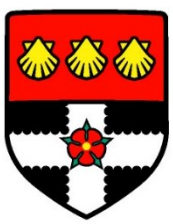

University of Reading

Towards genetic mapping of nutritional quality traits in faba bean (*Vicia faba* L.)

Ahmed Omar Warsame

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy to the School of Agriculture, Policy and Development

University of Reading

February 2021

Declaration of authorship

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

……

Ahmed O. Warsame

Abstract

Faba bean (*Vicia faba*) is a high-yielding and protein-rich legume crop which is grown globally for human and animal nutrition. Given the global importance of its protein content, the aim of this thesis was to address major gaps in our understanding of faba bean seed protein composition by investigating seed protein composition and its diversity, identifying genetic loci associated with total protein content and the abundance of major seed proteins, to understand the accumulation patterns of these proteins during seed development, and finally to fine-map the genetic loci responsible for pale seed hilum colour in faba bean.

Diversity in seed protein composition was assessed in 35 diverse faba bean genotypes using one-dimensional sodium dodecyl sulphate–polyacrylamide gel electrophoresis (1D SDS–PAGE) and size exclusion high performance liquid chromatography (SE-HPLC). On 1D SDS–PAGE gels, 25 protein bands were accurately identified by mass spectrometry and genotypes carrying rare legumin subunits were found. Also, SE-HPLC analysis of fractionated seed proteins showed that the main globulin seed storage proteins (legumin, vicilin and convicilin) consist of heterogeneous subunits with varying solubility in aqueous and salt solutions. In addition, SE-HPLC based quantification of the proportions of legumin and vicilin/convicilin fractions showed significant variation among genotypes with the ratio of legumin: vicilin/convicilin ranging from 1:1 to 1:3.

To reveal genetic loci associated with protein quality traits, a genome-wide association study was conducted using a multi-parent structured population that was grown in the field for two seasons (n=149 and 168 in 2019 and 2020, respectively). For the first time in faba bean, Marker Trait Associations (MTAs) were detected for total protein content (3 MTAs) and the abundance of 24 seed protein subunits (59 MTAs). Among protein subunits with novel MTAs were 4 legumins, 2 vicilins and several convicilin subunits. As revealed by examination of gene content of the associated regions using synteny with the *Medicago truncatula* genome, structural

and regulatory seed storage protein candidate causative genes were found. On the other hand, proteomic analysis of developing seeds of the inbred line Hedin/2 at 12 growth stages showed diverse accumulation patterns within and between major protein classes of legumin, vicilin and convicilin. For instance, certain legumin B-type and a vicilin protein were accumulated in the seed early from 45 days after pollination (DAP) while two proteins annotated as legumin J had biphasic pattern during seed development.

Finally, through combined bulk segregant analysis and linkage mapping, dihydroflavonol 4-reductase-like (DFR) was identified as the likely candidate gene responsible for the pale hilum colour in faba bean seeds. However, the nature of the mutation associated with the pale colour could not be elucidated further due to the presence of near-identical DFR copies in the candidate interval.

Acknowledgements

I would like to thank everyone who supported me during my journey towards completing my PhD thesis. Most gratitude is to my major supervisor, Donal O'Sullivan, for giving me the opportunity to join University of Reading and be part of his crop genetic improvement group. He provided to me all the support, guidance, and encouragement that I needed during my PhD study. I also thank Dr. Paola Tosi, my 2nd supervisor, for her valuable suggestions and support. Prof. Paul Hadley is greatly thanked for his guidance during PhD progress monitoring sessions.

My PhD journey would never have started without the scholarship from Islamic Development Bank who gave me the opportunity to study at a world-class university. I also thank Rank Prize Fund for helping me financially to continue my program after the original scholarship has finished.

I deeply thank my colleagues in crop genetic improvement group: Dr. Deepti Angra, Dr. Samer Amer, Dr. Vicky Tagkouli and Tom Robertson-Shersby-Harvie for their support and sharing their experiences.

Many thanks to the technical staff who never hesitated to help during my work at various departments of University of Reading, namely Dr. George Gibbings, Nicholas Michael, Richard Casebow, David Casebow, Caroline Hadley, Liam Doherty, Val Jasper and Kwok Cheung. I also thank Dr. Ihsan Ullah for generously sharing his experience in molecular genetics and laboratory skills.

I am deeply thankful to my family who supported me in every aspect during my study. Special thanks to my beloved wife (Amal Mohamoud) who took care of our children and endured the burden of being the mother and father. Thanks to my mother, father, brothers and sisters for their encouragements and prayers.

Table of contents

List of Tables

Supplementary Tables

List of figures

Supplementary Figures

Chapter 1 General introduction

The global human population is projected to reach ~9 billion person by 2050 (Roser *et al.*, 2013) which is expected to propel a significant increase in food demands. Proteins are one of the most important dietary requirements for humans for maintaining normal body function (WHO/FAO/UNU, 2007). Based on 2009–2011 trends, it is projected that an increase of at least 30-40% in the protein production is needed to meet the demands of 9.6 billion people by 2050 (Henchion *et al.*, 2017). In particular, required amounts of plant proteins is estimated to rise by 110% in 2050 from the 2005 demands (Tilman *et al.*, 2011). The main driver of this demand is the requirements of protein sources in the world, particularly in the developed countries, for animal feeding (Voisin *et al.*, 2014; Speedy, 2004) and manufacturing of plant-based meat alternatives and other healthy foods (Ismail *et al.*, 2020). However, in the face of climate change, any future increases in plant protein production will have to be achieved through sustainable production systems which adopt innovative production methods and high yielding, climate-resilient and protein-rich crops.

Faba bean (*Vicia faba* L., hereafter *Vf*) is one of the major legume crops in the world where it is grown in over 60 countries with annual production of ~4 million tons (FAOSTAT,2018). With an average of ~29% seed protein content on dry weight basis, this crop has the third highest protein content among legumes after soybean and lupin. It is unparalleled amongst major grain legumes for yield potential (Cernay *et al.*, 2016) and its biological nitrogen fixation capacity is much greater than that of soybean and pea (Baddeley *et al.*, 2013). Together, these characteristics put *Vf* in the forefront of candidate protein sources that are in the best position to meet the nutritional requirements of the current and future generations. Yet, the available literature shows that little effort has been made so far to enhance the protein quality of this important crop.

In legume crops, seed protein quality can generally refer to the percentage dry weight of the seed represented by crude protein. However, from an end-user perspective, quality also encompasses the proportion of certain protein classes with desired functional and nutritional properties and, therefore, both protein content and composition need to be improved simultaneously to ensure that future cultivars fit market demands. In *Vf*, a considerable genetic variation in crude protein content (19-40%) has been reported (Griffiths, 1984; Sjödin, 1982; Griffiths and Lawes, 1978). For protein composition, Gatehouse *et al.* (1980) and Martensson (1980) were first to report the possibility of genetically improving the ratio between the major protein classes–legumin and vicilin. However, to date, not a single study has been published on the genetic basis for either seed protein content or abundance of the major seed protein classes in *Vf*.

Therefore, this thesis project set out to thoroughly survey variation, both in amount and composition, of *Vf* seed protein and investigate genetic control of such compositional variation in order to contribute towards development of cultivars with desired protein content and composition.

1.1 Thesis objectives

1. Identify the most abundant seed proteins and provide a comprehensive list that can be used as reference for SDS-PAGE based protein analysis.
2. Survey genetic variation among *Vf* germplasm for seed protein composition.
3. Map genetic loci associated with total seed protein content and abundance of specific seed protein classes and subunits.
4. Evaluate patterns in the accumulation of seed proteins during seed development.
5. Fine-map the genetic locus underlying the pale seed hilum colour .

1.2 Thesis outline

This thesis contains seven chapters starting with this general introduction (**Chapter 1**) followed by the literature review (**Chapter 2**) which highlights the nutritional and ecological importance of *Vf* in meeting the 21st century protein demands. It also presents the current knowledge on the structure, synthesis and genetics of seed storage proteins and outlines a research agenda that could deliver concrete improvements in seed protein quantity and quality. **Chapter 3** focuses on providing an up-to-date survey on types of seed proteins using a panel of diverse germplasm including the founders of a number of mapping populations, including one that was used in the subsequent Chapter 4. In addition, the diversity in protein subunit composition and potential of SE-HPLC method for quantifying the relative abundance of the major storage proteins is explored. **Chapter 4** capitalises on the analytical methods refined and variation discovered in Chapter 3 to undertake a genome-wide association study, providing the first insights into genetic control of crude protein content and the abundance of major seed proteins. **Chapter 5** moves the investigation of variation in seed protein profile from a genetic to a developmental perspective by investigating the seed proteome profile across 12 time points throughout seed development, revealing a number of distinct patterns in the accumulation of major seed proteins during seed filling stages. **Chapter 6** is dedicated to a sensory quality trait, seed hilum colour, and describes the fine-mapping of the gene responsible for the pale hilum colour using bulk segregant analysis and linkage mapping methods. Finally, the general discussion contained in **Chapter 7** links and contextualises the main findings of the previous chapters and suggests some areas for future focus.

1.3 References

Baddeley, J. A., Jones, S., Topp, C. F. E., Watson, C. A., Helming, J. & Stoddard, F. L. (2013). Biological nitrogen fixation (BNF) by legume crops in Europe. *Legume Futures Report 1.5.*

Cernay, C., Pelzer, E. & Makowski, D. (2016). A global experimental dataset for assessing grain legume production. *Scientific Data,* **3**, 160084.

Food and Agriculture Organization of the United Nations. (2018). FAOSTAT Database. Rome, Italy: FAO. Retrieved July 20, 2018 from http://www.fao.org/faostat/en/#home

Gatehouse, J., Croy, R., McIntosh, R., Paul, C. & Boulter, D. (1980). Quantitative and qualitative variation in the storage proteins of material from the EEC joint field bean test. *Quantitative and qualitative variation in the storage proteins of material from the EEC joint field bean test.*, 173-188.

Griffiths, D. W. (1984). An Assesment of the Potential for Improving the Nutritive Value of Field Beans (*Vicia faba*)- A Progress Report. *In:* Hebblethwaite, P. D., Dawkins, T. C. K., Heath, M. C. & Lockwood, G. (eds.) *Vicia faba: Agronomy, Physiology and Breeding.* Brussels-Luxembourg: Springer-Science+Business Media, B.V. .

Griffiths, D. W. & Lawes, D. A. (1978). Variation in the crude protein content of field beans (*Vicia faba* L.) in relation to the possible improvement of the protein content of the crop. *Euphytica,* **27** (2), 487-495.

Henchion, M., Hayes, M., Mullen, A., Fenelon, M. & Tiwari, B. (2017). Future protein supply and demand: strategies and factors influencing a sustainable equilibrium. *Foods,* **6** (7), 53.

Ismail, B. P., Senaratne-Lenagala, L., Stube, A. & Brackenridge, A. (2020). Protein demand: review of plant and animal proteins used in alternative protein product development and production. *Animal Frontiers,* **10** (4), 53-63.

Martensson, P. (1980). Variation in legumin : vicilin ratio between and within cultivars of *Vicia faba* L. var. minor. The Hague: Martinus Nijhoff. World crops: production, utilization and description, volume 3.

Roser, M., Ritchie, H. & Ortiz-Ospina, E. (2013). World population growth. Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/world-population-growth' [Online Resource].

Sjödin, J. (1982). Protein Quantity and Quality in *Vicia faba*. In: Hawtin, G. & Webb, C. (eds.) Faba Bean Improvement: Proceedings of the Faba Bean Conference held in Cairo, Egypt, March 7–11, 1981. Dordrecht: Springer Netherlands.

Speedy, A. W. (Year). Overview of world feed protein needs and supply. In: FAO Animal Production and Health Proceedings (FAO), 2004. FAO.

Tilman, D., Balzer, C., Hill, J. & Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences,* **108** (50), 20260-20264.

Voisin, A.-S., Guéguen, J., Huyghe, C., Jeuffroy, M.-H., Magrini, M.-B., Meynard, J.-M., Mougel, C., Pellerin, S. & Pelzer, E. (2014). Legumes for feed, food, biomaterials and bioenergy in Europe: a review. *Agronomy for Sustainable Development,* **34** (2), 361-380.

Warsame, A. O., O'Sullivan, D. M. & Tosi, P. (2018). Seed storage proteins of faba bean (*Vicia faba* L): current status and prospects for genetic improvement. *Journal of Agricultural and Food Chemistry,* **66** (48), 12617-12626.

WHO/FAO/UNU (2007). Protein and amino acid requirements in human nutrition. *WHO Technical Report Series,* (935), 1-265, back cover.

Chapter 2 Literature review

Seed Storage Proteins of Faba Bean (*Vicia faba* L): Current Status and Prospects for Genetic Improvement

Ahmed O. Warsame, Donal M. O'Sullivan, and Paola Tosi

[Published in *J. Agric. Food Chem*. (2018), 66, 12617−12626]

2.1 Abstract

Faba bean (*Vicia faba*, L.) is one of the foremost candidate crops for simultaneously increasing both sustainability and global supply of plant protein. On a dry matter basis, its seeds contain about 29% protein of which more than 80% consists of globulin storage proteins (convicilin, vicilin and legumin). For optimum utilization for human and animal nutrition, both protein content and quality have to be improved. Though initial investigations on the heritability of these traits indicated the possibility for genetic improvement, little has been achieved so far partly due to the lack of genetic information coupled with the complex relationship between protein content and grain yield. This review reports on the current knowledge on *V. faba* seed storage proteins; their structure, composition and genetic control and highlights key areas for further improvement of the content and composition of *V. faba* seed storage proteins on the basis of recent advances in *V. faba* genome knowledge and genetic tools.

Key words: *Vicia faba*; sustainability; storage proteins; legumin and vicilin; genetic improvement

2.2 Introduction

2.2.1 Faba bean production and utilization

Nearly 60% of the global protein supply for human nutrition is sourced from plants (Henchion *et al.*, 2017; Young and Pellett, 1994) and about one third of this originates from grain legumes of the Fabaceae family (Smýkal *et al.*, 2015). Besides their nutritional significance, legume crops' ability to fix atmospheric nitrogen via rhizobial symbiosis makes them invaluable components of sustainable crop production systems (Foyer *et al.*, 2016). *Vicia faba (Vf)*, also known as fava bean, broad bean, horse bean or field bean (Duc, 1997) is one of the world's oldest legume crops, its cultivation dating back to the 10th millennium BC (Tanno and Willcox, 2006; Cubero, 1974). From its origin in the Near East, *Vf* spread to the rest of the globe (Cubero, 1974) and is currently cultivated in nearly 70 countries over the world (**Figure 2.1**A), occupying about 2.2 million ha that produce nearly 4 million tonnes annually (FAOstat, 2018). China is the leading *Vf* producer with 36% of the global output, followed by Ethiopia (20%), Australia (10%) and United Kingdom (6%) (**Figure 2.1**B). The wide geographical distribution of *Vf* implies not only a great adaptation to diverse environmental conditions, but also suitability for diverse end uses and trade across continents.

Seeds of *Vf* contain on average about 29% protein on a dry matter (DM) basis (https://www.feedipedia.org) which provides affordable nutrition for millions of people around the world, hence its denomination as "the poor man's meat". Beside proteins, *Vf* seeds contain a variety of other constituents (fibres, starches, vitamins, minerals) (Vilariño *et al.*, 2009) which give them additional value as foods, feeds and ingredients. While *Vf* has been traditionally utilized as dry grain for human consumption in developing countries, there is growing interest from food industries in developed countries to exploit its protein for the production of protein-rich vegan/vegetarian foods (Kaskinen *et al.*, 2018), the fortification of cereal-based food products such as bread and pasta without significantly affecting their structural and sensory

quality (Coda *et al.*, 2017; Rizzello *et al.*, 2017), or even the production of wholly *Vf*-based bread and pasta products (VTT, 2014). Additionally, *Vf* proteins can potentially fit into the growing demand for manufacturing plant-based meat alternatives and other healthy foods (Ismail *et al.*, 2020) including sport nutrition (Grebow, 2020). *Vf* also represents a significant resource for agro-ecosystem sustainability and provision of feed for the growing global livestock inventory. Overall, the global production area for *Vf* has been increasing in the last two decades (**Figure 2.1**C) and a recent meta-analysis of yield data from 39 legume species indicated that, in the right environment, *Vf* can be the highest yielding grain legume (Cernay *et al.*, 2016). *Vf* also has a high capacity for biological nitrogen fixation, to the extent that the amount of N fixed by *Vf* alone was estimated to be comparable to that of soybean and pea combined (Baddeley *et al.*, 2013). For further details on the role of *Vf* on sustainable cropping systems, readers are referred to Jensen *et al.* (2010); Köpke and Nemecek (2010).

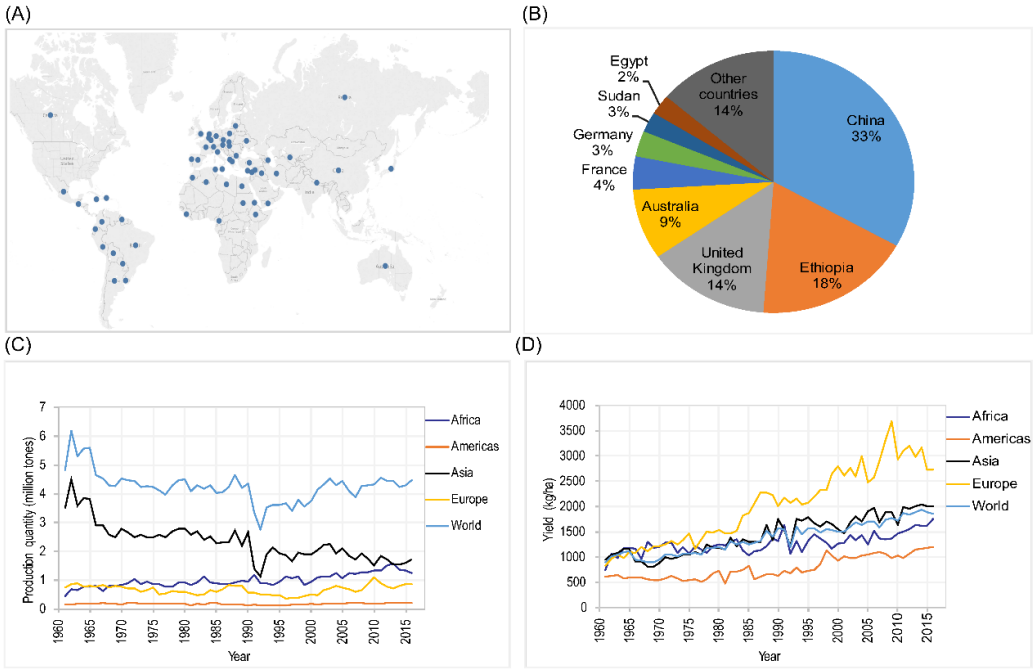

Figure 2.1. Global production and cultivation of *Vf*. (A) Map of world distribution of *Vf* cultivation; (B) Percentage of production by the major producing countries (C). Trends in the production of *Vf* in four continents from 1960-2016 (D) Global view of *Vf* productivity between 1960-2016. Data was sourced from FAOstat (2016) except UK which was obtained from Eurostat. The world distribution map was generated using Tableau Public 2018.1.

On the other hand, *Vf* is yet to be fully exploited as a feedstock for animal production due to the presence of some anti-nutrients which limit its optimal inclusion ratio (Perez-Maldonado *et al.*, 1999; Koivunen *et al.*, 2014 ; Lessire *et al.*, 2016). Removal of these anti-nutrients through the development of new low anti-nutrient cultivars or using simple processing techniques like fermentation (Coda *et al.*, 2017; Rizzello *et al.*, 2017) would make this crop a valuable protein resource for the animal production industry.

2.2.2 Faba bean as a sustainable global protein resource

One of the greatest challenges in the 21st century is feeding the growing world population which it has been estimated may necessitate a 70% increase in food production by 2050 (Foyer *et al.*, 2016). More than 30% of this increase has to be made via the production of protein-rich foods (Henchion *et al.*, 2017) to meet the expected rise in demands due to population growth, increased urbanization and improved incomes in many parts of the world (Kawashima *et al.*, 2002; Speedy, 2004; Henchion *et al.*, 2017; Chiari, 2017). Protein is a critical nutrient required in large quantity by humans (~ 50 g protein per adult per day) to maintain normal body function (WHO/FAO/UNU, 2007). However, about one-third of the world population, mainly in Asia, Africa and Latin America, suffers from inadequate intake of proteins, vitamins and minerals (Balyan *et al.*, 2013). On the other hand, in higher income countries, where daily animal-based protein intake is already high (Chiari, 2017; Henchion *et al.*, 2017), continued provision of nutritious feeds for the intensive animal production industry will pose a major challenge in the future. In particular, the livestock production sector in soybean non-producing countries will be burdened by the high price of imported soybean and soybean meal. For instance, EU countries have a huge deficit in protein-rich feeds with nearly 70% being imported (de Visser *et al.*, 2014). *Vf* is well-adapted to European climates, as testified by the high yields recorded in this continent for this legume (**Figure 2.1**D), and therefore it has the potential to contribute to bridging the gap in animal feed self-sufficiency as part of the EU's policies to increase protein production from

locally grown crops (de Visser *et al.*, 2014). *Vf* is also a candidate crop to meet the protein demands of an emerging consumer category, particularly in developed economies, who are opting for animal meat free lifestyle. For example, Statista (2017) reported that 13% of European citizens would consider avoiding red meat while nearly 50% of the respondents in another study were willing to replace meat with other sources of proteins (de Boer and Aiking, 2018). Thus, considering the projected impact of climate change on global crop production, meeting the nutritional requirements of the current and future generations would necessitate increased exploitation of the global genetic and natural resources for protein production systems based increasingly on biological nitrogen fixation. In this context, the fact that *Vf* is a high-yielding protein-rich crop with superior N fixation capability makes it a candidate crop for supporting increased protein production while maintaining sustainability of crop production systems.

2.2.3 Constraints to faba bean production and utilization

Despite the importance of *Vf* in sustainable crop production and the supply of plant proteins, there are several factors that limit the realization of its potential. For instance, faba bean yield and yield stability is affected by various biotic and abiotic stresses (see Duc *et al*., 2015; Torres *et al*., 2012). In general, faba bean is susceptible to fungal diseases such as *aschochyta* blight, chocolate spot and rust. In addition, in Mediterranean region, the plant-parasitic plant, Orobanche, causes a serious damage to faba bean. On the other hand, the partial allogamous nature and poor suitability for mechanization (for instance, indeterminate growth habit and relatively large seed sizes) are key challenges in faba bean breeding programs.

From end-use point, the main determinants of *Vf* utilization for human food and animal feed include: (i) protein concentration, (ii) protein quality, defined mainly by the content of sulphur-containing amino acids (S-AA) cysteine and methionine, and (iii) concentration of antinutrients in the seeds (Duc, 1997). Protein concentration of *Vf*, although it can vary greatly between different genotypes (19-39 %) (Griffiths and Lawes, 1978; Sjödin, 1982; Frauen *et al.*,

1984), is one of the highest among legumes. However, commercial varieties on the UK market contain about 27% protein on average, which is still far less than the protein density of soya meal, and so, further improvement in protein content would help *Vf* to displace imported soya in animal feed. The proportion of S-AA in the protein is another crucial quality criterion, particularly in animal feeding. However, like most pulse crop proteins, *Vf* is poor in certain essential amino acids, namely methionine, cysteine and tryptophan (Duc, 1997). Though relatively narrow, the genetic variation for S-AA reported in *Vf* indicates the possibility of improving its nutritional quality. So far, the major nutrition-related breeding objectives for *Vf*, have been the reduction or removal of vicine and convicine (V-C) and tannins; V-C causes favism in humans and have deleterious effects on animals (Crépon *et al.*, 2010; Yu *et al.*, 2017), while tannins lower protein digestibility (Makkar *et al.*, 1997). Although these compounds can be removed by processing techniques (Rizzello *et al.*, 2016; Coda *et al.*, 2015), the most effective approach is probably removing them by breeding. For instance, low V-C varieties with reduced risk of favism (Gallo *et al.*, 2018) can be utilized in breeding programs by utilizing molecular markers closely linked to the *VC* locus (Khazaei *et al.*, 2017). Similarly, identification of the zero-tannin gene (*zt-1*) (Webb *et al.*, 2016) can accelerate development of zero-tannin cultivars. Furthermore, the reduction of less studied antinutrients such as trypsin inhibitors, lectins and phytates would improve the nutritional value of *Vf* based feed products.

Understanding the genetic basis of the above limiting factors is a prerequisite for the development of new cultivars with desirable agronomic and nutritional attributes. Unfortunately, while scientific interest in *Vf* was high during the 1970's and 1980's , when it became the model species for studying plant cytogenetics and stomatal regulation, *Vf* can still be considered an orphan crop (O'Sullivan and Angra, 2016). For instance, less than 5% of the publications on legumes in the years 2004–2013 referred to *Vf* (Duc *et al.*, 2015). This is further reflected by the scarcity of information on the genetics of many important traits including protein content and

quality, for which not a single QTL (Quantitative Trait Loci) has been reported, compared to 160 QTLs from 35 independent studies on soybean protein content (Patil *et al.*, 2017).

2.3 Seed storage proteins of faba bean

The major storage proteins of legumes are mainly enzymatically inactive proteins deposited in seed cotyledons which provide the nutrients needed for seed germination and seedling growth and development (Liu *et al.*, 2017; Shewry and Casey, 1999). Some storage proteins such as albumins, lectins and some vicilins may play a role in plant defence system (de Souza Cândido *et al.*, 2011) while others, including albumins and trypsin inhibitors, have been identified as antinutritional or allergenic agents and therefore are targeted for removal in breeding programs (Joshi *et al.*, 2017). Seed storage proteins are classified according to the system developed by TB Osborne which is based on their solubility in different solvents (Shewry and Casey, 1999). Globulins and albumins are the major classes of storage proteins of legumes and are soluble in saline and water solutions, respectively. Globulins alone constitute more than 80% of total seed protein in *Vf* (Müntz *et al.*, 1999) and they are further classified based on their sedimentation coefficients into vicilin-type (7S) and legumin-type (11S) (Shewry and Casey, 1999). Both globulin proteins are found in nearly all legumes, but their denotations vary across species. For instance, globulins of *Vf* and pea are often referred as vicilin/convicilin and legumin, while they are denoted as conglycinin and glycinin in soybean, β and α conglutins in lupin, while phaseolin (a vicilin-like protein) is the only major globulin in common beans. Furthermore, decades of research on legume storage proteins have produced a database of annotated SDS-PAGE images of various species which facilitates faster comparison and identification of major globulin bands. When extracted under reducing conditions, the salt soluble fraction of *Vf* seed proteins can be separated on SDS-PAGE into distinct bands which, based on their molecular weights, are identified as: convicilin (Mr >~60 kDa), vicilin (Mr=~46-55 kDa) and two major legumin subunits (Mr= ~38-40 and 23 kDa) (**Table 2.1**).

Table 2.1. Major globulin polypeptides of *Vf* and related species as reported in the literature. The molecular mass measured in kilodalton (kDa) is calculated either based on relative migration distance (Rf) or sum of molecular weight of amino acids.

Species	11S legumin-like (kDa)		7S Vicilin-like (kDa)		References
	α chain	β chain	vicilin	Convicilin	
V. faba	38	22-24	31-65		De Pace *et al.* (1991)
	38-47	..	..	64	Liu *et al.* (2017)
	40	20	..	..	Gatehouse *et al.* (1980)
	35-39	23-25	42-48	66	Tucci *et al.* (1991)
	36-51	19-23	..	..	Utsumi *et al.* (1980)
M. truncatula	36-46	23-24	46-47	60-92	Le Signor *et al.* (2005)
	42-46	23	46-47		Gallardo *et al.* (2003)
	..	..	16-48	53-100	Le Signor *et al.* (2017)
*G. max**	37	20	52-72		Fontes *et al.* (1984)
	37	20	52-72		Boehm *et al.* (2017)
	37	20	52-72		Poysa *et al.* (2006)
	37−44	17−22	53-76		Krishnan *et al.* (2017)
P. sativum	40-45	18-25	53	60-88	Bourgeois *et al.* (2009)
	40	24.8	47.2	67.2	Mertens *et al.* (2012)
	40	..	..	>70	Rubio *et al.* (2014)
	37	25	43-53	70	Ladjal E *et al.* (2015)

*7S subunits of *G. max* consist of α`, α and β polypeptides.

Legumin and vicilin share notable sequence and structural homology and are believed to originate from a common ancestral gene (Kesari *et al.*, 2017). Mature legumin is hexameric with a mass of about 330 kDa (Müntz *et al.*, 1999) and is composed of two trimeric subunits (legumin A and B), while vicilin is a trimeric protein formed by the assembly of three monomers (**Figure 2.2**). In contrast to legumin, vicilin lacks cysteine and is usually glycosylated in its C-terminus (Kesari *et al.*, 2017). These structural variations may result in differences in the physiochemical properties of seed storage proteins which in turn determine their nutritional value and utilization. For instance, legumin and vicilin differ in their thermal properties (Meng and Ma, 2001; Kimura *et al.*, 2008), affinity to bind flavour compounds under varying pH conditions (Heng *et al.*, 2004) and emulsifying ability (Kimura *et al.*, 2008). Therefore, from a breeding point of view, legumin/vicilin ratio could be manipulated to meet certain end-user requirements for protein functionality.

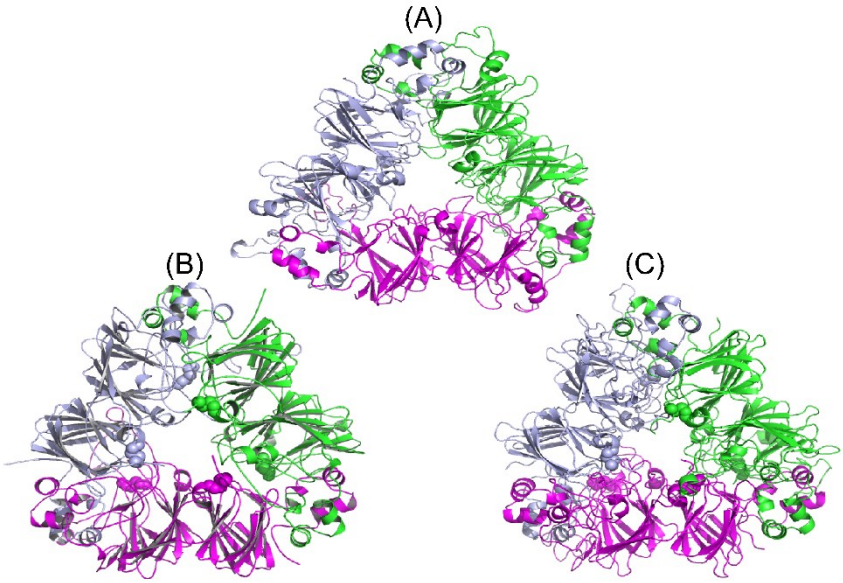

Figure 2.2. Predicted ribbon structures of *Vf* globulins. Vicilin (A) is trimeric consisting of 3 protomers (a=light blue, b= magenta and c= green) while legumin is hexameric consisting of legumin A (B) and legumin B (C). Spherical balls in legumin subunits represent disulphide bonds. The models were generated using SWISS-MODEL and processed with PyMOL software.

2.3.1 Structure and composition of *Vf* globulins

Legumin constitutes more than 50% of *Vf* globulins (Müntz *et al.*, 1999). It is a hexameric protein with two major subunits - the α and β chains - which are connected by disulphide bonds. Under reducing conditions, these subunits form two bands of molecular weights of about 40 and 24 kDa, respectively. These subunits are also referred to as acidic and basic subunits or simply legumin α and β. Polypeptides of both legumin A and B are highly homologous but notably distinguishable by the presence of more methionine residues in the peptide sequences of legumin A subunits (Baumlein *et al.*, 1986). *Vf* legumin A subunits appear to be more variable and show polymorphic bands between genotypes (Tucci *et al.*, 1991) as is also the case with Medicago legumin A (Le Signor *et al.*, 2005). On the other hand, vicilin-type proteins of *Vf* are trimeric (Müntz *et al.*, 1999) consisting predominantly of subunits of ~50 kDa while bands of ~66 kDa are referred as convicilin (Tucci *et al.*, 1991; Liu *et al.*, 2017). The classification of 7S proteins into vicilin and convicilin was first coined in pea and has been accepted in many legumes including *Vf* (**Table 2.1**). Nonetheless, further investigation into their possible structural and

functional differences have concluded that convicilin may be regarded as subunit of vicilin (O'Kane *et al.*, 2004). Such a denotation exists in soybean whereby subunits of 7S protein are categorized into α′ (~76 kDa), α (~72 kDa), and β (~53) kDa (Krishnan *et al.*, 2017; Boehm *et al.*, 2017).

Regarding amino acid composition, nearly 50% of *Vf* seed protein is accounted for by just a few non-essential amino acids such as glutamic acid, aspartic acid, arginine, and leucine while it is low in essential amino acids, particularly S-AA (**Figure 2.3**). The concentration of S-AA is a critical determinant of the nutritional value of plant proteins destined for human consumption and animal feeding. In humans, dependence on poor quality proteins can result in reduced immunity and underdeveloped mental and physical capacity among young children (Galili and Amir, 2013). Also, animal feeds deficient in critical amino acids can cost farmers in the form of animal feed supplements of industrially synthesized S-AA (Boehm *et al.*, 2017).

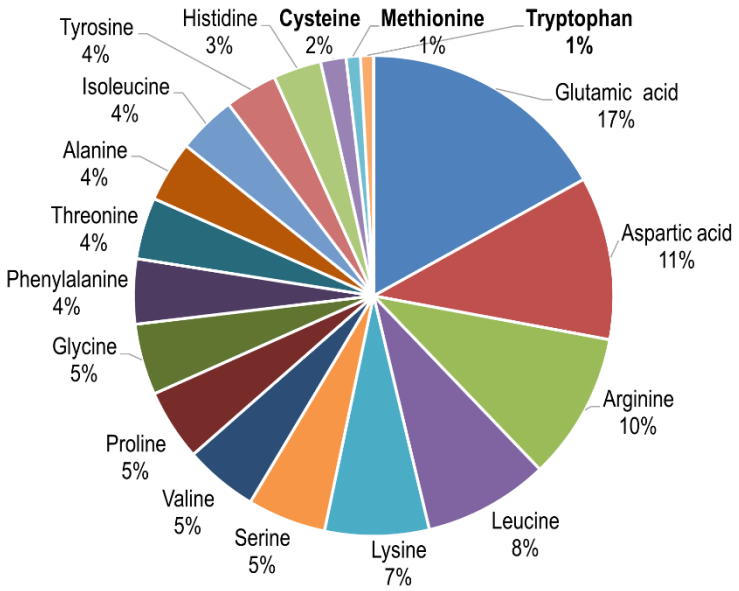

Figure 2.3. Amino acid composition (g/16 g N) of *Vf* seed protein (Makkar *et al.*, 1997; Grela *et al.*, 2017). Numbers on the figure represent percentage of amino acids in the protein.

Since the concentration of S-AA is strongly related to the relative proportions of S-AA rich proteins in the seeds, it is well accepted in *Vf* and other legumes that legumins contain a relatively higher S-AA concentration compared to vicilin (Kwanyuen *et al.*, 1997; Jackson *et al.*, 1969 ; Liu *et al.*, 2017; Joshi *et al.*, 2017). This is further confirmed by comparative analysis of coding sequences of vicilin and legumin subunits across legume species which clearly show that

legumin subunits contain more residues of cysteine and methionine (**Figure 2.4**). This observation leads to the hypothesis that increasing the proportion of legumin subunits relative to vicilin would improve nutritional content of plant proteins. However, considering that vicilin is accumulated in legume seeds earlier than legumin (De Pace *et al.*, 1991; Abirached-Darmency *et al.*, 2012; Gallardo *et al.*, 2003), their ratios could be easily offset by the prevailing environmental conditions, e.g., soil nutritional status and onset of biotic and abiotic stresses during the plant growth, and in particular, during grain filling.

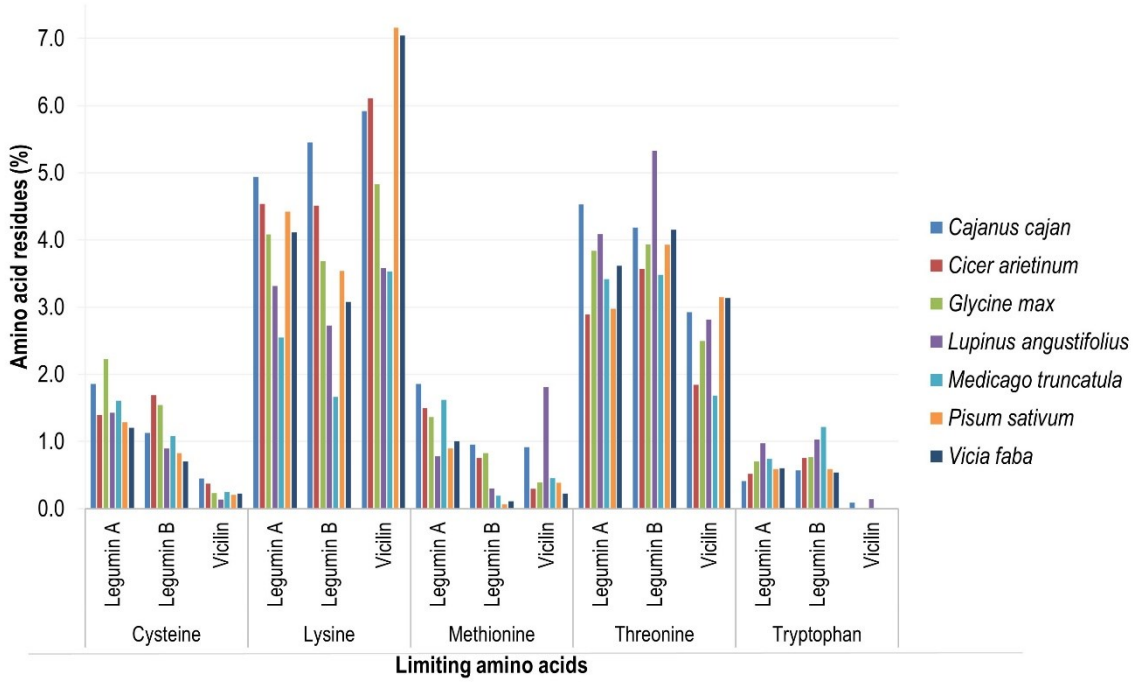

Figure 2.4. Relative abundances of limiting amino acids within legumin and vicilin coding sequences of seven legume species. Annotated protein accessions were obtained from Uniprot and the amino acid residues were counted using "seqinr" package in R.

In contrast to globulins, minor legume seed proteins such as elongation factor Tu, citrate synthase, albumin 2 (PA2), defensins 1 and 2 and Bowman–Birk inhibitors (BBI) contain higher S-AA (Liu *et al.*, 2017; Rubio *et al.*, 2014). According to Krishnan *et al.* (2005), under higher N availability through fertilizer application or symbiotic fixation, S-AA containing proteins like BBI were decreased in favour of β subunits of β-conglycinins of soybeans. Similarly, ectopic

overexpression of the *VfAAP1* gene in *P. sativum* and *V. narbonensis* resulted in a 30% increase in the globulin fraction but no significant effect on albumins, a S-AA rich protein subunit (Rolletschek *et al.*, 2005). Hence, it would appear that the negative correlation between high protein and S-AA content in *Vf* (Lafiandra *et al.*, 1981; Sjödin, 1982; Schumacher *et al.*, 2011) may be the result of preferential accumulation of low nutritional quality protein fractions in higher protein content lines.

2.3.2 Genetic control of globulins

Globulins are by far the most abundant seed proteins in legumes and, subsequently, their genetic control has been well investigated. In *Vf*, legumin is encoded by relatively few genes which are classified as legumin A and B genes. A single legumin A gene has been located on the telomeric region of chromosome V of *Vf* (Fuchs and Schubert, 1995). It is not clear, however, whether the legumin A2 gene (*LegA2)* reported in pea (Rerie *et al.*, 1991) also exists in *Vf,* as no up to date information is available. Conversely, there are at least five transcribed genes (*LeB2, LeB3, LeB4, LeB6, LeB7*) for legumin B subunits (Heim *et al.*, 1989; Fuchs and Schubert, 1995), of which *LeB3* and *LeB4* have been mapped to chromosome II and III, respectively (Fuchs and Schubert, 1995). The vicilin coding gene (Weschke *et al.*, 1988) was also located on chromosome II, near the centromere (Jiri *et al.*, 1993; Fuchs *et al.*, 1994). While the documented number of genes for *Vf* globulins is relatively small, numerous legumin and vicilin minor subunits with various molecular masses and isoelectric points can be observed in 2D gel electrophoresis analysis (Tucci *et al.*, 1991), suggesting that *Vf* globulins undergo extensive post-translational processing. A similar occurrence has been found in other legumes including *Medicago truncatula* (Le Signor *et al.*, 2017) *and Pisum sativum* (Bourgeois *et al.*, 2009).

There is considerable homology between *Vf* globulin subunits and those of other legumes (**Table S 2.1**), and where genome sequences are available, it is now possible to classify and associate seed storage subunits to specific genome locations (**Table S 2.2**). Considering the lack

of genome sequence for *Vf*, this information is critical for synteny-based mapping of globulin genes and QTLs. For instance, in *M. truncatula,* several genomic regions coding for globulins have been mapped on chromosome I and VII (Le Signor *et al.*, 2017), which are notably syntenic with *Vf* chromosomes III and V (Webb *et al.*, 2016; O'Sullivan and Angra, 2016) where certain legumin A and B genes were previously located, respectively (Fuchs and Schubert, 1995).

2.3.3 Expression of globulin genes

Seed protein content can be thought of as the final output of a number of biochemical and physiological processes occurring throughout the crop life cycle, each of which is under the control of a regulatory network. Abundance of globulin proteins is regulated by a network of genes involving transcriptional regulation, transport, and post-translation modifications of storage proteins (Le Signor *et al.*, 2017). Among these are numerous seed specific genes which play profound regulatory roles in the synthesis and accumulation of seed storage proteins (Le Signor *et al.*, 2017; Zhaoming *et al.*, 2018). Notably, seed specific transcription factors (TFs) such as *ABI5, LEC1, LEC2, ABI3, MYB#2, bHLH#1* and *FUS3* are key storage protein regulators (Verdier and Thompson, 2008; Le Signor *et al.*, 2017). ABA insensitive 5 (*ABI5*) is expressed during seed filling stages in plants (Verdier and Thompson, 2008) and has been found at the centre of the regulatory gene network for storage protein synthesis in *M. truncatula* (Le Signor *et al.*, 2017). Specifically, it is a major regulator for vicilin polypeptide abundance with *P. sativum abi5* mutants showing nearly 30% decrease in the abundance of vicilin-type globulin (Le Signor *et al.*, 2017). Similarly, *ABI3b* and *LEAFY COTYLEDON-1* (*LEC*-1) homologs in soybean have been located at the hub of 118 genes related to seed protein content (Zhaoming *et al.*, 2018). Given the microsynteny between *Vf* and the model crop *M. truncatula* (Webb *et al.*, 2016), these findings will provide a reference for further discoveries in the genetics of *Vf* globulins.

