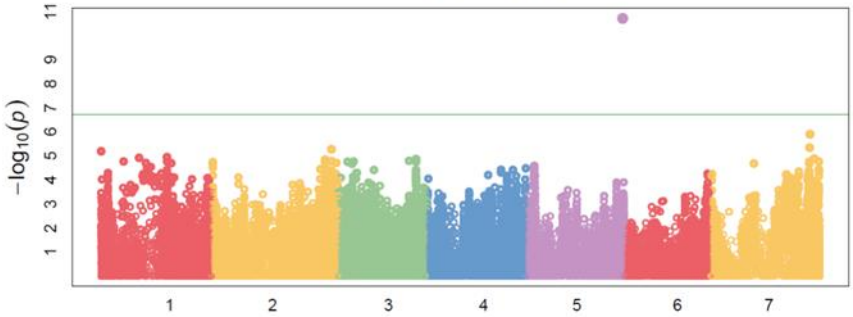


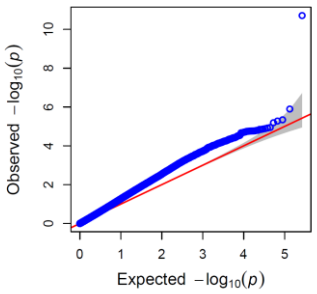
Blink.Lys.16SUT.



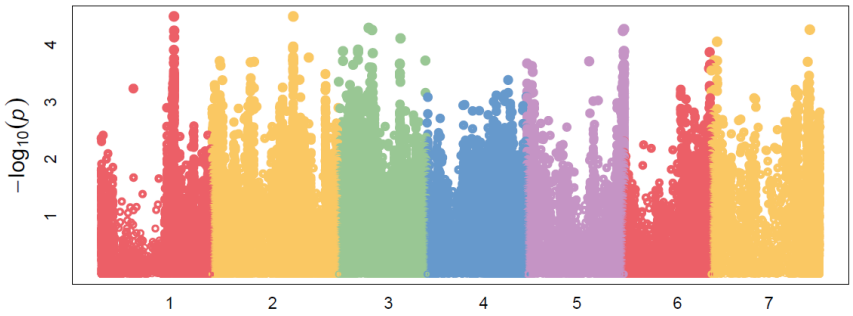


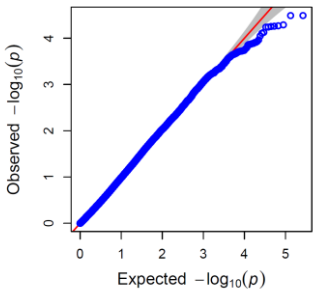
Blink.Lys.17ROS.

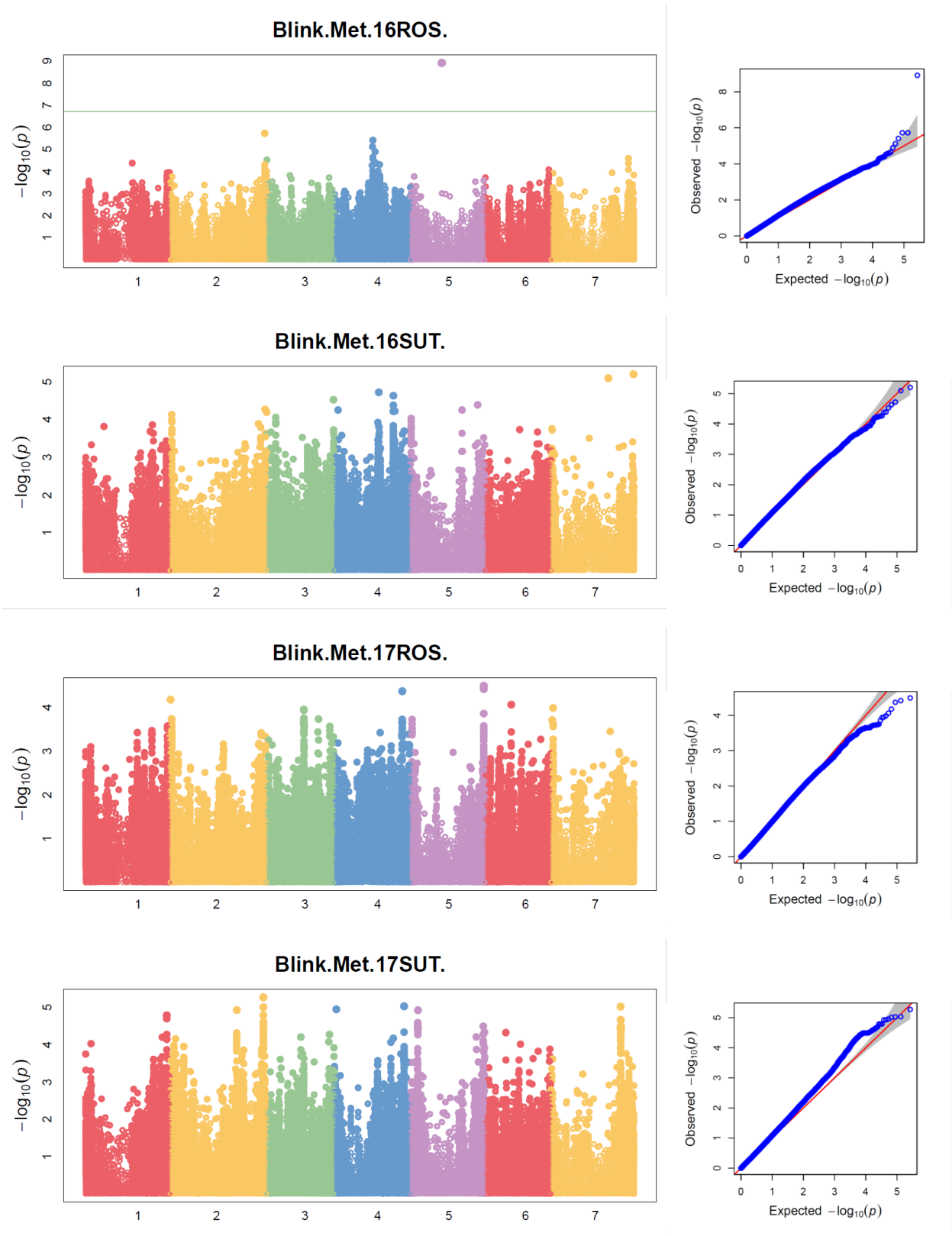




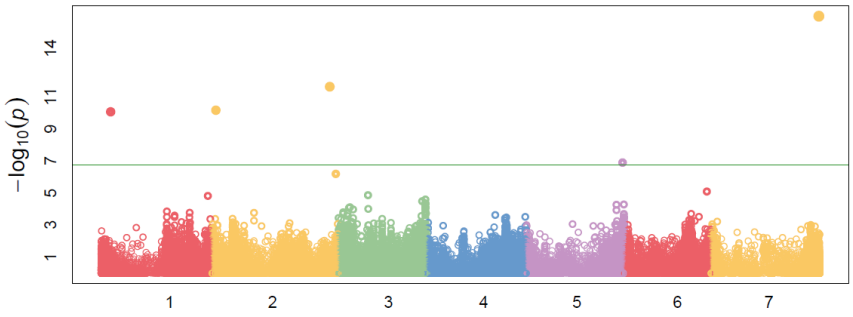
Blink.Lys.17SUT.

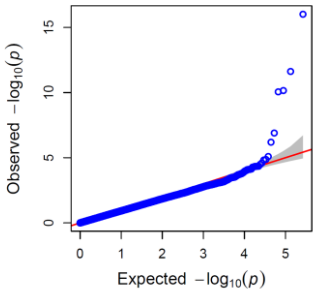




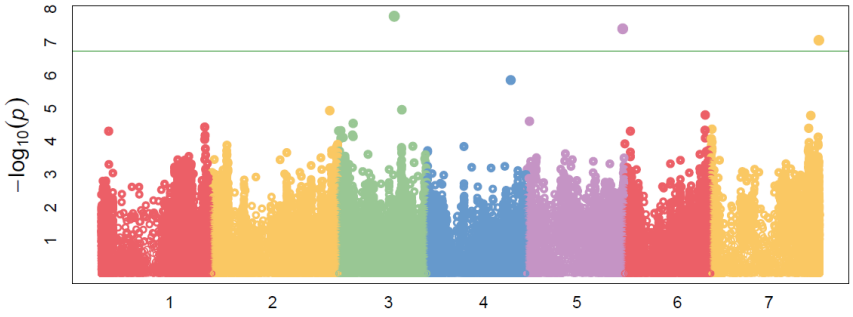

Blink.Met.16ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)
Blink.Met.16SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)
Blink.Met.17ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)
Blink.Met.17SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)

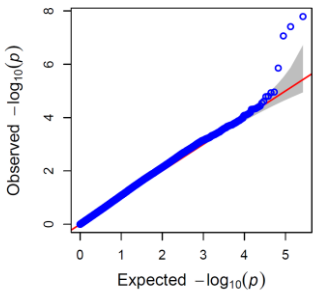
Blink.Val.16ROS.