2.3.4 Synthesis and accumulation of seed storage proteins

Globulins are synthesized in the endoplasmic reticulum (ER), sorted in the Golgi body, and transported to the protein storage vacuole (PSV) by vesicles (Mori *et al.*, 2004; Le Signor *et al.*, 2017). During *Vf* seed development, a diphasic pattern of protein accumulation exists in which proteins synthesized during early developmental stages are only transiently accumulated and subsequently degraded to sustain the growing embryo while proteins accumulated after the heart stage (~12 DAP) are mainly destined for storage into protein bodies in the cotyledon (Panitz *et al.*, 1995). During the latter stage, globulin proteins show distinct expression patterns in which vicilin synthesis and accumulation precedes that of legumin and A-type polypeptides of legumin appear earlier than B-type subunit (De Pace *et al.*, 1991). A similar pattern of vicilin and legumin gene expression has been reported in Medicago (Wang *et al.*, 2012) and soybean (Mori *et al.*, 2004).

The amount of protein accumulated during seed development can be attributed to various genetic and environmental factors acting on various plant processes ranging from nutrient uptake and transport, photosynthate production and remobilization to protein accumulation rate in the storage organs. However, there are strong indications that mechanisms underlying nitrogen (N) uptake, transport and assimilation could explain the variation in protein content more than any other factor. For instance, in pea, overexpression of the amino acid transporter gene amino acid permease (*AAP*), has been confirmed to play a critical role in increasing synthesis of seed storage proteins owing to increased leaf and pod phloem loading with free amino acids (Zhang *et al.*, 2015). A similar mechanism could be attributed to the observed 2-3 times higher free amino acids in the cotyledons of high-protein (HP) *Vf* genotypes as compared to low-protein genotypes (Golombek *et al.*, 2001). In rice, a major seed protein content QTL harbouring the *OsAAP* gene was associated with higher uptake of amino acids and their distribution across plant tissues (Peng *et al.*, 2014). In addition, QTL for N-fixation have been linked to QTL for total N accumulation

in common bean (Ramaekers *et al.*, 2013) and pea (Bourion *et al.*, 2010). Thus, improved capacity for N uptake can be a candidate trait to relax the yield-protein negative correlation. In fact, increased genetic capacity for N supply was associated with increased seed size in *Vf* (Rolletschek *et al.*, 2005) or seed number in pea (Zhang *et al.*, 2015). These results should be taken into consideration when screening for high protein content in *Vf*.

2.4 Genetic improvement of protein content and quality

2.4.1 Summary of the past work

Several studies have focused on the genetic variation for protein content (**Table 2.2**) and to what extent protein content was correlated with yield of *Vf*. One of the earliest insights into the genetics of protein content in *Vf* was provided by Picard (1977) who reported that protein content is highly variable (23-40%) with good heritability and additive genetic control as demonstrated by transgressive segregation for protein content in F2 progenies. In addition, according to this study, there was no evidence for negative correlation between protein and grain yield. Similar wide genetic variation was found between and within varieties (n=33) with broad sense heritability of 0.70 and no significant correlation with seed weight (Griffiths and Lawes, 1978). However, when a larger set of germplasm (n=600) was screened, a clear negative relationship between seed weight (g) and protein content (% dry weight) was detected, although some large-seeded genotypes with above average protein content were also found (Lafiandra *et al.*, 1981). Moreover, after four cycles of selection for protein content, Sjödin (1982) concluded that protein content in *Vf* could be improved by selection but tended to negatively correlate with number of seeds per plant regardless of thousand seed weight. More recently, (Skovbjerg *et al.*, 2020) found a moderate negative correlation (r=-0.6) between seed yield (g/m2) and protein content (%) among 17 commercial faba bean cultivars evaluated in four sites in Europe. Additionally, some early efforts have established the variability for S-AA content in *Vf* (**Table 2.3**) and several studies have found a negative correlation between protein and S-AA content

(Lafiandra *et al.*, 1981; Sjödin, 1982; Griffiths, 1984). Under circumstances where desirable traits of interest are negatively correlated, deeper understanding of the genetic basis of the trade-offs between the traits and availability of appropriate tools to dissect and recombine them is crucial.

Table 2.2. Genetic variability in seed protein content in *Vf*

No. genotypes	Range*	Mean	References
33	22-38	29.8	(Griffiths and Lawes, 1978)
600	19-34	29.6	(Lafiandra *et al.*, 1981)
125	22-36	-	(Sjödin, 1982)
125	29-38	29.8	(Frauen *et al.*, 1984)
30	23-39	-	(Griffiths, 1984)
12	26-30	27.5	(Makkar *et al.*, 1997)
74	25-37	31	(Duc *et al.*, 1999)
..	23-40	-	(Picard, 1977)

* Note that in the older literature, most studies have used 6.25 nitrogen-to-protein conversion factor while, in this thesis, we used 5.4 which is considered more accurate for faba bean (Mosse, 1990).

Table 2.3. Genetic variability in sulphur-containing amino acids in *Vf* (g/16 g N)

No. genotypes	Methionine	Cysteine	References
111	0.6-1.0	1.0-1.5	Lafiandra *et al.* (1981)*
125	0.8-1.4	1.3-1.4	Sjödin (1982)*
12	0.8-1.1	1.1-1.4	Makkar *et al.* (1997)
50	0.6 - 0.9	1.0 - 1.4	Schumacher *et al.* (2009)
46	0.6-0.9	0.9-1.2	Schumacher *et al.* (2011)

* S-AA reported as % protein

2.4.2 Areas for future focus

2.4.3 Uncoupling the negative correlation between yield and protein content

Correlation between traits can arise due to gene linkage or pleiotropy (Chen and Lübberstedt, 2010), with the latter being most common in plants, and its resolution requires

deeper understanding of both traits. Therefore, several possible mechanisms have been investigated in various crops in order to unlock the protein-yield association. It is hypothesized that the negative correlation between the two traits results when the high demand for N during seed filling stage coincides with decline in soil nutrients in the rhizosphere and nitrogen fixation, resulting in re-mobilization of nitrogen from leaves, which in turn shortens grain filling and reduces seed weight (Munier-Jolain *et al.*, 2008). This is in line with findings by Egle *et al.* (2015) who showed that the majority of N accumulated during seed filling in barley was remobilized from leaves and stems, but that ongoing N uptake could also contribute. Furthermore, wheat genotypes with a higher capability for post-anthesis N uptake deviate from the grain-protein negative relationship (Bogard *et al.*, 2010; Taulemesse *et al.*, 2016) and selection for this trait has been therefore proposed as a possible criterion for simultaneous improvement of protein content and grain yield. The genetic basis of post-flowering N uptake is not yet fully understood either in cereals or in legumes but could be related to root structure and/or N transport capacity. For instance, pea genotypes with higher mineral nitrogen absorption and symbiotic nitrogen fixation have shown enhanced seed N content and yield (Bourion *et al.*, 2010). Moreover, a faster rate and relatively longer duration of N accumulation during seed development has been reported as a possible mechanism for combining high protein and large seed size in soybean (Poeta *et al.*, 2017). The importance of N uptake capacity for protein content and yield was further demonstrated by Peng *et al.* (2014), who found major protein content QTL *qPC1* harbouring a putative amino acid transporter gene (*OsAAP6),* which they proposed as candidate QTL for simultaneous selection for yield and protein content in rice. These areas of enquiry are amenable for further investigation and can potentially point to QTLs that can be used to improve protein content in *Vf* without significant yield reduction.

2.4.4 Improving S-AA content by modifying legumin: vicilin ratio

Considering difficulties in genetic improvement of limiting amino acids through conventional breeding approaches, several genetic engineering approaches have been attempted in various crops over recent decades. Detailed information on these strategies and results obtained can be found in Galili and Amir (2013). These included (i) overexpression of genes encoding proteins rich in the limiting amino acid, (ii) *in vitro* modification of genes encoding proteins of interest by adding more residues of the desired amino acid, (iii) introduction of genes coding for protein rich in the limiting amino acid from one species to another target food crop, or by (iv) modification of biosynthetic and catabolic pathways to directly increase accumulation of target amino acid or indirectly by increasing accumulation of proteins containing the limiting amino acid. Yet, most of these attempts have not succeeded in producing new crop cultivars combining increased protein quality with desired agronomic traits. In rare cases where reasonable success was achieved, commercialization of the improved cultivars was hindered by legal restrictions on GMO release (Galili and Amir, 2013) and consumer resistance in Europe. Besides these challenges of consumer acceptability, the potential of transgenic approaches in *Vf* is limited by the inherently poor regenerating ability of *Vf* transgenics (Hanafy *et al.*, 2005).

Alternative strategies include direct selection on QTL for S-AA content or indirectly by selecting for greater relative expression of protein subunits rich in S-AA rich subunits. To our knowledge, soybean is the only legume crop in which QTLs for individual S-AA has been mapped (Wang *et al.*, 2014; Warrington *et al.*, 2015). Though total S-AA content of the seed *per se* would be a good indicator, it may not be sufficient when considering selection criteria, due to uncertainty about what percentage of the total S-AA detected is indeed embedded in the main storage proteins. In *Vf* and other legumes, since it is observed that the legumin protein subunit has a relatively higher S-AA content compared to vicilin (Kwanyuen *et al.*, 1997; Jackson *et al.*, 1969 ; Liu *et al.*, 2017; Joshi *et al.*, 2017), increasing amounts of the legumin subunit relative to

vicilin would be expected to enhance protein quality. In fact, the concept of manipulating legumin: vicilin (L/V) ratio to improve nutritional quality is not new in *Vf*. It was previously reported that variation in L/V ratio among varieties was consistent across years (Martensson, 1980) and environments (Gatehouse *et al.*, 1980) and concluded that the L/V ratio has a genetic basis and could be used as a selection criteria to improve nutritional quality in *Vf* (Gatehouse *et al.,* 1980; Martensson, 1980). To our knowledge, since the L/V ratio based approach was suggested as a practical breeding strategy for improving nutritional quality in soybean (Kwanyuen *et al.*, 1997), only one study has tried to map QTLs for L/V ratio and showed co-location between some QTLs for structural legumin and vicilin loci and L/V ratio in soybean (Ma *et al.*, 2016). Recent advances in *Vf* genetics tools such as the development of a 50K SNP genotyping array and high-density linkage map may offer an unprecedented opportunity to discover novel QTLs that could represent targets for improving nutritional quality.

2.4.5 Exploiting mutagenesis approaches

Large-scale mutagenesis using physical or chemical mutagenic agents is a well-established method of inducing novel variation to meet human requirements, but which is unlikely to be present in nature. This approach is more justified in the case of *Vf* where the primary gene pool lacks any known wild relatives. Indeed, several mutagenesis efforts have produced new sets of morphological phenotypes in *Vf* (Sjödin, 1971; Duc, 1995; O'Sullivan and Angra, 2016). However, no data is available on potential beneficial mutations in the seed composition of *Vf*. Although Sjödin (1971) identified some high protein content genotypes from a lot of seeds which had been mutagenized, he could not ascertain whether the selected plants were genuine mutants or randomly isolated extremes in the original seed lot. There are several potential ways of exploiting induced mutations for improving protein content and/or quality. First, desirable mutations involving photosynthetic and N provision mechanisms can improve protein content. From ethyl methane sulfonate (EMS) mutagenized seeds, Duc (1995) discovered a

supernodulating line with 3-4 times higher number of nodules compared to the parental line. However, this line had reduced stem diameter and fewer number of nodes. In general, considering the close relationship between N fixation and protein content, a desirable mutation related to nitrogen fixation capacity and root establishment would be a useful means in breeding for higher protein content. Secondly, knockdown/knockout or regulatory mutations leading to absence of major protein subunits such as vicilins can result in improved nutritional quality by increasing the ratio of S-AA rich subunits like legumin and albumins. Such mutations could be *cis*-linked to the structural loci themselves or *trans*-acting factors that would need to be mapped *de novo*. For instance, mutants of *PsABI5*, a major *trans*-acting regulator of vicilin abundance in pea, have shown an increased legumin abundance (Le Signor *et al.*, 2017). Thirdly, presence or absence of certain subunits can enable dissection of genetic control of individual protein subunits via a QTL mapping approach (Boehm *et al.*, 2017). Lastly, it is possible via a reverse genetic screen to select non-synonymous mutations that convert non-S-AA residues to S-AA residues in S-AA poor storage proteins such as vicilins, although only a proportion of codons are available for single base changes that would result in this outcome. Moreover, the physico-chemical properties of cysteine (disulphide bridge-forming) and methionine (hydrophobic) may cause undesired steric constraints (Brosnan and Brosnan, 2006). However, even a single well-placed additional methionine in each vicilin could give rise to a significant step up in S-AA levels and this approach is therefore worth trying. On a more practical level, full exploitation of mutagenesis for the above purposes requires high-throughput and cheap phenotyping methods to screen tens of thousands of plants for nutritional and agronomic traits.

In summary, *Vf* is one of the most important legumes crops with great potential to fulfil multiple nutritional and ecological services for the current and future generations. However, *Vf* can only play this role if it meets certain producer and end-user expectations which requires plant breeders and the research community to address both agronomic and nutritional constraints simultaneously. In drawing together a synthesis of the literature on *Vf* seed protein content, the

contribution of different storage protein classes to overall abundance and to varying relative amounts of essential amino acids, on globulin structure and globulin-encoding genes, we aim to provide an updated and comprehensive primer for researchers interested in the nutritional optimization of *Vf*. We discuss a range of approaches by which protein content could be increased (without compromising yield) and protein quality ameliorated, some of which have successful precedent in related legume species. These include: high resolution mapping of protein, L:V ration and S-AA QTL using powerful modern quantitative genetics methods and genomics technologies; manipulation of known or still-to-be-discovered structural and regulatory genes by transformation and screening of mutant libraries to reveal novel structural and regulatory variants not found in nature. In parallel, as genome sequencing become cheaper and more genomic resources for *Vf* are accumulated, all the above should become ever more efficient, enhancing the prospects of increasing protein content and quality in this strategic crop.

2.5 References

Abirached-Darmency, M., Dessaint, F., Benlicha, E. & Schneider, C. (2012). Biogenesis of protein bodies during vicilin accumulation in *Medicago truncatula* immature seeds. *BMC Research Notes,* **5**, 409-409.

Baddeley, J. A., Jones, S., Topp, C. F. E., Watson, C. A., Helming, J. & Stoddard, F. L. (2013). Biological nitrogen fixation (BNF) by legume crops in Europe. *Legume Futures Report 1.5.*

Balyan, H. S., Gupta, P. K., Kumar, S., Dhariwal, R., Jaiswal, V., Tyagi, S., Agarwal, P., Gahlaut, V. & Kumari, S. (2013). Genetic improvement of grain protein content and other health-related constituents of wheat grain. *Plant Breeding,* **132** (5), 446-457.

Baumlein, H., Wobus, U., Pustell, J. & Kafatos, F. C. (1986). The legumin gene family: structure of a B type gene of *Vicia faba* and a possible legumin gene specific regulatory element. *Nucleic Acids Research,* **14** (6), 2707-2720.

Boehm, J. D., Nguyen, V., Tashiro, R. M., Anderson, D., Shi, C., Wu, X., Woodrow, L., Yu, K., Cui, Y. & Li, Z. (2018). Genetic mapping and validation of the loci controlling 7S α' and 11S A-type storage protein subunits in soybean [*Glycine max* (L.) Merr.]. *Theoretical and Applied Genetics*, **131**, 659-671.

Bogard, M., Allard, V., Brancourt-Hulmel, M., Heumez, E., Machet, J.-M., Jeuffroy, M.-H., Gate, P., Martre, P. & Le Gouis, J. (2010). Deviation from the grain protein concentration–grain yield negative relationship is highly correlated to post-anthesis N uptake in winter wheat. *Journal of Experimental Botany,* **61** (15), 4303-4312.

Bourgeois, M., Jacquin, F., Savois, V., Sommerer, N., Labas, V., Henry, C. & Burstin, J. (2009). Dissecting the proteome of pea mature seeds reveals the phenotypic plasticity of seed protein composition. *Proteomics,* **9** (2), 254-271.

Bourion, V., Rizvi, S. M. H., Fournier, S., Larambergue, H. d., Galmiche, F., Marget, P., Duc, G. & Burstin, J. (2010). Genetic dissection of nitrogen nutrition in pea through a QTL approach of root, nodule, and shoot variability. *Theoretical and Applied Genetics,* **121**, 71-86.

Brosnan, J. T. & Brosnan, M. E. (2006). The sulfur-containing amino acids: an overview. *The Journal of Nutrition,* **136** (6 Suppl), 1636-1640.

Cernay, C., Pelzer, E. & Makowski, D. (2016). A global experimental dataset for assessing grain legume production. *Scientific Data,* **3**, 160084.

Chen, Y. & Lübberstedt, T. (2010). Molecular basis of trait correlations. *Trends in Plant Science,* **15** (8), 454-461.

Chiari, N. (2017). Food security. The challenge of nutrition in the New Century. *Relations. Beyond Anthropocentrism,* **5** (2), 145-156.

Coda, R., Melama, L., Rizzello, C. G., Curiel, J. A., Sibakov, J., Holopainen, U., Pulkkinen, M. & Sozer, N. (2015). Effect of air classification and fermentation by *Lactobacillus plantarum* VTT E-133328 on faba bean (*Vicia faba* L.) flour nutritional properties. *International Journal of Food Microbiology,* **193**, 34-42.

Coda, R., Varis, J., Verni, M., Rizzello, C. G. & Katina, K. (2017). Improvement of the protein quality of wheat bread through faba bean sourdough addition. *LWT - Food Science and Technology,* **82** (Supplement C), 296-302.

Crépon, K., Marget, P., Peyronnet, C., Carrouée, B., Arese, P. & Duc, G. (2010). Nutritional value of faba bean (*Vicia faba* L.) seeds for feed and food. *Field Crops Research,* **115** (3), 329-339.

Cubero, J. I. (1974). On the evolution of *Vicia faba* L. *Theoretical and Applied Genetics,* **45** (2), 47-51.

de Boer, J. & Aiking, H. (2018). Prospects for pro-environmental protein consumption in Europe: Cultural, culinary, economic and psychological factors. *Appetite,* **121**, 29-40.

De Pace, C., Delre, V., Mugnozza, G. T. S., Maggini, E., Cremonini, R., Frediani, M. & Cionini, P. G. (1991). Legumin of *Vicia faba* major: accumulation in developing cotyledons, purification, mRNA characterization and chromosomal location of coding genes. *Theoretical and Applied Genetics,* **83**, 17-23.

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

de Visser, C. L. M., Schreuder, R. & Stoddard, F. (2014). The EU's dependency on soya bean import for the animal feed industry and potential for EU produced alternatives. OCL, **21** (4), D407.

Duc, G. (1995). Mutagenesis of faba bean (*Vicia faba* L.) and the identification of five different genes controlling no nodulation, ineffective nodulation or supernodulation. *Euphytica,* **83** (2), 147-152.

Duc, G. (1997). Faba bean (*Vicia faba* L). *Field Crops Research,* **53**, 99-109.

Duc, G., Aleksić, J. M., Marget, P., Mikic, A., Paull, J., Redden, R. J., Sass, O., Stoddard, F. L., Vandenberg, A., Vishnyakova, M. & Torres, A. M. (2015). Faba Bean. *In:* Ron, A. M. D. (ed.) *Grain Legumes.* New York: Springer Science+Business Media.

Duc, G., Marget, P., Esnault, R., Le Guen, J. & Bastianelli, D. (1999). Genetic variability for feeding value of faba bean seeds (*Vicia faba*): Comparative chemical composition of isogenics involving zero-tannin and zero-vicine genes. *The Journal of Agricultural Science,* **133** (2), 185-196.

Egle, K., Beschow, H. & Merbach, W. (2015). Nitrogen allocation in barley: Relationships between amino acid transport and storage protein synthesis during grain filling. *Canadian Journal of Plant Science,* **95** (3), 451-459.

FAOstat. (2018). *United Nations Organization for Food and Agriculture.* [20 May 2018].

Fontes, E. P. B., Moreira, M. A., Davies, C. S. & Nielsen, N. C. (1984). Urea-elicited changes in relative electrophoretic mobility of certain glycinin and β-conglycinin subunits. *Plant Physiology,* **76** (3), 840-842.

Foyer, C. H., Lam, H. M., Nguyen, H. T., Siddique, K. H., Varshney, R. K., Colmer, T. D., Cowling, W., Bramley, H., Mori, T. A., Hodgson, J. M., Cooper, J. W., Miller, A. J., Kunert, K., Vorster, J., Cullis, C., Ozga, J. A., Wahlqvist, M. L., Liang, Y., Shou, H., Shi, K., Yu, J., Fodor, N., Kaiser, B. N., Wong, F. L., Valliyodan, B. & Considine, M. J. (2016). Neglecting legumes has compromised human health and sustainable food production. *Nature Plants,* **2**, 16112.

Frauen, M., Röbbelen, G. & Ebrneyer, E. (1984). Quantitative Measurement of Quality Determining Constituents in Seeds of Different Inbred Lines from A World Collection of *Vicia Faba*. *In:* Hebblethwaite, P. D., Dawkins, T. C. K., Heath, M. C. & Lockwood, G. (eds.) *Vicia faba: Agronomy, Physiology and Breeding.* Brussels-Luxembourg: Springer-Science+Business Media, B.V.

Fuchs, J., Joos, S., Licheter, P. & Schubert, I. (1994). Localization of vicilin genes on field bean chromosome II by fluorescent in situ hybridization. *Journal of Heredity,* **85** (6), 487-488.

Fuchs, J. & Schubert, I. (1995). Localization of seed protein genes on metaphase chromosomes of *Vicia faba* via fluorescence in situ hybridization. *Chromosome Research,* **3** (2), 94-100.

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

Gallardo, K., Le Signor, C., Vandekerckhove, J., Thompson, R. D. & Burstin, J. (2003). Proteomics of Medicago truncatula seed development establishes the time frame of

diverse metabolic processes related to reserve accumulation. *Plant Physiology,* **133** (2), 664-682.

Gallo, V., Skorokhod, O. A., Simula, L. F., Marrocco, T., Tambini, E., Schwarzer, E., Marget, P., Duc, G. & Arese, P. (2018). No red blood cell damage and no hemolysis in G6PD-deficient subjects after ingestion of low vicine/convicine *Vicia faba* seeds. *Blood,* **131** (14), 1621-1625.

Gatehouse, J., Croy, R., McIntosh, R., Paul, C. & Boulter, D. (1980). Quantitative and qualitative variation in the storage proteins of material from the EEC joint field bean test. *Quantitative and qualitative variation in the storage proteins of material from the EEC joint field bean test.*, 173-188.

Golombek, S., Rolletschek, H., Wobus, U. & Weber, H. (2001). Control of storage protein accumulation during legume seed development. *Journal of Plant Physiology,* **158** (4), 457-464.

Grebow, J. (2020). Sports Nutrition in 2020: Plant proteins and beyond. *Nutritional Outlook.* Viewed 25 May 2021, https://www.nutritionaloutlook.com/view/sports-nutrition-2020-plant-proteins-and-beyond.

Griffiths, D. W. (1984). An Assesment of the Potential for Improving the Nutritive Value of Field Beans (*Vicia faba*)- A Progress Report. *In:* Hebblethwaite, P. D., Dawkins, T. C. K., Heath, M. C. & Lockwood, G. (eds.) *Vicia faba: Agronomy, Physiology and Breeding.* Brussels-Luxembourg: Springer-Science+Business Media, B.V. .

Griffiths, D. W. & Lawes, D. A. (1978). Variation in the crude protein content of field beans (*Vicia faba* L.) in relation to the possible improvement of the protein content of the crop. *Euphytica,* **27** (2), 487-495.

Hanafy, M., Pickardt, T., Kiesecker, H. & Jacobsen, H.-J. (2005). Agrobacterium-mediated transformation of faba bean (*Vicia faba* L.) using embryo axes. *Euphytica,* **142** (3), 227-236.

Heim, U., Schubert, R., Baumlein, H. & Wobus, U. (1989). The legumin gene family: structure and evolutionary implications of *Vicia faba* B-type genes and pseudogenes. *Plant Molecular Biology,* **13** (6), 653-663.

Henchion, M., Hayes, M., Mullen, A., Fenelon, M. & Tiwari, B. (2017). Future protein spply and demand: strategies and factors influencing a sustainable equilibrium. *Foods,* **6** (7), 53.

Heng, L., van Koningsveld, G. A., Gruppen, H., van Boekel, M. A. J. S., Vincken, J. P., Roozen, J. P. & Voragen, A. G. J. (2004). Protein–flavour interactions in relation to development of novel protein foods. *Trends in Food Science & Technology,* **15** (3), 217-224.

Ismail, B. P., Senaratne-Lenagala, L., Stube, A. & Brackenridge, A. (2020). Protein demand: review of plant and animal proteins used in alternative protein product development and production. *Animal Frontiers,* **10** (4), 53-63.

Jackson, P., Boulter, D. & Thurman, D. A. (1969). A comparison of some properties of vicilin and legumin isolated from seeds of *Pisum sativum*, *Vicia faba* and *Cicer arietinum*. *New Phytologist,* **68** 25-33.

Jensen, E. S., Peoples, M. B. & Hauggaard-Nielsen, H. (2010). Faba bean in cropping systems. *Field Crops Research,* **115** (3), 203-216.

Jiri, M., Winfriede, W., Helmut, B., Uta, P., Andreas, H., Ulrich, W. & Ingo, S. (1993). Localization of vicilin genes via polymerase chain reaction on microisolated field bean chromosomes. *The Plant Journal,* **3** (6), 883-886.

Joshi, J., Pandurangan, S., Diapari, M. & Marsolais, F. (2017). Comparison of Gene Families: Seed Storage and Other Seed Proteins. *In:* Pérez de la Vega, M., Santalla, M. & Marsolais, F. (eds.) *The Common Bean Genome.* Cham: Springer International Publishing.

Kaskinen, T., Lähteenoja, S., Sokero, M. & Suomela, I. (2018). Strategic Business Examples from Finland: The Growth of the Smartup Industry. *In:* Lehmann, H. (ed.) *Factor X: Challenges, Implementation Strategies and Examples for a Sustainable Use of Natural Resources.* Cham: Springer International Publishing.

Kawashima, H., Bazin, M. J. & Lynch, J. M. (2002). A modelling study of world protein supply and nitrogen fertilizer demand in the 21st century. *Environmental Conservation,* **24** (1), 50-57.

Kesari, P., Sharma, A., Katiki, M., Kumar, P., R Gurjar, B., Tomar, S., K Sharma, A. & Kumar, P. (2017). Structural, functional and evolutionary aspects of seed globulins. *Protein and Peptide Letters,* **24** (3), 267-277.

Khazaei, H., Purves, R. W., Song, M., Stonehouse, R., Bett, K. E., Stoddard, F. L. & Vandenberg, A. (2017). Development and validation of a robust, breeder-friendly molecular marker for the *vc*-locus in faba bean. *Molecular Breeding,* **37** (11), 140.

Kimura, A., Fukuda, T., Zhang, M., Motoyama, S., Maruyama, N. & Utsumi, S. (2008). Comparison of physicochemical properties of 7S and 11S globulins from pea, fava bean, cowpea, and rench bean with those of soybean—French bean 7S globulin exhibits excellent properties. *Journal of Agricultural and Food Chemistry,* **56** (21), 10273-10279.

Koivunen, E., Tuunainen, P., Valkonen, E., Rossow, L. & Valaja, J. (2014). Use of faba beans (*Vicia faba* L.) in diets of laying hens. *Agricultural And Food Science,* **23**, 165–172.

Köpke, U. & Nemecek, T. (2010). Ecological services of faba bean. *Field Crops Research,* **115** (3), 217-233.

Krishnan, H. B., Bennett, J. O., Kim, W.-S., Krishnan, A. H. & Mawhinney, T. P. (2005). Nitrogen lowers the sulfur amino acid content of soybean (*Glycine max* [L.] Merr.) by regulating the accumulation of Bowman−Birk protease inhibitor. *Journal of Agricultural and Food Chemistry,* **53** (16), 6347-6354.

Krishnan, H. B., Natarajan, S. S., Oehrle, N. W., Garrett, W. M. & Darwish, O. (2017). Proteomic analysis of pigeonpea (*Cajanus cajan*) seeds reveals the accumulation of numerous stress-related proteins. *Journal of Agricultural and Food Chemistry,* **65** (23), 4572-4581.

Kwanyuen, P., Pantalone, V. R., Burton, J. W. & Wilson, R. F. (1997). A new approach to genetic alteration of soybean protein composition and quality. *Journal of the American Oil Chemists' Society,* **74** (8), 983-987.

Ladjal E, Y., Boudries, H., Mohamed, C. & Romero, A. (2015). *Pea, Chickpea and Lentil Protein Isolates: Physicochemical Characterization and Emulsifying Properties.*

Lafiandra, D., Polignano, G. B., Filippetti, A. & Porceddu, E. (1981). Genetic variability for protein content and S-aminoacids in broad-beans (*Vicia faba* L.). *Die Kulturpflanze,* **29** (1), 115-127.

Le Signor, C., Aimé, D., Bordat, A., Belghazi, M., Labas, V., Gouzy, J., Young, N. D., Prosperi, J.-M., Leprince, O., Thompson, R. D., Buitink, J., Burstin, J. & Gallardo, K. (2017). Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. *New Phytologist,* **214** (4), 1597-1613.

Le Signor, C., Gallardo, K., Prosperi, J. M., Salon, C., Quillien, L., Thompson, R. & Duc, G. (2005). Genetic diversity for seed protein composition in *Medicago truncatula*. *Plant Genetic Resources,* **3** (1), 59-71.

Lessire, M., Gallo, V., Prato, M., Akide-Ndunge, O., Mandili, G., Marget, P., Arese, P. & Duc, G. (2016). Effects of faba beans with different concentrations of vicine and convicine on egg production, egg quality and red blood cells in laying hens. *Animal*, **11**, 1270-1278.

Liu, Y., Wu, X., Hou, W., Li, P., Sha, W. & Tian, Y. (2017). Structure and function of seed storage proteins in faba bean (*Vicia faba* L.). *3 Biotech,* **7** (1), 74.

Ma, Y., Kan, G., Zhang, X., Wang, Y., Zhang, W., Du, H. & Yu, D. (2016). Quantitative trait loci (QTL) mapping for glycinin and beta-conglycinin contents in soybean (*Glycine max* L. Merr.). *Journal of Agricultural and Food Chemistry,* **64** (17), 3473-3483.

Makkar, H. P. S., Becker, K., Abel, H. & Pawelzik, E. (1997). Nutrient contents, rumen protein degradability and antinutritional factors in some colour- and white-flowering cultivars of Vicia faba beans. *Journal of the Science of Food and Agriculture,* **75** (4), 511-520.

Martensson, P. (1980). Variation in legumin : vicilin ratio between and within cultivars of *Vicia faba* L. var. minor. The Hague: Martinus Nijhoff. World crops: production, utilization and description, volume 3, pp. 169-172.

Meng, G. T. & Ma, C. Y. (2001). Thermal properties of Phaseolus angularis (red bean) globulin. *Food Chemistry,* **73** (4), 453-460.

Mertens, C., Dehon, L., Bourgeois, A., Verhaeghe-Cartrysse, C. & Blecker, C. (2012). Agronomical factors influencing the legumin/vicilin ratio in pea (*Pisum sativum* L.) seeds. *Journal of the Science of Food and Agriculture,* **92** (8), 1591-1596.

Mori, T., Maruyama, N., Nishizawa, K., Higasa, T., Yagasaki, K., Ishimoto, M. & Utsumi, S. (2004). The composition of newly synthesized proteins in the endoplasmic reticulum determines the transport pathways of soybean seed storage proteins. *The Plant Journal,* **40** (2), 238-249.

Mosse, J. (1990). Nitrogen-to-protein conversion factor for ten cereals and six legumes or oilseeds. A reappraisal of its definition and determination. Variation according to species and to seed protein content. *Journal of Agricultural and Food Chemistry,* **38** (1), 18-24.

Munier-Jolain, N., Larmure, A. & Salon, C. (2008). Determinism of carbon and nitrogen reserve accumulation in legume seeds. *Comptes Rendus Biologies,* **331** (10), 780-787.

Müntz, K., Horstmann, C. & Schlesier, B. (1999). Vicia globulins. *In:* Shewry, P. R. & Casey, R. (eds.) *Seed Proteins.* Dordrecht: Springer Netherlands.

O'Kane, F. E., Happe, R. P., Vereijken, J. M., Gruppen, H. & van Boekel, M. A. J. S. (2004). Characterization of pea vicilin. 1. Denoting convicilin as the α-subunit of the Pisum vicilin family. *Journal of Agricultural and Food Chemistry,* **52** (10), 3141-3148.

O'Sullivan, D. M. & Angra, D. (2016). Advances in faba bean genetics and genomics. *Frontiers in Genetics,* **7**, 150.

Panitz, R., Borisjuk, L., Manteuffel, R. & Wobus, U. (1995). Transient expression of storage-protein genes during early embryogenesis of *Vicia faba*: synthesis and metabolization of vicilin and legumin in the embryo, suspensor and endosperm. *Planta,* **196** (4), 765-774.

Patil, G., Mian, R., Vuong, T., Pantalone, V., Song, Q., Chen, P., Shannon, G. J., Carter, T. C. & Nguyen, H. T. (2017). Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. *Theoretical and Applied Genetics,* **130** (10), 1975-1991.

Peng, B., Kong, H., Li, Y., Wang, L., Zhong, M., Sun, L., Gao, G., Zhang, Q., Luo, L., Wang, G., Xie, W., Chen, J., Yao, W., Peng, Y., Lei, L., Lian, X., Xiao, J., Xu, C., Li, X. & He, Y. (2014). *OsAAP6* functions as an important regulator of grain protein content and nutritional quality in rice. *Nature Communications,* **5**, 4847.

Perez-Maldonado, R. A., Mannion, P. F. & Farrell, D. J. (1999). Optimum inclusion of field peas, faba beans, chick peas and sweet lupins in poultry diets. I. Chemical composition and layer experiments. *British Poultry Science,* **40** (5), 667-673.

Picard, J. (1977). Some results dealing with breeding for protein content in *Vicia faba* L. *In:* Protein quality from leguminous crops. Dijon (France). 3 Nov 1976.

Poeta, F., Ochogavia, A. C., Permingeat, H. R. & Rotundo, J. L. (2017). Storage-associated genes and reserves accumulation in soybean cultivars differing in physiological strategies for attaining high seed protein concentration. *Crop Science,* **57** (1), 427-436.

Poysa, V., Woodrow, L. & Yu, K. (2006). Effect of soy protein subunit composition on tofu quality. *Food Research International,* **39** (3), 309-317.

Ramaekers, L., Galeano, C. H., Garzón, N., Vanderleyden, J. & Blair, M. W. (2013). Identifying quantitative trait loci for symbiotic nitrogen fixation capacity and related traits in common bean. *Molecular Breeding,* **31** (1), 163-180.

Rerie, W. G., Whitecross, M. & Higgins, T. J. V. (1991). Developmental and environmental regulation of pea legumin genes in transgenic tobacco. *Molecular and General Genetics MGG,* **225** (1), 148-157.

Rizzello, C. G., Losito, I., Facchini, L., Katina, K., Palmisano, F., Gobbetti, M. & Coda, R. (2016). Degradation of vicine, convicine and their aglycones during fermentation of faba bean flour. *Scientific Reports,* **6**, 32452.

Rizzello, C. G., Verni, M., Koivula, H., Montemurro, M., Seppa, L., Kemell, M., Katina, K., Coda, R. & Gobbetti, M. (2017). Influence of fermented faba bean flour on the nutritional, technological and sensory quality of fortified pasta. *Food & Function,* **8** (2), 860-871.

Rolletschek, H., Hosein, F., Miranda, M., Heim, U., Gotz, K. P., Schlereth, A., Borisjuk, L., Saalbach, I., Wobus, U. & Weber, H. (2005). Ectopic expression of an amino acid transporter (*VfAAP1*) in seeds of *Vicia narbonensis* and pea increases storage proteins. *Plant Physiology,* **137** (4), 1236-1249.

Rubio, L. A., Perez, A., Ruiz, R., Guzman, M. A., Aranda-Olmedo, I. & Clemente, A. (2014). Characterization of pea (*Pisum sativum*) seed protein fractions. *Journal of the Science of Food and Agriculture,* **94** (2), 280-287.

Schumacher, H., Paulsen, H. M. & Gau, A. E. (2009). Phenotypical indicators for the selection of methionine enriched local legumes in plant breeding. *Agriculture and Forestry Research* **4**(59), 339-344.

Schumacher, H., Paulsen, H. M., Gau, A. E., Link, W., Jurgens, H. U., Sass, O. & Dieterich, R. (2011). Seed protein amino acid composition of important local grain legumes *Lupinus angustifolius* L., *Lupinus luteus* L., *Pisum sativum* L. and *Vicia faba* L. *Plant Breeding,* **130** (2), 156-164.

Shewry, P. R. & Casey, R. (1999). Seed Proteins. *In:* Shewry, P. R. & Casey, R. (eds.) *Seed Proteins.* Dordrecht: Springer Netherlands, pp. 1-10.

Sjödin, J. (1971). Induced morphological variation in *Vicia faba* L. *Hereditas,* **67** (2), 155-179.

Sjödin, J. (1982). Protein Quantity and Quality in *Vicia Faba*. *In:* Hawtin, G. & Webb, C. (eds.) *Faba Bean Improvement: Proceedings of the Faba Bean Conference held in Cairo, Egypt, March 7–11, 1981.* Dordrecht: Springer Netherlands, pp. 319-331.

Skovbjerg, C. K., Knudsen, J. N., Füchtbauer, W., Stougaard, J., Stoddard, F. L., Janss, L. & Andersen, S. U. (2020). Evaluation of yield, yield stability, and yield–protein relationship in 17 commercial faba bean cultivars. 2 (3), e39.

Smýkal, P., Coyne, C. J., Ambrose, M. J., Maxted, N., Schaefer, H., Blair, M. W., Berger, J., Greene, S. L., Nelson, M. N., Besharat, N., Vymyslický, T., Toker, C., Saxena, R. K., Roorkiwal, M., Pandey, M. K., Hu, J., Li, Y. H., Wang, L. X., Guo, Y., Qiu, L. J., Redden, R. J. & Varshney, R. K. (2015). Legume crops phylogeny and genetic diversity for science and breeding. Critical Reviews in Plant Sciences, 34 (1-3), 43-104.

Speedy, A. W. (2004). Overview of world feed protein needs and supply. *In:* FAO Animal Production and Health Proceedings (FAO), 2004. FAO, pp. 9-27.

Statista. (2017). *Meat consumption and vegetarianism in Europe - Statistics and Facts.* Viewed 25 December 2017, https://www.statista.com/topics/3345/meat-consumption-and-vegetarianism-in-europe/.

Tanno, K.-i. & Willcox, G. (2006). The origins of cultivation of Cicer arietinum L. and Vicia faba L.: early finds from Tell el-Kerkh, north-west Syria, late 10th millennium b.p. Vegetation History and Archaeobotany, 15 (3), 197-204.

Taulemesse, F., Le Gouis, J., Gouache, D., Gibon, Y. & Allard, V. (2016). Bread wheat (Triticum aestivum L.) grain protein concentration is related to early post-flowering nitrate uptake under putative control of plant satiety level. PLOS ONE, 11 (2), e0149668.

Torres, A. M., Avila, C. M., Stoddard, F. L. & Cubero, J. I. (2012). Faba Bean. In: Vega, M. P. d. l., Torres, A. M., Cubero, J. I. & Kole, C. (eds.) Genetics, genomics and breeding of cool season grain legumes. USA: Taylor & Francis Group, pp. 50-97.

Tucci, M., Capparelli, R., Costa, A. & Rao, R. (1991). Molecular heterogeneity and genetics of Vicia faba seed storage proteins. Theoretical and Applied Genetics, 81 (1), 50-58.

Utsumi, S., Yokoyama, Z.-i. & Mori, T. (1980). Comparative studies of subunit compositions of legumins from various cultivars of Vicia faba L. seeds. Agricultural and Biological Chemistry, 44 (3), 595-601.

Verdier, J. & Thompson, R. D. (2008). Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. *Plant and Cell Physiology,* **49** (9), 1263-1271.

Vilariño, M., Métayer, J. P., Crépon, K. & Duc, G. (2009). Effects of varying vicine, convicine and tannin contents of faba bean seeds (*Vicia faba* L.) on nutritional values for broiler chicken. *Animal Feed Science and Technology,* **150** (1-2), 114-121.

VTT. (2014). *Gluten-free faba bean for bread and pasta.* Viewed 25 July 2018, https://www.foodingredientsfirst.com/news/gluten-free-faba-bean-for-bread-and-pasta.html.

Wang, X., Jiang, G.-L., Song, Q., Cregan, P. B., Scott, R. A., Zhang, J., Yen, Y. & Brown, M. (2014). Quantitative trait locus analysis of seed sulfur-containing amino acids in two recombinant inbred line populations of soybean. *Euphytica,* **201** (2), 293-305.

Wang, X. D., Song, Y., Sheahan, M. B., Garg, M. L. & Rose, R. J. (2012). From embryo sac to oil and protein bodies: embryo development in the model legume *Medicago truncatula*. *New Phytologist,* **193** (2), 327-338.

Warrington, C. V., Abdel-Haleem, H., Hyten, D. L., Cregan, P. B., Orf, J. H., Killam, A. S., Bajjalieh, N., Li, Z. & Boerma, H. R. (2015). QTL for seed protein and amino acids in the Benning x Danbaekkong soybean population. *Theoretical and Applied Genetics,* **128** (5), 839-850.

Webb, A., Cottage, A., Wood, T., Khamassi, K., Hobbs, D., Gostkiewicz, K., White, M., Khazaei, H., Ali, M., Street, D., Duc, G., Stoddard, F. L., Maalouf, F., Ogbonnaya, F. C., Link, W., Thomas, J. & O'Sullivan, D. M. (2016). A SNP-based consensus genetic map for synteny-based trait targeting in faba bean (*Vicia faba* L.). *Plant Biotechnology Journal,* **14** (1), 177-185.

Weschke, W., Bassüner, R., Van Hai, N., Czihal, A., Baümlein, H. & Wobus, U. (1988). The structure of a *Vicia faba* vicilin gene. *Biochemie und Physiologie der Pflanzen,* **183** (2-3), 233-242.

WHO/FAO/UNU (2007). Protein and amino acid requirements in human nutrition. *WHO Technical Report Series,* (935), 1-265.

Young, V. R. & Pellett, P. L. (1994). Plant proteins in relation to human protein and amino acid nutrition. *The American Journal of Clinical Nutrition,* **59** (5), 1203S-1212S.

Yu, E.-M., Zhang, H.-F., Li, Z.-F., Wang, G.-J., Wu, H.-K., Xie, J., Yu, D.-G., Xia, Y., Zhang, K. & Gong, W.-B. (2017). Proteomic signature of muscle fibre hyperplasia in response to faba bean intake in grass carp. *Scientific Reports,* **7**, 45950.

Zhang, L., Garneau, M. G., Majumdar, R., Grant, J. & Tegeder, M. (2015). Improvement of pea biomass and seed productivity by simultaneous increase of phloem and embryo loading with amino acids. *The Plant Journal,* **81** (1), 134-146.