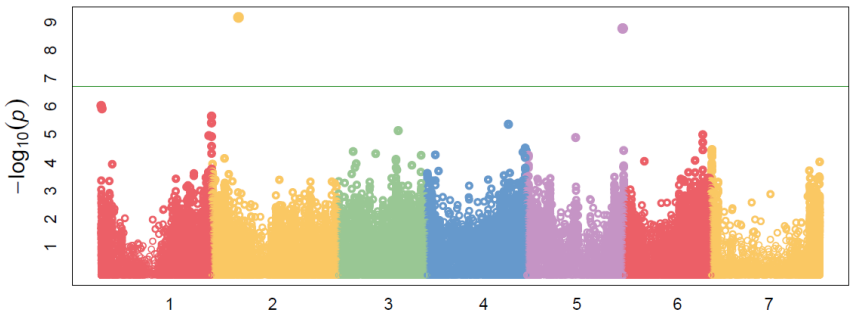


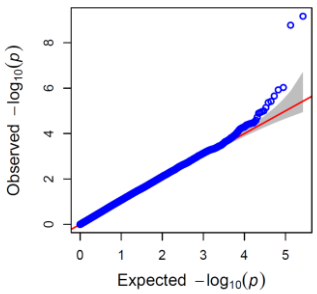
Blink.Val.16SUT.



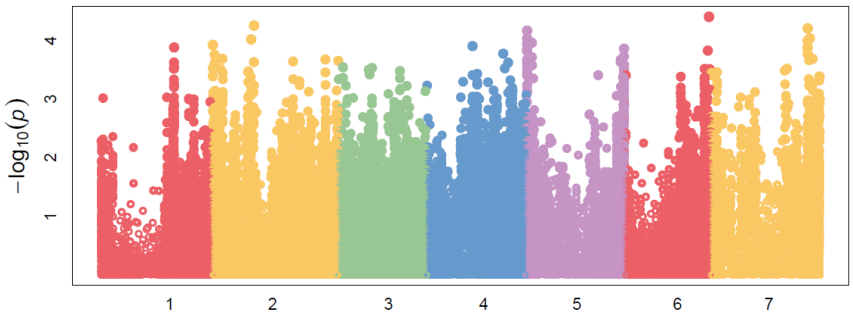


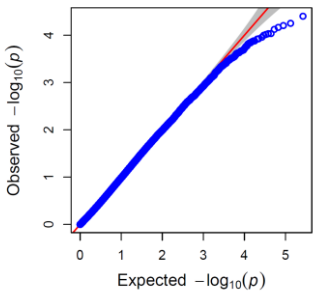
Blink.Val.17ROS.



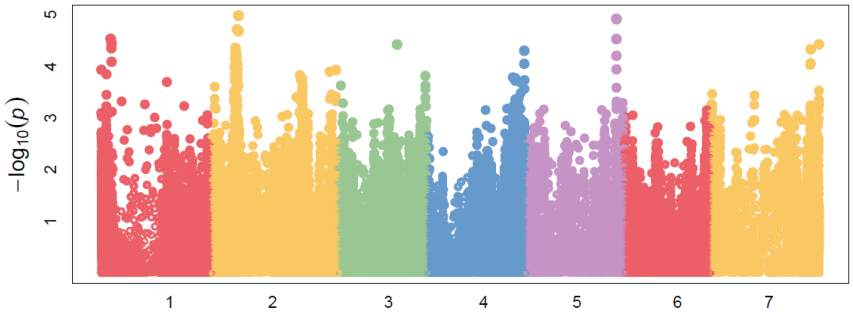


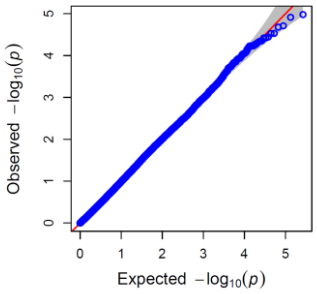
Blink.Val.17SUT.



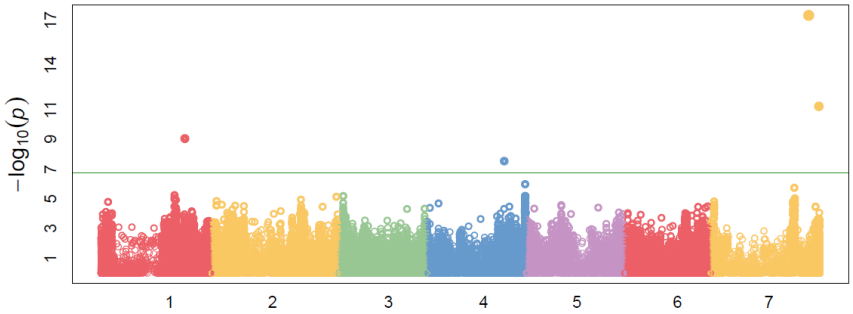


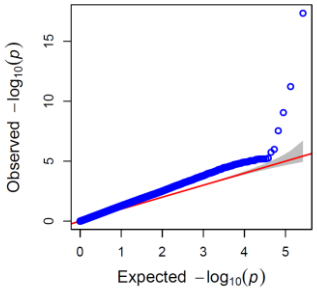
Blink.Ile.16ROS.



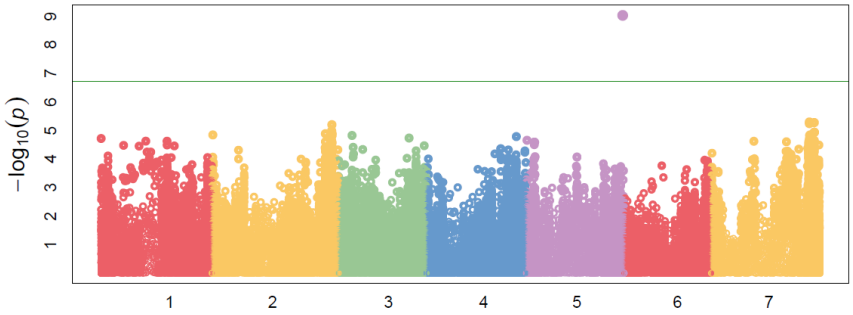


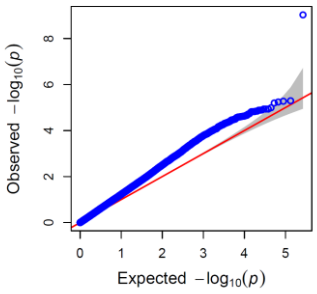
Blink.Ile.16SUT.



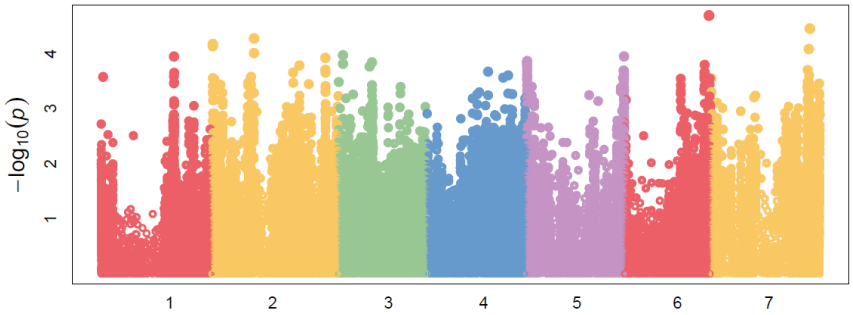


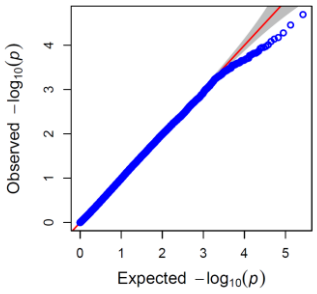
Blink.Ile.17ROS.