Zhaoming, Q., Zhanguo, Z., Zhongyu, W., Jingyao, Y., Hongtao, Q., Xinrui, M., Hongwei, J., Dawei, X., Zhengong, Y., Rongsheng, Z., Chunyan, L., Wei, Y., Zhenbang, H., Xiaoxia, W., Jun, L. & Qingshan, C. (2018). Meta‐analysis and transcriptome profiling reveal

hub genes for soybean seed storage composition during seed development. *Plant, Cell & Environment,* **41**, 2109–2127..

2.6 Supplementary

Table S 2.1. List of protein accessions of legumes with high similarity to *Vf* legumin/vicilin-like subunit

Legume species	Legumin-like		Vicilin-like
	Legumin A	Legumin B	
Cajanus cajan	KYP70740.1	KYP44257.1	A0A151S2A5
Cicer arietinum	XP_012569358.1	XP_004495100.1	A0A1S2XQR
Glycine max	P11828.1	G4_P02858.1	NP_001236872.2
Lupinus angustifolius	AEB33709.1	XP_019429051.1	F5B8V9
Medicago truncatula	XP_013449900.1	XP_003590689.1	Q2HW19
Pisum sativum	P15838.1	P05692.1	P13918.2
Vicia faba	CAA38758.1	P05190.1	P08438

Table S 2.2. Gene models associated with legumin/vicilin-like subunits in model legumes

Model crops	Legumin like subunits		Vicilin-like subunits
	Legumin A	Legumin B	
Cajanus cajan	C.cajan_09695 C.cajan_09691	C.cajan_34796	C.cajan_07496 C.cajan_28781
Cicer arietinum	Ca_12229	Ca_07751	Ca_06137 Ca_06139 Ca_06135
Glycine max	Glyma.19G164900.1 Glyma.03G163500.1	Glyma.10G037100.1	Glyma.20G148200.1 Glyma.10G246300.1 Glyma.20G148300.1 Glyma.20G148400.1
Lupinus angustifolius	Lup028353	Lup032393	Lup027356 Lup019231 Lup029350 Lup015052
Medicago truncatula	Medtr7g096970	Medtr1g072630 Medtr1g072600 Medtr1g072610	Medtr7g079770 Medtr7g079730 Medtr7g079780 Medtr7g079820

*Data source: https://legumeinfo.org

Table S 2.3. Description of the models used to predict structure of *Vf* globulin subunits

Vf globulins	Template accession	Description	Sequence similarity	Seq. identity	Seq. coverage	GMQE	QMEAN
Vicilin	1ipk.1.C	Beta-conglycinin, beta chain	0.46	56.76	0.89	0.71	-1.64
Legumin A-type	3ksc.1. A	LegA class	0.58	89.24	0.95	0.71	-1.42
Legumin B-type	2d5h.2.A	Glycinin A3B4 subunit	0.50	63.93	0.90	0.68	-2.20

Chapter 3 Identification and quantification of major faba bean seed proteins

Ahmed O. Warsame, Nicholas Michael, Donal M. O'Sullivan and Paola Tosi

[Published in *J. Agric. Food Chem*. (2020),32, 8535-8544]

3.1 Abstract

Faba bean (*Vicia faba* L.) holds great importance for human and animal nutrition for its high protein content. However, better understanding of its seed protein composition is required in order to develop cultivars that meet market demands for plant proteins with specific quality attributes. In this study, we screened 35 diverse *V. faba* genotypes by employing one-dimensional sodium dodecyl sulphate–polyacrylamide gel electrophoresis (1D SDS–PAGE) method, and 35 major protein bands obtained from three genotypes with contrasting seed protein profiles were further analysed by mass spectrometry (MS). Twenty-five of these protein bands (MW range: ~9-107 kDa) had significant ($p \leq 0.05$) matches to polypeptides in protein databases. MS analysis showed that most of the analysed protein bands contained more than one protein type and, in total, over 100 proteins were identified. These included major seed storage protein such as legumin, vicilin and convicilin, as well as other protein classes like lipoxygenase, heat shock proteins, sucrose-binding proteins, albumin, and defensin. Furthermore, seed protein extracts were separated by size-exclusion high-performance liquid chromatography (SE-HPLC), and percentages of the major protein classes were determined. On average, legumin and vicilin/convicilin accounted for 50 and 27% of the total protein extract, respectively. However, the proportions of these proteins varied considerably among genotypes, with the ratio of legumin:vicilin/convicilin ranging from 1:1 to 1:3. In addition, there was a significant ($p < 0.01$) negative correlation between the contents of these major fractions ($r = -0.83$). This study significantly extends the number of identified *V. faba* seed proteins and reveals new qualitative and quantitative variation in seed protein composition, filling a significant gap in the literature. Moreover, the germplasm and screening methods presented here are expected to contribute to selecting varieties with improved protein content and quality.

KEY WORDS: *Vicia faba*; legumin; vicilin; protein quantification; SE-HPLC

3.2 Introduction

Vicia faba (hereafter *Vf*) seeds contain about 29% protein (Warsame *et al.*, 2018) and the crop is well adapted to various climates and is grown for both human and animal nutrition (Duc *et al.*, 2015; Multari *et al.*, 2015). Given its high yield potential (Cernay *et al.*, 2016) and unparalleled nitrogen fixation capacity (Baddeley *et al.*, 2013), *Vf* is among the few crops with great potential to meet the dietary needs of the growing human population while maintaining sustainability of agricultural production systems (Foyer *et al.*, 2016). Much of the research on seed quality to date has focused on the reduction or removal of anti-nutrients, namely vicine and convicine (Khazaei *et al.*, 2019; Khazaei *et al.*, 2017; Khazaei *et al.*, 2015) and seed coat tannins (Gutierrez *et al.*, 2008; Webb *et al.*, 2016; Zanotto *et al.*, 2019), with surprisingly little effort dedicated to improving the protein composition.

Utilization of plant proteins for human or animal nutrition is largely determined by the nutritional and functional properties of their constituent protein classes. It is estimated that *Vf* seed proteins contain ~80% globulin which in turn comprises legumin and vicilin/convicilin, known in the older literature (Chakraborty *et al.*, 1979) by their ultracentrifugation sedimentation coefficients as 11S and 7S respectively. Globulins belong to the cupin superfamily (Dunwell *et al.*, 2004), and the legumin and vicilin types have a high degree of structural homology (Kesari *et al.*, 2017; Fukushima, 1991). Legumin is the major *Vf* seed protein and is estimated to represent about 50% of the storage proteins (Müntz *et al.*, 1999; Horstmann *et al.*, 1993). It is encoded by multiple genes belonging to type-A (Methionine-containing) and type-B (Methionine-lacking) subunits (Horstmann *et al.*, 1993). Only few genes encoding type-A (A1 and A2), type-B (LeB2, LeB4, LeB6, and LeB7) and one high-molecular mass legumin polypeptide (LeB3) have been described (Fuchs and Schubert, 1995; Horstmann *et al.*, 1993; Baumlein *et al.*, 1986). However, Tucci *et al.* (1991) reported 29 biochemically distinct disulphide-linked αβ legumin subunit pairs with molecular weights between 39-81 kDa,

suggesting that the number of legumin-encoding genes could be much higher than is currently known. Vicilin is also a heterogeneous protein in its native trimer form (Tucci *et al.*, 1991). Regarding convicilin subunits, at least two structural genes have been described (Sáenz de Miera *et al.*, 2008), though the question of whether convicilin can be considered a vicilin subunit or a distinct class of globulin is yet to be resolved in *Vf*.

The relationship between subunit composition of major storage proteins and the overall seed protein quality has been studied in other legumes like soybean (Poysa *et al.*, 2006), where molecular markers for specific legumin and vicilin-like subunit variants with desirable qualities have been developed (Boehm *et al.*, 2017). In *Vf*, it is generally accepted that selection for a higher legumin: vicilin ratio could enhance its nutritional quality (Warsame *et al.*, 2018; Martensson, 1980; Gatehouse *et al.*, 1980), since some major legumin subunits contain relatively higher proportions of sulphur-containing amino acids (S-AA) compared to vicilin. However, given the underlying genetic complexity of these broadly defined classes of storage proteins, concrete exploitation of genetic variation in seed protein composition for the development of cultivars with improved protein profiles would require identification of the genes encoding the major seed storage proteins, as well as understanding the genetic basis for their abundance in seeds. To date, studies have referred to just a few major protein subunits of legumin and vicilin (Tucci *et al.*, 1991; Utsumi *et al.*, 1980; Martensson, 1980; Gatehouse *et al.*, 1980; Müntz *et al.*, 1999), and although Liu *et al.* (2017) identified several additional non-globulin seed storage proteins from *Vf* by mass spectrometry, the identification of the full set of proteins that contribute to the nutritional and functional properties of the *Vf* seed is far from complete.

The one-dimensional sodium dodecyl sulphate–polyacrylamide gel electrophoresis (1D SDS-PAGE) method has been exploited in the qualitative and quantitative analysis of protein composition in various legume species (Panthee *et al.*, 2004; Tzitzikas *et al.*, 2006; Le Signor *et al.*, 2017; Boehm *et al.*, 2017). However, the main problem inherent in this method is that,

depending on the particular electrophoresis conditions used, unrelated proteins of similar mobility can partially or completely overlap, which can lead to over or underestimation of certain proteins. An alternative method of protein separation based on size-exclusion high-performance liquid chromatography (SE-HPLC) has been widely used in studying seed proteins of wheat, notably in determining the proportions of gliadin and glutenin fractions associated with certain quality attributes, including pasta-cooking and bread-making qualities (Ohm *et al.*, 2017; Ohm *et al.*, 2009; Larroque and Bekes, 2000; Oomah *et al.*, 1994). The advantage of this method is that proteins can be quantified in their native condition and the sample analysis is amenable to automation.

In this study, our aim was to (1) assess the diversity in subunit composition of major *Vf* seed proteins in genetically diverse germplasm; (2) accurately identify the most abundant seed proteins; and (3) quantify the proportions of legumin and vicilin/convicilin proteins using a panel of diverse *Vf* genotypes.

3.3 Materials and methods

3.3.1 Reagents

Sodium phosphate, calcium chloride, trichloroacetic acid, dithiothreitol, iodoacetamide, triethylammonium bicarbonate, and a Bradford assay reagent were obtained from Sigma-Aldrich (UK). PageBlue®, NuPAGE LDS sample buffer, and NuPAGE MES SDS running buffer were sourced from Thermo Fisher Scientific (UK). Acetonitrile, sulphuric acid, and HPLC grade water were from Fisher Scientific (UK). A sequence grade porcine trypsin enzyme was obtained from Promega (UK).

3.3.2 Plant materials

Thirty-five *Vf* genotypes, including inbred lines derived from breeding materials, landraces and cultivars from different locations around the world, were used in this study for 1D SDS-

PAGE and SE-HPLC protein subunit profiling (**Table 3.1**). This genetically diverse population contained genotypes collected by the University of Reading (UK), the Agricultural Research Centre (Egypt), Nordic Seeds (Denmark) and the University of Saskatchewan (Canada). The majority of these genotypes are parents of *Vf* mapping populations which already exist (Khazaei *et al.*, 2018) or are currently under development (**Table 3.1**). Except two genotypes, L170 and L43, all genotypes were previously grown in the same glasshouse at University of Reading.

3.3.3 Total protein extraction

Five to 10 seeds per genotype were dried in an oven at 80 °C for 48-hours and ground using a Laboratory Mill 3303 (Perten Instruments, Warrington, UK). The flour was then sieved through a 1 mm diameter sieve to obtain a homogenous sample. Total seed proteins were extracted according to the procedure reported by Mertens *et al.* (2012) with some modifications. Briefly, we used 0.1 M phosphate buffer (pH 7.2) containing 5 g L^{-1} of potassium sulphate with a sample/buffer ratio of 1:10 (w/v). Samples were vortexed briefly and stirred for 30 minutes at 300 rpm followed by centrifugation at 20,000 × g for 30 minutes at room temperature. The supernatant was then transferred to a new tube and stored at -20 ºC until further analysis. The protein concentration in protein extracts was measured using the Bradford method (Bradford, 1976) with a SpectraMax i3x microplate reader (Molecular Devices, UK).

3.3.4 Protein fractionation

The total seed protein extracts were fractionated by sequential extraction in aqueous and salt solutions to obtain fractions enriched for water-soluble and salt-soluble proteins (for details, see **Figure S 3.1**). The globulin precipitation step was conducted according to the procedure reported by Krishnan *et al.* (2009). A total of five protein fractions (hereafter F1-5) were obtained: water soluble (F1), globulin-depleted water soluble (F2), salt-soluble (F3), globulin-depleted salt-soluble (F4), and globulin-enriched fraction (F5). These fractions were then analysed by SE-HPLC and SDS–PAGE.

Table 3.1. List of *Vf* genotypes used for protein subunit diversity and quantification

Genotype	Original Source	Germplasm Category	Country*	Populations founded*
LG Cartouche	-	cultivar	UK	-
Lynx	-	cultivar	UK	-
Vertigo	-	cultivar	UK	RSBP
Wizard	-	cultivar	UK	-
Fanfare	-	cultivar		RSBP
Icarus	Icarus	ILC	Ecuador	7-way MAGIC; Icarus × Ascot
NV640	Maris Bead	ILC	UK	RSBP
NV643	Albus	ILC	Poland	Albus×BPL10; RSBP; 7-way MAGIC
NV672	Betty	ILC		RSBP
NV866	Disco/2	ILC	France	Hedin/2×Disco/2; 4WP; RSBP
NV639-2	Hedin	ILC	Germany	RSBP
RV501	Robin Hood	ILC	UK	-
RV502	The Sutton	ILC	UK	-
RV503	Casata Midwinter	ILC	UK	RSBP
RV504	Crimson Flowered-3	ILC	UK	RSBP
RV505	Diana	ILC	Canada	7-way MAGIC
RV506	Cuscan Super Yellow-1	PILC	Peru	RSBP
RV507	Iantos-3	PILC	Peru	RSBP
RV508	Mustard Yellow	PILC	Peru	RSBP
RV509	Sakha4	ILC	Egypt	RSBP
RV510	Nubaria3	ILC	Egypt	RSBP
RV511	Misr3	ILC	Egypt	RSBP
RV512	Giza716	ILC	Egypt	RSBP
NV735	Mélodie	ILC	France	Melodie×ILB938-2; RSBP
RV319-2	-	inbred line	UK	-
NV153	ig12658	ILL	Ethiopia	-
NV648-1	BPL10	inbred line	Unknown	Albus×BPL10; RSBP
NV734	ILB938-2	inbred line	Colombia	Melodie×ILB938-2; 4WP; 7-way MAGIC
NV657	INRA 29H	inbred line	France	RSBP
L170	ig132238	inbred line	China	4WP
NV651-3	BPL21	inbred line	Unknown	RSBP
NV658-2	CGN07715 cf-3	inbred line	Unknown	-
L43	ig114476	inbred line	Bangladesh	4WP
RV322	HEL170	inbred line	China	RSBP
NV873-13	F5 from NV644xNV153	RIL	Unknown	RSBP

*Country of release (for cultivars) or collection (landrace materials),**RSBP: Reading Spring Bean Population (currently under development); 7-way MAGIC: Multiparent advanced generation intercross (under development); 4WP: 4-Way cross population (Khazaei *et al*. 2018). ILC= inbred line from cultivar, ILL= inbred line from landrace, RIL= recombinant inbred line, PILC=pure inbred line from cultivar.

3.3.5 Protein and Sulphur content analysis

Nitrogen and sulphur contents (%) were determined using an isotope ratio mass spectrometer (DELTA V™ IRMS, Thermo Fisher, UK). The analysis was carried out in duplicate using oven-dried flours of ~1 mg. Nitrogen content data were then converted to protein content as: protein (%) = %N × 5.4 (Mosse, 1990).

3.3.6 1D SDS–PAGE Analysis

One-dimensional SDS–PAGE analysis of the total protein extract (~15 µg per well) was performed using NuPAGE 10% Bis–Tris precast gels. Before gel loading, the samples were mixed with the NuPAGE LDS sample buffer and sample reducing agent following manufacturer's instructions. Gels were run in NuPAGE MES SDS buffer in an XCell SureLock™ Mini-Cell at a constant current of 70 mA and a maximum voltage of 200 V for 1 hour. Before staining, the gels were fixed with 12% trichloroacetic acid for 15 minutes and washed twice with 250 mL of deionized water for another 15 minutes on a rocker (SSL3, Stuart, UK). Gels were then stained with 50 mL of a PageBlue® protein staining solution for two hours followed by destaining overnight with deionized water.

3.3.7 Identification of major seed protein subunits

3.3.8 In-gel protein digestion

Individual protein bands were carefully excised from gel lanes of the selected genotypes (LG Cartouche, NV657 and NV734) and were destained in 0.6 mL tubes with 400 µL 50% acetonitrile (MeCN) and 50% 10 mM triethylammonium bicarbonate (TEAB) overnight. Gel pieces were then reduced with 10 mM dithiothreitol (DTT) in 10 mM TEAB for 30 minutes at 50 °C, followed by alkylation with 50 mM iodoacetamide in 10 mM TEAB for 30 minutes in the dark. After washing three times with 400 µL of 10 mM TEAB and once with MeCN, the dehydrated gel samples were resuspended in 10 µL of 10 mM TEAB containing 200 ng of porcine trypsin and incubated at 25 °C overnight. The gel digests were placed on dry ice for five

minutes, then allowed to thaw, and 30 µL of 10% MeCN/5% formic acid was added. After 15 minutes of sonication, peptide extracts were transferred to 250 µL PCR tubes. This step was repeated twice, and the resultant extract was pooled and dried in a centrifugal vacuum concentrator.

3.3.9 Mass spectrometry analysis

The dried peptides were resuspended in 20 µL of LC–MS buffer A (0.1% formic acid in water) and 10 µL sample was injected into an Ace C18 column (150 x 2.1 mm, 5 µM particle size with 300 Å pore size) and analysed by LC-MS on a Thermo Scientific LTQ-Orbitrap XL interfaced with an Accela HPLC instrument. Buffer B was 0.1% formic acid in MeCN. The gradient was as follows: 0–2 minutes; 5% B, 20 mins; 60% B, 20.1–23 minutes; 80% B, 23.1–30 minutes; 5% B. The column oven was at 30 °C and at 15 °C for the autosampler. The first two minutes and the last six minutes of each run were excluded from the analysis.

A data-dependent acquisition (DDA) strategy was employed. In brief, ions were measured using the Orbitrap at 30,000 resolution, scanning from 400–2000 m/z. Three ions from each MS1 scan that were most abundant and multiply charged were chosen for MS2. MS2 was performed using collision-induced disassociation (CID) in the ion trap and scanned out at a unit resolution. The acquired data were analysed using an in-house version of MASCOT search engine (Matrix Science, UK) via Mascot Daemon with file conversion performed using ProteoWizard. The acquired MS spectra were searched against the NCBI non-redundant protein (https://www.ncbi.nlm.nih.gov) common Repository of Adventitious Proteins (ftp://ftp.thegpm.org/fasta/cRAP), and other contaminant databases.

The search parameters for MASCOT search were set as follows: type of search = MS/MS ion search, enzyme = trypsin, variable modifications = acetyl (protein N–term), carbamidomethyl (C), Gln- > pyro-Glu (N-term Q), oxidation (M), mass values = monoisotopic,

protein mass = unrestricted, peptide mass tolerance = ±10 ppm, fragment mass tolerance = ±1 Da, max missed cleavages = 2, and instrument type = ESI–TRAP.

3.3.10 Protein composition analysis by SE-HPLC

Size-exclusion HPLC analysis was conducted with the Waters Alliance 2695 Separations Module using a Phenomenex BioSep SEC S–2000 column with silica resin (300 × 7.8 mm, 5 µm particle size and 145 Å pore size). The same extraction buffer (0.1 M phosphate buffer containing 5 g L-1 of potassium sulphate, pH = 7.2) was used as a mobile phase with a flow rate of 0.5 mL min^{-1}. The injection volume of protein sample was 20 µl and was detected at 210 nm using the Waters® 2996 photodiode array (PDA) detector. Two technical replicates were analysed for each genotype and the raw chromatogram data were exported for peak integration in Origin software (OriginLab Corporation, Northampton, MA, USA).

3.4 Results and Discussion

3.4.1 A comprehensive survey of *Vf* seed proteins

In order to capture the most common seed protein variants, we first conducted a preliminary 1D SDS-PAGE screening of 35 diverse genotypes for their seed protein profiles (**Figure S 3.2**). From this analysis, we identified three genotypes—LG Cartouche, NV657 and NV734—with distinct protein profiles (**Figure 3.1**) and used them for protein band identification. Forty-six bands, with apparent molecular weights (MW) ranging from less than 10 to ~145 kDa on reducing 1D SDS-PAGE gels, were detected collectively from these three genotypes. Thirty-five of these bands were excised from the gel and subjected to mass spectrometry analysis with 25 of them reporting significant ($p \leq 0.05$) matches with proteins in the database, mainly from *Vf* and related legumes (**Table 3.2**). Failure to identify the remaining 10 bands can be explained in terms of their relatively lower abundance which made it technically challenging to elute enough protein for the MS analysis.

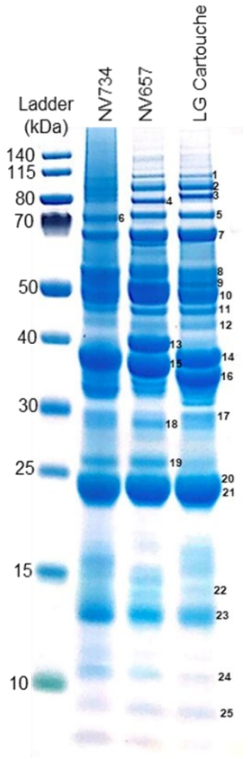

Figure 3.1. SDS-PAGE profile of three *Vf* genotypes with distinct seed protein profiles which were used for seed protein identification by mass spectrometry analysis.

Table 3.2. Major proteins identified by mass spectrometry analysis of protein bands excised from reducing SDS-PAGE gel of *Vf* seed proteins in figure 3.1.

SDS-PAGE band	Band apparent MW (kDa)	Gene bank accession	Score	Num. of significant sequences	emPAI	Description	Species
1	106.9	gi|126405	565	15	1.08	seed linoleate 9S-lipoxygenase-3	*Pisum sativum*
		gi|164512572	128	2	0.18	convicilin	*Vf*
2	96.3	gi|126405	508	15	1.18	seed linoleate 9S-lipoxygenase-3	*P. sativum*
		gi|164512572	178	4	0.39	convicilin	*Vf*
3	88.8	gi|164512572	120	2	0.18	convicilin	*Vf*
		gi|187766747	99	1	0.26	Gly m Bd 28K allergen	*Glycine max*
4	83	gi|164512572	165	6	0.68	convicilin	*Vf*
		gi|22053	154	9	1.34	vicilin: Precursor	*Vf*
5	75.2	gi|357480003	391	8	0.81	heat shock 70 kDa protein	*Medicago truncatula*
		gi|126162	94	4	0.74	legumin type B	*Vf*
6	73.1	gi|562006	364	12	1.26	PsHSP71.2	*P. sativum*
		gi|164512572	123	4	0.4	convicilin	*Vf*
7	64.7	gi|164512572	1145	25	6.89	convicilin	*Vf*
		gi|126164	101	3	0.3	legumin type B; Precursor	*Vf*
8	54.1	gi|164512572	1074	21	4.67	convicilin	*Vf*
		gi|403336	312	7	0.68	legumin-related high molecular weight polypeptide (LHMW)	*Vf*
		gi|3122060	123	6	0.78	elongation factor 1-alpha	*Vf*

Table 3.2. (continued)

9	50	gi|137584	1344	22	6.28	vicilin: Precursor	*Vf*
		gi|403336	589	11	1.25	LHMW	*Vf*
10	48.2	gi|137584	1374	22	6.28	vicilin: Precursor	*Vf*
		gi|403336	342	7	0.68	LHMW	*Vf*
		gi|12580894	176	6	0.69	putative sucrose binding protein	*Vf*
11	45.4	gi|12580894	1018	18	4.4	putative sucrose binding protein	*Vf*
		gi|22008	226	9	1.16	legumin A2 primary translation product	*Vf*
		gi|126166	178	8	1.84	legumin type B	*Vf*
12	43.5	gi|2578438	98	3	0.26	legumin (minor small)	*P. sativum*
		gi|403336	90	3	0.26	LHMW	*Vf*
13	40.2	gi|22008	662	14	2.51	legumin A2 primary translation product	*Vf*
		gi|259474	312	6	1.42	legumin propolypeptide alpha chain	*Vf*
14	38.4	gi|22008	875	14	2.61	legumin A2 primary translation product	*Vf*
		gi|126166	628	12	3.78	legumin type B	*Vf*
		gi|22053	392	11	1.75	vicilin: Precursor	*Vf*
15	37.6	gi|542002	823	9	2.67	legumin type B alpha chain; Precursor	*Vf*
		gi|137584	506	16	3.24	vicilin: Precursor	*Vf*
		gi|22008	312	10	1.31	legumin A2 primary translation product	*Vf*
16	36.2	gi|542002	926	8	2.28	legumin type B alpha chain: Precursor	*Vf*
		gi|137584	747	19	4.83	vicilin: Precursor	*Vf*
		gi|22008	253	7	0.83	legumin A2 primary translation product	*Vf*
17	31.5	gi|137584	277	11	1.71	vicilin: Precursor	*Vf*
		gi|137582	203	4	0.44	vicilin: Precursor	*Vf*
18	30.4	gi|137584	300	11	1.73	vicilin: Precursor	*Vf*
		gi|137582	157	4	0.45	vicilin: Precursor	*Vf*
19	26	gi|22008	76	2	0.18	legumin A2 primary translation product	*Vf*
		gi|29539109	54	3	0.35	allergen Len c	*Lens culinaris*
20	24	gi|12580894	53	1	0.09	putative sucrose binding protein	*Vf*
21	22.3	gi|259475	399	5	-	legumin propolypeptide beta chain	*Vf*
		gi|403336	369	5	-	LHMW	*Vf*
22	13.7	gi|51704211	97	2	0.98	albumin-1 E	*P. sativum*
23	12.4	gi|51704211	72	1	0.27	albumin-1 E	*P. sativum*
		gi|27466894	70	2	0.68	thioredoxin h	*P. sativum*
		gi|763805274	50	1	0.25	hypothetical protein	*Gossypium raimondii*
24	10.2	gi|51704209	60	1	0.29	albumin-1 C	*P. sativum*
25	9.5	gi|205277584	56	2	1.15	defensin-like protein	*Vf*
		gi|205277582	55	2	1.19	defensin-like protein	*Vf*

Nearly all analysed bands contained more than one type of protein and a total of 106 proteins were identified (for detailed list, see **Table S 3.1**). As expected, the most abundant proteins were globulins, with polypeptides belonging to legumin, vicilin, and convicilin identified in 13, 8 and 4 of the 25 bands, respectively (**Table 3.2**). This wide molecular mass distribution of legumin and vicilin subunits was previously reported (Tucci *et al.*, 1991) using antibodies specific to these proteins. However, in the case of convicilin, for which a single discrete band near 68 kDa has been so far reported in the literature (Müntz *et al.*, 1999; Warsame *et al.*, 2018; Liu *et al.*, 2017; Tucci *et al.*, 1991), we have identified multiple bands, including a major band at ~54 kDa (**Figure 3.1**, **Table 3.2**). Although this is a new observation for *Vf*, it is not surprising, considering that multiple convicilin subunits with wide MW range have been reported in the related species including *Medicago truncatula* (52–99 kDa) (Le Signor *et al.*, 2005; Le Signor *et al.*, 2017) and *Pisum sativum* (62-86 kDa) (Bourgeois *et al.*, 2011). To further investigate whether the two major convicilin bands identified (7, 8 in **Figure 3.1**) represent the *Vf* convicilin A and B proteins reported in the past (Sáenz de Miera *et al.*, 2008), we compared the protein sequences derived from these convicilin gene products with the MS peptide sequences from band 7 and 8 of LG Cartouche and ILB 938-2. While convicilin B-specific peptides were found in both bands 7 and 8 in both genotypes, three peptides unique to convicilin A were found only in band 8 of LG Cartouche (Table S3.2). Furthermore, peptide sequences from band 7 of LG Cartouche contained a 37 AA long peptide which aligned to a region with significant polymorphism between convicilin A and B. Interestingly, this unique peptide had nine and five mismatches with A and B genes, respectively, but had 100% similarity with a convicilin accession (CAP06324.1) from *Lathyrus ochrus.* Taken together, these results indicate that convicilin structural diversity in *Vf* is greater than previously thought, comprising of at least two B-type isoforms, as well as A and other unnamed convicilin polypeptides which appear to be expressed in a genotype-dependent manner.

Mass spectrometry analysis also identified several less abundant but nonetheless distinct and well-conserved protein bands. These include two distinct lipoxygenase bands (~96 and 107 kDa), a heat shock protein (~73–75 kDa, depending on the genotype), a sucrose-binding protein (~45 kDa), albumins (~10.2, 12.4 and 13.7 kDa) and defensins (**Table 3.2**, **Figure 3.1**). From a nutritional quality point of view, lipoxygenase is considered antinutritional due to its role in lipid oxidation, which also leads to undesirable flavours during food processing (Lampi *et al.*, 2020). The studied genotypes show noticeable variation in the intensity of lipoxygenase bands (**Figure S 3.2**) but establishing the significance of this variation requires further scrutiny. In other legumes such as soybean (Lee *et al.*, 2014) and pea (Forster *et al.*, 1999), efforts to develop genotypes lacking the major seed protein lipoxygenase have been successful. However, evidence from peanut suggest that these proteins could play important role in seed defence against pathogens (Müller *et al.*, 2014) and seed storage quality (Zhang and Zhang, 2020), which in the context of grain protein quality improvement, calls for better understanding of the potential consequences of modifying protein composition on seed biology.

3.4.2 Protein subunit diversity among *Vf*

In total, we identified 15 protein bands polymorphic among the *Vf* genotypes, with variation being concentrated in less abundant proteins with MW of more than 70 kDa or less than 20 kDa (**Figure S 3.2**). The most interesting protein variants were found in the α subunits of legumin, represented in the majority of *Vf* genotypes by a single major legumin band of about 38 kDa and by additional rare legumin α subunits of about 36 and 40 kDa in LG Cartouche and NV657, respectively (**Figure 3.1**). MS analysis showed that the higher MW legumin α subunit in NV657 is an A-type legumin while that of a lower mass in LG Cartouche is a B-type legumin α subunit (**Table 3.2**). Further evidence that these genotypes contain novel legumin subunits comes from the observation that unreduced proteins of these genotypes have two distinct major bands of αβ polypeptides (data not shown). These natural variants in subunit composition can be

exploited to address questions on the genetic architecture of seed protein composition and the impact of discrete protein subunit variants on the nutritional and processing quality of the overall seed protein.

Although the majority of the analysed bands contained one predominant protein type, the existence of some bands where there is an overlap between major bands of different protein classes underpins the need for an alternative method to the conventional SDS–PAGE based densitometric approach for quantifying protein composition. Nonetheless, this expanded and refined list of identified seed proteins can be utilized as a reference for qualitative SDS–PAGE–based screening for protein subunit variants of interest in breeding and research materials like mutant or mapping populations.

3.4.3 SE-HPLC analysis of seed proteins

1.1.1.1 Total seed protein extract

The total seed protein extract from the NV639-2 inbred line was separated using a Phenomenex BioSep SEC S-2000 column, producing chromatographic peaks between 10 and 28 minutes of the analysis time (**Figure 3.2**), and four major peaks (1, 2, 4, and 18) accounted for more than 70% of the total chromatogram peak area. To confirm the identity of proteins associated with these peaks, SE-HPLC peak fractions were collected at 1-minute intervals and separated by 1D SDS-PAGE. By comparing these gels with the annotated SDS-PAGE (on the basis of MS analysis), it was determined that peaks 2 and 4 were legumin and vicilin/convicilin aggregates with retention times of 12.4 and 14.0 minutes, respectively (**Figure 3.2**). Proteins with smaller molecular weights were eluted in the expected order, suggesting that the selected column was suitable for the separation of *Vf* proteins. However, despite having strong signals at 214, 254 and 280 nm, no detectable proteins were found in peak 1 and all other peaks eluted after ~21 minutes (**Figure 3.2**). We therefore hypothesized that peak 1 corresponds to protein–phenol complexes that could not be detected by SDS-PAGE. Sęczyk *et al.* (2019) found that some

phenolic compounds preferentially interact with globulins, leading to changes in their SE-HPLC and SDS-PAGE profiles. Regarding peak 18, Defaix *et al.* (2019), who used the same type of column used in our study, suggested that the strong signal near the end of the analysis was due to phenolic compounds. To further investigate this hypothesis, SE-HPLC profiles of protein extracts of dehulled and whole seeds were compared; since *Vf* seed coats have a higher phenol content, we would expect the proportion of peak 1 and 18 to be substantially reduced in the dehulled sample. Indeed, dehulled protein samples showed a nearly 50% decrease in both peaks (**Figure S 3.3**) On this basis, only the peak area between 11.5 and 21.5 minutes was considered for SE-HPLC protein composition analysis.

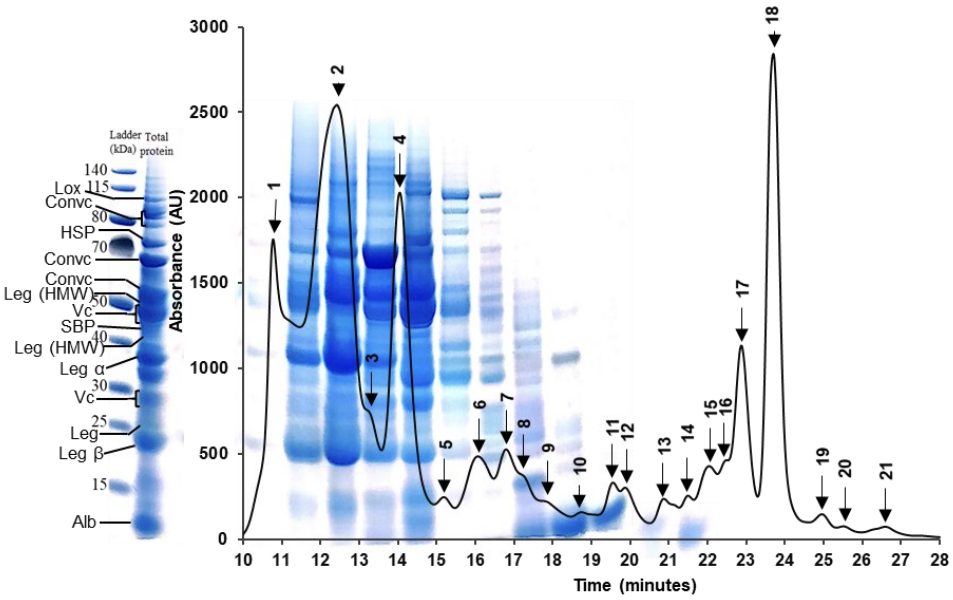

Figure 3.2. SE-HPLC chromatogram of *Vf* seed protein extract from NV639-2 which is overlaid on SDS-PAGE of protein fractions collected at 1-minute interval across the analysis time. Observable peaks are numbered from 1-21 and labels on the left refer to some of the major protein subunits identified in this study. Lox=lipoxygenase, HSP=heat shock protein, Convc=convicilin, Vc=vicilin, Leg=legumin, SBP=sucrose binding protein, Alb=albumin.

1.1.1.2 Fractionated seed proteins

To further confirm that the peaks resolved by SE-HPLC belong to the major seed proteins, we separated the protein fractions prepared by sequential extraction (denoted as F1-F5 in **Figure 3.3**) both by SE-HPLC and 1D SDS-PAGE. Comparison of the separation profiles obtained for F1-F5 protein fractions in the two systems (**Figure 3.3**A&B), showed that the functional proteins like lipoxygenase, heat shock protein and albumin have relatively higher solubility in water and they were enriched in F1 and F2, with an elution time between 15 and 20 minutes under the SE-HPLC conditions used in this work. However, since these peaks, unlike globulin peaks, were poorly resolved by SE-HPLC, they are referred collectively as "other proteins" as shown in **Figure 3.3**A.

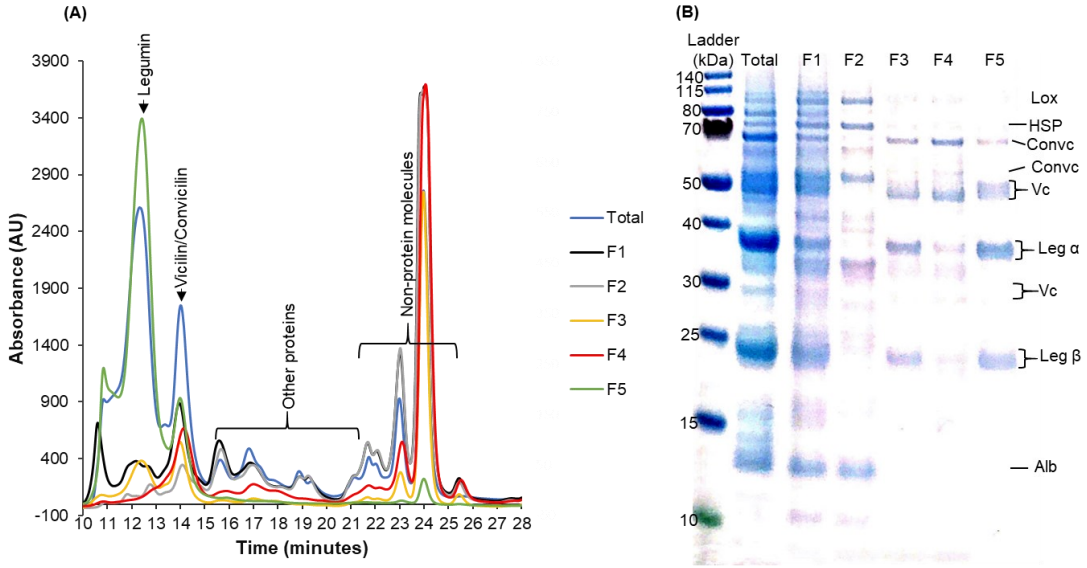

Figure 3.3. SE-HPLC chromatogram (A) and SDS-PAGE (B) of fractionated proteins of NV639-2 line. Fractions (F1-5) are water extractable proteins (F1), globulin-removed water-soluble fraction by addition of 10 mM CaCl2 (F2), pellet from F1 extracted with 0.1 mM phosphate buffer (pH=7.2) (F3), globulin-depleted salt-soluble. Lox=lipoxygenase, HSP=heat shock protein, Convc=convicilin, Vc=vicilin, Leg=legumin, Alb=albumin.

On the other hand, different globulin subclasses were enriched in specific extraction buffers, reflecting their different physicochemical properties. For instance, legumin subunits were soluble in water (F1) and to a higher degree in the phosphate buffer (F3), which could be

further enriched by the addition of Ca^{+2} (F5), producing a major HPLC peak with a retention time of 12.4 minutes. However, the vicilin/convicilin subclass of globulin appears to contain a mixture of subunits with varying properties. As shown in **Figure 3.3**B, protein bands corresponding to subunits of convicilin (~ 54 kDa) and vicilin (~ 37 kDa) were extractable in water (F1) and did not precipitate in the presence of Ca^{+2} (F2). Conversely, other subunits of convicilin (~65 kDa) and vicilin (~50 kDa) were soluble in the phosphate buffer (F3) and precipitated, to a certain degree, with the addition of Ca^{+2} (F4 and F5). Even though the fractions F1, 4 and 5 have a vicilin/convicilin peak of nearly the same magnitude, the SDS-PAGE profile of these different fractions showed distinct subsets of vicilin/convicilin (**Figure 3.3**B). It was therefore concluded that convicilin and vicilin polypeptides form heterogeneous subclasses of the globulin type protein with distinct physicochemical properties but eluted as a single peak with a retention time of 14 minutes under the SE-HPLC conditions used in our study. In pea, Bourgeois *et al.* (2011) who used an Anion Exchange Fast Protein Liquid Chromatography (FPLC) followed by 2D-PAGE of the fractions also found that both vicilin and convicilin were eluted in the same peak. This observation would explain why O'Kane *et al.* (2004), who conducted various fractionation and physicochemical characterization of vicilin and convicilin proteins in pea, concluded that convicilin is a α subunit of vicilin.

3.4.4 Quantification of legumin and vicilin/convicilin contents by SE-HPLC

Since one of the major indicators of protein quality is the content of S-AA, which in turn is determined by the relative proportions of the major protein classes, the SE-HPLC method was used to quantify legumin and vicilin/convicilin contents in a panel of 35 genetically diverse *Vf* genotypes. On the average of the 35 *Vf* genotypes, legumin and vicilin/convicilin accounted for 50% and 27% of the protein extract, respectively. Among the genotypes, legumin accounted for 35% to 62% of the quantified peak area while vicilin/convicilin for 20 –34% (**Figure 3.4**). These results are comparable with the findings of Utsumi *et al.* (1980) who reported ranges of 42% to

47% and 28% to 31% for 11S and 7S globulins in crude protein extracts of six *Vf* cultivars analysed by the sucrose density gradient fractionation technique. In another study, the *Vf* legumin and vicilin content reportedly varied between 40% to 45% and 20% to 25%, respectively (Multari *et al.*, 2015). Moreover, according to our study, globulin peaks represent 77% of the total protein peak area, which is very close to the estimated 70% to 80% globulin content in *Vf* seed proteins reported by other researchers (Verni *et al.*, 2019; Müntz *et al.*, 1999; Multari *et al.*, 2015). The present study appears to capture a wider variation in *Vf* protein composition than previously reported, likely reflecting the fact that the plant materials we used spanned a deliberately broad genetic base.

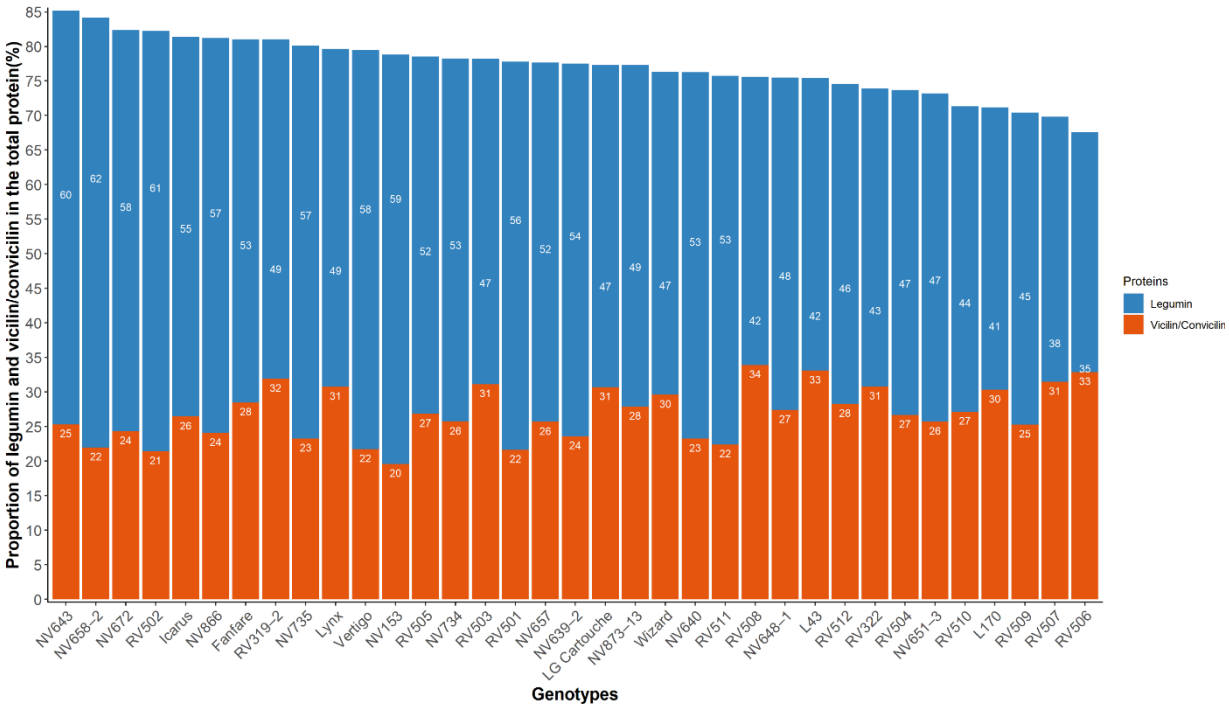

Figure 3.4. Bar graph showing proportions of legumin and vicilin/convicilin in the total seed protein extracts of 35 *Vf* genotypes. Protein percentages are determined from the relative area of SE-HPLC peaks belonging to each protein class in two technical replicates.

The legumin to vicilin/convicilin ratio (L/V) varied from 1:1 to 3:1, which is comparable to the 2.1 to 3.6 range reported previously (Gatehouse *et al.*, 1980). Among the genotypes of

special interest for their high L/V ratio are two inbred lines, NV153 and NV658-2, which have been previously used as parents in mapping populations (**Table 3.2**). However, it is important to mention that this ratio is highly sensitive and can be affected by many factors, including genotype, environment, protein extraction method, and quantification techniques. In this work, the reproducibility of the results was measured by comparing five replicates of a single genotype (NV639-2) that were independently extracted and analysed in different batches. The coefficient of variation between the five replicates was higher in the legumin fraction (6%) compared to vicilin/convicilin (3%). However, biological replicates of each genotype analysed in the same run were highly correlated (r^2>0.98) (**Figure S 3.4**). This indicates the importance of including the batch as a cofactor for statistical analysis.

Finally, we exploited this quantitative data from a wide spectrum of germplasm to examine possible limits and trade-offs between the two main classes of storage protein and overall sulphur and protein content. In fact, legumin content was significantly and strongly negatively correlated with vicilin/convicilin (r=-0.83, p<0.001) as well as with 'other' proteins (r=-0.87, p<0.001) (**Figure 3.5**). Seed sulphur content tended to correlate negatively with vicilin/convicilin, the S-AA poor fraction (Warsame *et al.*, 2018), but did not show a positive correlation with legumin as expected. Interestingly, protein content was independent of any of the protein fractions, suggesting that protein composition can be improved without penalizing protein content. Similar independence of total protein and globulin fractions has been observed in pea (Tzitzikas *et al.*, 2006) while a highly significant negative correlation between certain 7S fractions of protein and total seed protein content has been reported in soybeans (Oomah *et al.*, 1994).

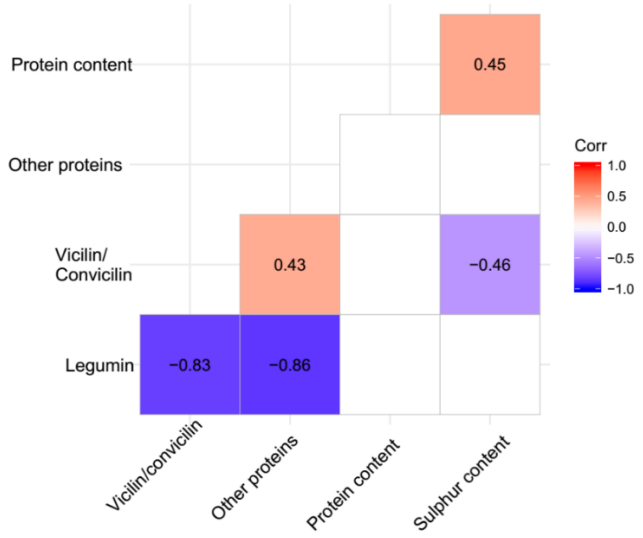

Figure 3.5. Correlation matrix between proportion of globulin fractions and other seed composition parameters at significance level p≤0.05.

In conclusion, seed proteins of *Vf* have been poorly understood in terms of their identities and quantities. This work provides a contemporary survey on the major seed proteins and their subunit composition over genetically diverse germplasm and a timely update linking a greater diversity of seed storage protein sequences to specific protein subunits which can be readily resolved on SDS-PAGE gels. Such information can facilitate screening for germplasm with unique protein profiles, such as naturally occurring or induced mutations resulting in a reduced content in undesirable proteins like lipoxygenase or S-AA poor globulins. Also, we have demonstrated the potential of SE-HPLC as a method to efficiently determine the contents of legumin and vicilin/convicilin from legume flours. This work paves the way for further understanding of genetic control of *Vf* seed protein composition and the development of cultivars with desired protein quality.

3.5 References

Baddeley, J. A., Jones, S., Topp, C. F. E., Watson, C. A., Helming, J. & Stoddard, F. L. (2013). Biological nitrogen fixation (BNF) by legume crops in Europe. *Legume Futures Report 1.5.*

Baumlein, H., Wobus, U., Pustell, J. & Kafatos, F. C. (1986). The legumin gene family: structure of a B type gene of *Vicia faba* and a possible legumin gene specific regulatory element. *Nucleic Acids Research,* **14** (6), 2707-2720.

Boehm, J. D., Nguyen, V., Tashiro, R. M., Anderson, D., Shi, C., Wu, X., Woodrow, L., Yu, K., Cui, Y. & Li, Z. (2017). Genetic mapping and validation of the loci controlling 7S α′ and 11S A-type storage protein subunits in soybean [*Glycine max* (L.) Merr.]. *Theoretical and Applied Genetics*, 1-13.

Bourgeois, M., Jacquin, F., Cassecuelle, F., Savois, V., Belghazi, M., Aubert, G., Quillien, L., Huart, M., Marget, P. & Burstin, J. (2011). A PQL (protein quantity loci) analysis of mature pea seed proteins identifies loci determining seed protein composition. *Proteomics*, 11 (9), 1581-94.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry,* **72**, 248-54.

Cernay, C., Pelzer, E. & Makowski, D. (2016). A global experimental dataset for assessing grain legume production. *Scientific Data,* **3**, 160084.

Chakraborty, P., Sosulski, F. & Bose, A. (1979). Ultracentrifugation of salt-soluble proteins in ten legume species. *Journal of the Science of Food and Agriculture,* **30** (8), 766-771.

Defaix, C., Aymes, A., Slabi, S. A., Basselin, M., Mathe, C., Galet, O. & Kapel, R. (2019). A new size-exclusion chromatography method for fast rapeseed albumin and globulin quantification. *Food Chemistry,* **287**, 151-159.

Duc, G., Aleksić, J. M., Marget, P., Mikic, A., Paull, J., Redden, R. J., Sass, O., Stoddard, F. L., Vandenberg, A., Vishnyakova, M. & Torres, A. M. (2015). Faba Bean. *In:* Ron, A. M. D. (ed.) *Grain Legumes.* New York: Springer Science+Business Media, pp. 141-178.

Dunwell, J. M., Purvis, A. & Khuri, S. (2004). Cupins: the most functionally diverse protein superfamily? *Phytochemistry,* **65** (1), 7-17.

Forster, C., North, H., Afzal, N., Domoney, C., Hornostaj, A., Robinson, D. S. & Casey, R. (1999). Molecular analysis of a null mutant for pea (*Pisum sativum* L.) seed lipoxygenase-2. *Plant Molecular Biology,* **39** (6), 1209–1220.

Foyer, C. H., Lam, H. M., Nguyen, H. T., Siddique, K. H., Varshney, R. K., Colmer, T. D., Cowling, W., Bramley, H., Mori, T. A., Hodgson, J. M., Cooper, J. W., Miller, A. J., Kunert, K., Vorster, J., Cullis, C., Ozga, J. A., Wahlqvist, M. L., Liang, Y., Shou, H., Shi, K., Yu, J., Fodor, N., Kaiser, B. N., Wong, F. L., Valliyodan, B. & Considine, M. J. (2016). Neglecting legumes has compromised human health and sustainable food production. *Nature Plants,* **2**, 16112.

Fuchs, J. & Schubert, I. (1995). Localization of seed protein genes on metaphase chromosomes of Vicia faba via fluorescence in situ hybridization. *Chromosome Res.,* **3** (2), 94-100.

Fukushima, D. (1991). Structures of plant storage proteins and their functions. *Food Reviews International,* **7** (3), 353-381.

Gatehouse, J., Croy, R., McIntosh, R., Paul, C. & Boulter, D. (1980). Quantitative and qualitative variation in the storage proteins of material from the EEC joint field bean test. *Quantitative and qualitative variation in the storage proteins of material from the EEC joint field bean test.*, 173-188.

Gutierrez, N., Avila, C. M., Moreno, M. T. & Torres, A. M. (2008). Development of SCAR markers linked to *zt-2*, one of the genes controlling absence of tannins in faba bean. *Australian Journal of Agricultural Research,* **59** (1), 62-68.

Horstmann, C., Schlesier, B., Otto, A., Kostka, S. & Muntz, K. (1993). Polymorphism of legumin subunits from field bean (*Vicia faba* L. var. minor) and its relation to the corresponding multigene family. *Theoretical and Applied Genetics,* **86** (7), 867-874.

Kesari, P., Sharma, A., Katiki, M., Kumar, P., R Gurjar, B., Tomar, S., K Sharma, A. & Kumar, P. (2017). Structural, functional and evolutionary aspects of seed globulins. *Protein and Peptide Letters,* **24** (3), 267-277.

Khazaei, H., M. O'Sullivan, D., Jones, H., Pitts, N., Sillanpää, M., Pärssinen, P., Manninen, O. & Stoddard, F. (2015). Flanking SNP markers for vicine–convicine concentration in faba bean (*Vicia faba* L.). *Molecular Breeding,* **35** (38).

Khazaei, H., Purves, R. W., Hughes, J., Link, W., O'Sullivan, D. M., Schulman, A. H., Björnsdotter, E., Geu-Flores, F., Nadzieja, M., Andersen, S. U., Stougaard, J., Vandenberg, A. & Stoddard, F. L. (2019). Eliminating vicine and convicine, the main anti-nutritional factors restricting faba bean usage. *Trends in Food Science and Technol.,* **91**, 549-556.

Khazaei, H., Purves, R. W., Song, M., Stonehouse, R., Bett, K. E., Stoddard, F. L. & Vandenberg, A. (2017). Development and validation of a robust, breeder-friendly molecular marker for the *vc*-locus in faba bean. *Molecular Breeding,* **37** (11), 140.

Khazaei, H., Stoddard, F. L., Purves, R. W. & Vandenberg, A. (2018). A multi-parent faba bean (*Vicia faba* L.) population for future genomic studies. *Plant Genetic Resources: Characterization and Utilization,* **16** (5), 419-423.

Krishnan, H. B., Oehrle, N. W. & Natarajan, S. S. (2009). A rapid and simple procedure for the depletion of abundant storage proteins from legume seeds to advance proteome analysis: A case study using *Glycine max*. *Proteomics,* **9** (11), 3174-3188.

Lampi, A.-M., Yang, Z., Mustonen, O. & Piironen, V. (2020). Potential of faba bean lipase and lipoxygenase to promote formation of volatile lipid oxidation products in food models. *Food Chemistry,* **311**, 125982.

Larroque, O. R. & Bekes, F. (2000). Rapid size-exclusion chromatography analysis of molecular size distribution for wheat endosperm protein. *Cereal Chemistry,* **77** (4), 451-453.

Le Signor, C., Aimé, D., Bordat, A., Belghazi, M., Labas, V., Gouzy, J., Young, N. D., Prosperi, J.-M., Leprince, O., Thompson, R. D., Buitink, J., Burstin, J. & Gallardo, K. (2017). Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. *New Phytologist,* **214** (4), 1597-1613.

Le Signor, C., Gallardo, K., Prosperi, J. M., Salon, C., Quillien, L., Thompson, R. & Duc, G. (2005). Genetic diversity for seed protein composition in *Medicago truncatula*. *Plant Genetic Resources,* **3** (1), 59-71.

Lee, K. J., Hwang, J. E., Velusamy, V., Ha, B. K., Kim, J. B., Kim, S. H., Ahn, J. W., Kang, S. Y. & Kim, D. S. (2014). Selection and molecular characterization of a lipoxygenase-free soybean mutant line induced by gamma irradiation. *Theoretical and Applied Genetics,* **127** (11), 2405–2413.

Liu, Y., Wu, X., Hou, W., Li, P., Sha, W. & Tian, Y. (2017). Structure and function of seed storage proteins in faba bean (*Vicia faba* L.). *3 Biotech,* **7** (1), 74.

Martensson, P. (1980). Variation in legumin : vicilin ratio between and within cultivars of *Vicia faba* L. var. minor. The Hague: Martinus Nijhoff. World crops: production, utilization and description, volume 3, pp. 159-171.

Mertens, C., Dehon, L., Bourgeois, A., Verhaeghe-Cartrysse, C. & Blecker, C. (2012). Agronomical factors influencing the legumin/vicilin ratio in pea (*Pisum sativum* L.) seeds. *Journal of the Science of Food and Agriculture,* **92** (8), 1591-1596.

Mosse, J. (1990). Nitrogen-to-protein conversion factor for ten cereals and six legumes or oilseeds. A reappraisal of its definition and determination. Variation according to species and to seed protein content. *Journal of Agricultural and Food Chemistry,* **38** (1), 18-24.

Müller, V., Amé, M. V., Carrari, V., Gieco, J. & Asis, R. (2014). Lipoxygenase activation in Peanut seed cultivars resistant and susceptible to *Aspergillus parasiticus* colonization. *Phytopathology,* **104** (12), 1340-8.

Multari, S., Stewart, D. & Russell, W. R. (2015). Potential of fava bean as future protein supply to partially replace meat intake in the human diet. *Comprehensive Reviews in Food Science and Food Safety,* **14** (5), 511-522.

Müntz, K., Horstmann, C. & Schlesier, B. (1999). Vicia globulins. *In:* Shewry, P. R. & Casey, R. (eds.) *Seed Proteins.* Dordrecht: Springer Netherlands, pp. 259-284

O'Kane, F. E., Happe, R. P., Vereijken, J. M., Gruppen, H. & van Boekel, M. A. J. S. (2004). Characterization of pea vicilin. 1. Denoting convicilin as the α-subunit of the Pisum vicilin family. *Journal of Agricultural and Food Chemistry,* **52** (10), 3141-3148.

Ohm, J.-B., Hareland, G., Simsek, S. & Seabourn, B. (2009). Size-exclusion HPLC of protein using a narrow-bore column for evaluation of breadmaking quality of hard spring wheat flours. *Cereal Chemistry,* **86** (4), 463-469.

Ohm, J.-B., Manthey, F. & Elias, E. M. (2017). Variation and correlation of protein molecular weight distribution and semolina quality parameters for durum genotypes grown in North Dakota. **94** (4), 780-788.

Oomah, B., Voldeng, H. & Fregeau-Reid, J. (1994). Characterization of soybean proteins by HPLC. *Plant Foods for Human Nutrition,* **45** (3), 251-263.

Panthee, D. R., Kwanyuen, P., Sams, C. E., West, D. R., Saxton, A. M. & Pantalone, V. R. (2004). Quantitative trait loci for β-conglycinin (7S) and glycinin (11S) fractions of soybean storage protein. *Journal of the American Oil Chemists' Society,* **81** (11), 1005-1012.

Poysa, V., Woodrow, L. & Yu, K. (2006). Effect of soy protein subunit composition on tofu quality. *Food Research International,* **39** (3), 309-317.

Sáenz de Miera, L. E., Ramos, J. & Pérez de la Vega, M. (2008). A comparative study of convicilin storage protein gene sequences in species of the tribe Vicieae. *Genome,* **51** (7), 511-523.

Sęczyk, Ł., Świeca, M., Kapusta, I. & Gawlik-Dziki, U. (2019). Protein-phenolic interactions as a factor affecting the physicochemical properties of white bean proteins. *Molecules* **24** (3), 408.

Tucci, M., Capparelli, R., Costa, A. & Rao, R. (1991). Molecular heterogeneity and genetics of *Vicia faba* seed storage proteins. *Theoretical and Applied Genetics,* **81** (1), 50-58.

Tzitzikas, E. N., Vincken, J.-P., de Groot, J., Gruppen, H. & Visser, R. G. F. (2006). Genetic variation in pea seed globulin composition. *Journal of Agricultural and Food Chemistry,* **54** (2), 425-433.

Utsumi, S., Yokoyama, Z.-i. & Mori, T. (1980). Comparative studies of subunit compositions of legumins from various cultivars of *Vicia faba* L. seeds. *Agricultural and Biological Chemistry,* **44** (3), 595-601.

Verni, M., Coda, R. & Rizzello, C. G. (2019). Chapter 37 - The Use of Faba Bean Flour to Improve the Nutritional and Functional Features of Cereal-Based Foods: Perspectives and Future Strategies. *In:* Preedy, V. R. & Watson, R. R. (eds.) *Flour and Breads and their Fortification in Health and Disease Prevention (Second Edition).* Academic Press, pp. 465-475.

Warsame, A. O., O'Sullivan, D. M. & Tosi, P. (2018). Seed storage proteins of faba bean (*Vicia faba* L): current status and prospects for genetic improvement. *Journal of Agricultural and Food Chemistry,* **66** (48), 12617-12626.

Webb, A., Cottage, A., Wood, T., Khamassi, K., Hobbs, D., Gostkiewicz, K., White, M., Khazaei, H., Ali, M., Street, D., Duc, G., Stoddard, F. L., Maalouf, F., Ogbonnaya, F. C., Link, W., Thomas, J. & O'Sullivan, D. M. (2016). A SNP-based consensus genetic map for synteny-based trait targeting in faba bean (*Vicia faba* L.). *Plant Biotechnology Journal,* **14** (1), 177-185.

Zanotto, S., Vandenberg, A. & Khazaei, H. (2019). Development and validation of a robust KASP marker for zt2 locus in faba bean (Vicia faba). Plant Breeding, 139, 375-380.

Zhang, Y. & Zhang, Y. (2020). Effect of lipoxygenase-3 on storage characteristics of peanut seeds. Journal of Stored Products Research, 87, 101589.

3.6 Supplementary

Table S 3.1. Detailed list of proteins identified by mass spectrometry analysis of protein bands from the seeds of three *Vf* genotypes. From MASCOT search results, proteins with peptide matches above identity threshold at p-value<0.05 are reported. The column containing SDS-PAGE band numbers refers to **Figure 2.1** in the main text.

SDS-PAGE Band	Accession	Score	No. of significant matches	No. of significant sequences	Description	Species
1	gi|126405	565	33	15	Seed linoleate 9S-lipoxygenase-3	*Pisum sativum*
	gi|164512572	128	3	2	Convicilin	*Vicia faba*
	gi|118573101	75	1	1	Putative poly [ADP-ribose] polymerase 3	*Medicago truncatula*
	gi|923709735	60	3	3	PREDICTED:Elongation factor 2-like	*Brassica napus*
	gi|126164	52	2	2	Legumin type B; Precursor	*Vicia faba*
2	gi|126405	508	30	15	Seed linoleate 9S-lipoxygenase-3	*Pisum sativum*
	gi|164512572	178	7	4	Convicilin	*Vicia faba*
	gi|118573101	80	1	1	Putative poly [ADP-ribose] polymerase 3	*Medicago truncatula*
	gi|187766747	59	1	1	Gly m Bd 28K allergen	*Glycine max*
	gi|976928307	56	2	1	Hypothetical protein	*Cynara cardunculus var. scolymus*
	gi|147766023	52	2	1	Hypothetical protein	*Vitis vinifera*
	gi|81988	50	1	1	Legumin B	*Vicia faba*
3	gi|164512572	120	3	2	Convicilin	*Vicia faba*
	gi|187766747	99	3	1	Gly m Bd 28K allergen	*Glycine max*
	gi|1021036037	88	4	4	Hypothetical protein DCAR_017065	*Daucus carota subsp. Sativus*
	gi|137582	73	2	2	Vicilin: Precursor	*Vicia faba*
	gi|1297070	69	2	1	Convicilin; Precursor	*Vicia narbonensis*
	gi|743859611	62	1	1	PREDICTED: BTB/POZ domain-containing protein At3g49900	*Elaeis guineensis*
	gi|902227102	54	1	1	Hypothetical protein	*Spinacia oleracea*
	gi|828330409	52	2	1	PREDICTED: globulin-1 S allele	*Cicer arietinum*
4	gi|164512572	165	10	6	Convicilin	*Vicia faba*
	gi|22053	154	11	9	Vicilin: Precursor	*Vicia faba*
	gi|12580894	151	9	7	Putative sucrose binding protein	*Vicia faba*
	gi|126162	108	5	3	Legumin type B	*Vicia faba*
	gi|187766747	105	2	1	Gly m Bd 28K allergen	*Glycine max*
	gi|168000434	68	1	1	Predicted protein	*Physcomitrella patens*
	gi|22008	61	4	4	legumin A2 primary translation product	*Vicia faba*

Table S3.1 continued

	gi|356495423	59	2	1	PREDICTED: Vicilin-like antimicrobial peptides 2-2	*Glycine max*
	gi|743755771	54	1	1	PREDICTED: Xyloglucan galactosyltransferase KATAMARI1 homolog	*Elaeis guineensis*
	gi|147800376	50	1	1	Hypothetical protein	*Vitis vinifera*
	gi|357480003	391	14	8	Heat shock 70 kDa protein	*Medicago truncatula*
	gi|126162	94	6	4	Legumin type B	*Vicia faba*
5	gi|137582	84	2	1	Vicilin: Precursor	*Vicia faba*
	gi|22008	84	2	1	Legumin A2 primary translation product	*Vicia faba*
	gi|2578438	55	2	2	Legumin (small minor)	*Pisum sativum*
	gi|562006	364	17	12	PsHSP71.2	*Pisum sativum*
	gi|164512572	123	5	4	Convicilin	*Vicia faba*
	gi|126162	81	2	1	Legumin type B	*Vicia faba*
6	gi|22008	62	3	2	Legumin A2 primary translation product	*Vicia faba*
	gi|950935871	60	1	1	PREDICTED: Probable 2-oxoglutarate-dependent dioxygenase	*Vigna radiata*
	gi|164512572	1145	54	25	Convicilin	*Vicia faba*
	gi|126164	101	5	3	Legumin type B; Precursor	*Vicia faba*
	gi|22008	65	2	1	Legumin A2 primary translation product	*Vicia faba*
7	gi|12580894	64	2	2	Putative sucrose binding protein	*Vicia faba*
	gi|1009154921	56	2	2	PREDICTED: phosphoglucomutase (cytoplasmic)	*Ziziphus jujuba*
	gi|920692855	53	1	1	Hypothetical protein	*Vigna angularis*
	gi|164512572	1074	47	21	Convicilin	*Vicia faba*
	gi|403336	312	15	7	Legumin-related high molecular weight polypeptide	*Vicia faba*
	gi|3122060	123	6	6	Elongation factor 1-alpha	*Vicia faba*
	gi|12580894	99	5	4	putative sucrose binding protein	*Vicia faba*
	gi|126162	94	3	2	Legumin type B	*Vicia faba*
	gi|137582	86	1	1	Vicilin: Precursor	*Vicia faba*
8	gi|6094228	80	3	3	Adenosylhomocysteinase	*Mesembryanthemum crystallinum*
	gi|303287803	66	2	1	Predicted protein	*Micromonas pusilla*
	gi|137584	53	2	2	Vicilin: Precursor	*Vicia faba*
	gi|22008	56	1	1	Legumin A2 primary translation product	*Vicia faba*
	gi|702309265	54	2	1	PREDICTED: Anthocyanidin 3-O-glucosyltransferase 7-like	*Eucalyptus grandis*

Table S3.1 continued

	gi|7688419	52	2	1	ATP synthase beta subunit	*Viburnum opulus*
	gi|727544242	52	1	1	PREDICTED: Actin cytoskeleton-regulatory complex protein PAN1-like	*Camelina sativa*
9	gi|137584	1344	51	22	Vicilin: Precursor	*Vicia faba*
	gi|403336	589	23	11	Legumin-related high molecular weight polypeptide	*Vicia faba*
	gi|965671203	203	5	3	Hypothetical protein	*Vigna angularis*
	gi|685277624	69	2	1	PREDICTED: Uncharacterized protein	*Brassica rapa*
	gi|22008	64	2	2	Legumin A2 primary translation product	*Vicia faba*
	gi|302836123	58	2	1	Hypothetical protein	*Volvox carteri f. nagariensis*
	gi|18403402	57	2	1	AGC2 kinase 3	*Arabidopsis thaliana*
	gi|58618856	56	1	1	S-RNase S8	*Prunus armeniaca*
	gi|326506984	54	3	1	Predicted protein	*Hordeum vulgare*
	gi|1550740	54	1	1	GDP-associated inhibitor	*Arabidopsis thaliana*
	gi|902227102	53	1	1	Hypothetical protein	*Spinacia oleracea*
	gi|922329067	53	1	1	PIF1-like helicase	*Medicago truncatula*
	gi|590139302	53	1	1	salt overly sensitive 1	*Cardamine hirsuta*
10	gi|137584	1374	53	22	Vicilin: Precursor	*Vicia faba*
	gi|403336	342	15	7	Legumin-related high molecular weight polypeptide	*Vicia faba*
	gi|12580894	176	8	6	Putative sucrose binding protein	*Vicia faba*
	gi|828300518	93	6	2	PREDICTED: Legumin J	*Cicer arietinum*
	gi|22008	80	3	3	Legumin A2 primary translation product	*Vicia faba*
	gi|303844	72	2	1	Eukaryotic initiation factor 4A	*Oryza sativa*
	gi|685277624	68	2	1	PREDICTED: Uncharacterized protein	*Brassica rapa*
	gi|593606611	65	2	1	Hypothetical protein	*Phaseolus vulgaris*
	gi|242092744	63	2	1	Hypothetical protein	*Sorghum bicolor*
	gi|672176717	61	1	1	PREDICTED: probable beta-1,3-galactosyltransferase 2	*Phoenix dactylifera*
	gi|514815832	60	2	1	PREDICTED: valine--tRNA ligase, mitochondrial 1-like	*Setaria italica*
	gi|356513082	59	1	1	PREDICTED: Argininosuccinate synthase, chloroplastic isoform X2	*Glycine max*

Table S3.1 continued

	gi|27805602	58	2	1	Maturase K	*Medicago sativa*
	gi|698461818	56	1	1	PREDICTED: Uncharacterized protein LOC104230689 isoform	*Nicotiana sylvestris*
	gi|702268352	56	1	1	PREDICTED: Glucose and ribitol dehydrogenase homolog 1-like	*Eucalyptus grandis*
	gi|926776784	55	1	1	Acetaldehyde dehydrogenase / alcohol dehydrogenase	*Monoraphidium neglectum*
	gi|302836123	54	1	1	Hypothetical protein	*Volvox carteri f. nagariensis*
	gi|302802235	52	1	1	Hypothetical protein	*Selaginella moellendorffii*
	gi|502178554	50	1	1	PREDICTED: Transcription factor GTE8-like isoform X1	*Cicer arietinum*
11	gi|12580894	1018	44	18	Putative sucrose binding protein	*Vicia faba*
	gi|126166	178	11	8	Legumin type B	*Vicia faba*
	gi|22008	226	14	9	Legumin A2 primary translation product	*Vicia faba*
	gi|9230771	203	5	4	Cytosolic phosphoglycerate kinase	*Pisum sativum*
	gi|164512572	162	6	4	Convicilin	*Vicia faba*
	gi|137582	156	4	2	Vicilin: Precursor	*Vicia faba*
	gi|147785051	71	1	1	hypothetical protein	*Vitis vinifera*
	gi|303844	58	1	1	Eukaryotic initiation factor 4A	*Oryza sativa*
	gi|971520411	56	1	1	Hypothetical protein	*Klebsormidium flaccidum*
12	gi|2578438	98	4	3	Legumin (minor small)	*Pisum sativum*
	gi|403336	90	4	3	Legumin-related high molecular weight polypeptide	*Vicia faba*
	gi|720035184	84	2	1	PREDICTED: Alcohol dehydrogenase-like 7 isoform X1	*Nelumbo nucifera*
	gi|12580894	76	4	3	Putative sucrose binding protein	*Vicia faba*
	gi|9230771	66	1	1	Cytosolic phosphoglycerate kinase	*Pisum sativum*
	gi|527209526	59	2	1	Hypothetical protein	*Genlisea aurea*
	gi|259475	59	1	1	Legumin propolypeptide beta chain	*Vicia faba*
	gi|168000434	51	1	1	Predicted protein	*Physcomitrella patens*
	gi|22008	50	3	2	Legumin A2 primary translation product	*Vicia faba*
13	gi|22008	662	37	14	Legumin A2 primary translation product	*Vicia faba*
	gi|164512572	208	10	7	Convicilin	*Vicia faba*
	gi|259474	312	14	6	Legumin propolypeptide alpha chain	*Vicia faba*
	gi|113366	190	11	6	Alcohol dehydrogenase 1	*Trifolium repens*

Table S3.1 continued

	gi|357477179	214	8	5	Glyceraldehyde-3-phosphate dehydrogenase	*Medicago truncatula*
	gi|1168410	198	9	5	Fructose-bisphosphate aldolase, cytoplasmic isozyme 2	*Pisum sativum*
	gi|720035184	172	5	2	PREDICTED: Alcohol dehydrogenase-like 7 isoform X1	*Nelumbo nucifera*
	gi|22053	95	4	2	Vicilin: Precursor	*Vicia faba*
	gi|697155095	89	2	2	PREDICTED: Auxin-induced protein PCNT115-like	*Nicotiana tomentosiformis*
	gi|1026078724	88	2	1	PREDICTED: glucose and ribitol dehydrogenase homolog 1-like isoform X2	*Capsicum annuum*
	gi|403336	76	2	1	Legumin-related high molecular weight polypeptide	*Vicia faba*
	gi|685382988	74	2	1	PREDICTED: Uncharacterized protein	*Brassica rapa*
	gi|12580894	61	1	1	Putative sucrose binding protein	*Vicia faba*
	gi|703070949	51	1	1	Putative fructose-bisphosphate aldolase 1	*Morus notabilis*
	gi|22008	875	41	14	Legumin A2 primary translation product	*Vicia faba*
	gi|126166	628	28	12	Legumin type B	*Vicia faba*
	gi|22053	392	17	11	Vicilin: Precursor	*Vicia faba*
	gi|950971152	189	6	3	PREDICTED: Annexin-like protein RJ4	*Vigna radiata*
	gi|459649445	76	2	1	Annexin AnxGb5	*Gossypium barbadense*
	gi|590643655	68	2	2	NAD(P)-binding Rossmann-fold superfamily protein	*Theobroma cacao*
14	gi|720035184	67	1	1	PREDICTED: Alcohol dehydrogenase-like 7 isoform X1	*Nelumbo nucifera*
	gi|1710585	63	1	1	60S acidic ribosomal protein P0	*Lupinus luteus*
	gi|77553217	53	1	1	Hypothetical protein	*Oryza sativa Japonica*
	gi|10945633	53	1	1	ribulose 1,5-bisphosphate carboxylase	*Aralidium pinnatifidum*
	gi|242092744	50	1	1	Hypothetical protein	*Sorghum bicolor*
	gi|542002	823	26	9	Legumin type B alpha chain; Precursor	*Vicia faba*
	gi|137584	506	29	16	Vicilin: Precursor	*Vicia faba*
	gi|22008	312	20	10	Legumin A2 primary translation product	*Vicia faba*
15	gi|168000434	125	4	2	Predicted protein	*Physcomitrella patens*
	gi|828335547	116	6	4	PREDICTED: Annexin-like protein RJ4	*Cicer arietinum*
	gi|403334	102	2	1	Legumin-related high molecular weight polypeptide	*Vicia faba*

Table S3.1 continued

	gi|147785051	67	2	1	Hypothetical protein	*Vitis vinifera*
	gi|403336	60	2	1	Legumin-related high molecular weight polypeptide	*Vicia faba*
	gi|685277624	57	2	1	PREDICTED: Uncharacterized protein	*Brassica rapa*
	gi|302802235	52	1	1	Hypothetical protein	*Selaginella moellendorffii*
16	gi|542002	926	27	8	Legumin type B alpha chain: Precursor	*Vicia faba*
	gi|137584	747	38	19	Vicilin: Precursor	*Vicia faba*
	gi|22008	253	12	7	Legumin A2 primary translation product	*Vicia faba*
	gi|828302237	138	7	4	PREDICTED: Glucose and ribitol dehydrogenase homolog 1-like	*Cicer arietinum*
	gi|403334	115	2	1	Legumin-related high molecular weight polypeptide	*Vicia faba*
	gi|363806816	91	2	1	Uncharacterized protein	*Glycine max*
	gi|12580894	65	1	1	Putative sucrose binding protein	*Vicia faba*
	gi|2827084	58	1	1	Malate dehydrogenase precursor	*Medicago sativa*
	gi|971520411	55	1	1	Hypothetical protein	*Klebsormidium flaccidum*
	gi|703143032	54	2	1	Hypothetical protein	*Morus notabilis*
17	gi|137584	277	17	11	Vicilin: Precursor	*Vicia faba*
	gi|137582	203	6	4	Vicilin: Precursor	*Vicia faba*
	gi|828302237	171	6	4	PREDICTED: Glucose and ribitol dehydrogenase homolog 1-like	*Cicer arietinum*
	gi|1032298838	103	5	4	Hypothetical protein	*Arabidopsis thaliana*
	gi|1026078722	94	3	3	PREDICTED: glucose and ribitol dehydrogenase-like isoform X1	*Capsicum annuum*
	gi|734320738	66	3	2	Glucose and ribitol dehydrogenase like 1	*Glycine soja*
	gi|147785051	58	2	1	Hypothetical protein	*Vitis vinifera*
18	gi|137584	300	16	11	Vicilin: Precursor	*Vicia faba*
	gi|137582	157	7	4	Vicilin: Precursor	*Vicia faba*
	gi|29539111	116	3	2	Allergen Len c	*Lens culinaris*
	gi|828302237	98	5	3	PREDICTED: Glucose and ribitol dehydrogenase homolog 1-like	*Cicer arietinum*
	gi|527190463	85	4	3	Hypothetical protein	*Genlisea aurea*
	gi|22008	59	2	2	Legumin A2 primary translation product	*Vicia faba*
	gi|403334	55	1	1	Legumin-related high molecular weight polypeptide	*Vicia faba*
	gi|1021023636	53	1	1	Hypothetical protein DCAR_029037	*Daucus carota subsp. Sativus*
	gi|18403402	50	1	1	AGC2 kinase 3	*Arabidopsis thaliana*

	Table S3.1 continued					
19	gi|22008	76	2	2	Legumin A2 primary translation product	*Vicia faba*
	gi|920689716	64	1	1	Hypothetical protein	*Vigna angularis*
	gi|1346672	63	2	2	Nucleoside diphosphate kinase 1	*Pisum sativum*
	gi|29539109	54	3	3	Allergen Len c	*Lens culinaris*
	gi|1012355391	51	2	1	DNA mismatch repair protein mutL	*Cajanus cajan*
20	gi|12580894	53	2	1	putative sucrose binding protein	*Vicia faba*
21	gi|259475	399	5	5	Legumin propolypeptide beta chain	*Vicia faba*
	gi|403336	369	4	5	Legumin-related high molecular weight polypeptide	*Vicia faba*
22	gi|51704211	97	4	2	Albumin-1 E	*Pisum sativum*
23	gi|51704211	72	2	1	Albumin-1 E	*Pisum sativum*
	gi|27466894	70	2	2	Thioredoxin h	*Pisum sativum*
	gi|763805274	50	1	1	Hypothetical protein	*Gossypium raimondii*
24	gi|51704209	60	1	1	Albumin-1 C	*Pisum sativum*
25	gi|205277584	56	3	2	Defensin-like protein	*Vicia faba*
	gi|205277582	55	4	2	Defensin-like protein	*Vicia faba*

Table S 3.2. Unique peptides encoded by convicilin genes (A&B) identified in major convicilin bands 7 and 8 in **Figure 2.1** (see main text). The two genes were previously reported by Sáenz de Miera *et al.* (2008). After obtaining database search results, non-redundant peptide sequences in each protein band of NV734 and LG Cartouche were aligned with protein sequences of convicilin genes; A (Accession: CAP06334.1) and B (Accession: CAP06335.1). Then, sequences with 100% alignment with distinct regions of either of the genes were identified.

Convicilin genes	Gene specific peptides	Region	LG Cartouche		ILB 398-2	
			Band 7	Band 8	Band 7	Band 8
Convicilin A (Accession: CAP06334.1)	AKPHTIFLPQHIDADLILVVFSGK	188-211	-	+	-	-
	KYPQLQDLDIFVSFSEISEGALLLPHYNSR	356-385	-	+	-	-
	AIVVLVVNEGQGNLELVGFKNEQQEQSLKEDEQQER	386-421	-	+	-	-
Convicilin B (Accession: CAP06335.1)	AKPHTIFLPQHIEADLILTVLSGK	188-211	+	+	+	+
	VVDLAISVNRPGKVESFNLYGNK	254-276	+	+	+	+
	KYPQLQDLDIFISSVEIK	383-400	+	+	+	+
	GNLELVGIQNEQQEQQER	424-441	+	+	+	+
	LSPGDVVIIPAGHPVAVSASSNLNLFAFGINAENNQR	452-488	+	+	+	+
Unique convicilin peptide	LSPGDVVVIPAGHPVAITASSNLNLLGFGINAENNQR	432-468 (Convicilin A) 452-488 (Convicilin B)	+	-	-	-

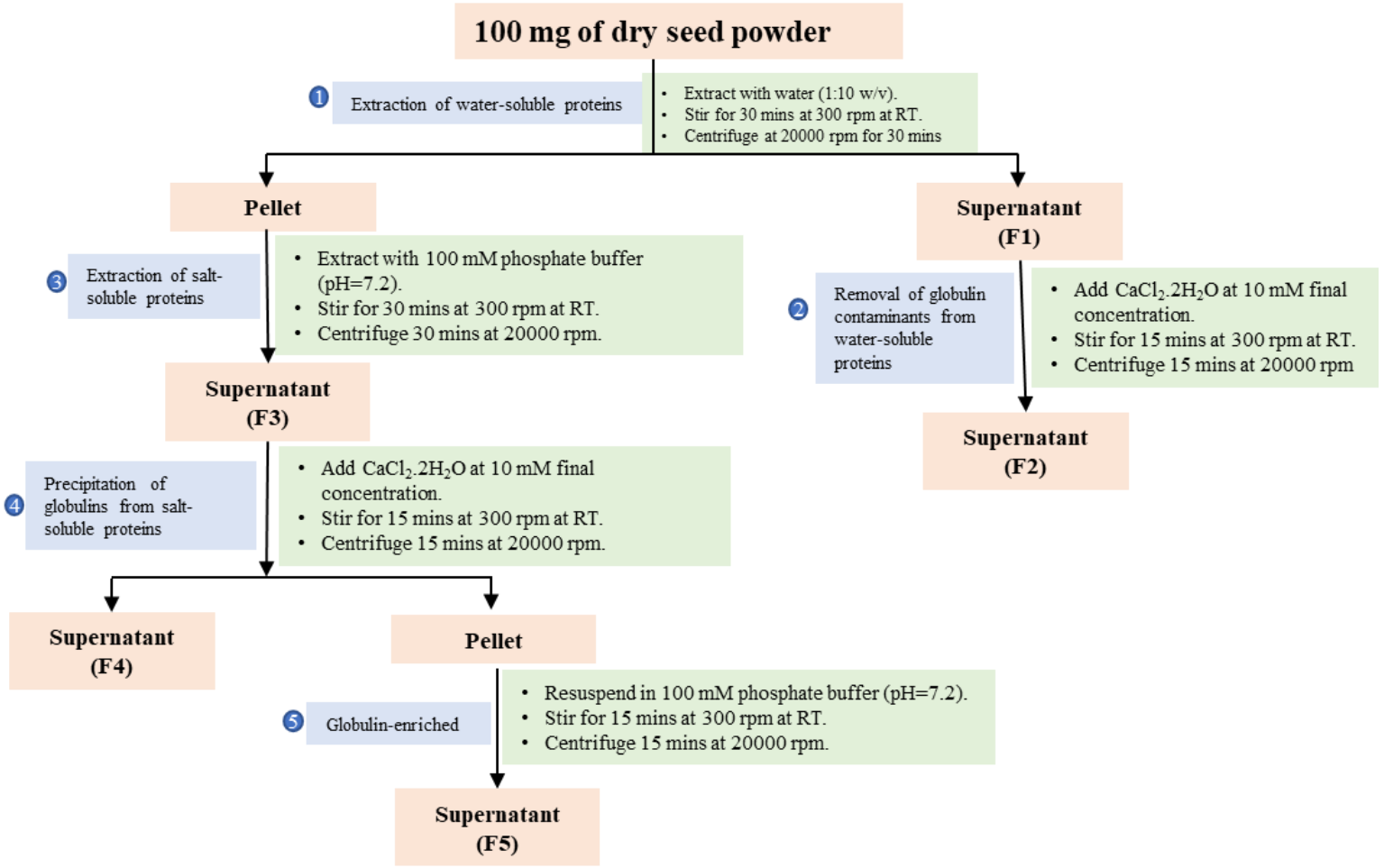

Figure S 3.1. Summary of the procedure used to fractionate *Vf* seed proteins based on their solubility in aqueous and salt solutions.