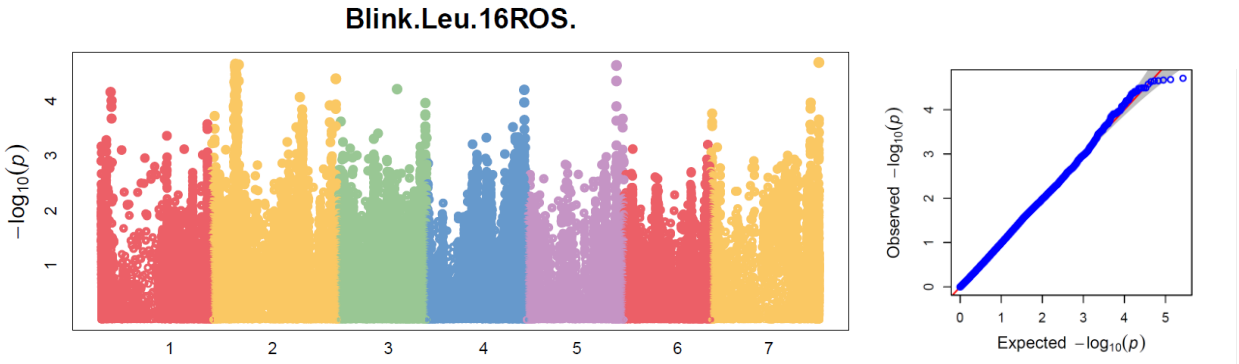


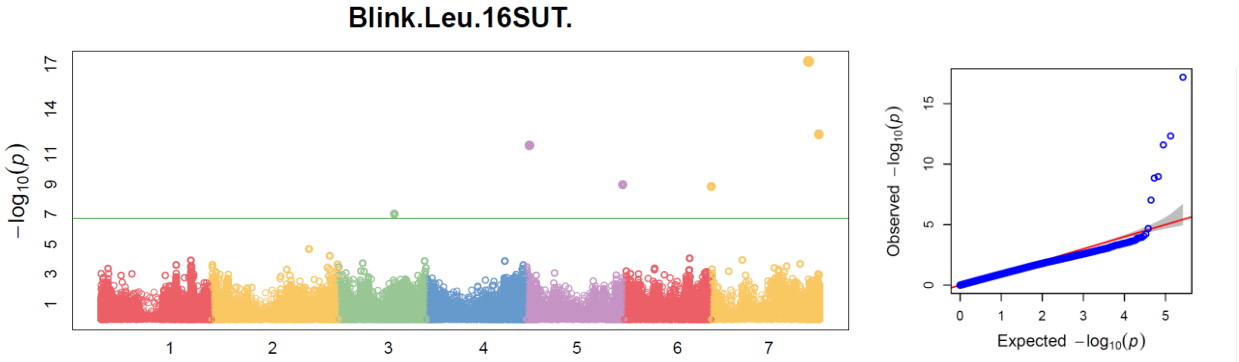


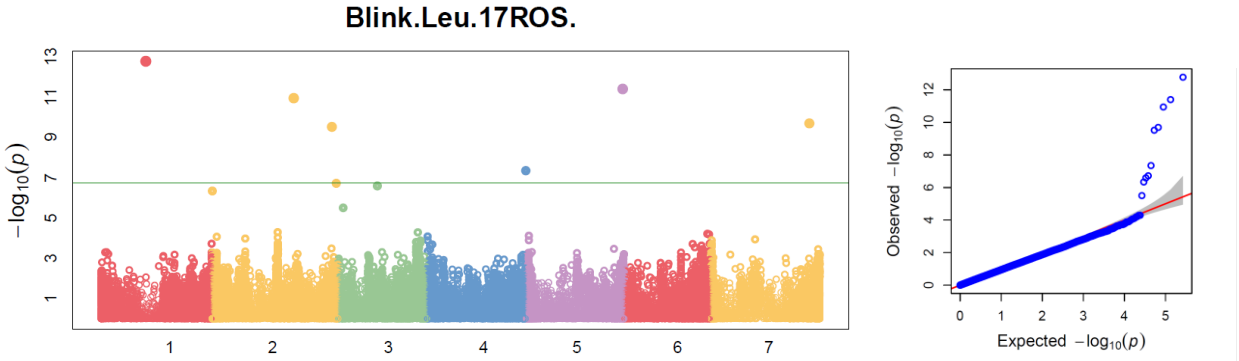
Blink.Ile.17SUT.

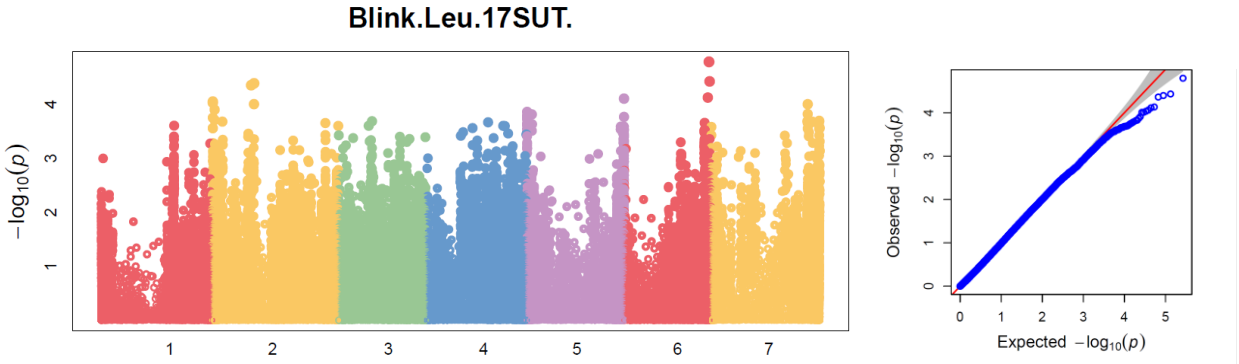




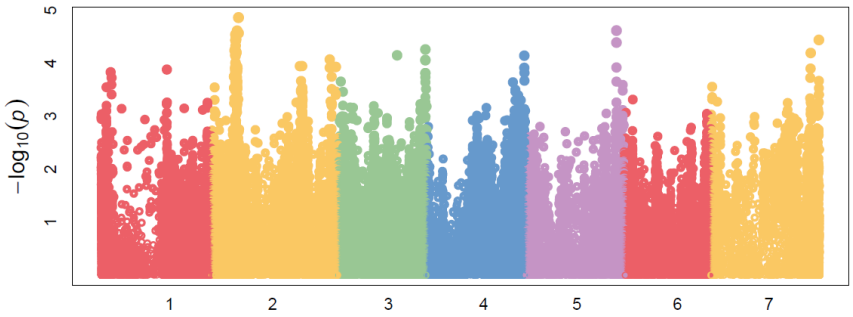

Blink.Leu.16ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)

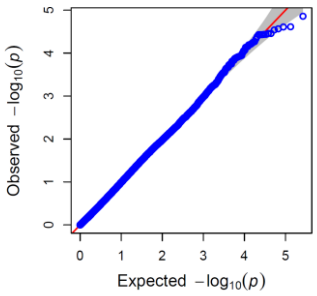

Blink.Leu.16SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)


Blink.Leu.17ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)

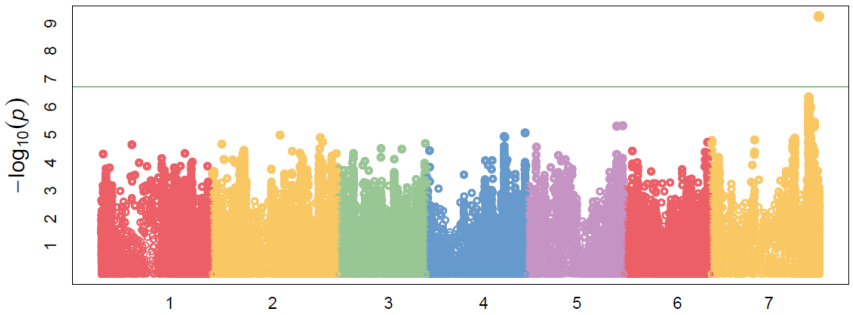

Blink.Leu.17SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)

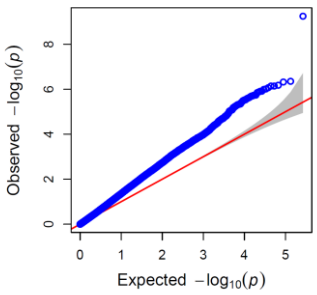
Blink.Phe.16ROS.



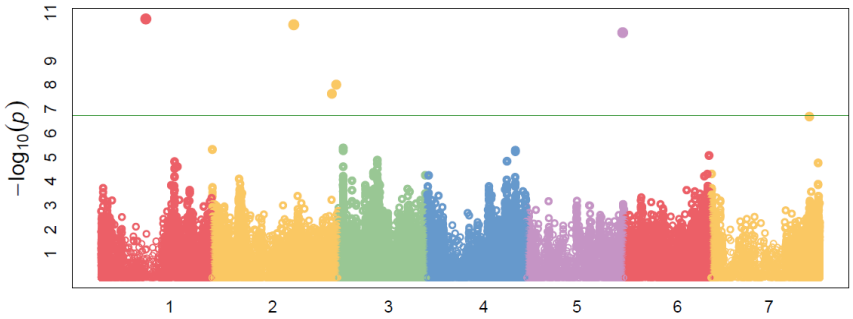


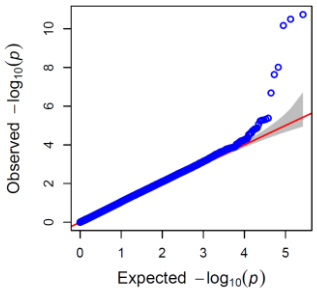
Blink.Phe.16SUT.



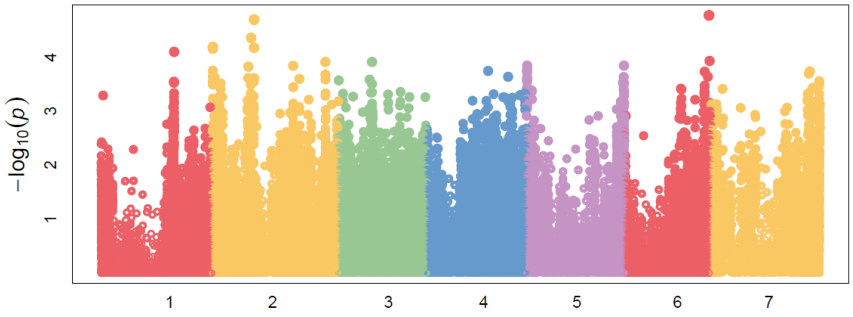


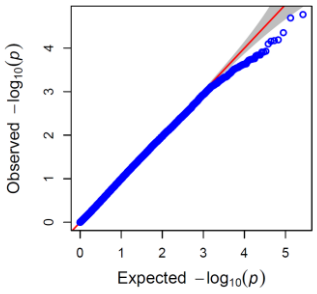
Blink.Phe.17ROS.

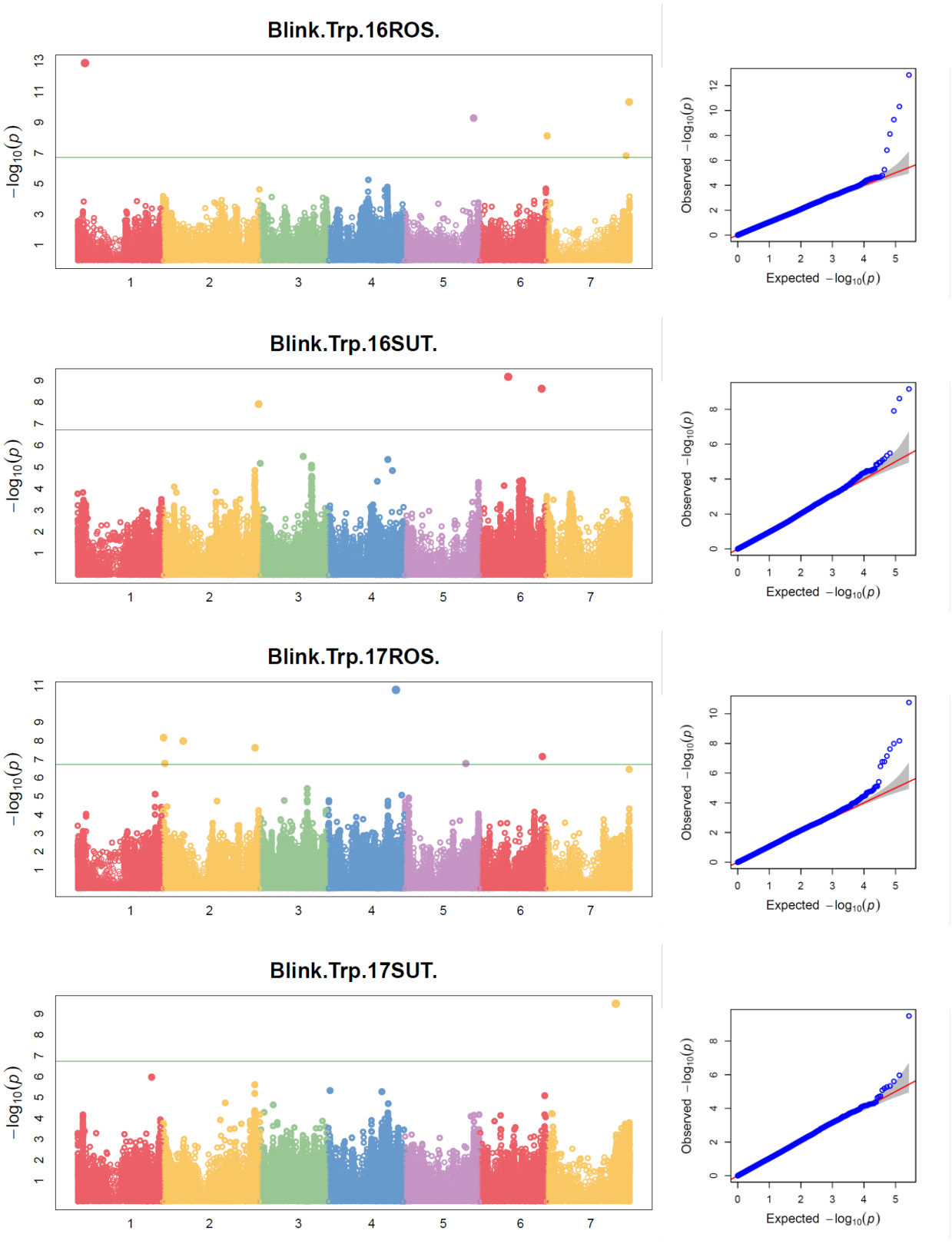


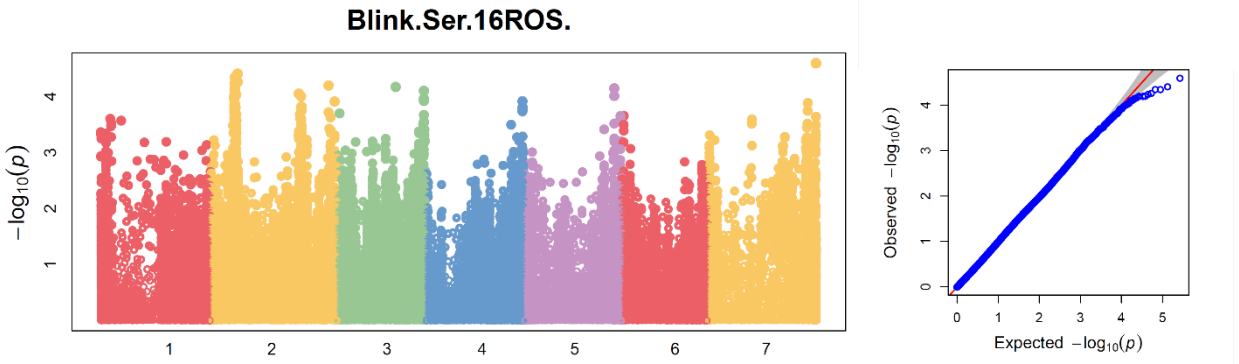


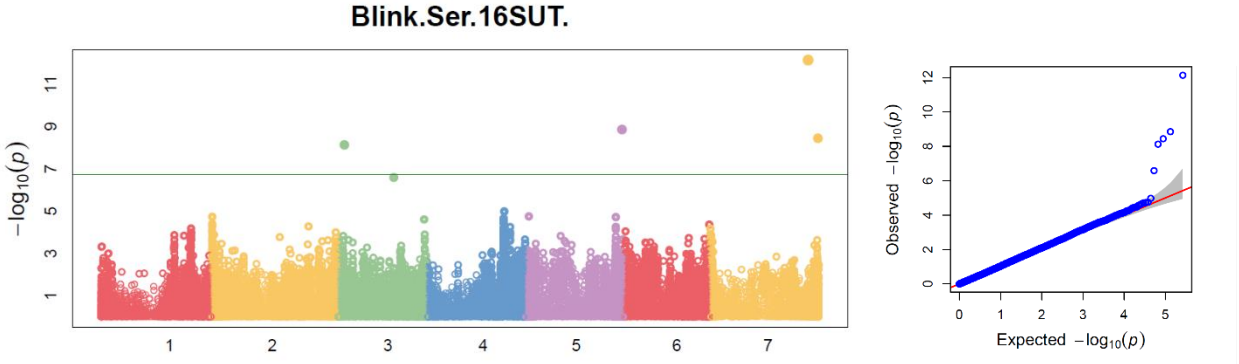
Blink.Phe.17SUT.