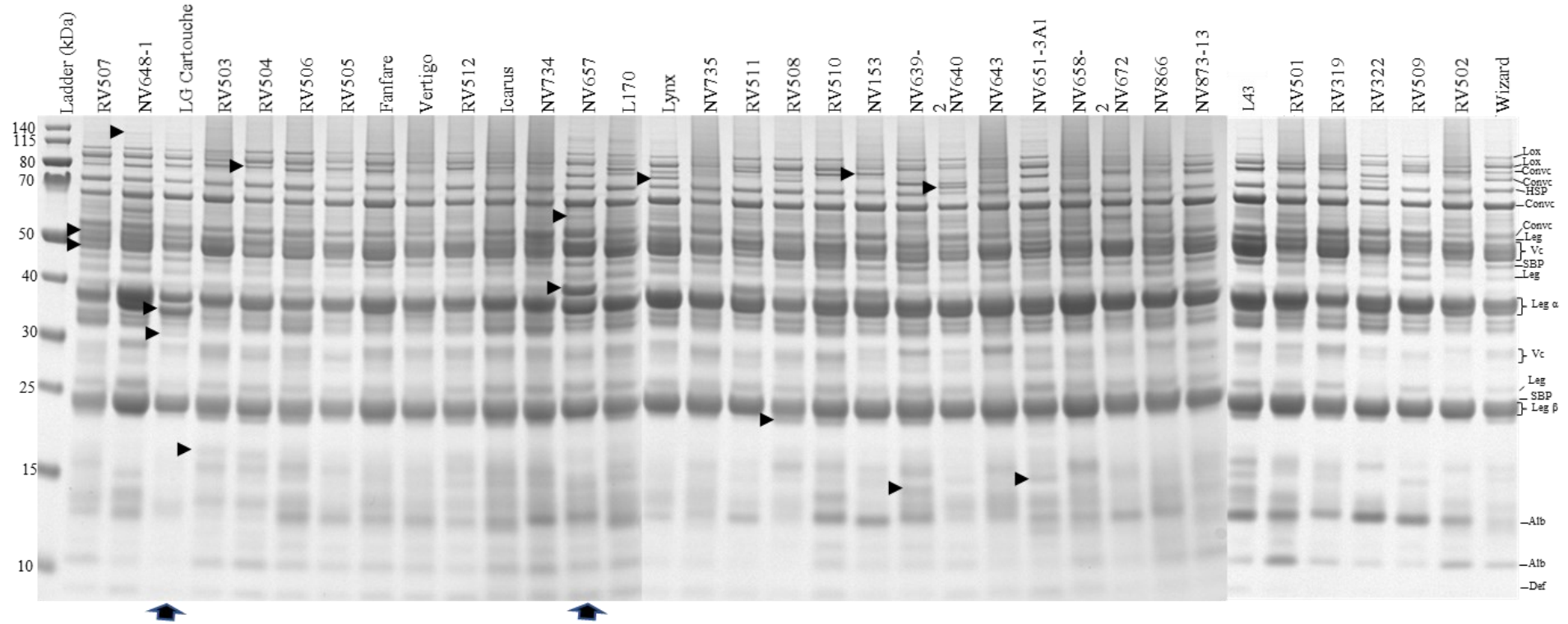

Figure S 3.2. SDS-PAGE profiles seed protein samples from 35 genetically diverse *Vf* genotypes. Arrows indicate polymorphic bands. Genotypes Cartouche and INRA 657 with the most prominent variants of legumin α are indicated by arrows. Lox= lipoxygenase; Convc=convicilin; vc=vicilin; HSP=heat shock protein; SBP=sucrose binding protein; Leg=legumin; Alb=albumin; Def=defensin.

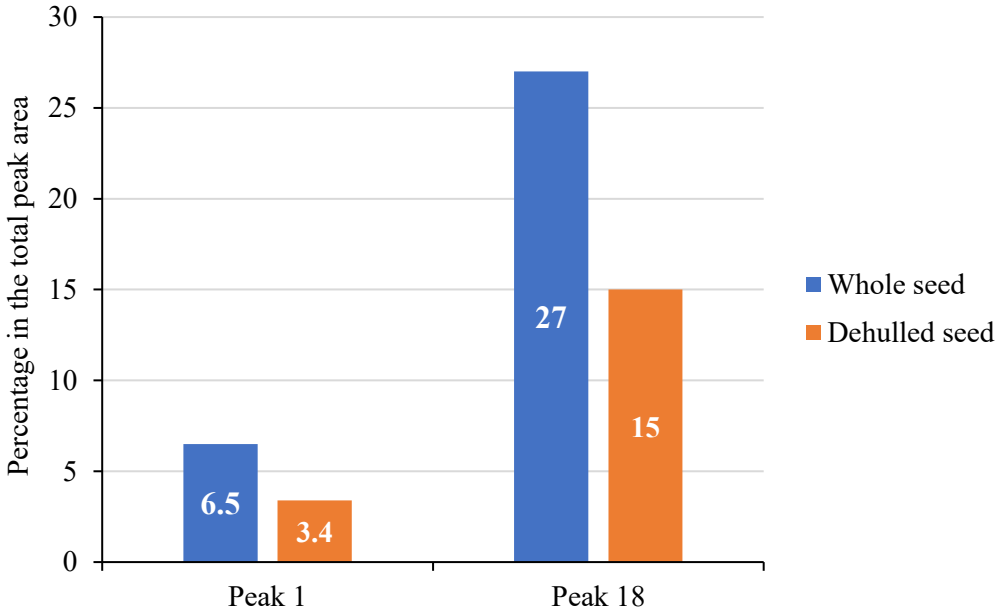

Figure S 3.3. A comparison between whole and dehulled seeds for the proportion of two SE-HPLC peaks (1 & 18) in which fractions did not contain proteins.

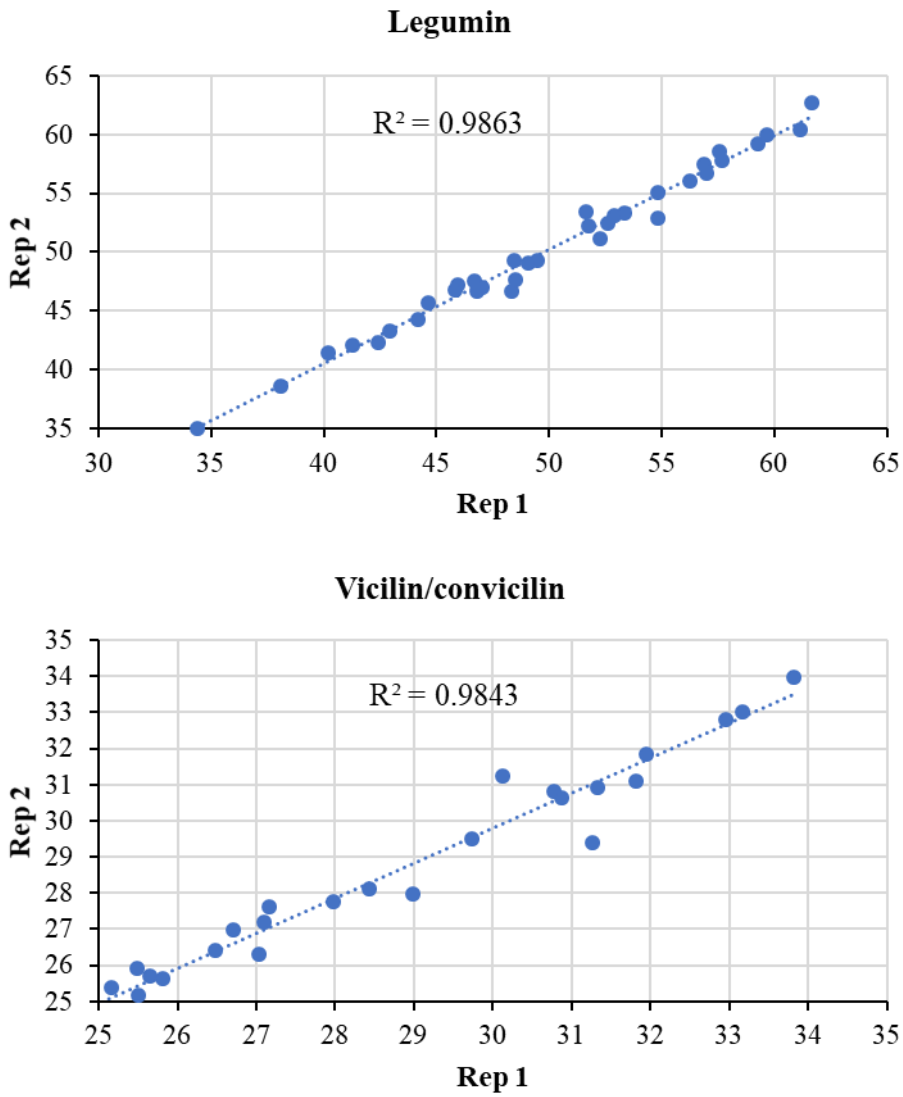

Figure S 3.4. Correlation between the values of two biological replicates in the quantification of legumin and vicilin/convicilin using SE-HPLC.

Chapter 4 Genetic control of protein content and composition in faba bean

Ahmed O. Warsame, Donal M. O'Sullivan, Deepti Angra and Vasiliki Tagkouli

4.1 Abstract

Faba bean (*Vicia faba*, L.) is a protein-rich grain legume which is considered one of the candidate crops to meet the increasing global demands for plant protein. In this study, using a panel of inbred lines from a multi-parent population and a high-density SNP array, we conducted a genome-wide association study (GWAS) for total seed protein content and the abundance of 24 seed protein subunits quantified on one-dimensional sodium dodecyl sulphate–polyacrylamide gel electrophoresis (1D SDS–PAGE) gels. For protein content, we identified three significant Marker Trait Associations (MTAs), of which two were identified by GWAS meta-analysis and together explained ~9% of the total phenotypic variation. For protein composition, 59 significant MTAs were detected for the abundance of 18 protein bands with some of the loci collocating at certain genetic regions. The identified MTAs included loci associated with the abundance of four legumin, two vicilin and multiple convicilin subunits. Also, genetic regions underlying the ratio between subunits of these globulin storage proteins were identified. Finally, exploiting the synteny and collinearity with *Medicago truncatula* genome, we have identified candidate structural and regulatory genes related to the significant GWAS associations. Our results are the first of their type in faba bean and lay the foundations for further genetic dissection of its seed protein quality. The study also demonstrated the power of synteny-guided GWAS scans in advancing the understanding of genetic underpinnings of traits in crops lacking genome sequence.

Key words: faba bean; protein content; legumin; convicilin; vicilin, GWAS.

4.2 Introduction

Vicia faba (hereafter, *Vf*) has one of the highest yielding (Cernay *et al.*, 2016) and nitrogen fixation capacity (Baddeley *et al.*, 2013) among crop legumes. It also has the second highest protein content after soybean with ~29% on average (Warsame *et al*., 2018). Globally, it is widely grown for human consumption, mainly in developing countries, and animal feed *(Duc et al*., 2015). Due to the increasing demand for plant-based protein driven by population growth and changing diets (Ismail *et al*., 2020), *Vf* will likely be a valuable protein resource in non-soybean producing regions, including the European Union and the UK, where this crop is well-adapted.

In a previous literature survey of genetic variation for total protein content in *Vf*, we found a wide diversity that ranged between 20% to 40% on a dry matter basis (Warsame *et al.*, 2018). This diversity not only presents a great opportunity for studying the genetic control of seed protein accumulation but also indicates the potential for developing cultivars with improved protein quality. However, to date, no progress had been made towards understanding the genetic control of this trait with not a single quantitative trait loci (QTL) for protein quantity or quality reported so far in *Vf* compared to over 250 QTL reported in soybean (https://soybase.org/). On the other hand, in the context of the development of protein-rich plant-based food products, understanding the genetic basis of seed protein constituents is a prerequisite for the development of cultivars with the right protein composition needed by the food processing industry. For instance, legumin-type proteins are known to contain more sulphur-containing amino acids compared to vicilin-type globulins (Warsame *et al.*, 2018; Martensson, 1980; Gatehouse *et al.*, 1980). Also, the relative proportions of specific protein subunits can influence the functional properties of proteins properties like gelation, solubility and emulsifying ability (Kesari *et al.*, 2017; Kimura *et al.*, 2008), which in turn may affect certain properties of food products (Poysa *et al.*, 2006). In *Vf*, we recently reported that protein composition analysed on SDS-PAGE and

SE-HPLC varies considerably between *Vf* lines (Warsame *et al.*, 2020), opening up the opportunity to map genetic loci controlling this variation. Such investigations have been carried out in other legumes. For example, several QTLs for major seed protein subunits have been mapped in soybean (Boehm *et al*., 2017; Ma *et al*., 2016; Panthee *et al*., 2004), while in *Medicago truncatula*, a close relative of *Vf*, Le Signor *et al.* (2017) have reported detailed information on the genetic regulation of the abundance of globulin proteins.

Until recently, *Vf* has been considered an 'orphan' crop (O'Sullivan and Angra, 2016) with little genetic information available on many important traits including protein quality. In this study, we provide a global view of the genetic control of seed protein content and composition in *Vf* by exploiting a newly developed high density SNP genotyping array, and the power of the genome-wide association approach to detect genomic regions underlying protein-related traits in segregating multi-parent population derived from genetically diverse founding lines. Here, we report for the first time, the genomic loci associated with total seed protein content and the abundance of protein subunits belonging to legumin, vicilin and convicilin and other seed proteins. Additionally, using the close synteny with the model plant *Medicago truncatula,* we report putative candidate genes for the abundance of some of these proteins.

4.3 Materials and methods

4.3.1 Plant material

The study material was initially constructed from 21 genetically diverse *Vf* genotypes (see Table 3.1 in chapter 3) that were cross-pollinated by bees for two cycles in which the most heterozygous and highest yielding plants were selected (**Figure S 4.1**). Then, from each of the 54 selected plants, four seeds were randomly selected to constitute 216 individual lines that were advanced via single seed descent (SSD) until the 4th selfing generation (S_4) (**Figure S 4.1**). In this way, line names consisted of a unique integer number ID of the 54 unique progenitor individuals concatenated with the sub-line number 1 to 4 e.g. 1135_1, 1135_2, 1135_3 and

1135_4. During population development by SSD, the winter generations were grown in a heated glasshouse in 3-litre pots filled with homogeneously mixed compost (John Innes No. 2, Clover Peat, UK) with drip irrigation and supplementary light using high pressure sodium lamps, while the spring/summer generations were grown outdoors in the soil under pollinator-proof cages (with supplementary irrigation).

4.3.2 Field experiments

The first trial was conducted from 22 March to 23 August 2019 and consisted of a subset of 149 S_3 lines (i.e. that had gone through three cycles of SSD) grown in 2-metre single rows consisting of 20 plants at the University of Reading's farm in Sonning, UK (51°28'47"N, 0°54'05.9"W). The experiment was laid out in an augmented design where test lines were replicated once and 8 times for five check varieties. The design was generated using the *agricolae* R package (Mendiburu and Yaseen, 2020). The second experiment was grown in from 2nd April to 14 August 2020 and comprised 168 S_4 lines (i.e. that had been advanced through four generations of SSD) were grown in a partially replicated design which was generated using *DiggeR*, an R package for generating efficient field experiment designs (Coombes, 2009). In this design, 20 lines with insufficient seeds were replicated once and 148 lines were replicated twice, while three *Vf* checks were replicated four times. In this trial, each plot consisted of two rows of a test line and two border rows belonging to the same check cultivar (LG Cartouche) sown between all plots in the experiment. Plot rows consisted of 10 plants spaced 10 cm apart with an inter-row spacing of 15 cm, giving an overall target population density of ~44 plants/m^2, which is in line with current recommendations for planting density of spring beans in the UK. In both seasons, data were collected on agronomic traits including flowering time (FT), 1000-seed weight (TSW) and grain yield. Also, sowing and harvesting was done manually. For yield data, it was based on single row in the first season while a 0.5 × 0.25 cm of the middle two rows were harvested in the second season.

4.3.3 Protein content and composition analysis

For protein content analysis, about ~25 g sample of oven-dried seed was finely ground in a milling machine (Perten LM 3100) and stored in air-tight 15 mL Falcon tubes until further analysis. Then, nitrogen content was measured in ~200 mg of seed flour using LECO Carbon/Hydrogen/Nitrogen Determinator (628 Series, LECO, USA). Subsequently, the nitrogen content was converted to protein content using a 5.4 conversion factor (Mosse, 1990).

In order to determine protein composition, total seed proteins were extracted according to Warsame *et al.* (2020) with minor modifications. To obtain a uniform ~20 µg/µL protein concentration in all samples, the sample to buffer ratio was adjusted based on the protein content in each seed sample. Then, ~10 µg of protein sample was separated on one-dimensional sodium dodecyl sulphate−polyacrylamide gel electrophoresis (1D SDS−PAGE) using a 15-well 4-12% Bolt™ Bis−Tris precast gels (NW04125BOX, ThermoFisher Scientific, UK). The test samples were randomly assigned to the middle 12 wells while wells 2 and 15 were reserved for a control sample (of inbred line Hedin/2) used as reference for protein bands across the gels. In addition, a protein ladder (10-140 kDa) was loaded on the 1st lane in all gels (**Figure 4.1**). Gels were run at 200 V for 75 minutes until proteins with molecular weight lower than ~30 kDa had run out of the gel. Although this meant losing all proteins under 30 kDa, it was necessary in order to improve the separation of higher molecular weight bands and increase the accuracy of quantification by minimizing the extent of overlap of similarly-sized protein bands. After gel electrophoresis, we used a microwave-based gel staining and distaining procedure described for the PageBlue Protein Staining Solution (ThermoFisher Scientific, UK). Gels were then put on an LED light pad and images were taken with the UVP Doc-it system (UVP LLC, USA). To quantify protein subunits, the band intensity of 24 protein subunits in each lane was measured using GelAnalyzer software (http://www.gelanalyzer.com/) and the percentage of each band was then calculated from the total intensity in each gel lane. From this data, the ratio between major

seed storage proteins were calculated with bands 10, 17 and 21+22 representing convicilin, vicilin and legumin subunits, respectively. Further details on the protein bands are in **Table S 4.1**.

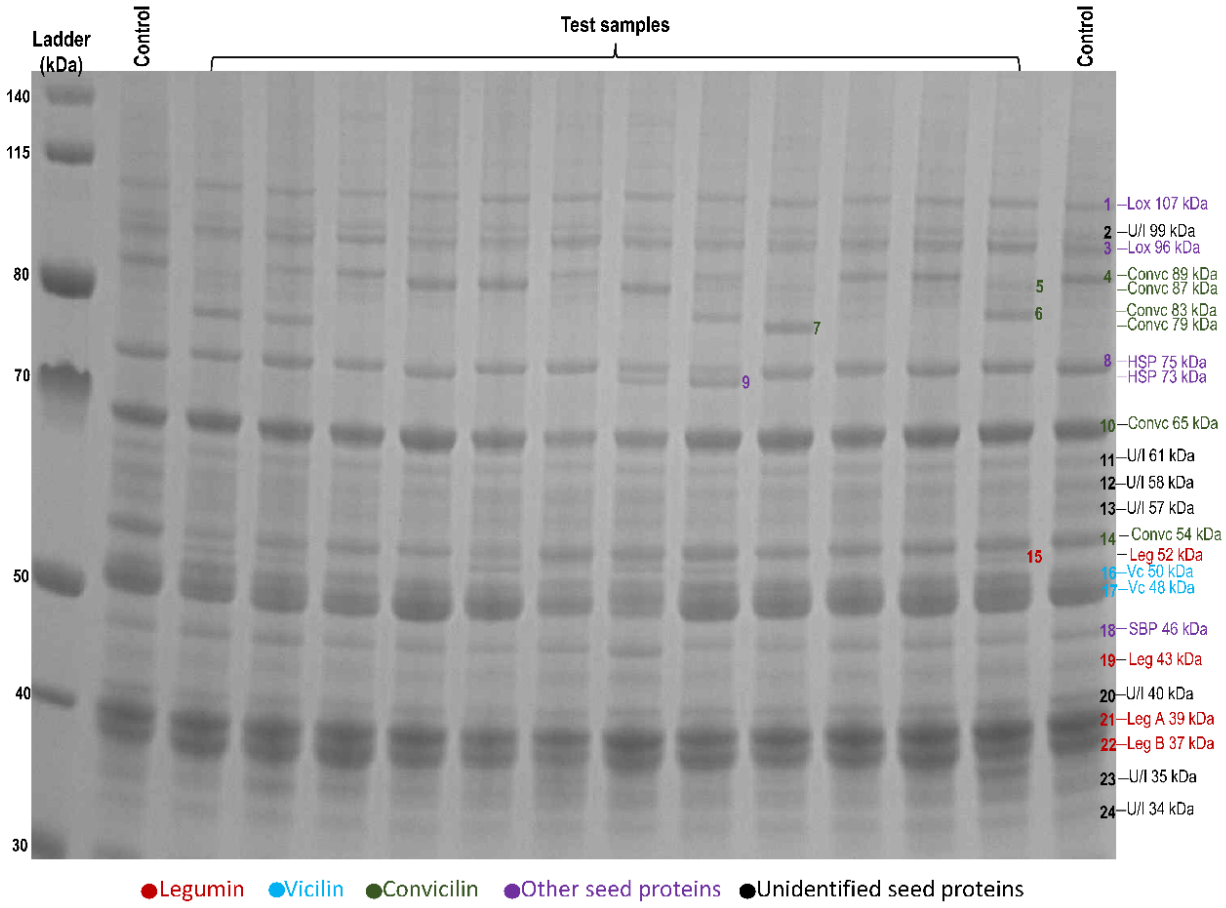

Figure 4.1. A representative 1D SDS-PAGE gel showing protein bands belonging to different protein classes that were used for GWAS analysis. Convc=convicilin, Leg=legumin, Lox= lipoxygenase, HSP= heat shock protein, SBP=sucrose binding protein, U/I=unidentified.

4.3.4 Statistical data analysis

Considering that field trials differed in the number of entries and experimental design used, the data from each year was initially analysed separately. The field and SDS-PAGE data was subjected spatial variation analysis using *SpATS* R package (Rodríguez-Álvarez *et al.*, 2018) which implements a smooth bivariate surface model that accounts variation in two dimensions as follows:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{u}, \boldsymbol{v}) + \boldsymbol{z_r c_r} + \boldsymbol{z_c c_c} + \boldsymbol{\varepsilon}$$

Where $\boldsymbol{y}$ is the phenotypic data at $\boldsymbol{i}$ plot, $\boldsymbol{f}(\boldsymbol{u}, \boldsymbol{v})$ is a smooth bivariate function with a smooth trend along the rows $\boldsymbol{f_u}(\boldsymbol{u})$ and columns $\boldsymbol{f_v}(\boldsymbol{v})$, $\boldsymbol{c_r}$ and $\boldsymbol{c_c}$ are random coefficients for the rows and columns, respectively, with $\boldsymbol{z_r}$ and $\boldsymbol{z_c}$ being the associated design matrices. The $\boldsymbol{\varepsilon}$ is the random error component of the model. For the field data, plots and blocks were the rows and columns, while the lanes in each gel and the different runs of SDS-PAGE represented the rows and columns, respectively. In both analyses, genotypes were considered as random and the best linear unbiased predictors (BLUPs) were extracted and used for GWAS analysis. On the other hand, to assess the effects of genotype, year and genotype × year on the abundance of protein subunits, we used data from a subset of 130 genotypes that were present in both years and fitted a General Linear Mixed (GLM) model in which different SDS-PAGE gels and lanes were considered random effects. This was done using *lme4* R package (Bates *et al.*, 2015). Finally, Pearson's r correlation coefficients were calculated with *Hmisc* R package (https://github.com/harrelfe/Hmisc/).

4.3.5 Genotyping and SNP calling

Genomic DNA was extracted from ~50 mg of young leaves from S_3 plants (i.e. that had undergone 3 selfing generations by SSD) using DNeasy 96 Plant Kit (QIAGEN Ltd, UK). DNA quality was assessed on agarose gel electrophoresis while concentration was assessed using Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher Scientific, UK) following manufacturer's guidelines. DNA samples at concentration of 35 ng/ul in 40 µl aliquots were genotyped using the Axiom 'Vfaba_v2' 60K array that has been developed from meta transcriptome data (Unpublished). Axiom Analysis Suite software (version 4.0, Thermofisher) was used for allele calling following best-practices workflow with default settings. SNPs where then filtered for missing data (>10%) and minor allele frequency (<5%). Also, SNPs and individuals with a heterozygosity rate greater than the mean + 1.5 standard deviations were excluded. Finally, a

custom R script was used to remove highly correlated SNPs ($r \geq 0.9$). This resulted in 12.5k good quality SNPs used for downstream analysis.

Linkage disequilibrium and population structure

A linkage map (unpublished) created from a 645 RILs of a 4-way cross described by Khazaei *et al.* (2018) was used for linkage disequilibrium (LD) and GWAS analysis. The map was built using GAPL software (Qu *et al.*, 2020) and contained ~10.8k SNPs of which 7.9k were informative in the study population. The LD (r^2) between markers was calculated using a subset of 4,050 SNPs using *sommer* R package (Covarrubias-Pazaran, 2016). Then, for each chromosome, significantly linked markers ($p < 0.001$) were retained and mean r^2 was calculated at 1 cM intervals.

Population structure was assessed using STRUCTURE software (Pritchard *et al.*, 2000) in which the number of populations (K) was set from 1 to 10, and 50,000 burn-in time and 100,000 MCMC iterations. The output from STRUCTURE was analysed in STRUCTURE HARVESTER (Earl and vonHoldt, 2012) and the K value from 1 to 10. The estimated number of subpopulations was determined based on the rate of change (ΔK) in mean log likelihood (LnP(K)) across different K numbers between 1 to 10. The STRUCTURE clustering was cross-checked with two model-free approaches: a distance-based hierarchical clustering using *hclust* function in R (R Core Team, 2020) and multivariate analysis method called Discriminant Analysis of Principal Components (DAPC) (Jombart *et al.*, 2010) implemented in *adegenet* R package (Jombart and Ahmed, 2011).

4.3.6 Genome-wide association analysis

For GWAS analysis, in addition to the 7.9k mapped SNPs, an additional 4,671 good quality unmapped SNPs were included. These additional markers were inserted in the genetic map in the syntenic order based on previously described blocks of collinearity between *V. faba* and *M. truncatula* (Webb *et al.*, 2016) and assigned with arbitrary loci positions for marker ordering

purposes. The fixed and random model circulating probability unification (FarmCPU) model (Liu *et al.*, 2016) implemented in *GAPIT* R package (Wang and Zhang, 2018) was used for the GWAS of all data. Furthermore, considering that phenotypic data was obtained from 137 and 156 lines in 2019 and 2020 trials, respectively, the FarmCPU GWAS analysis was performed for each year separately. Then, sample size weighted (p-values based) GWAS meta-analysis was conducted using METAL software (Willer *et al.*, 2010).

4.3.7 Estimating variance explained by significant loci

Since the above GWAS models did not produce phenotypic variances associated with the significant loci, a separate mixed linear model described by (Tang *et al.*, 2019) was used to estimate the genetic and phenotypic variances explained by the identified QTL region which is defined by the estimated LD decay distance of 2 cM on either side of the significant SNP. When the SNP is not mapped, only that SNP is considered for the analysis. The linear mixed model can be expressed as follows:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Z_i u_i} + \boldsymbol{\varepsilon}$$

Where $\boldsymbol{y}$ is the phenotypic BLUPs, $\boldsymbol{b}$ is a matrix containing fixed effects, $\boldsymbol{u}$ is effect of the SNPs in the significant GWAS region treated as random effect. The $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices for $\boldsymbol{b}$ and $\boldsymbol{u}$, respectively, while $\boldsymbol{\varepsilon}$ represents the residual error component of the model.

4.4 Results and Discussion

4.4.1 GWAS panel characteristics

After three generations of inbreeding by SSD, the overall rate of heterozygosity of the 12.5k SNPs used in GWAS analysis was 3.3 % while the mean minor allele frequency was 0.29 (**Figure S 4.2** A&B, respectively). The pattern of LD between SNPs was similar across all 6 individual chromosomes, decaying to below r^2=0.2 at ~2 centimorgans (cM) (**Figure S 4.2** C). This intermediate level of LD decay is typical of structured multi-parent populations where a

defined number of outcrossing generations separate the progenitor haplotypes from the inbred progeny but is likely to be lower than the levels of decay expected in a GWAS panel of unrelated inbreds of a partially allogamous species like *Vf*. With about 16 mapped markers per 1 cM on average, the marker density deployed is therefore not likely to be the limiting factor in determining power to detect associations.

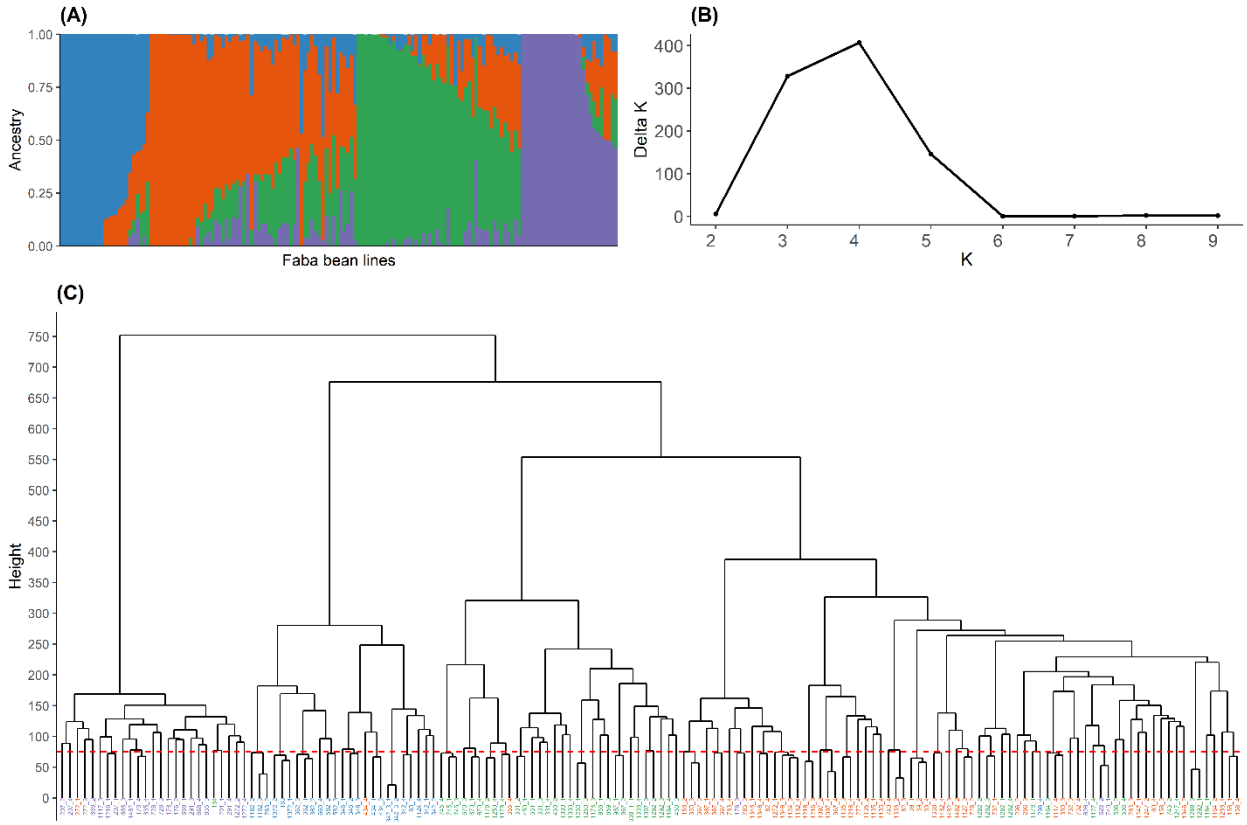

Figure 4.2. The structure of the Reading Spring Bean population. Four subgroups inferred by STRUCTURE (A) based on ΔK at K=1 to 10 (B). The dendrogram (C) shows the distance-based hierarchical clustering in which individuals are coloured based on the subgroups determined by STRUCTURE software. The height axis shows the distance between genotypes and/or clusters while the dashed horizontal line indicates the approximate genetic distance below which most siblings are clustered.

Population structure analysis using STRUCTURE software revealed four main subpopulations (**Figure 4.2** A&B) with significant levels of admixture however between subpopulations. A similar pattern was also shown by the distance-based hierarchical clustering of the individuals (**Figure 4.2** C) and Discriminant Analysis of Principal Components (data not shown). Strong kinship relationships were expected between (up to four) outcrossed progeny

lines tracing back to a single selected individual, and these sibling relationships represented by shared line IDs explain most of the branch lengths of genetic distance less than 75. The existence of the higher order structure in the population suggests that the cross-pollination was not uniform, likely due to differences in maternal and paternal outcrossing success as recently described by Brünjes and Link (in press). Also, this non-equal chance of outcrossing could be attributed to factors including differences in flowering times among genotypes or floral traits which may vary among *Vf* genotypes that can affect bee preferences for certain genotypes. For instance, according to Bailes *et al.* (2018), floral traits like nectar sugar concentration and its accessibility in *Vf* flower is a key driver of bee preferences and, as a result, the success of the cross-pollination. In GWAS analysis, population stratification is considered a confounding factor that can lead to erroneous trait-marker associations (Sul *et al.*, 2018; Price *et al.*, 2010) and therefore it is included as covariate in the model. According to Liu *et al.* (2016), inclusion of PCs in the FarmCPU as covariates can potentially control false positives that are associated with population structure.

4.4.2 Phenotypic data and trait correlations

Distribution of the raw data for protein content and proportion of protein bands is shown in **Figure 4.3**. The protein content of the study lines varied significantly ($p \leq 0.001$) between the years and ranged from 16 to 33% and 21 to 34% with a mean of ~24 % and ~27% in 2019 and 2020, respectively (**Figure 4.3**A). The major environmental factor that adversely affected *Vf* performance in 2019 was the infestation by bean aphid (*Aphis fabae*), which is a major *Vf* pest in the UK. For the seed protein composition, while the effect of year was significant ($p \leq 0.05$) for more than half the protein bands, some major bands including leg A and B, and convc 65 kDa were relatively stable across years (**Figure 4.3**B, **Table S 4.2**). Curiously, both vicilin subunits had higher abundance in 2019 when protein content was lower. Therefore, the underlying trait trade-offs were further investigated by conducting pairwise trait correlations for agronomic and quality attributes (**Table 4.1**).

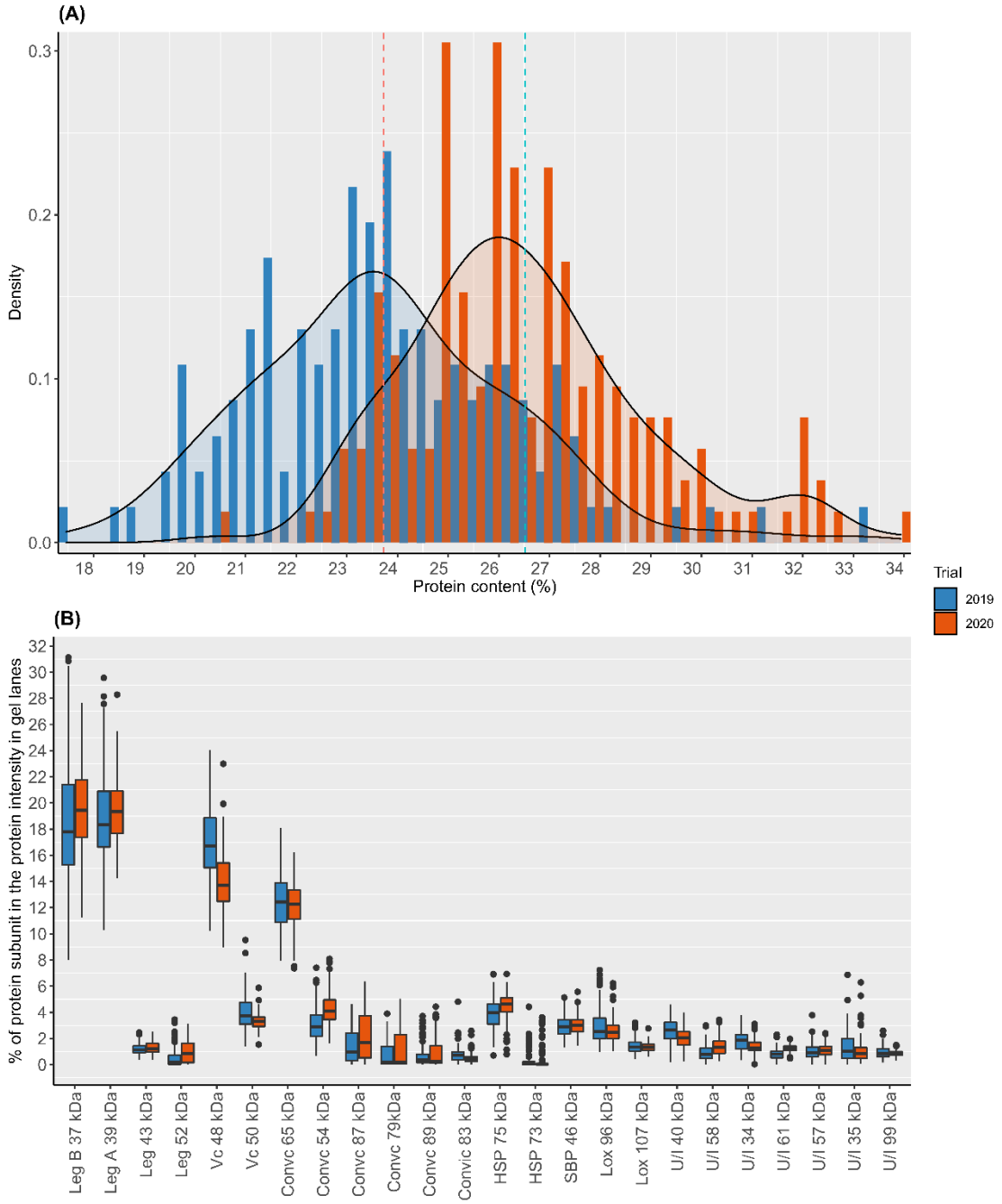

Figure 4.3. Distribution of raw phenotypic data of protein content and composition of study lines in 2019 and 2020 trials. (A) Histogram of protein content (%) where the vertical dashed lines indicate mean of each year. (B) Boxplot showing the variation in the abundance of 24 protein subunits in which the y-axis is the percentage of the intensity for each subunit calculated from the total protein intensity in SDS-PAGE gel lane.

Table 4.1. Correlations between agronomic and protein quality traits during 2019 and 2020 field trials.

	Traits	FT	TSW	GY	Protein (%)	Leg B 37 kDa	Leg A 37 kDa	Vc 48 kDa	Convc 65 kDa	HSP 75 kDa	Convc 54 kDa	
2019	FT	0.52***	-0.38***	0.28**	-0.50***	-0.21*	0.02	0.02	-0.03	0.26**	-0.19*	2020
	TSW	-0.29***	0.70***	0.26***	-0.01	-0.16	-0.08	0.29***	-0.03	0.04	0.35***	
	GY	-0.14	0.41***	0.38***	-0.39***	-0.20*	0.03	0.19*	0.01	0.14	0.01	
	Protein (%)	-0.32***	0.31***	0.18*	0.33***	0.37***	0.50***	-0.22***	0.15	-0.48***	-0.13	
	Leg B 37 kDa	-0.09	0.08	-0.12	0.20*	0.43***	0.23**	-0.43***	-0.35***	-0.40***	0.06	
	Leg A 37 kDa	-0.15	-0.01	0.01	0.19*	0.18*	0.30***	-0.43***	-0.01	-0.33***	-0.23***	
	Vc 48 kDa	0.15	0.07	0.07	0.00	-0.29***	-0.45***	0.48***	0.37***	-0.02	-0.11	
	Convc 65 kDa	-0.14	-0.03	0.07	0.25**	-0.16	-0.09	0.35***	0.57***	-0.21*	-0.53***	
	HSP 75 kDa	0.37***	-0.09	0.03	-0.42***	-0.27**	-0.24**	-0.01	-0.21*	0.49***	0.12	
	Convc 54 kDa	0.01	0.24**	0.00	-0.11	-0.01	-0.31***	0.15	-0.21*	0.11	0.65***	

The lower and upper diagonals are correlations in 2019 and 2020, respectively, while the middle thick-border cells are the correlation between years for the same trait. The colour gradient (green-yellow-red) indicates the direction and strength of the correlation (strong positive-no correlation-strong negative). *, **, *** statistically significant at $p \leq 0.05$, 0.01, 0.001, respectively. FT=flowering time; TSW=1000-seed weight; GY=grain yield.

The agronomic traits included in the correlation analysis were flowering time (FT), 1000-seed weight (TSW) and grain yield (GY). The relationship between protein content and yield attributes was strongly modulated by the environment. For instance, in 2019, when protein content was generally lower, the correlation with GY was weak but positive compared to 2020 where it was significantly negative and stronger. Other traits were consistently correlated across both years, such as the consistently negative correlation of FT with protein content (**Table 4.1**). Interestingly, while late flowering genotypes were associated with lower protein content overall, they tended to accumulate more HSP 75 kDa which, given the role of Heat Shock Proteins as chaperones induced by abiotic stress (Banerjee and Roychoudhury, 2018), could be as a response to heat stress conditions experienced by seeds whose filling is shifted into the hottest weeks of the season.

In terms of correlations involving the globulin seed storage proteins, heavier seeds had more Convc 54 kDa and both Leg A and B had a significant positive correlation with seed protein content in both years. In contrast, the abundance of HSP 75 kDa was negatively correlated with both seed protein and legumin content (**Table 4.1**). Similarly, Vc 48 kDa was negatively correlated with legumin subunits. Previously, we found a strong negative correlation ($r = -0.83$) between vicilin/convicilin and legumin fractions quantified by SE-HPLC (Warsame *et al.*, 2020). Together, these results suggest that increasing overall protein content and at the same time the relative content of legumin, which contains more sulphur-containing amino acids and is therefore could be a predictor of protein quality, could be achieved. However, the proportion of certain proteins can be modulated by the environment such as soil fertility (Krishnan *et al.*, 2005), an therefore, such variation should be accounted by using an appropriate experimental design and statistical methods when selecting for protein composition.

4.4.3 Genome-wide association analysis

Several statistical models to test trait-marker associations have been developed and are widely used in GWAS studies. According to Kaler *et al.* (2020), fixed and random model circulating probability unification (FarmCPU) is considered more powerful than other common models in controlling both type I & II errors. To confirm this conclusion, we first compared FarmCPU with General linear Model (GLM) and Mixed Linear Model (MLM) using two qualitative traits (flower and seed hilum colour) which have relatively known genetic control Using morphological data gathered only in the 2020 trial, FarmCPU identified two significant SNP associations (FDR≤0.05) for flower colour and three SNPs associated with hilum colour (**Figure S 4.3**). The identified SNPs for both traits were located within already known regions containing causative genes. For flower colour, the SNP AX.181489312.Medtr3g092090.1 on Chr 2 collocates with the *zt-1* gene responsible for zero tannin and white flower colour (Webb *et al.*, 2016) while AX.416742604.Medtr1g070380.1 on Chr 3 collocates with another complementary *zt-2* gene (Gutierrez *et al.*, 2020). Similarly, the significant SNP, AX.416810010.Medtr2g013580.1, on Chr 1 is mapped within the hilum colour locus region (Chapter 6). However, both GLM and MLM could not detect the *zt-1* locus for flower colour while many significant associations were detected in *zt-2* and hilum colour loci (data not shown). On this basis, the FarmCPU model was used for further GWAS analysis for protein content and composition.