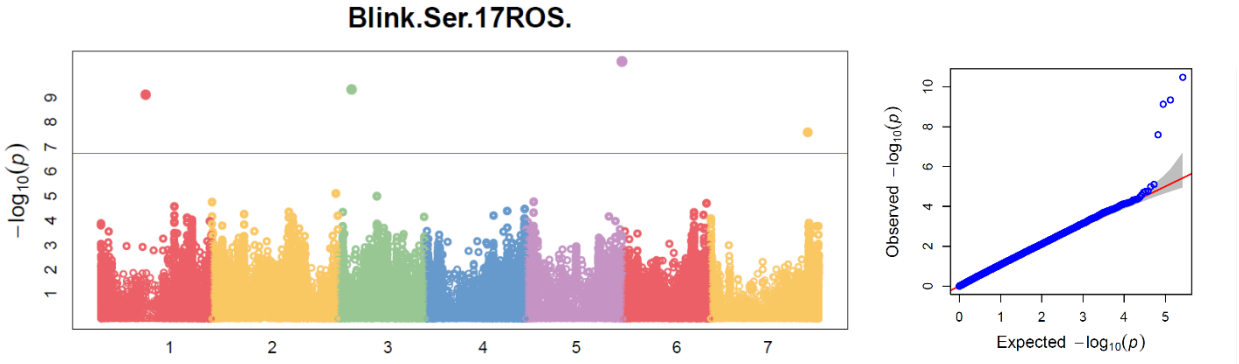


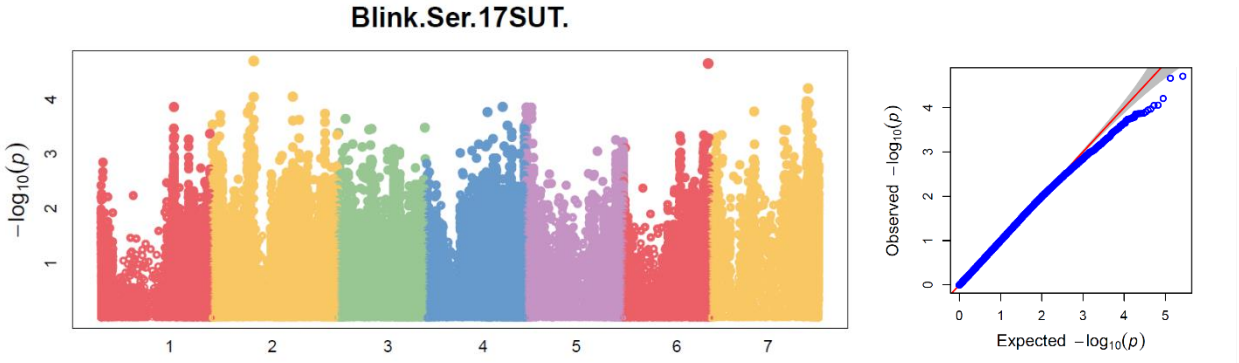


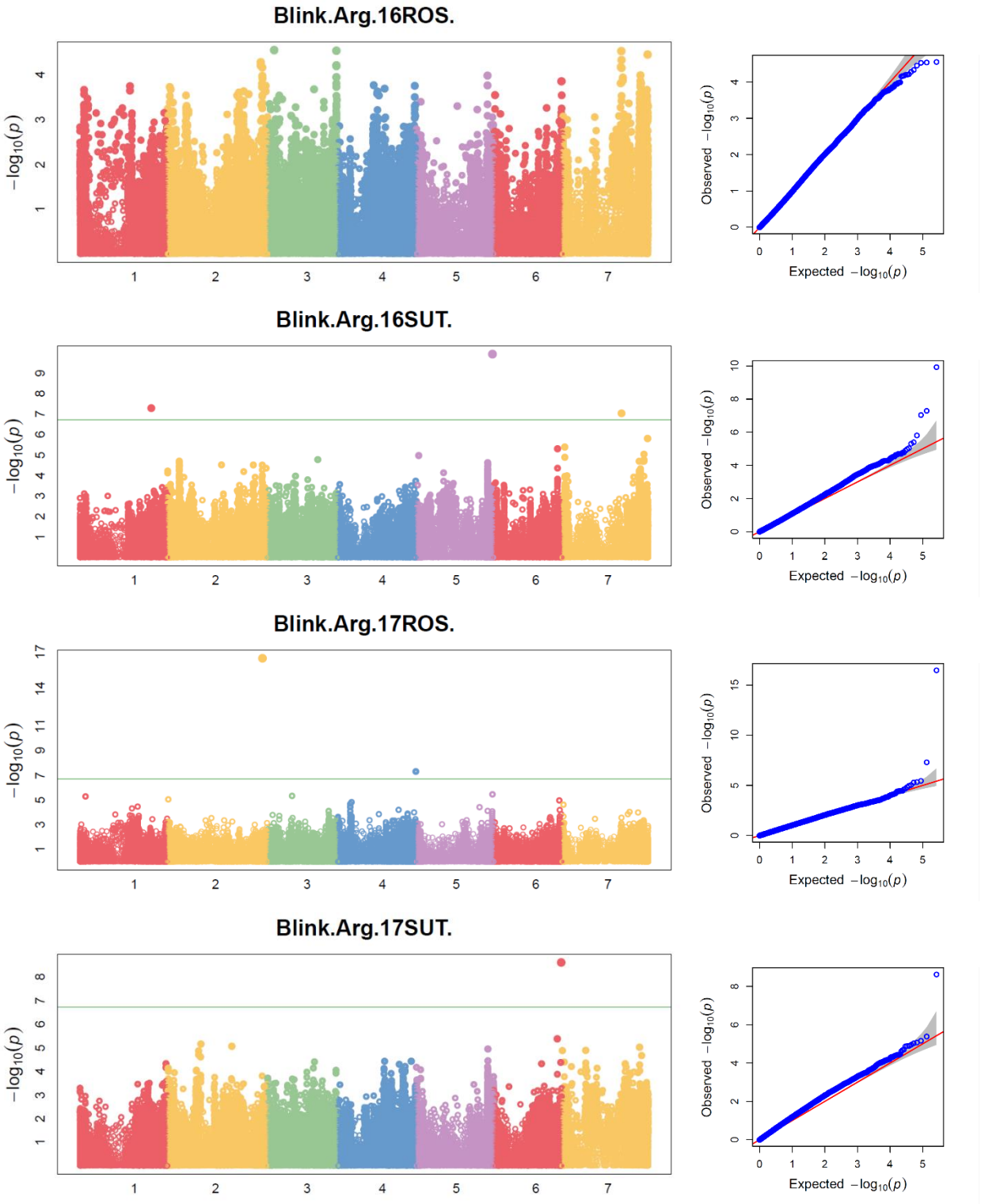

Blink.Trp.16ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)
Blink.Trp.16SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)
Blink.Trp.17ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)
Blink.Trp.17SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)

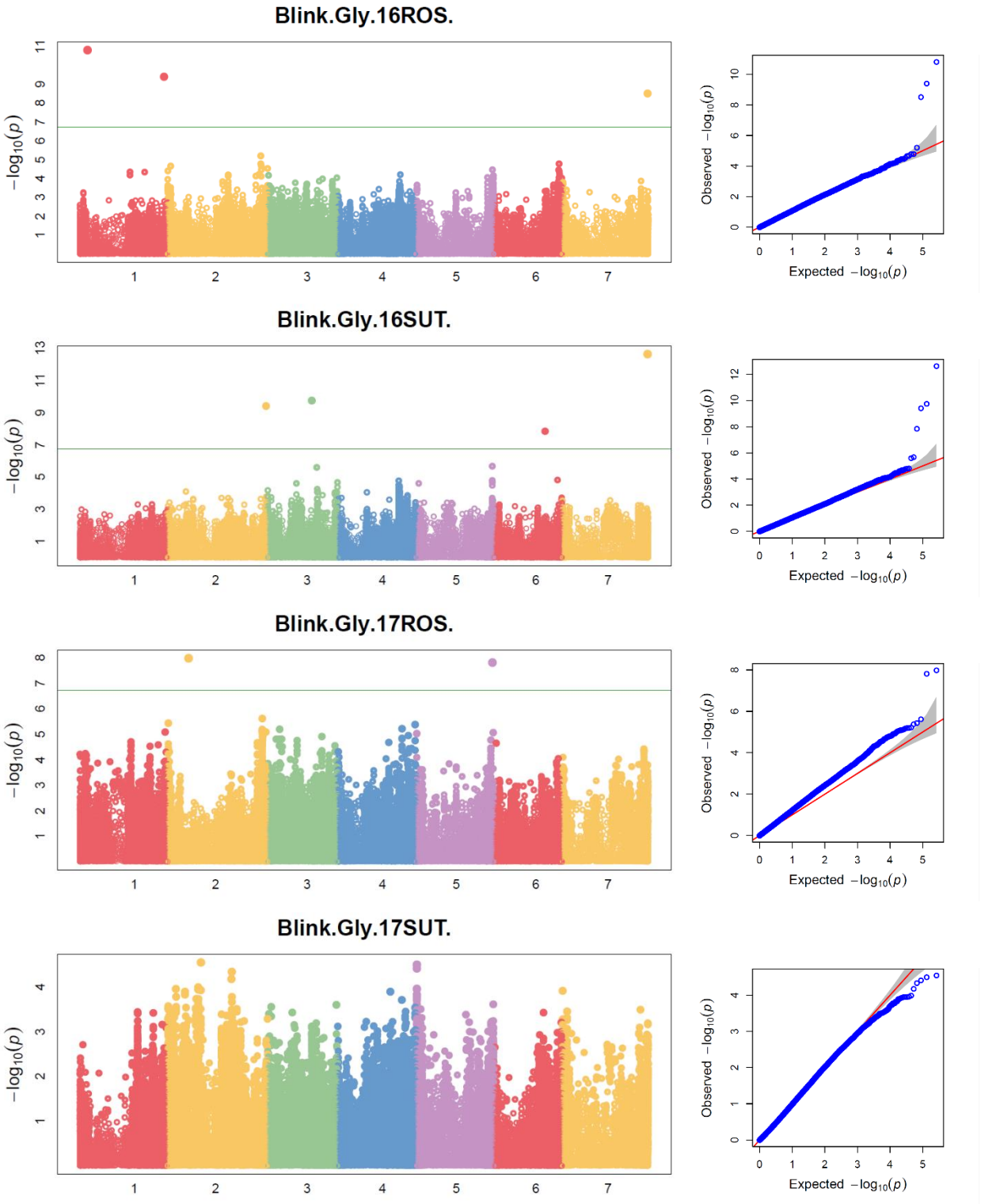

Blink.Ser.16ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)