4.4.4 Seed protein content

FarmCPU analysis of the 2020 trial and the GWAS meta-analysis of the across-year BLUPs identified two significant SNPs on Chr 1 and 4 (**Figure 4.4**A) that explained 5.6% and 3% of the total variation for seed protein content, respectively. Interestingly, a third locus on Chr 6 that was detected only with FarmCPU analysis in 2020 could explain up to 11% of the phenotypic variation. The inability to detect any significant associations in 2019 data could be

attributed to the smaller population size and the large environmental effects, compounded by a lack of replication, which may have masked genotype-dependent differences in protein content. As shown in **Figure 4.4**B, the identified loci show a clear SNP allele effect on protein content in both years. In *M. truncatula*, a SNP in the same gene as the SNP AX.181200178.Medtr7g013770.1 was previously found to be significantly associated with one of the convicilin subunits (Le Signor *et al.*, 2017). This is the first report on the genetic control of protein content in *Vf* and therefore these are considered novel QTL regions. However, considering that protein content is a polygenic trait which is controlled by many genes involved in various plant processes including nutrient uptake and assimilate transport (Egle *et al.*, 2015; Peng *et al.*, 2014; Rolletschek *et al.*, 2005; Poeta *et al.*, 2017) and photosynthate supply (Weichert *et al.*, 2010), it is likely that there are many other genomic regions associated with small effects on this trait that may be below the limits of detection of this study. In other crops, the most prominent QTL for protein content has been related to a loci harbouring an amino acid transport gene as has been reported in rice (Zhong *et al.*, 2011), soybean (Zhang *et al.*, 2017) and oilseed rape (Gacek *et al.*, 2018). In addition, several QTLs for root or nodule traits and seed nitrogen accumulation have been mapped in overlapping genomic locations in pea (Bourion *et al.*, 2010).

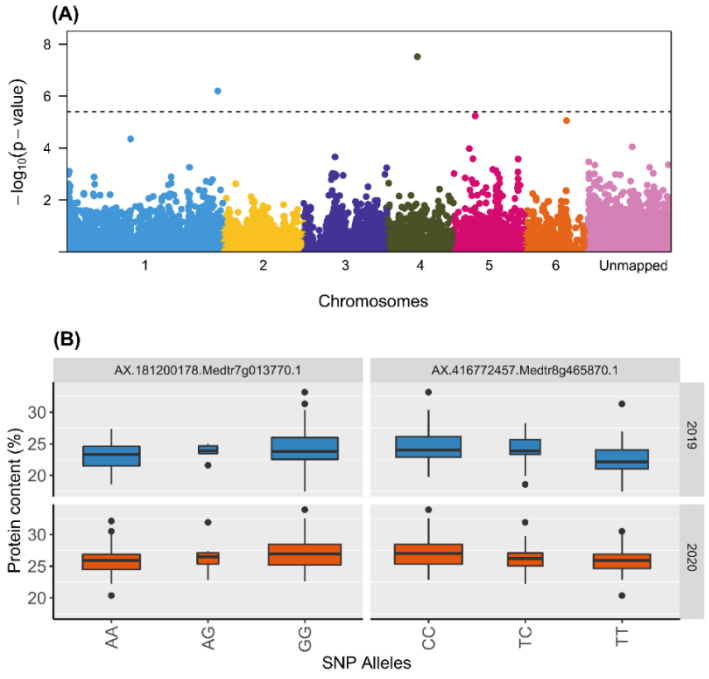

Figure 4.4. (A) Manhattan plot showing probability of marker associations with protein content from GWAS meta-analysis. (B) The allelic effects of the significant SNPs in 2019 and 2020. The width of the SNP allele boxes is scaled by the square root of the number of individuals carrying that allele.

Figure 4.5. Manhattan plots of GWAS meta-analysis showing significant associations for eleven protein subunits and ratio between three major protein subunits (coloured in red). For plotting purposes, the unmapped SNPs were arranged based on their order in *M. truncatula* genome and given arbitrary positions. For Convc 79 kDa and HSP 73 kDa, the associated SNPs exceeded 30 -$\log_{10}$(*p*-value) and their data points are off-scale in this figure but can be read from **Table S 4.3**.

4.4.5 QTL for seed protein composition

In total, 58 significant SNPs ($FDR < 0.05$) associated with 18 of the 24 protein subunits were detected by FarmCPU year-by-year and across years meta-analysis (**Table S 4.3**, **Figure 4.5**). Some of these loci associated with more than one protein subunit or class. In **Figure 4.5**, only associations identified in the across-years meta-analysis are shown. Of the major legumin subunits, Leg B 37 kDa was the most abundant and had two significant associations of which the one on Chr 3 explained up to 15% of the phenotypic variation. For Leg A 39 kDa, meta-analysis revealed an unmapped locus which was significantly linked to its abundance. Furthermore, it was found that the proportion between the two legumin alpha subunits (Leg A and B) deviated from the expected 1:1 ratio assuming a simple model where both subunits are required in equal proportions to form the mature hexameric legumin. Variation in the Leg A:B ratio was related to a region on Chr 5 which was neighbouring another QTL for Leg 52 kDa (**Table S 4.3**). Besides containing a legumin structural gene, this region is also syntenic and colinear with a *M. truncatula* region that contains the bZIP-ABI5 transcription factor gene which has been reported to be at the centre of network of genes regulating abundance of globulin proteins (Le Signor *et al.*, 2017). These results suggest a complex genetic control of subunit composition of *Vf* legumins. In this species, at least two A-type and several B-type legumins are known (Fuchs and Schubert, 1995; Horstmann *et al.*, 1993; Baumlein *et al.*, 1986) but existence of more complex combinations in the final hexamer legumin and the involvement of multiple structural genes and regulatory loci cannot be ruled out. In fact, Tucci *et al.* (1991) reported 29 disulphide-linked αβ legumin subunit pairs with molecular weights between 39-81 kDa with little known about the genetic control of subunit composition in *Vf*.

On the other hand, a locus on Chr 1 (46.1 cM) strongly associated with Leg:Vc ratio was also found to underlie abundance of Vc 48 kDa (**Figure 4.5**, **Table S 4.3**). This overlap gives some confidence in the accuracy of the protein subunit quantification process considering that the Leg:Vc ratio was result of independent quantification of three separate bands ((Leg A 39 kDa

+ Leg B 37 kDa)/Vc 48 kDa). Another region near the telomere of Chr1 (1.6-2.4 cM) was a hotspot of QTL for different protein subunits. These included Leg 43 kDa, Vc 50 kDa, Convc 65 kDa, Convc 54 kDa and U/I 58 kDa (**Figure 4.5**, **Table S 4.3**). Moreover, SNPs significantly associated with Leg:Convc and Vc:Convc ratio were also located in this region. Although co-location of QTL governing different protein subunits is not uncommon (Ma *et al.*, 2016; Zhang *et al.*, 2017), its interpretation would depend on the nature of relationships between these proteins. For instance, a common genetic locus for certain protein classes may result from causative alleles having an opposite effect on the abundance of negatively correlated proteins, as in the case for Convc 65 kDa and Convc 54 kDa which are negatively correlated with each other. In pea, the role of *ABI5* in globulin abundance was shown to be related to its regulation of legumin:vicilin ratio in which a mutation in the *ABI5* gene leads to a significant increase in legumin at the expense of vicilin (Le Signor *et al.*, 2017).

Convicilin proteins were of particular interest in terms of their diversity and genetic regulation. The most abundant subunit was Convc 65 kDa which, together with two major QTL in Chr 1 (1.6 cM), was associated with loci in Chr 2,3&4 (**Table S 4.3**). The QTL on Chr 3 at ~59 cM coincides with the location of a convicilin structural gene at 58 cM while the significant SNP in Chr 2 belongs to a gene annotated as a transmembrane amino acid transporter family protein. Regarding the latter, the role of some amino acid transporters in seed protein accumulation has been previously reported in *Vf* (Rolletschek *et al.*, 2005) and soybean (Zhang *et al.*, 2017; Cheng *et al.*, 2016). In addition, loci on Chr 1 (~45-48 cM) were associated with the variation in three protein bands of 89, 87 and 79 kDa (**Figure 4.5**, **Table S 4.3**). This locus contained four significant SNPs, including one within a gene belonging to a cupin family protein (**Figure 4.6**). This protein, also referred to as an RmlC-like cupin protein, is found in plant seeds and functions as a storage protein (Gábrišová *et al.*, 2016; Sghaier-Hammami *et al.*, 2020; Yobi *et al.*, 2020). On this basis, it could be hypothesized that this gene is the likely coding gene for these high molecular weight seed proteins and that certain mutations within this gene are

responsible for the protein band variations. This finding, however, is at odds with the identification by mass spectrometry of some of this cluster of protein bands as convicilin in *V. faba* (Warsame *et al.*, 2020) and *M. truncatula* where similar annotation is given to protein subunits of comparable molecular weights (Le Signor *et al.*, 2017; Le Signor *et al.*, 2005). To further investigate the relationship between the cupin-like storage protein and other globulin seed proteins (vicilin, convicilin and legumin), we retrieved sequences of 54 seed storage proteins belonging to 11 legume species and conducted phylogenetic analysis. As expected, two major clusters–legumin versus others–were identified (**Figure S 4.4**). Within the non-legumin cluster, all "cupin-like" storage proteins formed a distinct subgroup that diverged early from the vicilin/convicilin types. Although this is an intriguing observation and gives this poorly characterized protein a clear position among legume seed storage proteins, it does not resolve the relationship between identity of the protein variants and the candidate gene identified in this study and suggests the need for further in-depth investigation.

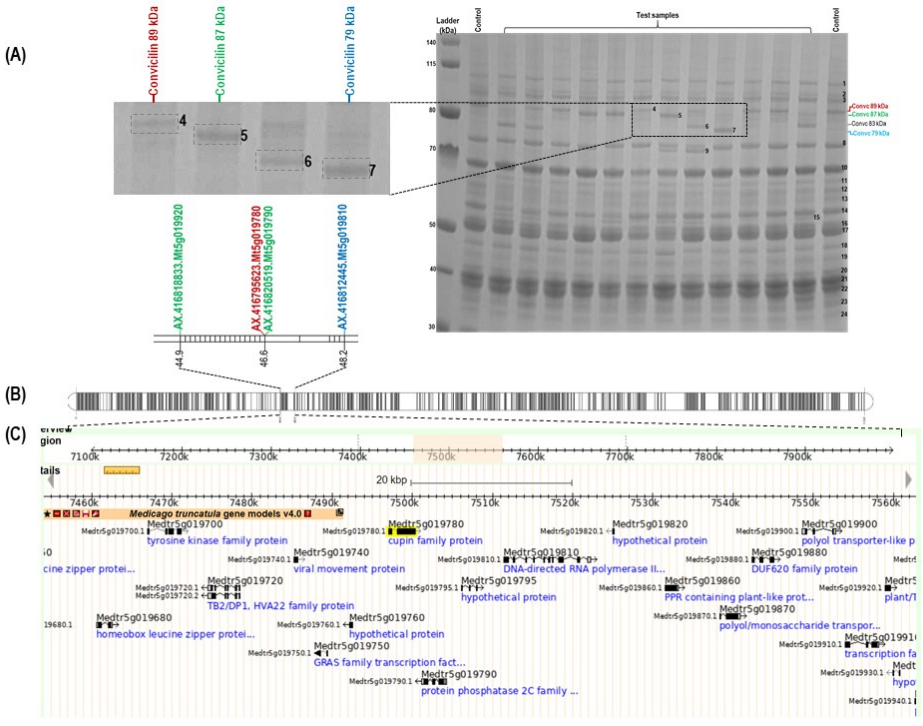

Figure 4.6. Genetic interval in *V. faba* which harbours loci associated with variation in convicilin-like bands >78 kDa and gene content of syntenic *M. truncatula*-region. (A) SDS-PAGE gel showing seed protein band variants. (B) SNPs on *V. faba* Chr 1 significantly associated with the abundance of three of the protein variants. The protein band and SNP labels with the same colour indicate that they are associated. (C) The *Mt* Chr 5 region containing a candidate gene annotated as cupin-like protein.

Heat shock proteins (HSP) and sucrose binding protein (SBP) that represented on average about 4% and 3% of the total protein in the gel lanes, respectively, were among the other seed proteins that showed significant GWAS signals. One locus associated with both HSP bands (75 and 73 kDa) was mapped on Chr 1 (158.6 cM), less than 1 cM distant from a gene annotated as heat shock protein 70 (HSP70)-interacting protein (Hip). Although the literature on the role of Hip in stabilizing and enhancing the function HSP comes from mammalian organisms (Webb *et al.*, 2001), they are predicted to play similar regulatory roles in plants (Nelson *et al.*, 2004). Moreover, although the HSP 75 kDa was present in all genotypes but quantitatively varied, the abundance of the HSP 73 kDa band was absent in some genotypes which may indicate a differential stress response among the genotypes. This hypothesis is supported by the finding that certain HSP70 isoforms from different pea tissues were expressed only as response to heat shock while others were transcribed constitutively (DeRocher and Vierling, 1995).

As for SBP, which, besides being storage protein, is thought to have sucrose transport and seed desiccation-related functions (Heim *et al.*, 2001), GWAS meta-analysis identified a locus on Chr 3 (~32 cM), near the location of the structural gene for this protein (~31 cM). From a nutritional point, these proteins are not considered as important as globulins, but as they account for several percent of total protein, it is of interest nonetheless to establish the effect of their relative abundance on the overall functional and quality properties of seed protein extracts.

4.4.6 Translational validation of faba bean seed protein genes and QTLs

Vicia faba has one of the largest genomes among legumes (~13 Gb) which has not yet been sequenced. Fortunately, the considerable synteny and collinearity of *Vf* with the model legume, *Medicago truncatula,* has been an instrumental tool in understanding its genome and genetics (Gutierrez and Torres, 2019; Webb *et al.*, 2016; O'Sullivan and Angra, 2016). Here, using a high density map with about 86% of the SNPs tagged by *M. truncatula* genomic positions and seed protein composition GWAS data from Medicago (Le Signor *et al.*, 2017), we show that it is

possible to predict the candidate loci for *V. faba* protein composition QTL within regions of uninterrupted micro-synteny and collinearity between the two genomes.

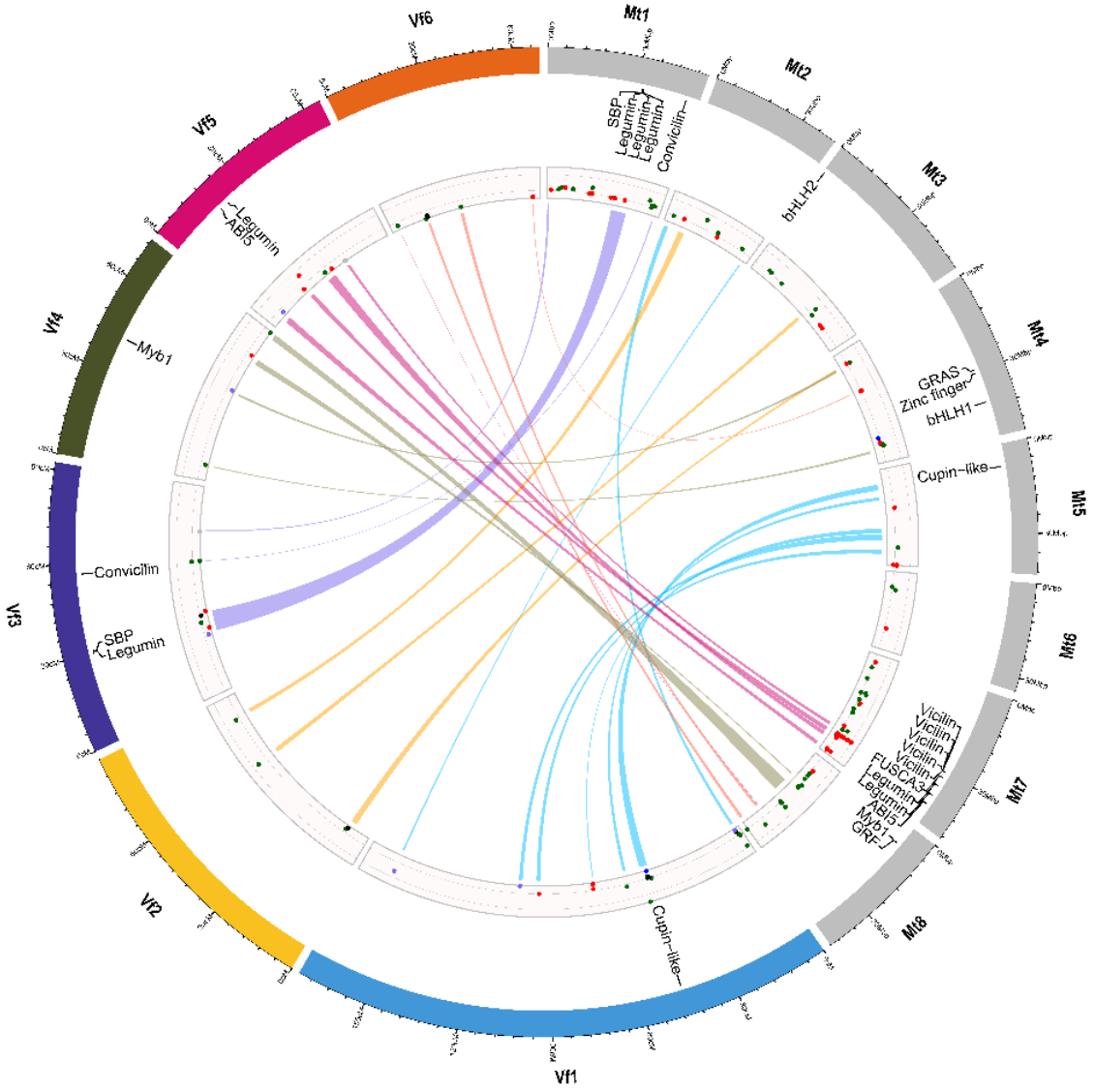

Figure 4.7. Circos plot showing synteny between *V. faba* and *M. truncatula* genomes and the positions of structural genes and QTL of seed storage proteins. The outer circle represents the six *V. faba* and eight *M. truncatula* chromosomes. The locus labels on the inside of the chromosome track are structural genes for major seed proteins and transcription factor genes with known roles in seed protein regulation while the track plot data is -$\log_{10}$(*p.value*) of GWAS results (GEMMA at FDR≤ 0.05) from Le Signor *et al.* (2017) for *M. truncatula* and GWAS results (FarmCPU and METAL at FDR≤ 0.05) from this study for *V. faba*. The colour of the points represents associations with different protein classes: red=legumin, dark green=convicilin, vicilin=blue, legumin:vicilin/convicilin ratio=orange, other seed proteins=slate blue, grey=unidentified proteins. The links are between SNPs in *V. faba* genome containing significant GWAS associations and collinear regions in *M. truncatula*.

As shown in **Figure 4.7**, the overlap in the seed protein QTL region between the two genomes is considerable with many of the identified *V. faba* QTL residing at or near regions where Medicago seed protein QTL and structural or regulatory genes are mapped. For instance, the two legumin structural genes mapped in *V. faba* were located at regions in Chr 3 and 5 which were colinear with locations on Chr 1 and 7 of *M. truncatula*, respectively, that harbour both structural and transcriptional regulators that are highly associated with legumin abundance (**Figure 4.7**). It is possible that the locus on *V. faba* Chr 5 (~ 23.6 cM) which accounted for 37-87% and 13-17% of the genetic and phenotypic variances in Leg A:B ratio, respectively, is indeed related to the legumin gene located at 22.6 cM on the same chromosome. This syntenic framework also show the locations of putative genes coding for Convc 65 kDa, SBP 46 kDa, and the transcription factor ABI5 (**Figure 4.7**). On the other hand, the synteny-based approach to mine candidate QTL regions could not provide any clue about the major seed protein composition QTL near the end of Chr 1 (1.5-2.4 cM), which could be unique to *Vf* or has yet to be discovered in Medicago.

In conclusion, despite the importance of protein content and composition in the utilization of a food and feed crop like *Vf*, the genetic understanding of these traits was lacking. In this study, we provided the first insights into the genetic control of the total seed protein content and the abundance of different seed proteins. Some of the identified regions for these traits explained large proportions of the genetic and phenotypic variation which suggested that they should be regarded as large effect QTL suitable for fine-mapping through more formally structured mapping populations and the development of diagnostic markers that can be used in marker-assisted breeding. Moreover, with the expected completion of the *Vf* genome in the near future, this study paves the way for further understanding of the genomic regulation of seed proteins in this important crop.

4.5 References

Baddeley, J. A., Jones, S., Topp, C. F. E., Watson, C. A., Helming, J. & Stoddard, F. L. (2013). Biological nitrogen fixation (BNF) by legume crops in Europe. *Legume Futures Report* 1.5.

Bailes, E. J., Pattrick, J. G. & Glover, B. J. (2018). An analysis of the energetic reward offered by field bean (*Vicia faba*) flowers: Nectar, pollen, and operative force. *Ecology and Evolution,* **8** (6), 3161-3171.

Banerjee, A. & Roychoudhury, A. (2018). Chapter 19 - Small heat shock proteins: structural assembly and functional responses against heat stress in plants. In: Ahmad, P., Ahanger, M. A., Singh, V. P., Tripathi, D. K., Alam, P. & Alyemeni, M. N. (eds.) *Plant metabolites and regulation under environmental stress*. Academic Press, pp. 367-376.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. 67 (1), JSSv067i01.

Baumlein, H., Wobus, U., Pustell, J. & Kafatos, F. C. (1986). The legumin gene family: structure of a B type gene of *Vicia faba* and a possible legumin gene specific regulatory element. *Nucleic Acids Research*, **14** (6), 2707-2720.

Boehm, J. D., Nguyen, V., Tashiro, R. M., Anderson, D., Shi, C., Wu, X., Woodrow, L., Yu, K., Cui, Y. & Li, Z. (2017). Genetic mapping and validation of the loci controlling 7S α′ and 11S A-type storage protein subunits in soybean [*Glycine max* (L.) Merr.]. *Theoretical and Applied Genetics*, 1-13.

Bourion, V., Rizvi, S. M. H., Fournier, S., Larambergue, H. d., Galmiche, F., Marget, P., Duc, G. & Burstin, J. (2010). Genetic dissection of nitrogen nutrition in pea through a QTL approach of root, nodule, and shoot variability. *Theoretical and Applied Genetics*, 121, 71-86.

Brünjes, L. & Link, W. (2021). Paternal outcrossing success differs among faba bean genotypes and impacts breeding of synthetic cultivars. *Theoretical and Applied Genetics*, in press.

Cernay, C., Pelzer, E. & Makowski, D. (2016). A global experimental dataset for assessing grain legume production. *Scientific Data*, **3**, 160084.

Cheng, L., Yuan, H.-Y., Ren, R., Zhao, S.-Q., Han, Y.-P., Zhou, Q.-Y., Ke, D.-X., Wang, Y.-X. & Wang, L. (2016). Genome-wide identification, classification, and expression analysis of amino acid transporter gene family in *Glycine max*. *Frontiers in Plant Science*, **7**, 515.

Coombes, N. (2009). DiGGer design search tool in R. http://nswdpibiom.org/austatgen/software.

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PLOS ONE*, **11** (6), e0156744.

DeRocher, A. & Vierling, E. (1995). Cytoplasmic HSP70 homologues of pea: differential expression in vegetative and embryonic organs. *Plant Molecular Biology*, **27** (3), 441-456.

Duc, G., Aleksić, J. M., Marget, P., Mikic, A., Paull, J., Redden, R. J., Sass, O., Stoddard, F. L., Vandenberg, A., Vishnyakova, M. & Torres, A. M. (2015). Faba Bean. In: Ron, A. M. D. (ed.) *Grain Legumes*. New York: Springer Science+Business Media, pp. 141-178.

Earl, D. A. & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4** (2), 359-361.

Egle, K., Beschow, H. & Merbach, W. (2015). Nitrogen allocation in barley: Relationships between amino acid transport and storage protein synthesis during grain filling. *Canadian Journal of Plant Science*, **95** (3), 451-459.

Fuchs, J. & Schubert, I. (1995). Localization of seed protein genes on metaphase chromosomes of *Vicia faba* via fluorescence in situ hybridization. *Chromosome Research*, **3** (2), 94-100.

Gábrišová, D., Klubicová, K., Danchenko, M., Gömöry, D., Berezhna, V. V., Skultety, L., Miernyk, J. A., Rashydov, N. & Hajduch, M. (2016). Do cupins have a function beyond being seed storage proteins? *Frontiers in Plant Science*, **6,** 1215.

Gacek, K., Bartkowiak-Broda, I. & Batley, J. (2018). Genetic and molecular regulation of seed storage proteins (SSPs) to improve protein nutritional value of oilseed rape (*Brassica napus* L.) Seeds. *Frontiers in Plant Science*, **9**, 890.

Gatehouse, J., Croy, R., McIntosh, R., Paul, C. & Boulter, D. (1980). Quantitative and qualitative variation in the storage proteins of material from the EEC joint field bean test. *Quantitative and qualitative variation in the storage proteins of material from the EEC joint field bean test*., 173-188.

Gutierrez, N. & Torres, A. M. (2019). Characterization and diagnostic marker for *TTG1* regulating tannin and anthocyanin biosynthesis in faba bean. *Scientific Reports*, **9** (1), 16174.

Gutierrez, N., Avila, C. & Torres, A. (2020). The bHLH transcription factor VfTT8 underlies *zt2*, the locus determining zero tannin content in faba bean (*Vicia faba* L.). *Scientific Reports*, **10**, 14299.

Heim, U., Wang, Q., Kurz, T., Borisjuk, L., Golombek, S., Neubohn, B., Adler, K., Gahrtz, M., Sauer, N., Weber, H. & Wobus, U. (2001). Expression patterns and subcellular localization of a 52 kDa sucrose-binding protein homologue of Vicia faba (VfSBPL) suggest different functions during development. *Plant Molecular Biology,* **47** (4), 461-474.

Horstmann, C., Schlesier, B., Otto, A., Kostka, S. & Muntz, K. (1993). Polymorphism of legumin subunits from field bean (*Vicia faba* L. var. minor) and its relation to the corresponding multigene family. *Theoretical and Applied Genetics*, **86** (7), 867-874.

Ismail, B. P., Senaratne-Lenagala, L., Stube, A. & Brackenridge, A. (2020). Protein demand: review of plant and animal proteins used in alternative protein product development and production. *Animal Frontiers*, **10** (4), 53-63.

Jombart, T. & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27** (21), 3070-3071.

Jombart, T., Devillard, S. & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11** (1), 94.

Kaler, A. S., Gillman, J. D., Beissinger, T. & Purcell, L. C. (2020). Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. *Frontiers in Plant Science*, **10**,1794.

Kesari, P., Sharma, A., Katiki, M., Kumar, P., R Gurjar, B., Tomar, S., K Sharma, A. & Kumar, P. (2017). Structural, functional and evolutionary aspects of seed globulins. *Protein and Peptide Letters*, **24** (3), 267-277.

Khazaei, H., Stoddard, F. L., Purves, R. W. & Vandenberg, A. (2018). A multi-parent faba bean (*Vicia faba* L.) population for future genomic studies. *Plant Genetic Resources: Characterization and Utilization*, **16** (5), 419-423.

Kimura, A., Fukuda, T., Zhang, M., Motoyama, S., Maruyama, N. & Utsumi, S. (2008). Comparison of physicochemical properties of 7S and 11S globulins from pea, fava bean, cowpea, and french bean with those of soybean—French bean 7S globulin exhibits excellent properties. *Journal of Agricultural and Food Chemistry*, **56** (21), 10273-10279.

Krishnan, H. B., Bennett, J. O., Kim, W.-S., Krishnan, A. H. & Mawhinney, T. P. (2005). Nitrogen lowers the sulfur amino acid content of soybean (*Glycine max* [L.] Merr.) by regulating the accumulation of Bowman−Birk protease inhibitor. *Journal of Agricultural and Food Chemistry,* **53** (16), 6347-6354.

Le Signor, C., Aimé, D., Bordat, A., Belghazi, M., Labas, V., Gouzy, J., Young, N. D., Prosperi, J.-M., Leprince, O., Thompson, R. D., Buitink, J., Burstin, J. & Gallardo, K. (2017).

Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. *New Phytologist*, **214** (4), 1597-1613.

Le Signor, C., Gallardo, K., Prosperi, J. M., Salon, C., Quillien, L., Thompson, R. & Duc, G. (2005). Genetic diversity for seed protein composition in *Medicago truncatula*. *Plant Genetic Resources*, **3** (1), 59-71.

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.

Ma, Y., Kan, G., Zhang, X., Wang, Y., Zhang, W., Du, H. & Yu, D. (2016). Quantitative trait loci (QTL) mapping for glycinin and beta-conglycinin contents in soybean (*Glycine max* L. Merr.). *Journal of Agricultural and Food Chemistry*, **64** (17), 3473-3483.

Martensson, P. (1980). Variation in legumin : vicilin ratio between and within cultivars of *Vicia faba* L. var. minor. The Hague: Martinus Nijhoff. *World crops: production, utilization and description*, volume 3, pp. 159-171.

Mendiburu, F. d. & Yaseen, M. (2020). agricolae: Statistical Procedures for Agricultural Research. 1.4.0 ed.

Mosse, J. (1990). Nitrogen-to-protein conversion factor for ten cereals and six legumes or oilseeds. A reappraisal of its definition and determination. Variation according to species and to seed protein content. *Journal of Agricultural and Food Chemistry*, **38** (1), 18-24.

Nelson, G. M., Prapapanich, V., Carrigan, P. E., Roberts, P. J., Riggs, D. L. & Smith, D. F. (2004). The Heat Shock Protein 70 cochaperone Hip enhances functional maturation of glucocorticoid receptor. *Molecular Endocrinology*, **18** (7), 1620-1630.

O'Sullivan, D. M. & Angra, D. (2016). Advances in faba bean genetics and genomics. *Frontiers in Genetics*, **7**, 150.

Panthee, D. R., Kwanyuen, P., Sams, C. E., West, D. R., Saxton, A. M. & Pantalone, V. R. (2004). Quantitative trait loci for β-conglycinin (7S) and glycinin (11S) fractions of soybean storage protein. *Journal of the American Oil Chemists' Society*, **81** (11), 1005-1012.

Peng, B., Kong, H., Li, Y., Wang, L., Zhong, M., Sun, L., Gao, G., Zhang, Q., Luo, L., Wang, G., Xie, W., Chen, J., Yao, W., Peng, Y., Lei, L., Lian, X., Xiao, J., Xu, C., Li, X. & He, Y. (2014). OsAAP6 functions as an important regulator of grain protein content and nutritional quality in rice. *Nature Communications*, **5**, 4847.

Poeta, F., Ochogavia, A. C., Permingeat, H. R. & Rotundo, J. L. (2017). Storage-associated genes and reserves accumulation in soybean cultivars differing in physiological strategies for attaining high seed protein concentration. *Crop Science*, **57** (1), 427-436.

Poysa, V., Woodrow, L. & Yu, K. (2006). Effect of soy protein subunit composition on tofu quality. *Food Research International*, **39** (3), 309-317.

Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature reviews. Genetics*, **11** (7), 459-463.

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

Qu, P., Shi, J., Chen, T., Chen, K., Shen, C., Wang, J., Zhao, X., Ye, G., Xu, J. & Zhang, L. (2020). Construction and integration of genetic linkage maps from three multi-parent advanced generation inter-cross populations in rice. *Rice*, **13** (1), 13.

Rodríguez-Álvarez, M. X., Boer, M. P., van Eeuwijk, F. A. & Eilers, P. H. C. (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics*, **23**, 52-71.

Rolletschek, H., Hosein, F., Miranda, M., Heim, U., Gotz, K. P., Schlereth, A., Borisjuk, L., Saalbach, I., Wobus, U. & Weber, H. (2005). Ectopic expression of an amino acid

transporter (VfAAP1) in seeds of *Vicia narbonensis* and pea increases storage proteins. *Plant Physiology*, **137** (4), 1236-1249.

Sghaier-Hammami, B., B.M. Hammami, S., Baazaoui, N., Gómez-Díaz, C. & Jorrín-Novo, J. V. (2020). Dissecting the seed maturation and germination processes in the non-orthodox *Quercus ilex* species based on protein signatures as revealed by 2-DE coupled to MALDI-TOF/TOF proteomics strategy. *International Journal of Molecular Sciences*, **21** (14), 4870.

Sul, J. H., Martin, L. S. & Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. *PLoS genetics*, 14 (12), e1007309-e1007309.

Tang, Z., Xu, J., Yin, L., Yin, D., Zhu, M., Yu, M., Li, X., Zhao, S. & Liu, X. (2019). Genome-wide association study reveals candidate genes for growth relevant traits in pigs. *Frontiers in Genetics*, **10**, 302.

Team, R. C. (2020). R base: A language and environment for statistical computing.

Tucci, M., Capparelli, R., Costa, A. & Rao, R. (1991). Molecular heterogeneity and genetics of *Vicia faba* seed storage proteins. *Theoretical and Applied Genetics*, **81** (1), 50-58.

Wang, J. & Zhang, Z. (2018). GAPIT Version 3: An Interactive Analytical Tool for Genomic Association and Prediction. Retrieved from https://github.com/jiabowang/GAPIT3

Warsame, A. O., Michael, N., O'Sullivan, D. M. & Tosi, P. (2020). Identification and quantification of major faba bean seed proteins. *Journal of Agricultural and Food Chemistry,* **68** (32), 8535-8544.

Warsame, A. O., O'Sullivan, D. M. & Tosi, P. (2018). Seed storage proteins of faba bean (*Vicia faba* L): current status and prospects for genetic improvement. *Journal of Agricultural and Food Chemistry*, **66** (48), 12617-12626.

Webb, A., Cottage, A., Wood, T., Khamassi, K., Hobbs, D., Gostkiewicz, K., White, M., Khazaei, H., Ali, M., Street, D., Duc, G., Stoddard, F. L., Maalouf, F., Ogbonnaya, F. C., Link, W., Thomas, J. & O'Sullivan, D. M. (2016). A SNP-based consensus genetic map for synteny-based trait targeting in faba bean (*Vicia faba* L.). *Plant Biotechnology Journal*, **14** (1), 177-185.

Webb, M. A., Cavaletto, J. M., Klanrit, P. & Thompson, G. A. (2001). Orthologs in *Arabidopsis thaliana* of the Hsp70 interacting protein Hip. *Cell stress & chaperones*, **6** (3), 247-255.

Weichert, N., Saalbach, I., Weichert, H., Kohl, S., Erban, A., Kopka, J., Hause, B., Varshney, A., Sreenivasulu, N., Strickert, M., Kumlehn, J., Weschke, W. & Weber, H. (2010). Increasing sucrose uptake capacity of wheat grains stimulates storage protein synthesis. *Plant Physiology*, **152** (2), 698-710.

Willer, C. J., Li, Y. & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26** (17), 2190-2191.

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

Zhang, D., Lü, H., Chu, S., Zhang, H., Zhang, H., Yang, Y., Li, H. & Yu, D. (2017). The genetic architecture of water-soluble protein content and its genetic relationship to total protein content in soybean. *Scientific Reports*, **7** (1).

Zhong, M., Wang, L.-q., Yuan, D.-j., Luo, L.-j., Xu, C.-g. & He, Y.-q. (2011). Identification of QTL affecting protein and amino acid contents in rice. *Rice Science*, *18* (3), 187-195.

4.6 Supplementary

Table S 4.1. Details of seed protein bands used in GWAS analysis. Bands in bold letters are those with significant GWAS hits (FDR≤0.05).

This study			*Vf* seed protein annotation (Warsame et al. 2020)	
Band No.	Protein full names	Protein abbreviated name	Band No.	Protein names
1	**Lipoxygenase 107 kDa**	Lox 107 kDa	1	Lipoxygenase-3
2	Unidentified 99 kDa	U/I 99 kDa		
3	Lipoxygenase 96 kDa	Lox 96 kDa	2	Lipoxygenase-3
4	**Convicilin 89 kDa**	Convc 89 kDa	3	Convicilin
5	**Convicilin 87 kDa***	Convc 87 kDa		
6	Convicilin 83 kDa	Convc 83 kDa	4	Convicilin
7	**Convicilin 79kDa***	Convc 79kDa		
8	**HSP 75 kDa**	HSP 75 kDa	5	Heat shock protein
9	**HSP 73 kDa**	HSP 73 kDa	6	Heat shock protein
10	**Convicilin 65 kDa**	Convc 65 kDa	7	Convicilin
11	Unidentified 61 kDa	U/I 61 kDa		
12	**Unidentified 58 kDa**	U/I 58 kDa		
13	Unidentified 57 kDa	U/I 57 kDa		
14	**Convicilin 54 kDa**	Convc 54 kDa	8	Convicilin
15	**Legumin 52 kDa**	Leg 52 kDa	9	HMW Legumin
16	**Vicilin 50 kDa**	Vc 50 kDa	10	Vicilin
17	**Vicilin 48 kDa**	Vc 48 kDa	10	Vicilin
18	**SBP 46 kDa**	SBP 46 kDa	11	sucrose binding protein
19	**Legumin 43 kDa**	Leg 43 kDa	12	HMW Legumin
20	Unidentified 40 kDa	U/I 40 kDa		
21	**Leg A 39 kDa**	Leg A 39 kDa	13,14	Alpha leg A
22	**Leg B 37 kDa**	Leg B 37 kDa	15,16	Alpha leg B
23	**Unidentified 35 kDa**	U/I 35 kDa		
24	**Unidentified 34 kDa**	U/I 34 kDa		

* these proteins were not identified by Warsame *et al.* 2020, but their identity were inferred from their close genetic control to Convc 89 kDa and Convc 83 kDa (see Figure 4.6).

Table S 4.2. ANOVA *p*-values of the effects of genotype, year and genotype × year on the protein composition. The protein bands are sorted by their abundance and protein class.

Band No. on SDS-PAGE gels	Protein subunits	Genotype	Year	Genotype × Year
22	Leg B 37 kDa	1.3E-10 ***	0.10	1.0E-03 **
21	Leg A 39 kDa	3.8E-05 ***	0.25	0.01 *
19	Leg 43 kDa	1.6E-06 ***	0.48	0.01 *
15	Leg 52 kDa	2.2e-16 ***	3.6E-04 ***	5.7E-03 **
	Leg A/B	2.0E-05 ***	0.39	0.26
	Leg/Vc	1.9E-09 ***	4.3E-05 ***	6.3E-03 **
	Leg/Convc	1.2E-13 ***	0.20	4.4E-03 **
17	Vc 48 kDa	8.9E-10 ***	4.3E-07 ***	0.01 *
16	Vc 50 kDa	2.5E-05 ***	9.9E-03 **	1.3E-03 **
	Vc/Convc	3.2E-08 ***	3.4E-06 ***	0.09
10	Convc 65 kDa	4.9E-13 ***	0.52	0.02 *
14	Convc 54 kDa	2.2E-16 ***	1.5E-09 ***	8.9E-04 ***
5	Convc 87 kDa	2.3E-12 ***	5.5E-04 ***	0.03 *
7	Convc 79kDa	2e-16 ***	0.06	0.03 *
4	Convc 89 kDa	7.1E-03 **	0.39	0.56
6	Convc 83 kDa	0.07	2.0E-03 **	0.04
8	HSP 75 kDa	8.3E-12 ***	1.4E-03 **	3.1E-03 **
9	HSP 73 kDa	3.6E-08 ***	7.0E-03 **	0.03 *
18	SBP 46 kDa	2.2E-14 ***	0.40	3.3E-03 **
3	Lox 96 kDa	6.2E-14 ***	0.19	8.1E-03 **
1	Lox 107 kDa	2.6E-10 ***	0.27	1.7E-04 ***
20	U/I 40 kDa	5.8E-04 ***	4.4E-03 **	2.9E-04 ***
12	U/I 58 kDa	2.9E-05 ***	7.1E-05 ***	0.10
24	U/I 34 kDa	3.1E-03**	6.9E-03 **	0.29
11	U/I 61 kDa	0.01 **	2.6E-06 ***	0.42
13	U/I 57 kDa	0.02 *	0.44	0.05 *
23	U/I 35 kDa	0.03 *	0.12	0.12
2	U/I 99 kDa	1.2E-05 ***	0.23	0.03 *

*, **, *** statistically significant at p≤0.05, 0.01, 0.001, respectively.

Table S 4.3. List of genomic loci that are significantly (FDR≤0.05) associated with the abundance of seed protein subunits detected by FarmCPU and METAL analysis. The proteins are in the order of their abundance and where applicable followed by its relative proportion in relation other major proteins.

Band No.	Protein name	Analysis	SNP	Chr	Position	*P*.value	GVE		PVE	
							2019	2020	2019	2020
22	Leg B 37 kDa	Meta	AX.416816132	1	70.8	9E-07	0.0	0.0	0.0	0.0
			AX.181161898	3	36.5	1E-06	93.0	15.9	15.6	8.4
21	Leg A 39 kDa	Meta	AX.416742431.Medtr5g008840.1	Unmapped	NA	2E-06	25.4	18.6	5.4	8.0
19	Leg 43 kDa	2020	AX.416756967.Medtr2g009320.1	1	1.9	2E-08		23.9		12.9
			AX.181186624.Medtr8g038990.1	4	59.9	3E-07		6.3		4.1
			AX.181182906.Medtr1g069715.1	3	28.9	2E-06		12.5		7.4
		Meta	AX.416756967.Medtr2g009320.1	1	1.9	1E-07	54.2	12.5	4.0	7.4
			AX.416820212.Medtr4g052460.1	6	65.2	1E-06	86.8	7.2	10.2	4.5
15	Leg 52 kDa	2019	AX.416784623.Medtr5g085790.1	1	94.8	4E-09	56.2		17.1	
		2020	AX.181469103.Medtr7g102940.1	Unmapped	NA	6E-09		97.7		79.3
			AX.181190969.Medtr7g100780.1	5	21.2	7E-08		87.5		63.9
Leg A:B		2020	AX.416738681	5	23.6	4E-13		36.8		17.5
			AX.181168504.Medtr3g028170.1	Unmapped	NA	7E-07		26.3		16.3
		Meta	AX.416738681	5	23.6	3E-12	86.9	36.8	13.3	17.5
Leg:Vc		2020	AX.416816615	1	46.1	6E-09		55.0		43.0
			AX.416816132	1	70.8	1E-08		13.2		8.7
		Meta	AX.416816615	1	46.1	1E-08	3.8	55.0	1.0	43.0
			AX.416816132	1	70.8	3E-07	0.0	13.2	0.0	8.7
Leg:Convc		2019	AX.416748066.Medtr2g008740.1	1	1.6	2E-09	17.0		12.8	
			AX.416723137	3	34.9	1E-08	18.0		13.0	
			AX.416741849.Medtr4g020640.1	2	10.4	5E-07	12.5		8.5	
			AX.416727365.Medtr2g102060.1	Unmapped	NA	5E-07	13.0		8.8	
		2020	AX.181194522.Medtr2g008750.1	1	2.4	1E-07		45.5		32.0
			AX.416759320.Medtr7g102560.1	5	36.4	7E-07		8.7		5.5
17	Vc 48 kDa	Meta	AX.416722749.Medtr5g021760.1	1	46.1	2E-06	2.9	29.4	1.0	17.5
16	Vc 50 kDa	2020	AX.181149609	1	2.3	4E-06		53.9		14.0
Vc:Convc		2020	AX.416728459.Medtr2g023850.1	Unmapped	NA	6E-10		58.9		48.4
		2020	AX.181152370.Medtr8g105780.1	6	17.2	4E-07		2.6		1.7
		Meta	AX.416728459.Medtr2g023850.1	Unmapped	NA	7E-07	0.0	58.9	0.0	48.4
			AX.181194522.Medtr2g008750.1	1	2.4	7E-07	100.0	45.5	20.1	32.0
			AX.181152370.Medtr8g105780.1	6	17.2	1E-06	0.0	2.6	0.0	1.7
10	Convc 65 kDa	2019	AX.181194534	1	1.6	8E-14	24.8		19.9	
			AX.416782744.Medtr3g069960.1	2	57.8	5E-13	13.8		14.8	
			AX.416789546.Medtr3g435170.1	2	79.0	1E-10	3.4		3.5	
			AX.416724445.Medtr5g030430.1	1	56.2	3E-10	20.3		23.9	
			AX.416770555.Medtr1g107490.1	3	58.8	1E-09	11.1		12.8	
			AX.181496690.Medtr5g090660.1	Unmapped	NA	8E-09	0.6		0.6	
			AX.181489427.Medtr4g114210.1	4	6.9	2E-07	6.8		7.6	

Table S 4.3 continued

		2020	AX.416748066.Medtr2g008740.1	1	1.6	2E-14		57.0		26.0
			AX.181194534	1	1.6	1E-09	24.8	57.8	19.9	26.0
			AX.416748066.Medtr2g008740.1	1	1.6	2E-08	24.7	57.3	19.9	25.8
		Meta	AX.416782744.Medtr3g069960.1	2	57.8	9E-08	0.6	7.3	0.6	5.2
			AX.416789546.Medtr3g435170.1	2	79.0	2E-07	3.4	0.0	3.5	0.0
			AX.416770555.Medtr1g107490.1	3	58.8	5E-07	11.1	0.0	12.8	0.0
14	Convc 54 kDa	2020	AX.416756967.Medtr2g009320.1	1	1.9	8E-09		36.1		19.7
			AX.416731667	2	10.8	9E-08		2.4		1.7
			AX.416740251.Medtr6g057750.1	7	69.2	1E-07		5.1		3.7
			AX.416796750.Medtr7g105830.1	5	32.9	2E-07		25.6		18.6
			AX.181457859.Medtr8g107510.1	6	16.3	4E-06		10.1		7.0
		Meta	AX.416756967.Medtr2g009320.1	1	1.9	3E-10	27.7	36.1	17.3	19.7
			AX.181482304.Medtr1g073790.1	3	31.7	3E-09	28.6	15.3	19.3	10.5
			AX.416796750.Medtr7g105830.1	5	32.9	3E-06	36.5	25.6	25.6	18.6
5	Convc 87 kDa	2019	AX.416820519.Medtr5g019790.1	1	46.6	1E-08	93.3		46.6	
		2020	AX.416818833.Medtr5g019920.1	1	44.9	2E-09		100.0		74.2
			AX.181460958.Medtr8g469310.1	4	73.0	3E-06		0.0		0.0
		Meta	AX.416734467.Medtr1g080910.1	Unmapped	NA	8E-07	11.2	4.6	5.0	2.7
			AX.416818833.Medtr5g019920.1	1	44.9	2E-06	94.4	100.0	52.5	74.2
7	Convc 79kDa	2019	AX.416812445.Medtr5g019810.1	1	48.2	2E-18	100.0		0.5	
			AX.416727301	3	59.0	2E-06	0.0		0.0	
		2020	AX.416812445.Medtr5g019810.1	1	48.2	7E-15		100.0		64.2
			AX.416812445.Medtr5g019810.1	1	48.2	2E-31	100.0	100.0	45.5	64.2
		Meta	AX.416727301	3	59.0	2E-07	0.0	0.0	0.0	0.0
			AX.416728780.Medtr5g017630.1	Unmapped	NA	2E-06	68.5	62.1	53.9	56.2
4	Convc 89 kDa	2019	AX.416791528.Medtr8g081510.1	6	3.3	8E-08	5.1		2.8	
		2020	AX.416795623.Medtr5g019780.1	1	46.6	5E-09		85.0		45.0
			AX.416759185.Medtr8g086380.1	6	32.7	1E-06		7.5		4.2
8	HSP 75 kDa	2019	AX.181460602	1	158.6	4E-06	59.2		25.8	
		2020	AX.181460602	1	158.6	4E-06		67.7		39.7
9	HSP 73 kDa		AX.181460602	1	158.6	8E-23	100.0		89.4	
		2019	AX.416821806.Medtr7g116660.1	5	7.5	1E-07	0.0	0.0		
			AX.181496843.Medtr8g020390.1	4	42.1	4E-07	0.0	0.0		
		2020	AX.181460602	1	158.6	1E-11		67.7		40.0
			AX.181446967.Medtr1g073930.1	3	31.8	2E-06		0.0		0.0
		Meta	AX.181460602	1	158.6	2E-31	100.0	67.7	89.4	40.0
18	SBP 46 kDa	2019	AX.181474569	3	26.0	1E-07	55.3		24.1	
			AX.416761203.Medtr1g069855.1	Unmapped	NA	1E-06	100.0		53.7	
			AX.181151011.Medtr5g071250.1	1	103.3	2E-06	14.9		7.7	
		2020	AX.416787302.Medtr1g073700.1	3	31.7	6E-14		69.7		36.1
		Meta	AX.416787302.Medtr1g073700.1	3	31.7	4E-09	68.9	69.7	22.1	36.1
1	Lox 107 kDa	Meta	AX.181165449.Medtr4g114900.1	Unmapped	NA	2E-06	24.1	14.4	4.5	6.3

Table S 4.3 continued

12	U/I 58 kDa	2020	AX.181194534	1	1.6	3E-07		53.0		19.2
			AX.416816157	3	72.1	3E-06		9.6		5.2
24	U/I 34 kDa	2019	AX.181450440.Medtr7g093390.1	5	43.2	3E-06		10.6		1.9
			AX.181183662.Medtr7g092460.1	5	43.6	4E-06		9.0		1.6
23	U/I 35 kDa	2019	AX.416776516.Medtr7g113730.1	5	4.9	2E-06	0.0	0.0	0.0	0.0

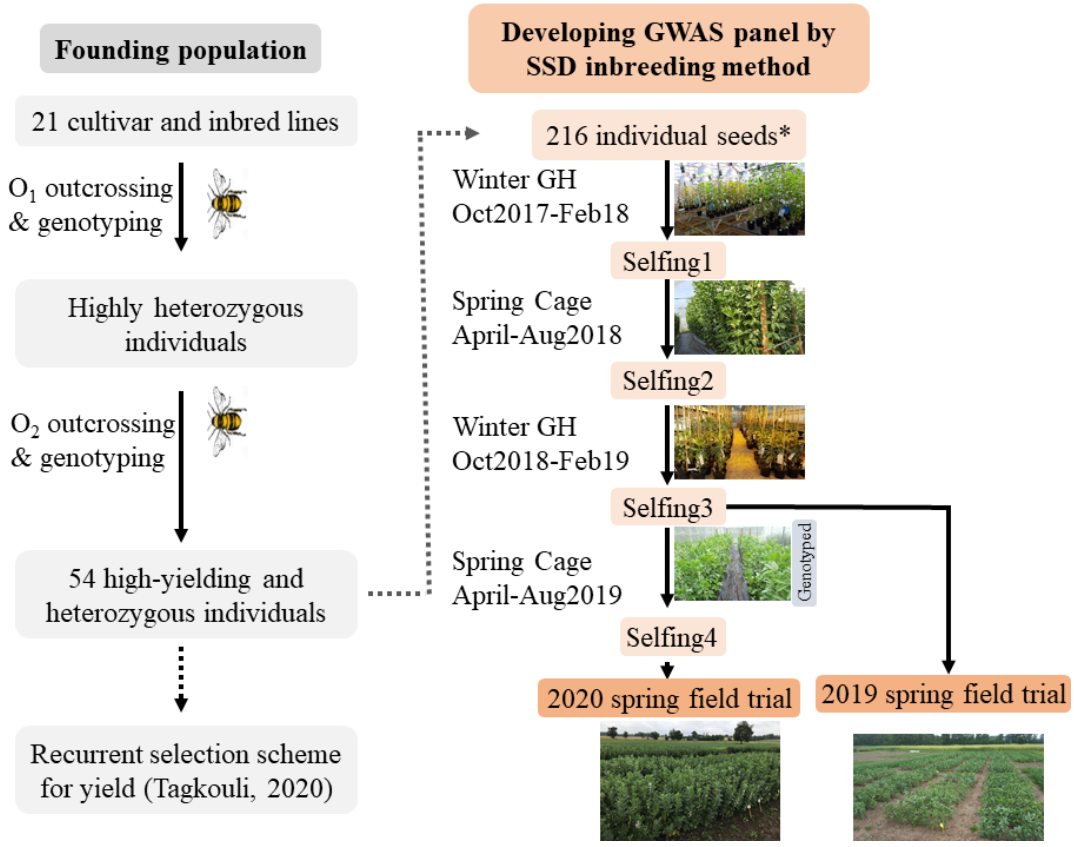

Figure S 4.1. Schematic summary of the development of the Reading Spring Bean (RSB) population. *Each four individual seeds were randomly selected from one plant which makes the population constituted of 54 families.

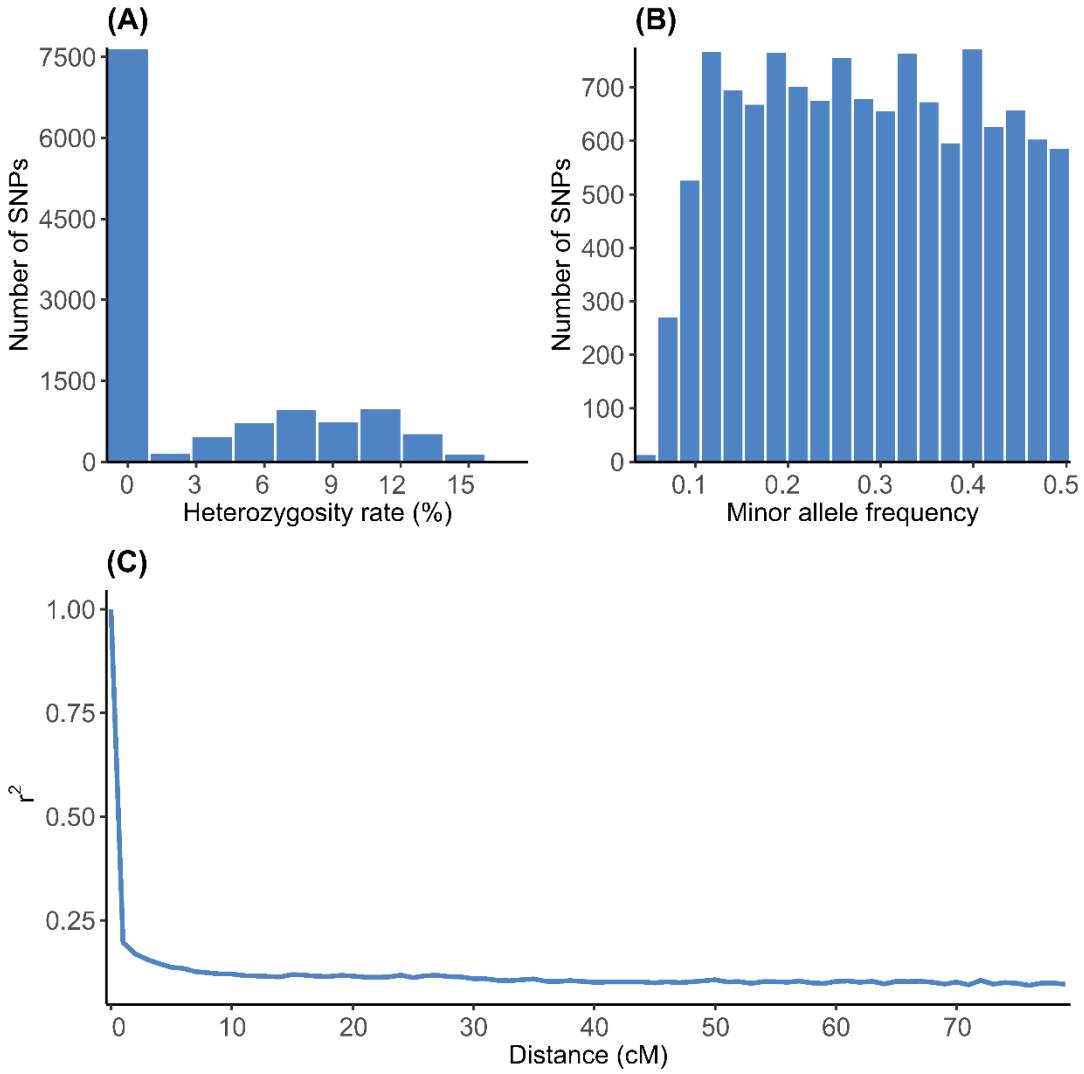

Figure S 4.2. Population genetic characteristics of the S_3 GWAS population genotyped in the study. Distribution of heterozygosity rate (A), minor allele frequency among SNPs (B) and mean genome-wide linkage disequilibrium between markers (C).

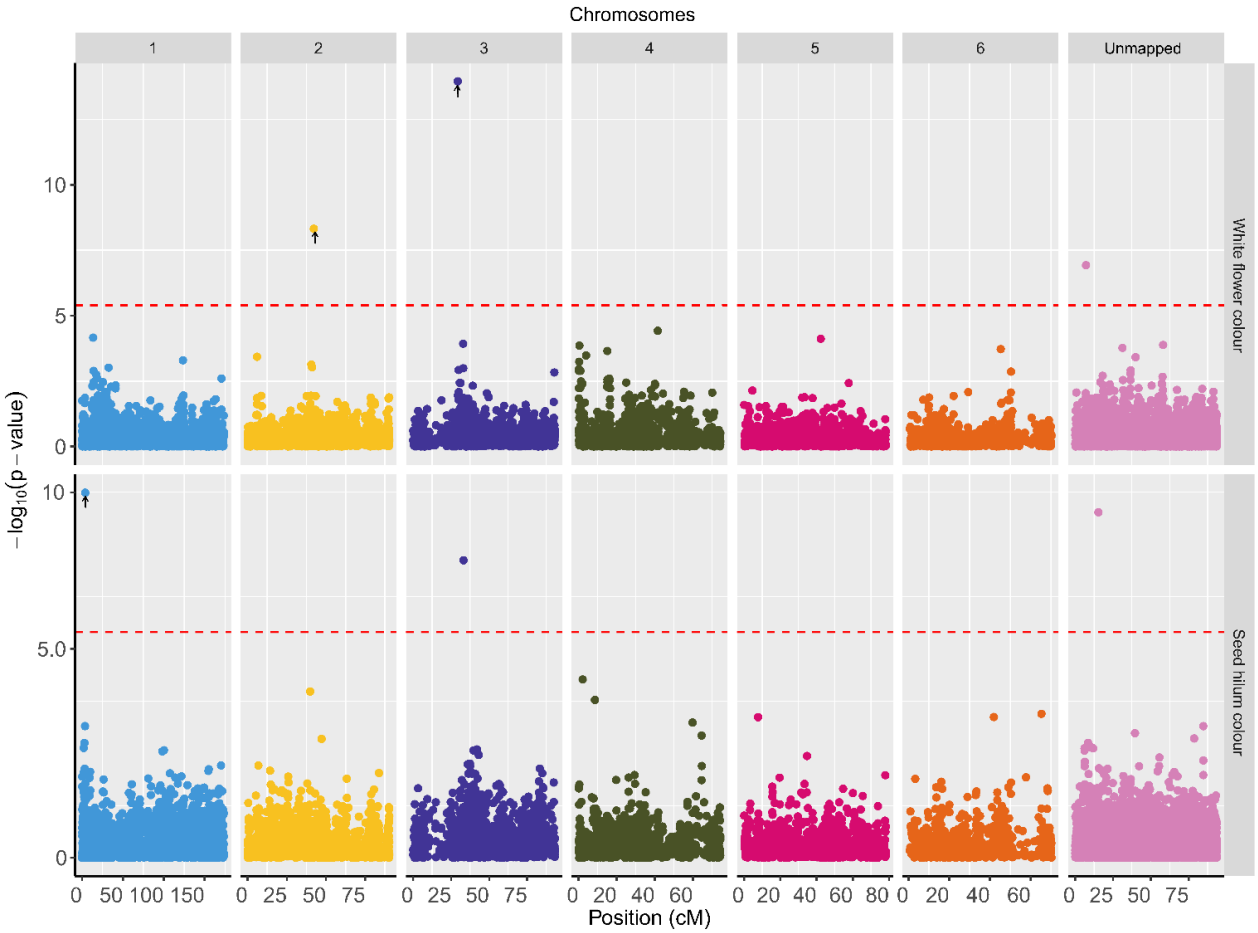

Figure S 4.3. Manhattan plot showing significant GWAS associations for white flower colour and hilum seed colour. SNPs flanking already known loci are indicated by arrows. For flower colour, the significant SNP (AX.181489312.Medtr3g092090.1) on Chr 2 is close to ZT-1 gene while the other AX.416742604.Medtr1g070380.1 near the ZT-2 gene on Chr 3. The significant hit for hilum colour on Chr 1 is AX.416810010.Medtr2g013580.1 which is near the region thought to contain the gene responsible for pale seed hilum colour in *Vf*.

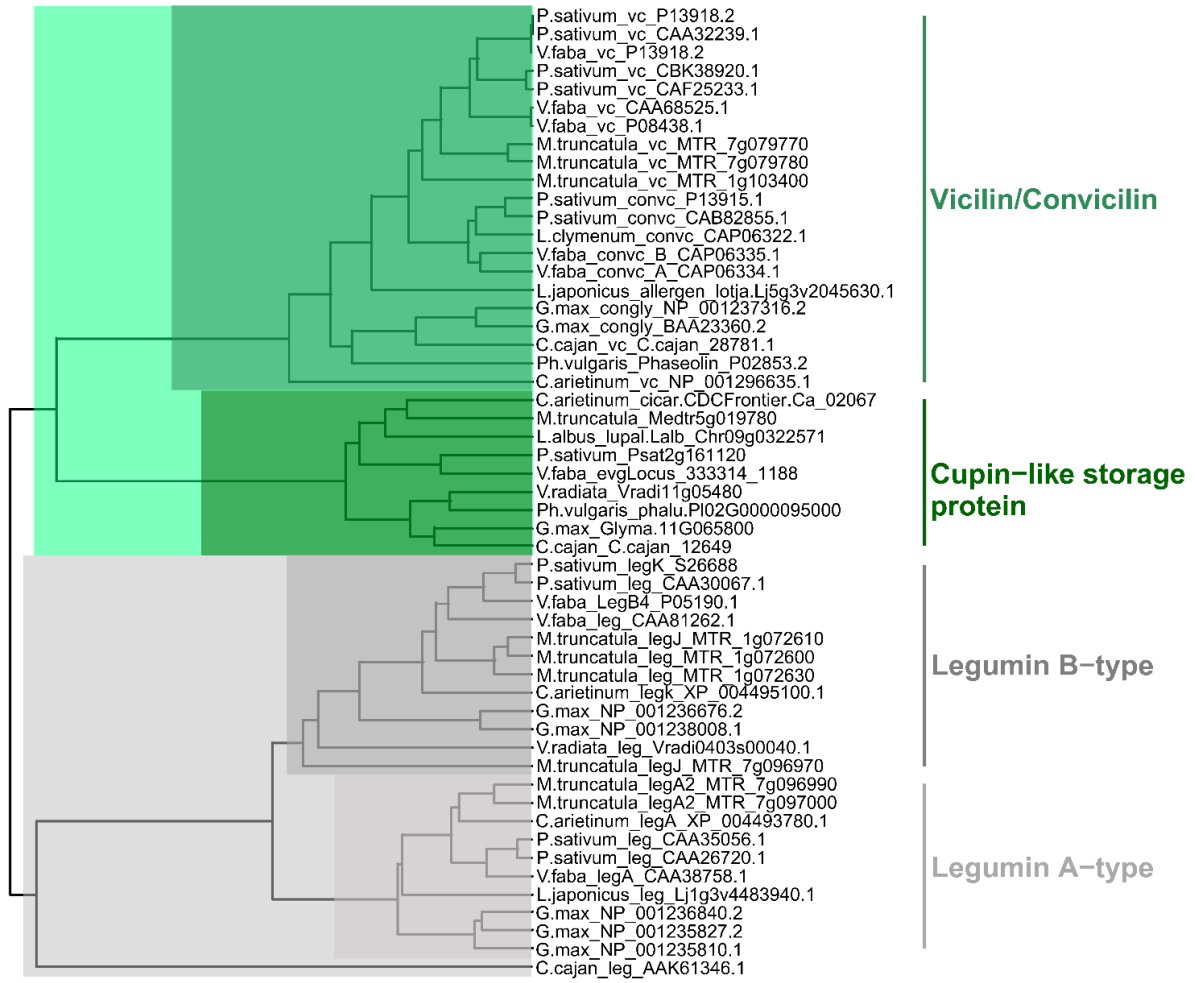

Figure S 4.4. Phylogenetic tree of seed storage protein genes among major legumes. Protein sequences for vicilin, convicilin and legumin genes were obtained from NCBI protein database while sequences of the "cupin-like" genes were retrieved from Legume Information System database using *M. truncatula* gene (Medtr5g019780.1) as reference. The tree was constructed using MEGA X 10 with UPGMA method and 5000 replications.

Chapter 5 Proteomic characterization of developing seeds of faba bean (*Vicia faba*, L.)

5.1 Abstract

Accumulation of proteins and other nutritional components of the mature seeds are determined by complex biological processes which are modulated by environmental factors during seed development. To better understand seed development and the dynamics of seed protein accumulation, we have investigated them in seeds of the inbred line Hedin/2 across 12 growth stages, from 20 days after pollination (DAP) to full maturity. To investigate the proteomic profile of seeds during development, trypsin digested total protein extracts from the 12 stages were analysed by micro-flow LC–MS/MS which, in total, identified 1217 proteins. The functional clusters of these proteins showed that, in early growth stages, proteins related to cell growth and division, and metabolism were most abundant, while seed storage proteins accumulated heavily from 50 DAP. Moreover, the relative abundance of 344 proteins, including 9 legumins, 3 vicilins and 4 convicilins, was quantified using a label-free quantification approach. This revealed several distinct temporal accumulation trends amongst the storage protein classes, which suggested that these proteins are regulated differently. These results lay the foundations for improved understanding of *Vf* seed protein synthesis and accumulation in relation to environmental stresses occurring during grain fill and the possible implication on harvesting time and nutritional quality.

Key words: faba bean; seed development; protein accumulation; legumin; vicilin

5.2 Introduction

The major reservoirs of nutritional compounds contained in seeds, whose biological role is to support seed germination and seedling establishment, also play a major role in human and animal nutrition. Legumes are most valued for their high protein content, which, in *Vicia faba* (*Vf*), predominantly consist of globulins, namely legumin, vicilin, and convicilin (Warsame *et al.*, 2018; Warsame *et al.*, 2020; Müntz *et al.*, 1999). Since the final amounts and relative abundance of specific proteins in the seeds is the result of a complex sequence of developmental processes and regulated by many genes which are expressed at different times during seed development, a full understanding of seed protein composition and quality must take into account the developmental dynamics of the seed proteome.

In *Vf* and other legume species, divergent temporal patterns of accumulation of some seed storage proteins have been reported. According to De Pace *et al.* (1991), accumulation of vicilin protein in *Vf* seeds precedes that of legumin by 4 days while legumin A-type was observed before legumin B-type. Differences in the timing of expression of major storage proteins during seed growth has been also reported in Medicago (Gallardo *et al.*, 2003; Verdier *et al.*, 2008) and pea (Kreplak *et al.*, 2019). From a nutritional standpoint, such temporal differences in protein composition during seed development will not only affect the nutritional profile of beans consumed as immature pods or seeds, but also means that changes in the environmental conditions at certain seed filling stages would have a modulating impact on seed protein composition and quality. For instance, expression of two specific legumin genes was preferentially downregulated in developing seeds of peas grown under sulphur deficient and/or water stress conditions (Henriet *et al.*, 2019). In soybean, abundance of several seed storage proteins was reported to vary considerably depending on whether genotypes were grown in the field or in glasshouse (John *et al.*, 2017). Therefore, deeper understanding of accumulation

patterns of different storage proteins during seed development may help predict protein composition based on the growth stage and the prevailing growing conditions.

In the past, De Pace *et al.* (1991) used one dimensional sodium dodecyl sulphate−polyacrylamide gel electrophoresis (1D SDS-PAGE) to investigate the accumulation of vicilin and legumin of *Vf* at seed developmental stages between 20-36 days after flower desiccation. Panitz *et al.* (1995), on the other hand, studied vicilin and legumin accumulation during early embryogenesis stages through an immunohistological technique. Although these studies provided an early indication of the dynamic nature of storage protein synthesis and accumulation, they do not capture the full picture as they were limited by the number of proteins studied and the sensitivity of the techniques. Now, it is known that each major seed storage protein class is coded by multiple gene families, which could each be differentially accumulated during seed development. Mass spectrometry-based proteomics have the potential to identify and quantify hundreds or thousands of proteins, with the ability to separately quantify even closely related members of protein families. Therefore, this study was aimed to characterize *Vf* seed development and the associated temporal trends in protein composition with particular emphasis on storage proteins belonging to legumin, vicilin and convicilin.

5.3 Materials and methods

5.3.1 Plant material and growth conditions

Hedin/2, which is a small-seeded, high purity inbred line that has been used as parent in multiple study populations and its genome being sequenced, was used in this study. Thirty single plants were grown in the glasshouse using 3-litre pots containing homogeneously mixed compost (John Innes No. 2, Clover Peat, UK) at the Crop and Environment Laboratory (CEL) of University of Reading, UK. Plants were well-watered and received supplementary lighting of about 600 μmol m^{-2} s^{-1} PPFD to achieve 16 hrs of light per day using high pressure sodium lamps. At flowering stage, individual flowers were tagged, and the presumed date of pollination

was recorded as the day when the standard petal curved back, and wing spots were visible from the front without undue effort. During tagging, flowers were also tripped by hand to encourage the pollination process. Then, pods were sampled at 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70 days after pollination (DAP). The mature seed sample was obtained from pods that naturally dried on the plants. Pods were carefully removed from plants, the number of the sampled node was recorded, and then immediately frozen in liquid nitrogen and stored at -80 °C until further analysis.

Table 5.1. Details of developing *Vf* seed samples for proteomic analysis

Developmental stage (DAP)	Number of seeds combined for each developmental stage	Obtained ground dry sample (g)
20*	148	0.111
25	108	0.185
30	106	0.486
35	63	0.358
40	26	0.509
45	9	0.392
50	12	0.719
55	10	1.523
60	10	1.927
65	10	2.619
70	10	2.357
Mature	10	2.361

* This was excluded from crude protein content analysis as there was insufficient sample.

5.3.2 Pod and seed measurements

At the end of the sampling process, pods belonging to each developmental stage were combined after visually examining for any apparent off-types such as pods with outlier sizes for the certain growth stage. Before removing seeds, pods were photographed along a 15 cm long ruler with 1 mm scale. The same was done for seeds before and after freeze-drying. Images where then processed with ImageJ software (Schneider *et al.*, 2012) where pod and seed length,

and seed size (based on the area of seeds laying on one of the cotyledons) were measured. Sample details are summarized in **Table 5.1**.

5.3.3 Crude protein content analysis

Seeds were freeze-dried until constant weight was obtained (~78 hours) and ground to homogeneous powder using mortar and pestle. Then, depending on the sample availability, ~60-100 mg of seed flour was analysed in a LECO Carbon/Hydrogen/Nitrogen Determinator (628 Series, LECO, USA). Subsequently, nitrogen content was converted to protein content using 5.4 conversion factor (Mosse, 1990).

5.3.4 Total protein extraction

Protein extraction was conducted as described by (Scollo *et al.*, 2018). Briefly, to remove phenolic compounds, a cold (~4° C) aqueous acetone (80%, v/v) containing 5 mM sodium ascorbate was added to the samples at 1:20 sample to buffer ratio. Then, the suspension was vortexed for 1 minute and then centrifuged at 4000 rpm for 10 minutes at 4º C. This step was done twice, and the supernatant was discarded each time. The sample was then washed with cold acetone and then air-dried. From this product, total protein was extracted with a solution containing 7 M urea, 2 M thiourea and 20 mM dithiothreitol (DTT). After stirring the suspension at 300 rpm for 1 hour at room temperature, samples were centrifuged at 4000 rpm for 15 minutes at 4° C and the supernatant was collected. These samples were checked on SDS-PAGE gels and protein concentration was assessed with the Bradford method (Bradford, 1976) using a SpectraMax i3x microplate reader (Molecular Devices, UK).

5.3.5 Trypsin digestion

Before the digestion, technical triplicates of each sample were created with each tube containing approximately 0.5 mg protein. Then, to lower urea concentration to below 1 M, an appropriate volume of 50 mM ammonium bicarbonate was added to each tube. Also, to ensure all proteins were completely reduced, further DTT solution was added to each tube to obtain a

final concentration of 10 mM and then samples were incubated for 30 minutes at 37°C. This was followed by alkylation of sulfhydryl groups with iodoacetamide (IAA) at a final concentration of 20 mM. For protein digestion, trypsin (Promega, UK) was added at 1:100 protease : protein ratio and the solutions were incubated overnight at 37° C. The reaction was terminated by freezing samples at -20°C. To confirm that all proteins in the samples were completely digested, aliquots of the digested samples were loaded on 1D SDS-PAGE gel. Finally, samples were dried in a centrifugal vacuum concentrator. The mass spectrometry analysis was conducted as described in chapter 3 (Warsame *et al.*, 2020) with each of the three technical replicates for each sample being analysed three times. In this way, nine MS/MS data files were obtained for each developmental stage with the exception of 60 DAP that was analysed in duplicate and had six data files.

5.3.6 MS data analysis

The raw MS/MS data was searched in the NCBI database using an in-house version of MASCOT search engine (Matrix Science, UK) via Mascot Daemon with file conversion performed using ProteoWizard. The MASCOT search parameters are described in chapter 3 (Warsame *et al.*, 2020). Then, for each developmental stage, proteins identified in the replicates were merged and filtered to obtain a final list of protein accessions for that stage. For functional clustering, protein sequences were functionally annotated using the MapMan4 framework and its associated online tool Mercator4 (Schwacke *et al.*, 2019). Considering the relatively large number of the resultant protein functional clusters, proteins were further grouped as described by Bevan *et al.* (1998).

To assess the relative abundance of proteins across seed developmental stages, a label-free quantification was conducted using MaxLFQ algorithm (Cox *et al.*, 2014) implemented in MaxQuant software (Cox and Mann, 2008). For this analysis, the list of proteins identified by MASCOT analysis was used as reference database. The details of the quantification parameters

are in **Table S 5.1**. The downstream analysis including normalization, imputation, ANOVA test for differentially abundant proteins and their clustering were conducted in Perseus software (Tyanova *et al.*, 2016).

5.4 Results and discussion

5.4.1 Faba bean seed development

In this study, we used the small-seeded reference inbred line Hedin/2, which carried flowers from nodes 8 to 27 on the main (and sole) stem with seven flowers per node on average. It is important to note that *Vf* genotypes can vary greatly in number of flowers per node, number of pods and seeds per pod (Duc, 1997; Suso *et al.*, 1996), and therefore, aspects of the described developmental stages should be regarded as genotype- and environment-dependent. It should also be noted that due to the extremely small size (<2 mm) of the seeds, it was not feasible to collect sufficient weight of seed before 20 DAP, and thus the very early stages of post-fertilization development are not investigated in this study.

Figure 5.1A& B show morphometric changes in the pod and seeds between 20 and 70 DAP and give an important context to parallel changes in biochemical composition. The early phase of growth is marked by rapid expansion of pods, which nearly doubled in length between 20 and 35 DAP and followed by more gradual growth to reach maximum length at 50 DAP (**Figure 5.1**A). In contrast, seeds increased in length gradually and in an almost linear fashion until 65 DAP (**Figure 5.1**A). This early rapid pod development before the onset of seed filling has been also observed in soybean (Li *et al.*, 2015) and serves to give ample space for the seed to develop. On a dry seed basis, seed growth and development was characterized by a gradual increase in weight and area until 50 DAP and a rapid increase in both parameters between 50 and 60 DAP, after which seed area plateaued at 60 DAP while dry weight continued to increase substantially between 60 and 70 DAP (**Figure 5.1**B).

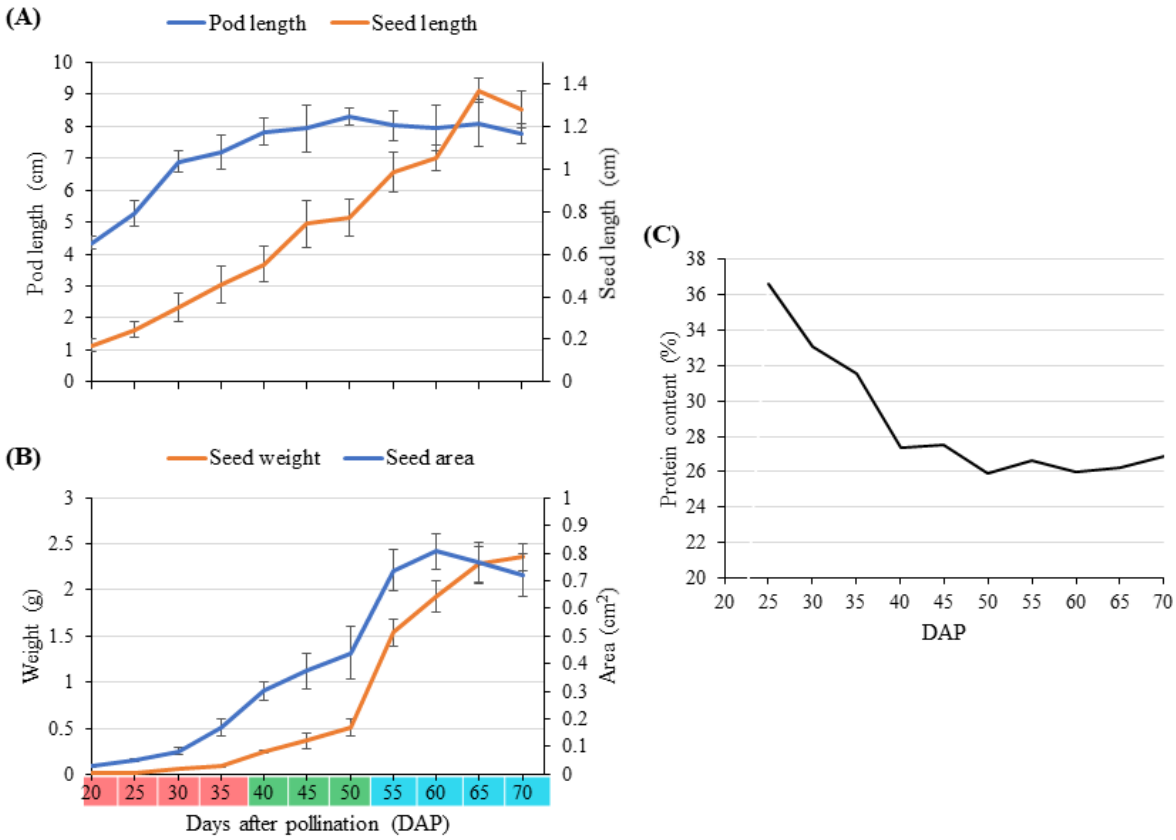

Figure 5.1. Characteristics of developing *Vf* seed. (A) Pod and seed fresh lengths between 20-70 days after pollination (DAP) as measured on 5-10 pods and 10 seeds. (B) Weight and area of freeze-dried seeds. (C) Protein content on dry weight basis for 11 growth stages. The error bars represent mean ± SD. The colour codes denote the main grain filling (GF) stages: early GF (light red), mid GF (light green), late GF (light blue)

On the other hand, percent crude protein content on a dry weight basis was highest during the earliest stage at which it was possible to measure (25 DAP) but dropped rapidly from >36% to <28% by 40 DAP remaining in the 26-28% range throughout the period where most seed weight was gained (40-70 DAP - **Figure 5.1**C). Bulk mature seed protein content at ~24% was comparable to the average protein content recorded in Hedin/2 plants grown in the field (data not shown). A similar protein content pattern was reported in developing seeds of mung bean (Sital *et al.*, 2011). However, Li *et al.* (2015) found no difference in protein content across growth stages of soybean seeds. The higher protein content (i.e. nitrogen content) we observed in the early developmental stages is probably due to the higher concentration of structural proteins, enzymes, free amino acids and other nitrogen-containing compounds associated with rapid cell

division during these early stages. Also, at this stage, other seed components including starch, which normally dilute percent protein content in final stages, may not yet have accumulated in the seeds.

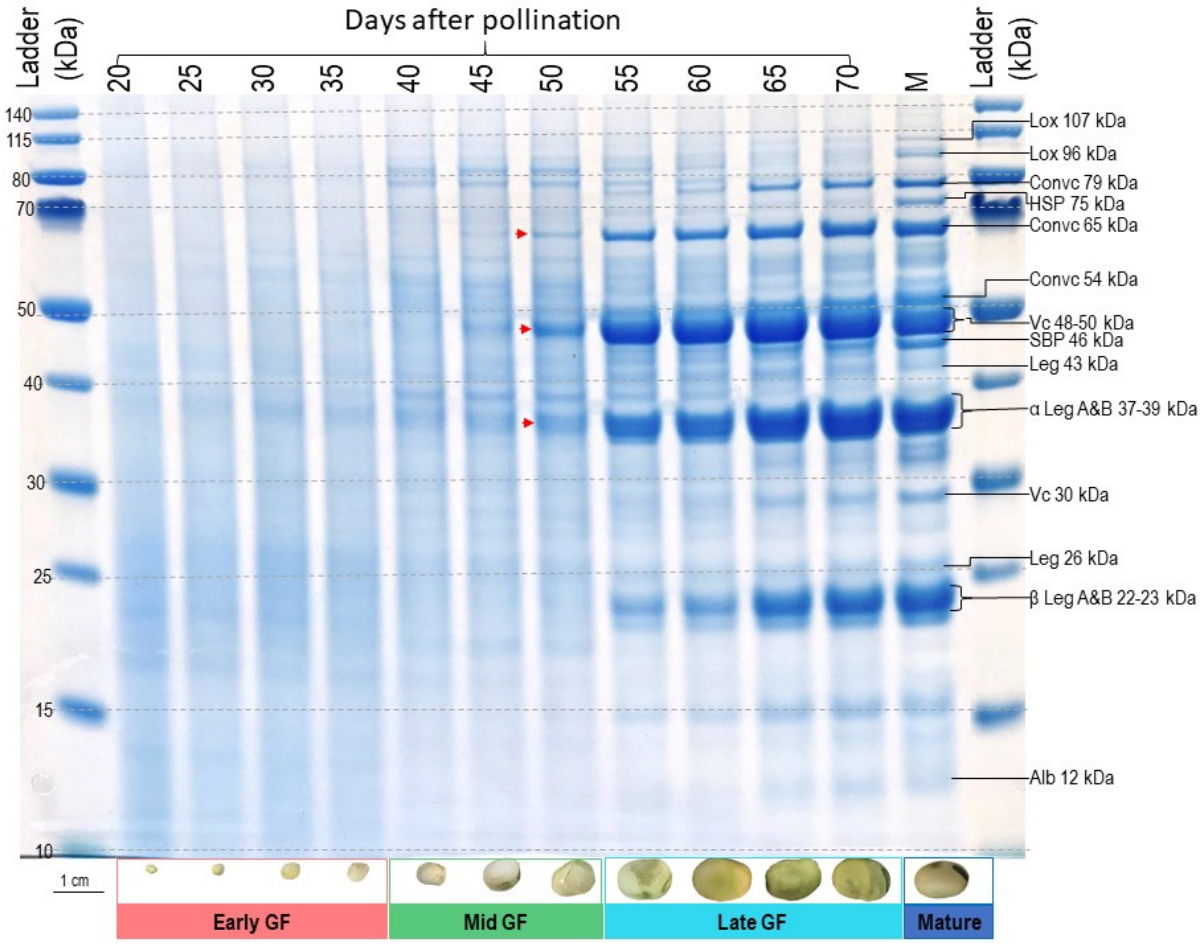

Figure 5.2. 1D SDS-PAGE gel showing protein profile of *Vf* seeds harvested at 12 developmental stages. Molecular weights of individual bands are estimated with respect to the bands of the MW ladder in the leftmost lane, with sizes of the marker bands given in kDa. Abbreviated names of discrete and most abundant protein bands are based on mass spectroscopic identification of proteins in Warsame *et al*. (2020) and are listed in Supplementary **Table S 5.3**. At the bottom of each lane, sample images of freeze-dried seeds belonging to the pool representing each growth stage are shown; coloured bars along the bottom of the gel denote the main phases of grain filling (GF); a 1 cm scale bar for seed images is given on the bottom left.

Using 1D SDS-PAGE gels, we next looked at changes in the abundance of specific protein classes in protein extracts from the 11 growth stages plus the mature seeds from bulk seed of pods whose flowering dates were not recorded. There were considerable differences in the protein composition profiles of growth stages (**Figure 5.2**), where the 20 to 35 DAP period was

characterized by greater apparent abundance of lower molecular weight proteins (<40 kDa) compared with later stages and a relative lack of highly abundant species that stand out as discrete bands against the background. By 45 DAP, some high molecular weight proteins have begun to stand out against the background with subunits of some seed storage proteins including convicilin 65 kDa, vicilin 48-50 kDa and alpha-legumin A&B 37-39 kDa [marked with red arrows in **Figure 5.2**] are clearly visible by 50 DAP. Finally, from 55 DAP seeds enter a phase of heavy protein accumulation throughout which the protein profile is quite comparable to that of the mature seed (**Figure 5.2**).

5.4.2 The faba bean seed proteome

A total of 1217 non-redundant proteins were identified by mass spectrometry analysis in the seed samples from 20 DAP to mature seed**.** This list included 36 of 104 protein accessions we previously identified in major protein bands of mature *Vf* seeds of three different accessions (Warsame *et al.*, 2020). This relatively lower overlap between the two experiments can be attributed to differences in the sample characteristics where LC-MS analysis of individual protein bands from 1D SDS-PAGE gels is more sensitive to detect low abundant proteins compared to whole protein extract in which storage proteins can mask the less abundant types. Traditionally, for total protein extracts from such high complexity samples, a fractionation or depletion of abundant proteins is performed before MS analysis, in order to capture the less abundant proteins. However, since the aim of this study was to compare between developmental stages and there was a clear gradient in the protein profiles, as shown in **Figure 5.2**, no complexity reduction was performed. In other studies on seed proteomics where nano-flow LC–MS/MS was used, 1,168 proteins were identified in mature barley seeds (Mahalingam, 2017) and 704 in total protein extract of cocoa beans (Scollo *et al.*, 2018a). Furthermore, the total number of proteins identified in this study was far less than the 4,172 identified in rice seeds across three developmental stages using a nano-LC-MS/MS system (Lee and Koh, 2011).