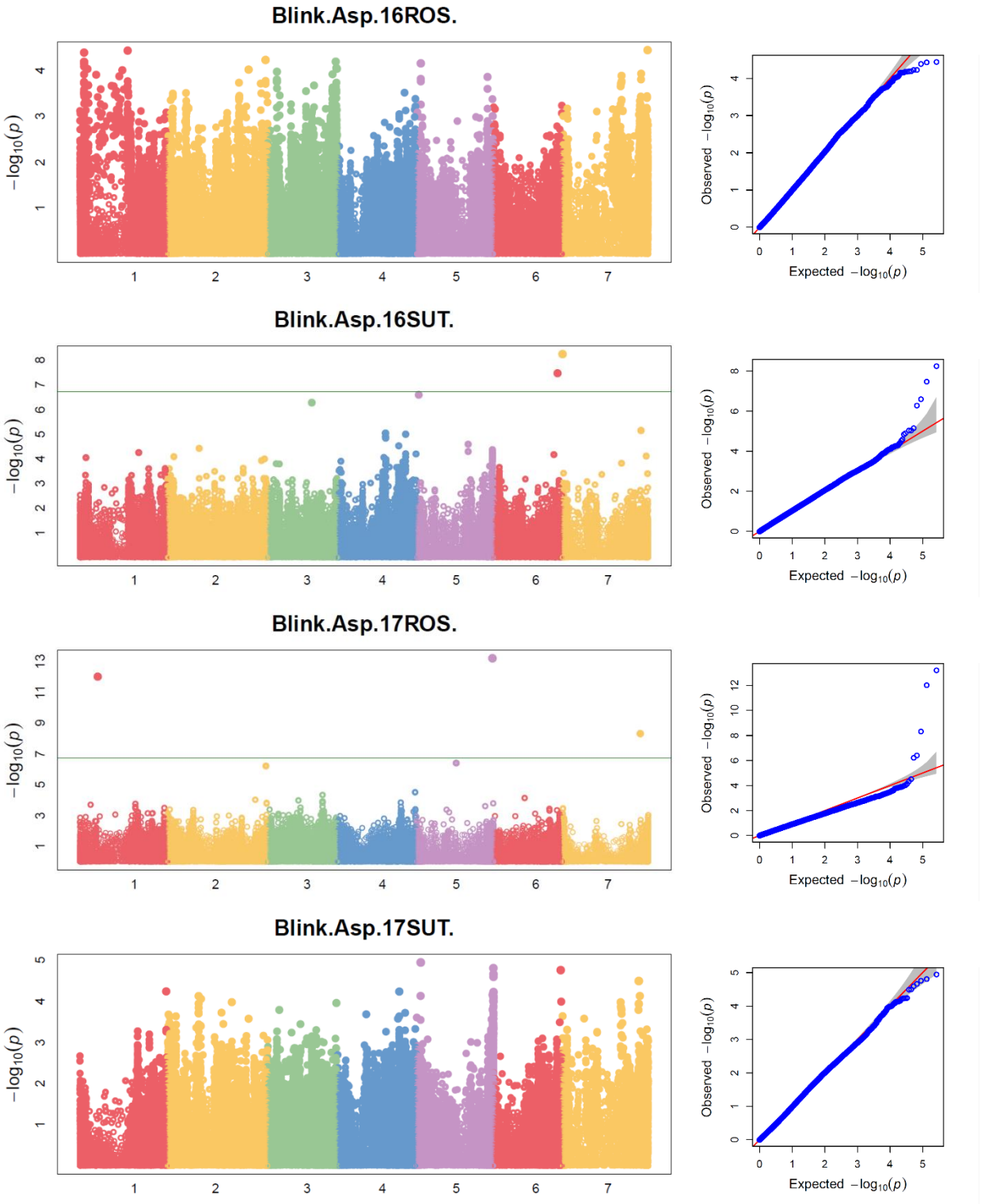

Blink.Ser.16SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)

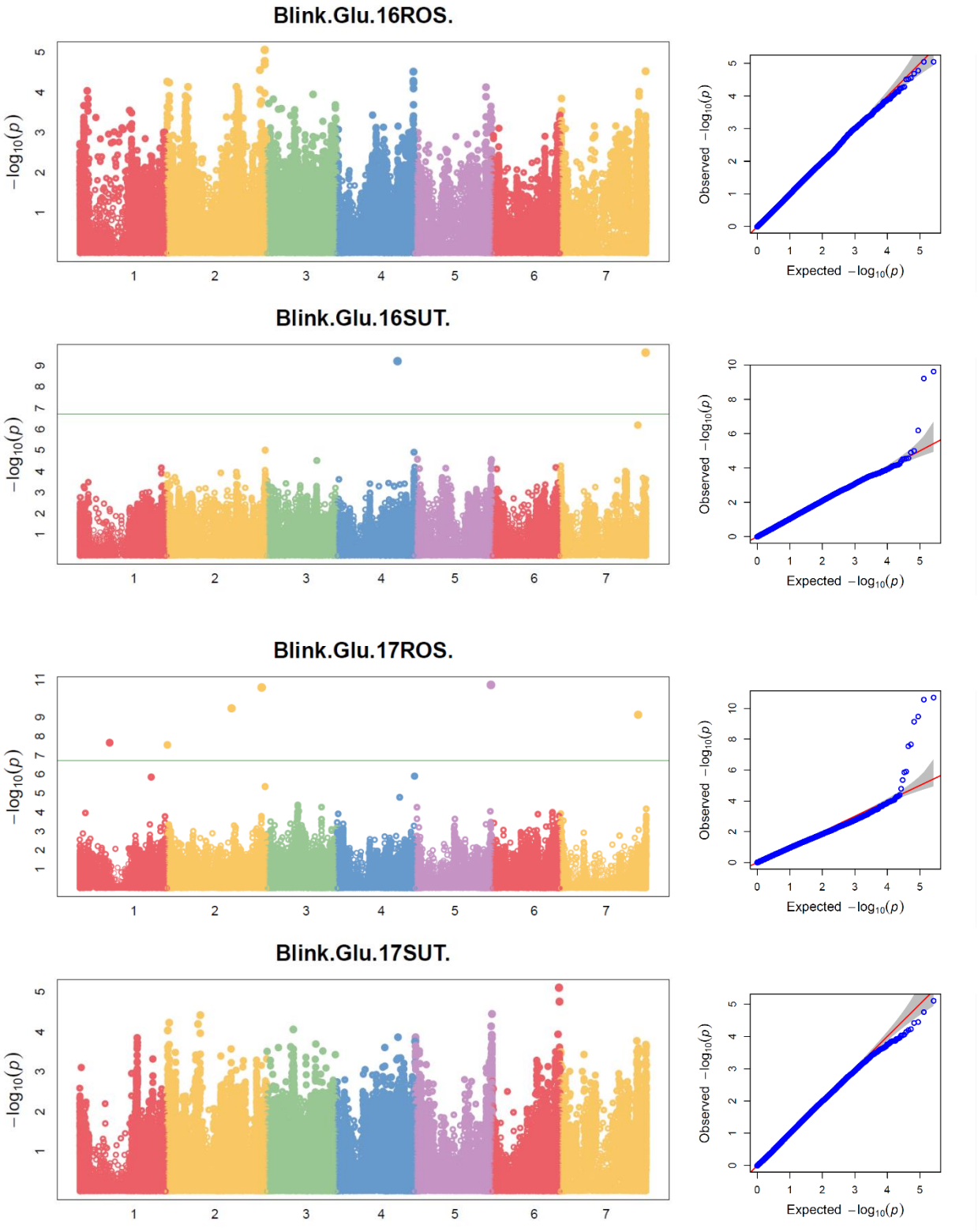

Blink.Ser.17ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)

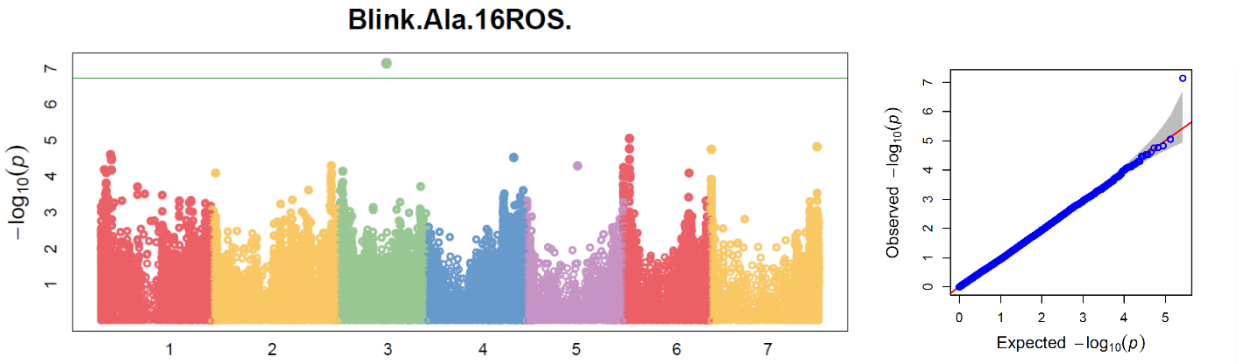

Blink.Ser.17SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)

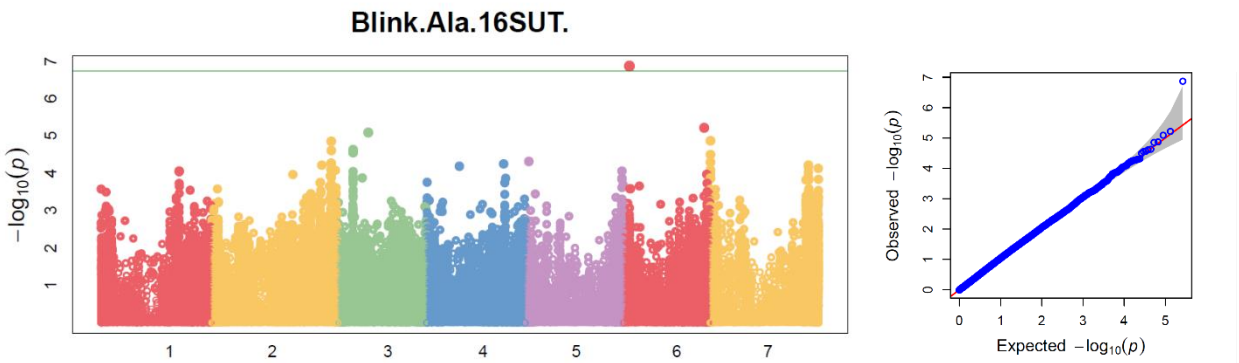

Blink.Arg.16ROS.
Blink.Arg.16SUT.
Blink.Arg.17ROS.
Blink.Arg.17SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)

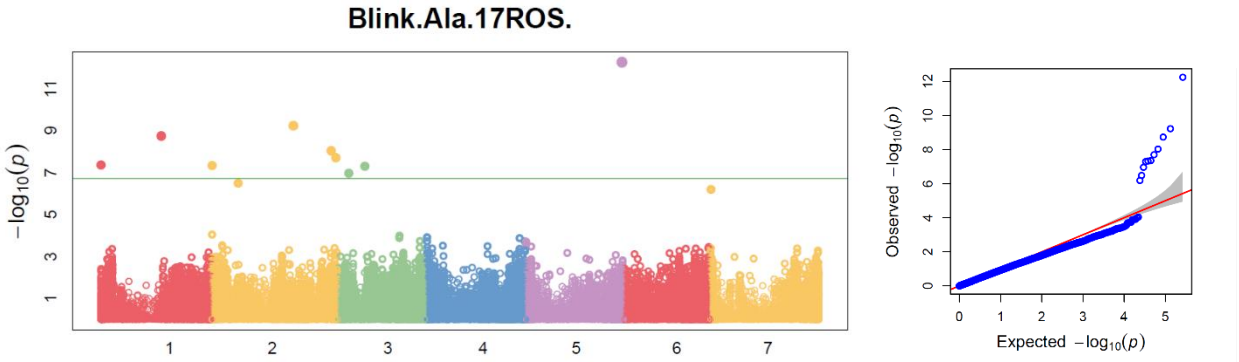

Blink.Gly.16ROS.
Blink.Gly.16SUT.
Blink.Gly.17ROS.
Blink.Gly.17SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)

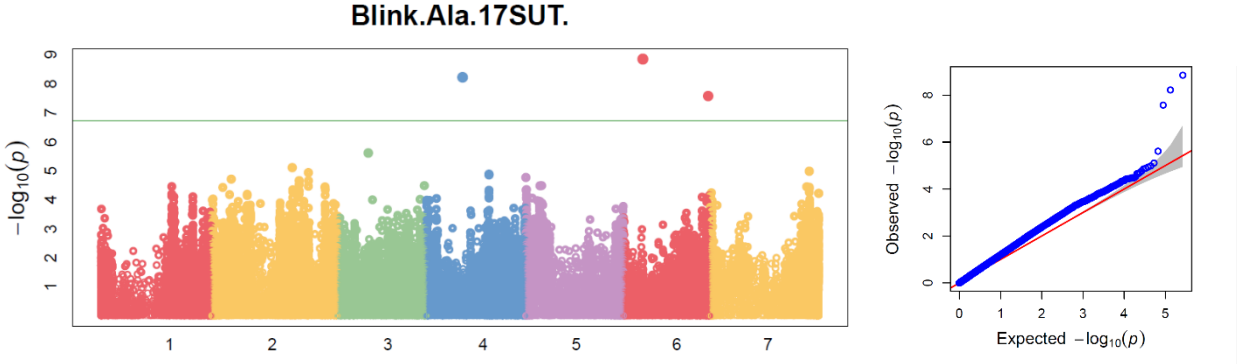

Blink.Asp.16ROS.
Blink.Asp.16SUT.
Blink.Asp.17ROS.
Blink.Asp.17SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)


Blink.Glu.16ROS.
Blink.Glu.16SUT.
Blink.Glu.17ROS.
Blink.Glu.17SUT.
$-\log_{10}(p)$
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

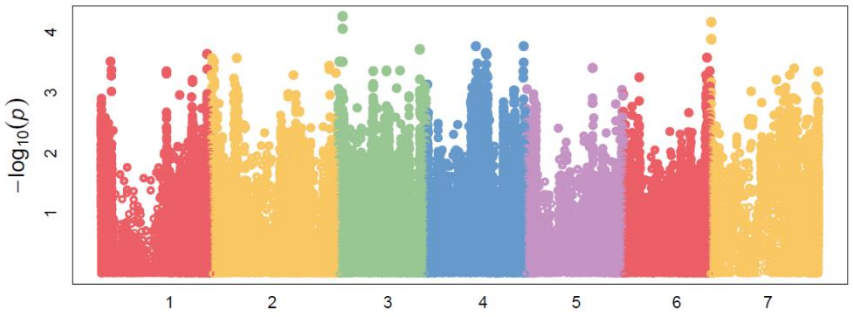

Blink.Ala.16ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)

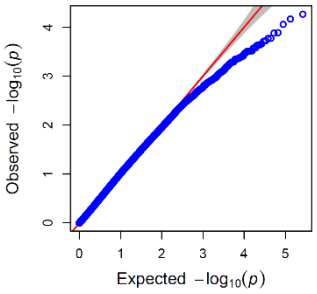

Blink.Ala.16SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)


Blink.Ala.17ROS.
−log10(p)
Observed −log10(p)
Expected −log10(p)

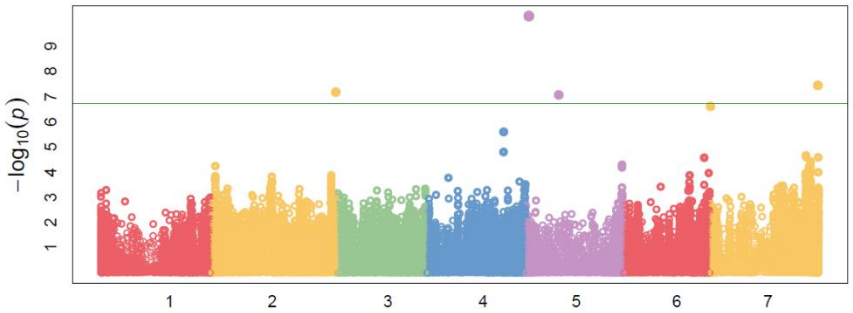

Blink.Ala.17SUT.
−log10(p)
Observed −log10(p)
Expected −log10(p)

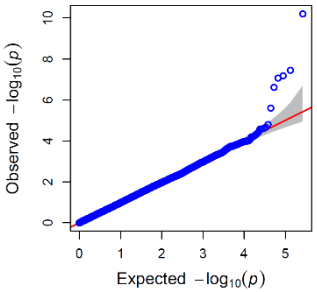
Blink.Pro.16ROS.



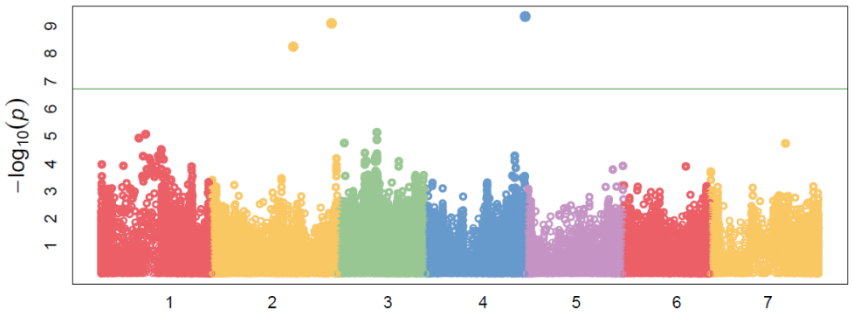


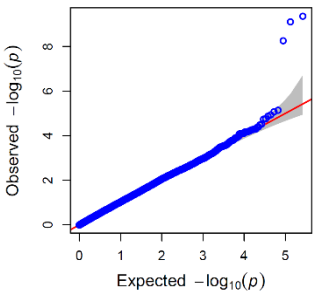
Blink.Pro.16SUT.



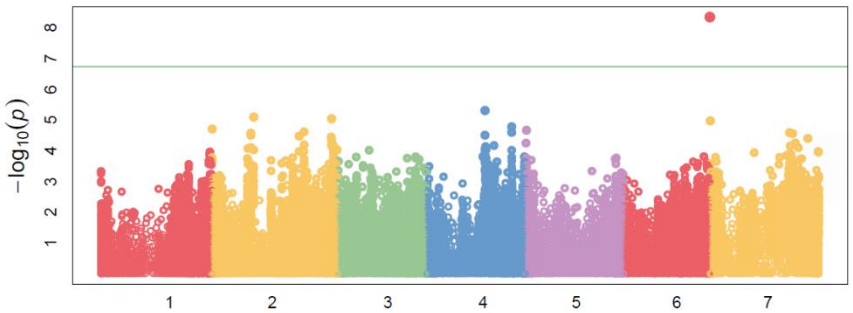


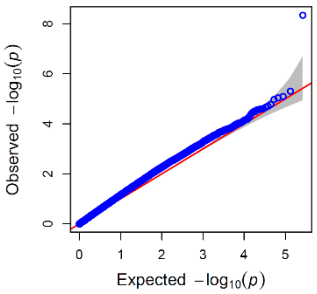
Blink.Pro.17ROS.

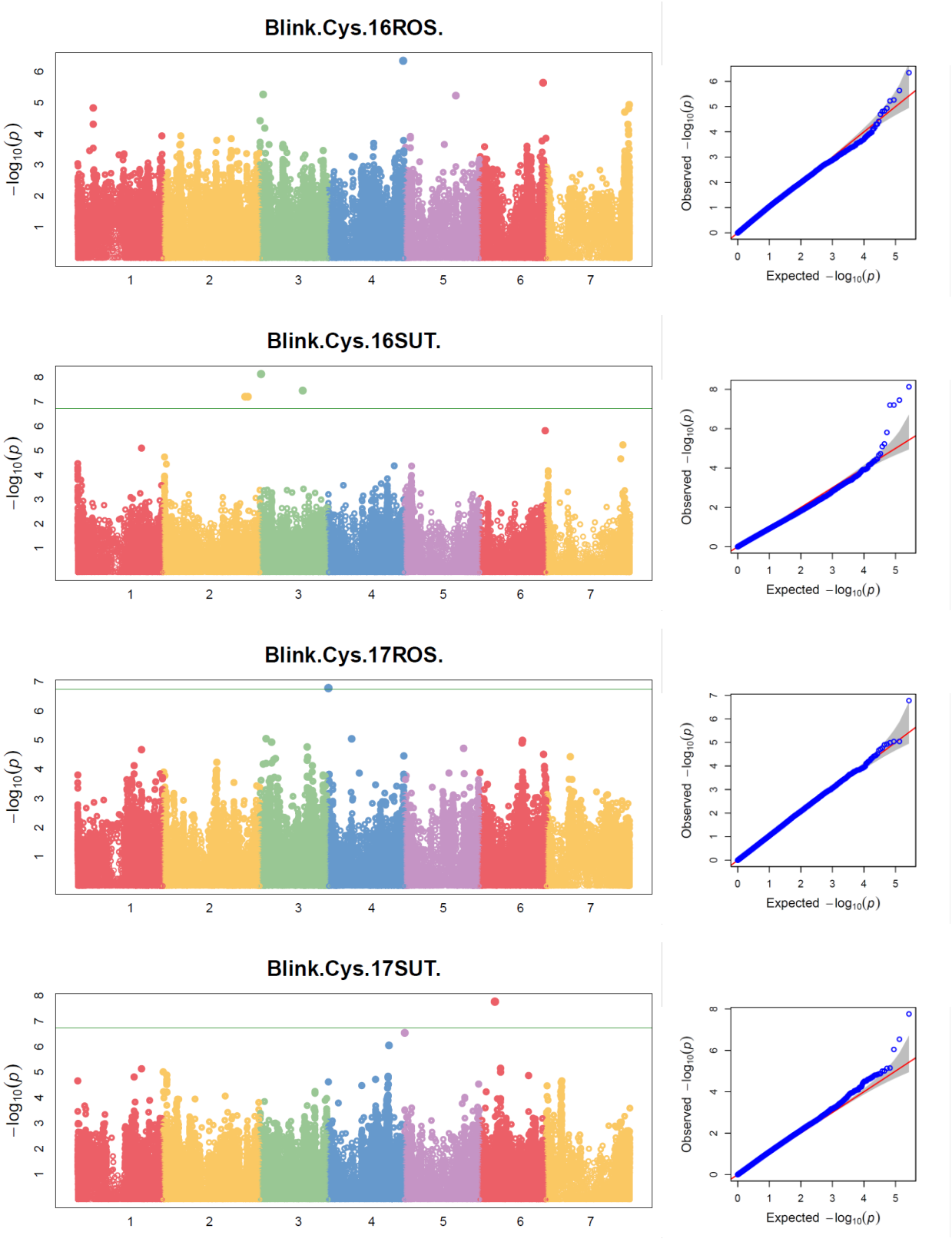


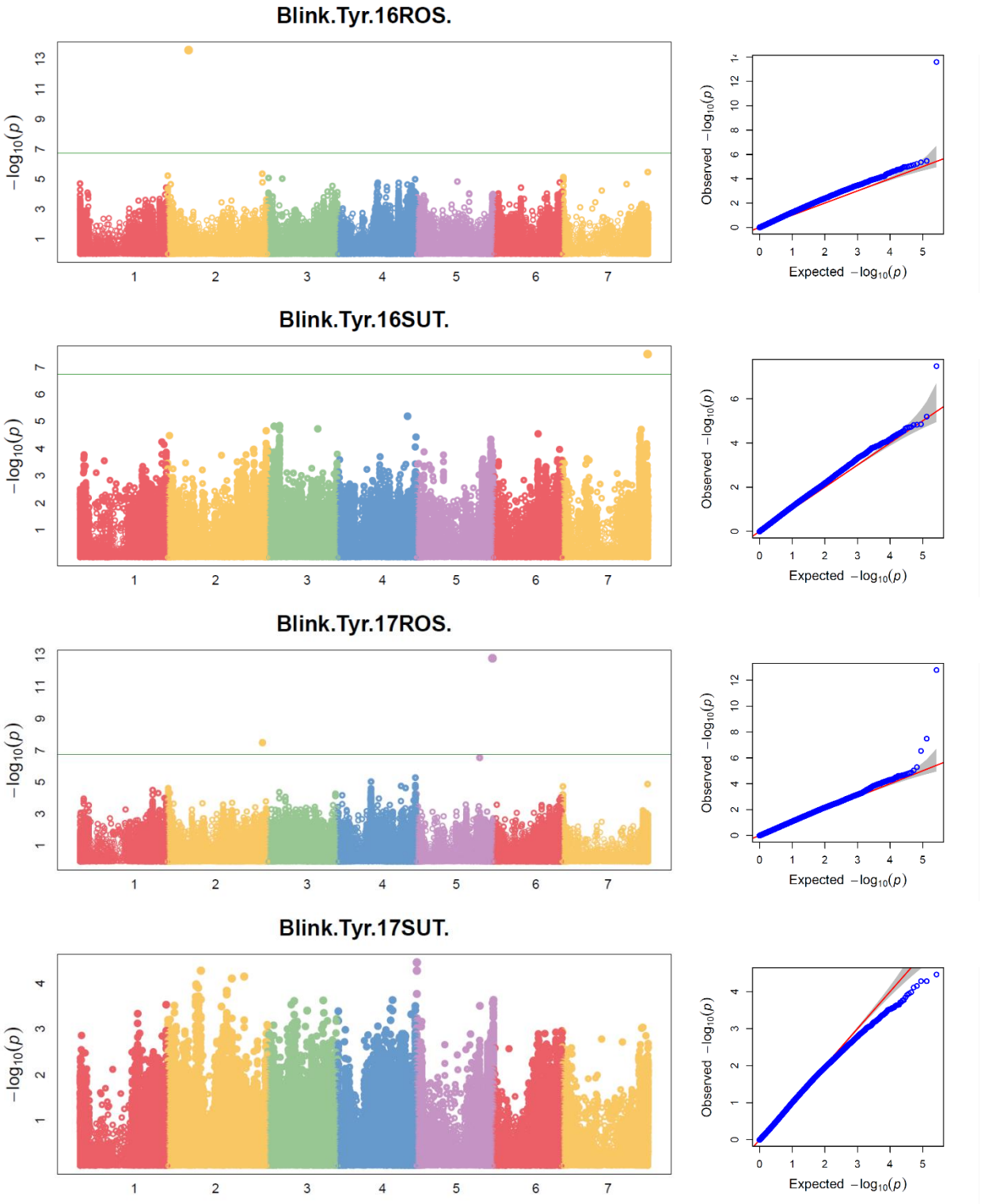


Blink.Pro.17SUT.






Blink.Cys.16ROS.
Blink.Cys.16SUT.
Blink.Cys.17ROS.
Blink.Cys.17SUT.
$-\log_{10}(p)$
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$


Blink.Tyr.16ROS.
Blink.Tyr.16SUT.
Blink.Tyr.17ROS.
Blink.Tyr.17SUT.
$-\log_{10}(p)$
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

Manhattan plots (left) and QQ-plots (right) for GWAS of the 324 soybean accessions for protein and 18 amino acid contents using Bayesian information and LD iteratively nested (BLINK) model. The trait associations for 266,164 SNPs were plotted by four different environments (16ROS, 16SUT, 17ROS, 17SUT). Manhattan plots showed -log10(p) values of each SNP marker along 7 chromosomes. Different colors represent different chromosomes. The green horizontal lines in the Manhattan plots represent the default Bonferroni correction threshold (-log10(p) = 6.73). In the QQ-plots, the red lines represent the diagonal lines, and shaded regions represent a 95% confidence interval.

Appendix I

Number of identical and significant SNP markers among protein and 18 amino acids in the same environment

	His	Ser	Arg	Gly	Asp	Glu	Thr	Ala	Pro	Cys	Lys	Tyr	Met	Val	Ile	Leu	Phe	Trp
Protein		2	1	1		1			1			1		1	1	2	1	
His																		
Ser			1	2	1	2	1	1	1		1	1		2	2	4	3	
Arg						1			2				1	1		3	1	1
Gly					1	2	2	1	1		1	1		5	2	3	2	3
Asp						1	1	1			1			1	1	1	1	1
Glu							1	2	2		1	2		1	3	5	4	1
Thr								1			1			2	1	2	1	
Ala											1			1	1	2	3	
Pro												2			1	4	2	1
Cys																		
Lys														1	1	1	1	
Tyr															1	2	2	1
Met																		
Val															1	3	1	3
Ile																2	2	
Leu																	5	1
Phe																		1