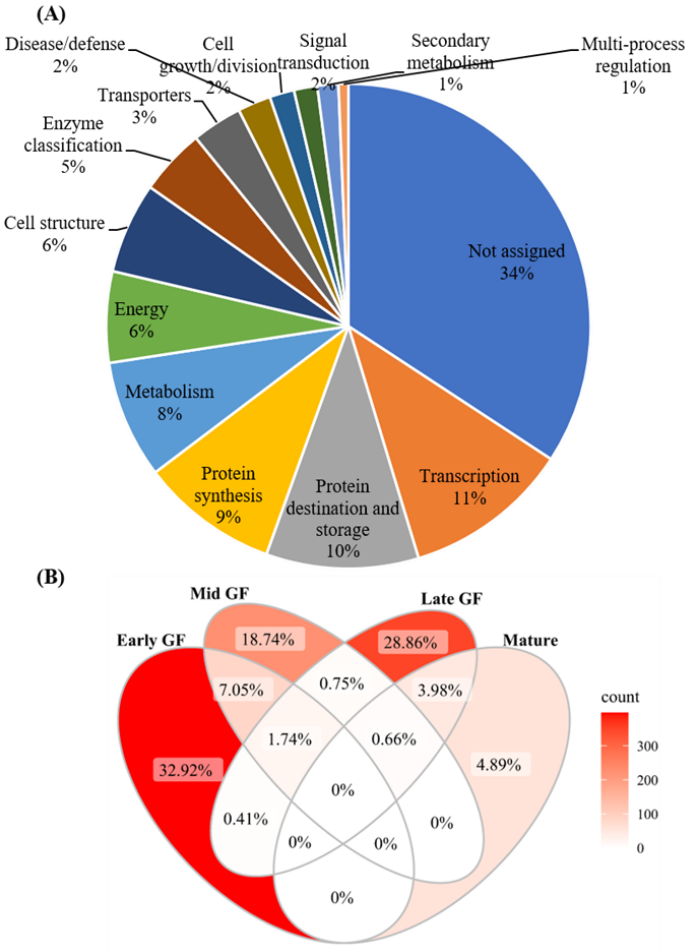

Figure 5.3. (A) Functional categories of the total list of 1217 proteins identified in *Vf* seeds anywhere from 20 DAP to mature stage. (B) Venn diagram showing percentage of the total number of identified proteins that is specific or common among four developmental stages: Early grain fill (GF) (20-35 DAP), Mid GF (40-50 DAP), Late GF (55-70 DAP) and Mature seed. The relative importance of each sector is indicated by a white to red heat scale.

Identified proteins could be assigned to 14 functional categories (**Figure 5.3**A). Although the MapMan4 framework used for functional categorisation was specifically designed for plant protein classification (Schwacke *et al.*, 2019), 34% of the protein sequences could not be assigned to any protein category. However, considering the proteins that were functionally categorised, the three largest classes, totalling 30% of all proteins identified, were proteins related to transcription, protein destination and storage, and protein synthesis at 11%, 10% and 9%, respectively. This is consistent with the nature of the developing seed, in particular the cotyledon, as a storage organ undertaking active protein synthesis and deposition.

Looking at the developmental stage specificity of the identified proteins, what was striking was the low overall level of overlap, in that a large majority of proteins identified at each stage

were found only in that stage with 33%, 19%, 29% and 5% of the total proteins were specific to early, mid, late and mature seed stages, respectively (**Figure 5.3**B). The far greater proportion of the proteins identified in the early stages (c. 33%) in the total list of proteins compared to the mature grain (4.9%) is consistent with the SDS-PAGE patterns in **Figure 5.2** that showed a transition from a protein smear caused by a multitude of proteins of every possible size at 20 DAP and a clear banding pattern in the mature seed featuring a small number of very abundant protein species. The maximum number of proteins in common between different stages was ~8.8%, shared between early and mid-stages. Growth stages also differed in the dominant protein functional classes. For instance, proteins related to cell growth and division, and to metabolism were most abundant in early and mid grain filling stages, constituting 8% and 10% of the proteins (**Figure 5.4**).

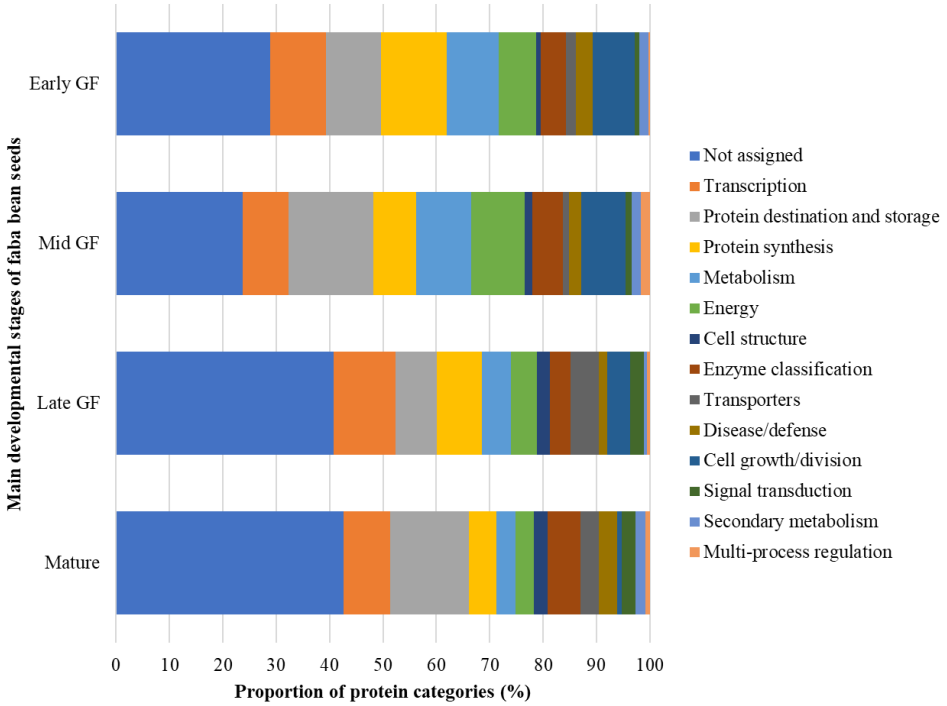

Figure 5.4. Changes through development of the relative importance of functional categories of proteins identified in *Vf* seeds at four major developmental stages.

These earlier stages also had higher percentages of proteins involved in protein synthesis and energy production pathways (**Figure 5.4**), which agrees with expression patterns reported in

soybean seed proteins (Li *et al.*, 2015) and protein gene cDNAs (Jones *et al.*, 2010). It is worth noting though that some of the differences in the representation of some protein clusters at certain growth stages could be an artefact of sample properties, particularly in the later developmental stages, where storage proteins are overwhelmingly more abundant than other protein classes.

5.4.3 Protein abundance profiles during seed development

As previously shown in **Figure 5.2**, individual proteins could be seen to vary progressively in abundance across seed growth and development stages. Thus, to quantify these trends, label-free quantification was performed, resulting in calculating the relative abundances of 344 proteins across the developmental timecourse. Generally, accurate quantification of hundreds or thousands of proteins in a sample is one of the major challenges in proteomic studies with each method having its limitations and advantages in terms of cost, sample preparation requirements and sensitivity (Brewis and Brennan, 2010). Here, considering the relative simplicity of the approach taken, the number of proteins identified and quantified was considered promising. Additionally, as mass spectrometry is becoming the gold-standard method for proteomic work, micro-flow LC-MS/MS has been regarded as a potential alternative to the more sensitive but less robust nano-LC-MS/MS systems (Bian *et al.*, 2020; Distler *et al.*, 2019).

The relative abundances of quantified proteins have shown characteristic differences in the patterns across seed developmental stages and could be clustered accordingly into six groups (**Figure 5.5**). For simplicity, the proteins will be referred by their generic names and/or accession numbers as they are not yet properly annotated in *Vf*. Details of the 344 proteins are summarized in **Table S 5.2**.

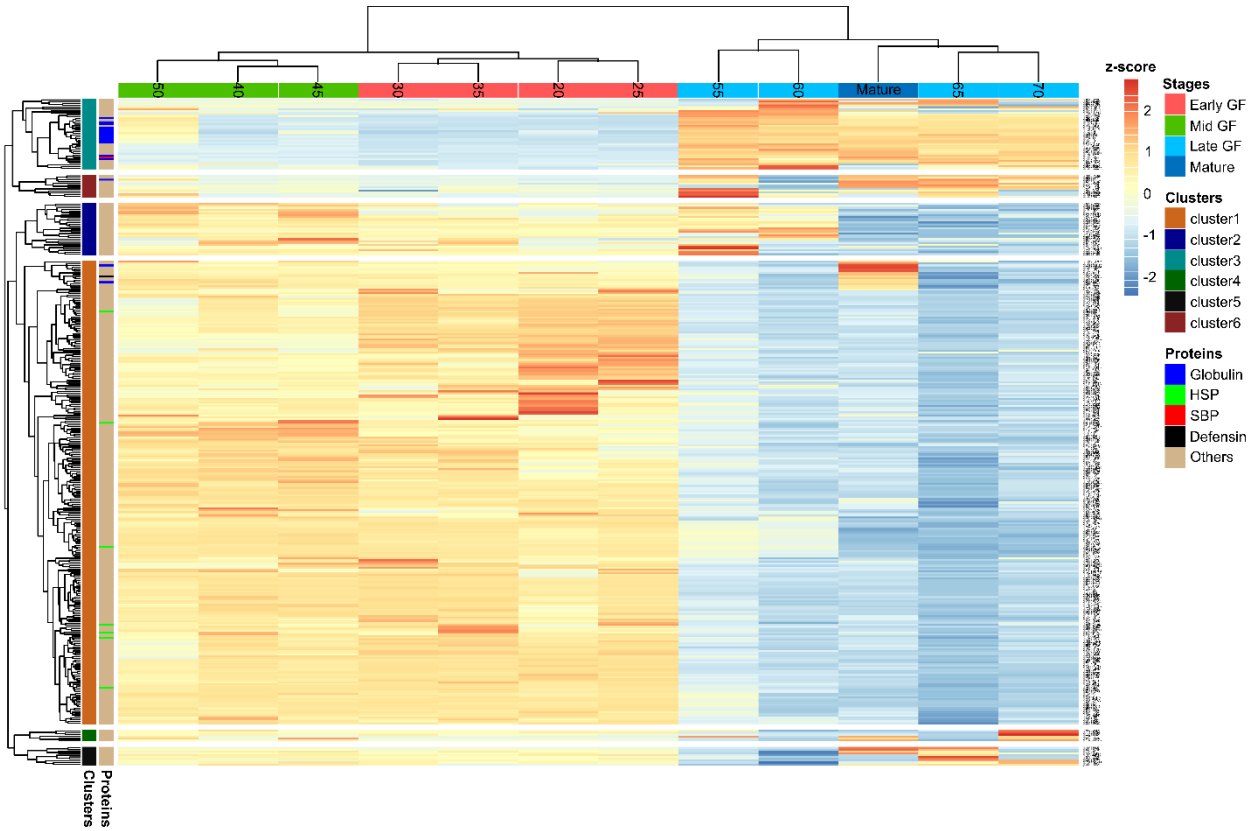

Figure 5.5. A heatmap showing relative abundances of 344 proteins (rows) at 12 seed growth stages (columns) between 20 DAP to maturity. The coloured bars at the top of the figure are early GF (light red), mid GF (light green), late GF (light blue) and Mature seed (dark blue). Individual proteins are colour-coded according to expression pattern cluster and belonging to protein families of interest. Colour coded protein families are globulins (including legumins, vicilins and convicilins), heat shock proteins (HSP), sucrose-binding protein (SBP), defensin and others. The data is a log transformed and normalized average abundances.

At the early growth stages (20-35 DAP), protein composition was marked by high relative abundance of large number of non-storage proteins (cluster 1) including those involved in transcription like histones–H2A & H2B (gi|593699727 & gi|470116864, respectively) and energy production such as Glyceraldehyde-3-phosphate dehydrogenase (gi|462138). Also, among seven chaperones identified in this analysis, the heat shock protein (gi|473217) was consistently upregulated until 45 DAP. This protein is 87% identical to the other HSP71.2 (gi|562006) which was previously found abundant in the dry mature *Vf* seeds (Warsame *et al.*, 2020). In pea, isoforms of these HSP proteins (PsHSP71.2, PsHSC71.0, and PsHSP70b) were

identified in different plant tissues and thought to play distinct biological functions (DeRocher and Vierling, 1995).

As seeds developed, cluster 2 proteins reached their highest intensity, mainly between 45 and 60 DAP. Though the majority of proteins in this group were not functionally annotated, the group included pectin acetylesterase 8-like (gi|356558882), annexin-like (gi|459649445, gi|828335547) and putative copper-transporting ATPase 3 (gi|734387082). On the other hand, proteins in cluster 3 mainly consisted of seed storage proteins, which started to accumulate around 45 DAP and continued steadily until maturity. Other non-storage proteins which were differentially accumulated during late grain filling also included a sodium/hydrogen exchanger (gi|29539109), which has been reported to play a critical role in protein trafficking and the biogenesis of protein storage vacuoles (PSV) in *Arabidopsis* (Wu *et al.*, 2016). Furthermore, a putative sugar phosphate transporter (gi|302854600), lectin-glucose complex (gi|82408030) and other protein with protein maintenance functions (gi|971508673) were in high abundance concomitantly to the phase of high protein accumulation. The remaining clusters (4-6) contained fewer proteins with distinct accumulation patterns. For instance, cluster 5 was upregulated mainly at late stages (65 DAP to maturity) and included proteins that may be related to defence/stress like ascorbate peroxidase (gi|731359393) and mitochondrial chaperonin (gi|461736). Another cluster 5 member, a serine protease inhibitor (gi|308800626), which is part of a gene family involved in the regulation of endogenous proteolysis in seeds and cell death during plant development and senescence (Clemente *et al.*, 2019), was also upregulated at 70 DAP and maturity.

5.4.4 Diverse accumulation patterns among storage proteins

Although as previously noted, the globulins were largely (though not exclusively) found in cluster 3 and were characterised by high accumulation in the late stages of grain filling, subtle differences in the onset and peak time of an individual protein could be highly significant given

the very high absolute abundance of these storage proteins. With this in mind, seventeen globulins, including nine legumins, four convicilins, three vicilins and a 7S vicilin-related storage protein were examined more closely (**Figure 5.6**). Within the legumin class, one legumin B, legumin A, two high molecular weight legumins and a minor-type legumin had similar patterns, where they were shown to accumulate from 50 DAP and remained in high abundance till maturity (**Figure 5.6**). On the other hand, a second legumin B (gi|126164) was deposited earlier from 45 DAP, while legumin K (gi|126171) was notably delayed and accumulated from 55 DAP. The most intriguing temporal trend amongst the globulins was a biphasic accumulation of two proteins annotated as legumin J, which were relatively abundant until 45 DAP, decreased to reach their lowest relative abundance between 55-70 DAP before peaking again in the mature seed. As suggested by their phylogenetic relationship (**Figure S 5.1**), these are closely related legumins, but which nonetheless may be under different regulatory mechanisms. Considering the diversity in the legumin genes and the possible modulating effect of environment and other genetic factors, it may not be possible to describe a deterministic and reproducible expression pattern for each protein. Nonetheless, temporal differences in the expression of these proteins has been documented in *Vf* and other legumes and some consistency does emerge. As was briefly mentioned in the introduction, De Pace *et al.* (1991), reported that legumin A subunits appeared to accumulate 2 days before the legumin B-type. In pea, gene expression analysis of developing seeds showed that legumin B-type gene (Psat3g055960) is highly expressed early (16 DAP) compared to legumin A-type (Psat3g058800) which had maximum expression at 19 to 23 DAP (Kreplak *et al.*, 2019). The phylogenetic relationships of pea and *Vf* seed proteins are in **Figure S 5.1**. Similarly, the expression pattern of legumin K in *Medicago truncatula* indicated that it was synthesized earlier (~16 DAP) compared to 24 DAP for legumin A (Verdier *et al.*, 2008).

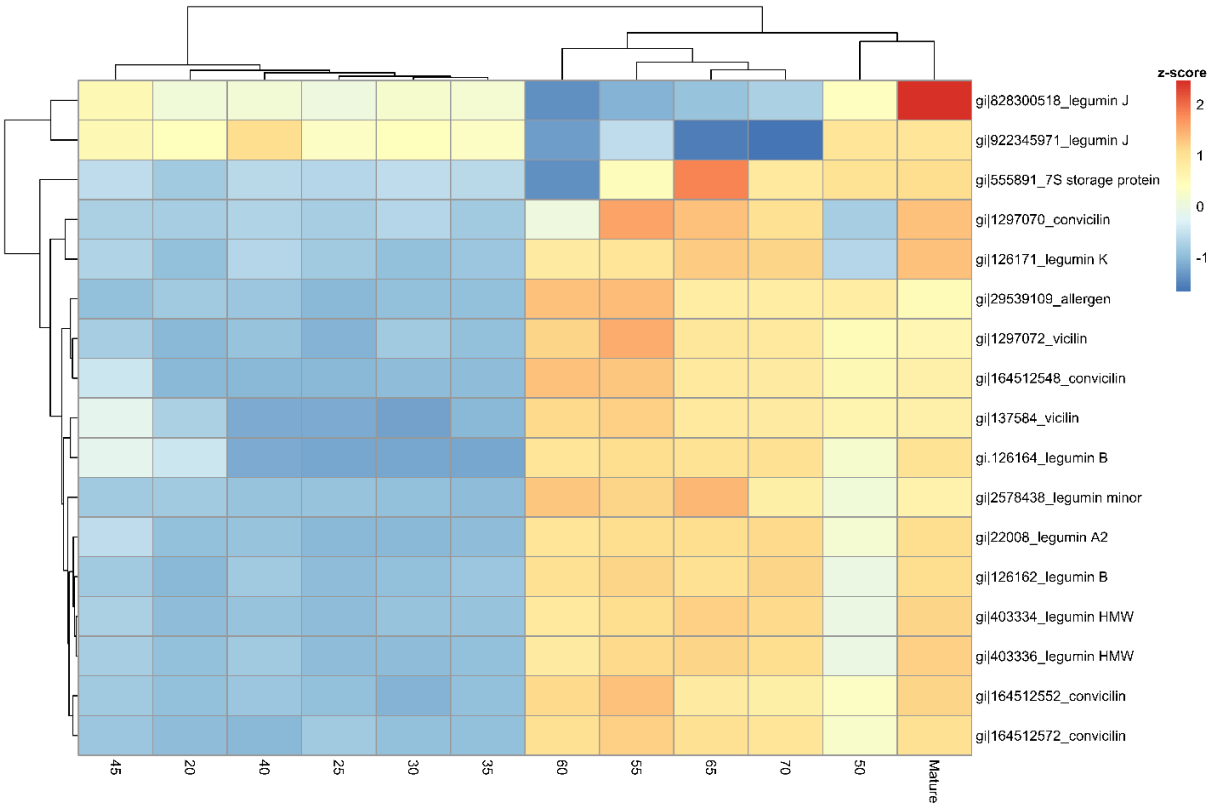

Figure 5.6. A heatmap showing the relative abundances of 17 globulins across 12 seed growth stages between 20 DAP to maturity. These proteins are a subset of cluster 3 described in **Figure 5.5**.

Regarding the three vicilin-type proteins quantified in this study, one of them was relatively abundant by 45 DAP (gi|137584_vicilin), 5 days earlier than others, but all had nearly uniform accumulation patterns peaking at 55 DAP and then decreasing towards maturity. The earlier accumulating vicilin (gi|137584) was the most abundant in major protein bands of 48-50 kDa in *Vf* seeds (Warsame *et al.*, 2020). Based on its apparent molecular weight, this protein is likely to be the vicilin band reported to be synthesized 4 days earlier than legumins (De Pace *et al.*, 1991). Convicilins generally became expressed at 50 DAP and reached maximum abundance between 55-60 DAP, after which their relative abundance declined slightly. However, one convicilin (gi|1297070) was an exception with a diphasic pattern like that of the 7S storage protein (gi|555891). Overall, the observed diversity in the timing of expression within the classes of storage proteins may indicate a complex regulatory system which needs further fine-tuning of its genetic basis and how it impacts the nutritional quality of the seeds.

In conclusion, the different accumulation patterns within and among protein classes revealed in this study suggest that a corresponding diversity may exist in transcriptional, translational and post-translational regulation of protein expression. This opens new avenues for further investigations into identification of master regulatory factors driving the developmental switch from cell division and growth to protein deposition, characterisation of promoter sequences that mediate differential responses to these master regulators and selection of the best targets for genetic improvement of nutritional composition. As by far the most comprehensive survey of *Vf* seed protein expression to date, the identities and expression pattern of the list of seed proteins reported here, and the first survey of the seed proteome of the reference inbred line Hedin/2, can contribute usefully to the annotation of the Hedin/2 *Vf* genome assembly which is currently in production.

In the following chapter 6, the thesis will focus on a sensory quality trait, hilum colour, with the aim of fine-mapping the locus containing the mutation responsible for the pale seed hilum. The importance of this trait stems from the fact that pale seed coat and hilum is preferred in the main faba bean export market in the Middle East and North Africa. Thus, the genetic information on seed protein content and quality as well as seed hilum colour is expected to contribute towards development of faba bean cultivars with desired nutritional and sensory quality traits.

5.5 References

Bevan, M., Bancroft, I., Bent, E., Love, K., Goodman, H., Dean, C., Bergkamp, R., Dirkse, W., Van Staveren, M., Stiekema, W., Drost, L., Ridley, P., Hudson, S. A., Patel, K., Murphy, G., Piffanelli, P., Wedler, H., Wedler, E., Wambutt, R., Weitzenegger, T., Pohl, T. M., Terryn, N., Gielen, J., Villarroel, R., De Clerck, R., Van Montagu, M., Lecharny, A., Auborg, S., Gy, I., Kreis, M., Lao, N., Kavanagh, T., Hempel, S., Kotter, P., Entian, K. D., Rieger, M., Schaeffer, M., Funk, B., Mueller-Auer, S., Silvey, M., James, R., Montfort, A., Pons, A., Puigdomenech, P., Douka, A., Voukelatou, E., Milioni, D., Hatzopoulos, P., Piravandi, E., Obermaier, B., Hilbert, H., Düsterhöft, A., Moores, T., Jones, J. D., Eneva, T., Palme, K., Benes, V., Rechman, S., Ansorge, W., Cooke, R., Berger, C., Delseny, M., Voet, M., Volckaert, G., Mewes, H. W., Klosterman, S., Schueller, C. & Chalwatzis, N. (1998). Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. *Nature,* **391** (6666), 485-8.

Bian, Y., Zheng, R., Bayer, F. P., Wong, C., Chang, Y.-C., Meng, C., Zolg, D. P., Reinecke, M., Zecha, J., Wiechmann, S., Heinzlmeir, S., Scherr, J., Hemmer, B., Baynham, M., Gingras, A.-C., Boychenko, O. & Kuster, B. (2020). Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC–MS/MS. *Nature Communications,* **11** (1), 157.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry,* **72**, 248-54.

Brewis, I. A. & Brennan, P. (2010). Proteomics technologies for the global identification and quantification of proteins. *Advances in Protein Chemistry and Structural Biology,* **80**, 1-44.

Clemente, M., Corigliano, M. G., Pariani, S. A., Sánchez-López, E. F., Sander, V. A. & Ramos-Duarte, V. A. (2019). Plant serine protease inhibitors: biotechnology application in agriculture and molecular farming. *International Journal of Molecular Sciences,* **20** (6), 1345.

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N. & Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics : MCP,* **13** (9), 2513-2526.

Cox, J. & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology,* **26** (12), 1367-1372.

De Pace, C., Delre, V., Mugnozza, G. T. S., Maggini, E., Cremonini, R., Frediani, M. & Cionini, P. G. (1991). Legumin of *Vicia faba* major: accumulation in developing cotyledons, purification, mRNA characterization and chromosomal location of coding genes. *Theoretical and Applied Genetics,* **83**, 17-23.

DeRocher, A. & Vierling, E. (1995). Cytoplasmic HSP70 homologues of pea: differential expression in vegetative and embryonic organs. *Plant Molecular Biology,* **27** (3), 441-456.

Distler, U., Łącki, M. K., Schumann, S., Wanninger, M. & Tenzer, S. (2019). Enhancing sensitivity of microflow-based bottom-up proteomics through postcolumn solvent addition. *Analytical Chemistry,* **91** (12), 7510-7515.

Duc, G. (1997). Faba bean (*Vicia faba* L). *Field Crops Research,* **53**, 99-109.

Gallardo, K., Le Signor, C., Vandekerckhove, J., Thompson, R. D. & Burstin, J. (2003). Proteomics of *Medicago truncatula* seed development establishes the time frame of

diverse metabolic processes related to reserve accumulation. *Plant Physiology,* **133** (2), 664-682.

Henriet, C., Aimé, D., Térézol, M., Kilandamoko, A., Rossin, N., Combes-Soia, L., Labas, V., Serre, R.-F., Prudent, M., Kreplak, J., Vernoud, V. & Gallardo, K. (2019). Water stress combined with sulfur deficiency in pea affects yield components but mitigates the effect of deficiency on seed globulin composition. *Journal of Experimental Botany,* **70** (16), 4287-4304.

John, K. M. M., Khan, F., Luthria, D. L., Matthews, B., Garrett, W. M. & Natarajan, S. (2017). Proteomic and metabolomic analysis of minimax and Williams 82 soybeans grown under two different conditions. *Journal of Food Biochemistry,* **41** (6), e12404.

Jones, S. I., Gonzalez, D. O. & Vodkin, L. O. (2010). Flux of transcript patterns during soybean seed development. *BMC Genomics,* **11** (1), 136.

Kreplak, J., Madoui, M.-A., Cápal, P., Novák, P., Labadie, K., Aubert, G., Bayer, P. E., Gali, K. K., Syme, R. A., Main, D., Klein, A., Bérard, A., Vrbová, I., Fournier, C., D'Agata, L., Belser, C., Berrabah, W., Toegelová, H., Milec, Z., Vrána, J., Lee, H., Kougbeadjo, A., Térézol, M., Huneau, C., Turo, C. J., Mohellibi, N., Neumann, P., Falque, M., Gallardo, K., McGee, R., Tar'An, B., Bendahmane, A., Aury, J.-M., Batley, J., Le Paslier, M.-C., Ellis, N., Warkentin, T. D., Coyne, C. J., Salse, J., Edwards, D., Lichtenzveig, J., Macas, J., Doležel, J., Wincker, P. & Burstin, J. (2019). A reference genome for pea provides insight into legume genome evolution. *Nature Genetics,* **51** (9), 1411-1422.

Lee, J. & Koh, H.-J. (2011). A label-free quantitative shotgun proteomics analysis of rice grain development. *Proteome Science,* **9** (1), 61.

Li, L., Hur, M., Lee, J.-Y., Zhou, W., Song, Z., Ransom, N., Demirkale, C. Y., Nettleton, D., Westgate, M., Arendsee, Z., Iyer, V., Shanks, J., Nikolau, B. & Wurtele, E. S. (2015). A systems biology approach toward understanding seed composition in soybean. *BMC Genomics,* **16** (Suppl 3), S9.

Mahalingam, R. (2017). Shotgun proteomics of the barley seed proteome. *BMC Genomics,* **18** (1), 44.

Mosse, J. (1990). Nitrogen-to-protein conversion factor for ten cereals and six legumes or oilseeds. A reappraisal of its definition and determination. Variation according to species and to seed protein content. *Journal of Agricultural and Food Chemistry,* **38** (1), 18-24.

Müntz, K., Horstmann, C. & Schlesier, B. (1999). Vicia globulins. *In:* Shewry, P. R. & Casey, R. (eds.) *Seed Proteins.* Dordrecht: Springer Netherlands. pp 259-284.

Panitz, R., Borisjuk, L., Manteuffel, R. & Wobus, U. (1995). Transient expression of storage-protein genes during early embryogenesis of *Vicia faba*: synthesis and metabolization of vicilin and legumin in the embryo, suspensor and endosperm. *Planta,* **196** (4), 765-774.

Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods,* **9**, 671.

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M. E. & Usadel, B. (2019). MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular Plant,* **12** (6), 879-892.

Scollo, E., Neville, D., Oruna-Concha, M. J., Trotin, M. & Cramer, R. (2018). Characterization of the proteome of *Theobroma cacao* beans by nano-UHPLC-ESI MS/MS. *Proteomics,* **18** (3-4), 1700339.

Sital, J. S., Malhotra, J. S., Sharma, S. & Singh, S. (2011). Comparative studies on biochemical components in mung bean [*Vigna radiata* (L.) Wilczek] varieties cultivated in summer and Kharif seasons. *Indian Journal of Agricultural Biochemistry,* **24**, 68-72.

Suso, M. J., Moreno, M. T., Mondragao-Rodrigues, F. & Cubero, J. I. (1996). Reproductive biology of *Vicia faba*: role of pollination conditions. *Field Crops Research,* **46** (1-3), 81-91.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M. & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods,* **13** (9), 731-740.

Verdier, J., Kakar, K., Gallardo, K., Le Signor, C., Aubert, G., Schlereth, A., Town, C. D., Udvardi, M. K. & Thompson, R. D. (2008). Gene expression profiling of *M. truncatula* transcription factors identifies putative regulators of grain legume seed filling. *Plant Molecular Biology,* **67** (6), 567-580.

Warsame, A. O., Michael, N., O'Sullivan, D. M. & Tosi, P. (2020). Identification and Quantification of major faba bean seed proteins. *Journal of Agricultural and Food Chemistry,* **68** (32), 8535-8544.

Warsame, A. O., O'Sullivan, D. M. & Tosi, P. (2018). Seed storage proteins of faba bean (*Vicia faba* L): current status and prospects for genetic improvement. *Journal of Agricultural and Food Chemistry,* **66** (48), 12617-12626.

Wu, X., Ebine, K., Ueda, T. & Qiu, Q.-S. (2016). AtNHX5 and AtNHX6 are required for the subcellular localization of the SNARE complex that mediates the trafficking of seed storage proteins in *Arabidopsis*. *PLOS ONE,* **11** (3), e0151658.

5.6 Supplementary

Table S 5.1. Summary of the parameter settings used in MaxQuant software for label-free quantification of protein abundances

MaxQuant version	1.6.17.0
Group specific parameters	
Type	Standard
Multiplicity	1
Labels	-
Digestion	
Enzyme mode	Specific
Enzyme	Trypsin/P
Max. missed cleavages	2
Separate enzyme for first search	False
Modifications	
Fixed modifications	Carbamidomethyl (C)
Variable modifications	Oxidation (M);Deamidation (NQ);Carbamyl (N-term)
Max. number of modifications per peptide	5
Label-free quantification	
LFQ min.ratio count	1
Normalization type	classic
Fast LFQ	True
Global parameters	
Sequences	
Fasta file	1218 proteins identified in faba bean seeds
Include contaminants	True
Decoy mode	revert
Special AAs	KR
Min. peptide length	6
Max. peptide mass	4600 Da
Min. peptide length for unspecific search	8
Max. peptide length for unspecific search	25
Identification	
PSM FDR	0.05
Protein FDR	0.05
Site FDR	0.01
Min. unique peptides	1
Min. razor + unique peptides	0
Min. peptides	0
Min. score for unmodified peptides	0
Min. score for modified peptides	40
Min. delta score for unmodified peptides	0
Min. delta score for modified peptides	6
Base FDR calculations on delta score	False
Razor protein FDR	False
Split protein groups by taxonomy ID	False
Second peptides	True
Match between runs	True
Dependent peptides	False

Table S 5.1. continued

Protein quantification	
Label min. ratio count	1
Peptides used for protein quantification	Unique+razor
Modifications included in protein quantification	Oxidation (M);Carbamyl (N-term)
Discard unmodified counterpart peptides	True
Min. ratio count	1
Advanced ratio estimation	True
Label-free quantification	
LFQ min. ratio count	1
Separate LFQ in parameter groups	True
Stabilize large LFQ ratios	True
Require MS/MS for LFQ comparisons	True
iBAQ	False
iBAQ log fit	False
Advanced site intensities	True

Table S 5.2. List of the 344 identified by mass spectrometer and MASCOT search with peptide identity threshold at $p < 0.05$ and quantified by MaxQuant across 12 developmental stages of *Vf* seeds.

Accession	Species	Description	Function
gi|15241168	*A. thaliana*	tubulin alpha-3	Cell structure
gi|17402467	*N. tabacum*	alpha-tubulin	Cell structure
gi|393715734	*B. rapa*	tubulin	Cell structure
gi|464849	*P. dulcis*	Tubulin alpha chain	Cell structure
gi|267072	*P. sativum*	Tubulin beta-1 chain	Cell structure
gi|267075	*P. sativum*	Tubulin beta-2 chain	Cell structure
gi|1021033319	*D. carota*	hypothetical protein DCAR_021532	Cell structure
gi|217072994	*M. truncatula*	unknown	Cell structure
gi|697114463	*N. tomentosiformis*	PREDICTED: probable 125 kDa kinesin-related protein	Cell structure
gi|1012356034	*C. cajan*	Protein transport protein Sec16B	Cell structure
gi|593699246	*P. vulgaris*	hypothetical protein PHAVU_005G125500g	Cell structure
gi|593780701	*P. vulgaris*	hypothetical protein PHAVU_003G031200g	Cell structure
gi|1035916500	*M. esculenta*	hypothetical protein MANES_07G115900	Cell structure
gi|20329	*O. sativa*	actin	Cell structure
gi|568868036	*C. sinensis*	PREDICTED: kinesin-related protein 11 isoform X1	Cell structure
gi|217073868	*M. truncatula*	unknown	Cell structure
gi|902550499	*D. officinale*	beta-1 tubulin	Cell structure
gi|356558882	*G. max*	PREDICTED: pectin acetylesterase 8-like	Cell structure
gi|302816083	*S. moellendorffii*	hypothetical protein SELMODRAFT_235864	Cell structure
gi|719995224	*N. nucifera*	PREDICTED: profilin-like	Cell structure
gi|292630923	*M. sativa*	Beta-xylosidase/alpha-L-arabinofuranosidase 2	Cell structure
gi|661899298	*C. canephora*	unnamed protein product	Cell growth/division
gi|475588355	*A. tauschii*	DNA polymerase epsilon catalytic subunit A	Cell growth/division
gi|1025206662	*N. tabacum*	PREDICTED: uncharacterized protein LOC107799874	Cell growth/division
gi|470116151	*F. vesca*	PREDICTED: proliferating cell nuclear antigen	Cell growth/division
gi|643730814	*J. curcas*	hypothetical protein JCGZ_04889	Cell growth/division
gi|548848479	*A. trichopoda*	hypothetical protein AMTR_s00154p00016840	Cell growth/division
gi|168006432	*P. patens*	predicted protein	Cell growth/division
gi|947071886	*G. max*	hypothetical protein GLYMA_13G199800	Cell structure
gi|363807732	*G. max*	uncharacterized protein LOC100806472	Cell structure
gi|922329195	*M. truncatula*	annexin D8	Cell structure
gi|1012357882	*C. cajan*	Annexin-like protein RJ4	Cell structure
gi|357514975	*M. truncatula*	annexin D8	Cell structure
gi|357514983	*M. truncatula*	annexin D8	Cell structure
gi|217072212	*M. truncatula*	unknown	Disease/defense
gi|643716085	*J. curcas*	hypothetical protein JCGZ_18938	Disease/defense

Table S 5.2. continued

gi|401108	*P. sativum*	Superoxide dismutase	Disease/defense
gi|920708746	*P. angularis*	hypothetical protein LR48_Vigan08g157000	Disease/defense
gi|1026094862	*C. annuum*	PREDICTED: putative respiratory burst oxidase homolog protein H isoform X1	Disease/defense
gi|727434280	*C. sativa*	PREDICTED: glutathione S-transferase DHAR1, mitochondrial	Disease/defense
gi|1012032416	*A. duranensis*	PREDICTED: L-ascorbate peroxidase, cytosolic	Disease/defense
gi|629109876	*E. grandis*	hypothetical protein EUGRSUZ_E03801, partial	Disease/defense
gi|734409612	*G. soja*	Putative disease resistance RPP13-like protein 1	Disease/defense
gi|731359393	*B. vulgaris*	PREDICTED: L-ascorbate peroxidase, cytosolic-like	Disease/defense
gi|502121419	*C. arietinum*	PREDICTED: glutathione S-transferase DHAR2-like	Disease/defense
gi|462138	*P. sativum*	Glyceraldehyde-3-phosphate dehydrogenase, cytosolic	Energy
gi|120666	*A. majus*	Glyceraldehyde-3-phosphate dehydrogenase, cytosolic	Energy
gi|828305948	*C. arietinum*	PREDICTED: ATP synthase subunit beta, mitochondrial	Energy
gi|1168410	*P. sativum*	Fructose-bisphosphate aldolase, cytoplasmic isozyme 2	Energy
gi|922407810	*M. truncatula*	F0F1-type ATP synthase, beta subunit	Energy
gi|373432589	*G. max*	malate dehydrogenase, cytoplasmic-like	Energy
gi|920704907	*P. angularis*	hypothetical protein LR48_Vigan07g183600	Energy
gi|1012211918	*A. duranensis*	PREDICTED: LOW QUALITY PROTEIN: ruBisCO large subunit-binding protein subunit beta, chloroplastic-like	Energy
gi|502145480	*C. arietinum*	PREDICTED: ruBisCO large subunit-binding protein subunit alpha, chloroplastic	Energy
gi|11133373	*M. sativa*	Malate dehydrogenase, cytoplasmic	Energy
gi|115278596	*T. dactyloides*	ATPase subunit 1 (mitochondrion)	Energy
gi|2827080	*M. sativa*	malate dehydrogenase precursor	Energy
gi|9230771	*P. sativum*	cytosolic phosphoglycerate kinase	Energy
gi|1710807	*P. sativum*	RuBisCO large subunit-binding protein subunit alpha, chloroplastic	Energy
gi|1026078724	*C. annuum*	PREDICTED: glucose and ribitol dehydrogenase homolog 1-like isoform X2	Energy
gi|55785631	*G. orbicularis*	ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit	Energy
gi|2116558	*P. sativum*	F1 ATPase	Energy
gi|470114187	*F. vesca*	PREDICTED: putative lactoylglutathione lyase	Energy
gi|145347850	*O. lucimarinus*	predicted protein	Energy
gi|37721023	*L. comosa*	ATP synthase beta subunit, partial (chloroplast)	Energy
gi|548855274	*A. trichopoda*	hypothetical protein AMTR_s00040p00201950	Energy

Table S 5.2. continued

gi|695044076	*M. acuminata*	PREDICTED: uncharacterized protein LOC103991948 isoform X1	Energy
gi|951027555	*V. radiata*	PREDICTED: ATP synthase subunit beta, mitochondrial	Energy
gi|502081910	*C. arietinum*	PREDICTED: triosephosphate isomerase, cytosolic	Energy
gi|527190463	*G. aurea*	hypothetical protein M569_13104, partial	Energy
gi|297826989	*A. lyrata*	pentatricopeptide repeat-containing protein	Energy
gi|1004145884	*G. pectorale*	hypothetical protein GPECTOR_6g779	Energy
gi|922346354	*M. truncatula*	phosphopyruvate hydratase	Enzyme classification
gi|674241778	*A. alpina*	hypothetical protein AALP_AA5G159700	Enzyme classification
gi|584292402	*P. granatum*	glyceraldehyde-3-phosphate dehydrogenase, partial	Enzyme classification
gi|541135687	*M. graminea*	NAD-dependent glyceraldehyde-3-phosphate dehydrogenase short paralog, partial	Enzyme classification
gi|976921045	*C. cardunculus*	hypothetical protein Ccrd_015939	Enzyme classification
gi|1012189741	*A. duranensis*	PREDICTED: enolase 2-like	Enzyme classification
gi|566169982	*P. trichocarpa*	hypothetical protein POPTR_0005s07250g	Enzyme classification
gi|976899454	*C. cardunculus*	disulfide isomerase	Enzyme classification
gi|523917083	*D. aberdeenense*	glyceraldehyde-3-phosphate dehydrogenase, partial	Enzyme classification
gi|133902308	*G. arboreum*	putative protein disulfide isomerase	Enzyme classification
gi|728838840	*G. arboreum*	disulfide-isomerase	Enzyme classification
gi|567187205	*E. salsugineum*	hypothetical protein EUTSA_v10010288mg	Enzyme classification
gi|543176810	*P. vulgaris*	mitochondrial aldehyde dehydrogenase	Enzyme classification
gi|571465933	*G. max*	PREDICTED: receptor like protein kinase S.2-like	Enzyme classification
gi|49387695	*O. sativa*	putative receptor kinase	Enzyme classification
gi|567201649	*E. salsugineum*	hypothetical protein EUTSA_v10020418mg	Enzyme classification
gi|1011927160	*H. annuus*	cytoplasmic enolase	Enzyme classification
gi|166418	*M. sativa*	putative endomembrane protein precursor	Enzyme classification
gi|44921641	*P. sativum*	NADP-dependent isocitrate dehydrogenase	Enzyme classification
gi|357463411	*M. truncatula*	myo-inositol 1-phosphate synthase	Metabolism
gi|401142	*V. faba*	Sucrose synthase	Metabolism
gi|658309958	*M. domestica*	bifunctional riboflavin biosynthesis protein RIBA 1, chloroplastic-like	Metabolism
gi|1346672	*P. sativum*	Nucleoside diphosphate kinase 1	Metabolism
gi|3913031	*M. sativa*	Beta-amylase	Metabolism
gi|148907091	*P. sitchensis*	unknown	Metabolism
gi|902174272	*S. oleracea*	hypothetical protein SOVF_162980	Metabolism
gi|356528022	*G. max*	PREDICTED: bifunctional riboflavin biosynthesis protein RIBA 1, chloroplastic-like	Metabolism
gi|357445919	*M. truncatula*	3,4-dihydroxy-2-butanone 4-phosphate synthase	Metabolism
gi|84468300	*T. pratense*	putative adenosylhomocysteinase	Metabolism

Table S 5.2. continued

gi|297829512	*A. lyrata*	hypothetical protein ARALYDRAFT_478305	Metabolism
gi|168014627	*P. patens*	predicted protein	Metabolism
gi|388501008	*L. japonicus*	unknown	Metabolism
gi|1012354795	*C. cajan*	Fructokinase-2	Metabolism
gi|147819622	*V. vinifera*	hypothetical protein VITISV_010090	Metabolism
gi|502139270	*C. arietinum*	PREDICTED: GMP synthase	Metabolism
gi|525345100	*C. arietinum*	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase-like	Metabolism
gi|662225959	*P. sativum*	vitamin B-12-independent methionine synthase	Metabolism
gi|357487801	*M. truncatula*	alpha-1,4-glucan-protein synthase	Metabolism
gi|13160142	*P. sativum*	sucrose synthase isoform 3	Metabolism
gi|168023302	*P. patens*	predicted protein	Metabolism
gi|1710838	*M. sativa*	Adenosylhomocysteinase	Metabolism
gi|75148854	*V. unguiculata*	Inosine-5'-monophosphate dehydrogenase	Metabolism
gi|15239735	*A. thaliana*	thiazole biosynthetic enzyme	Metabolism
gi|356525856	*G. max*	PREDICTED: bifunctional riboflavin biosynthesis protein RIBA 1, chloroplastic	Metabolism
gi|356574282	*G. max*	PREDICTED: D-3-phosphoglycerate dehydrogenase 1, chloroplastic-like	Metabolism
gi|166872	*A. thaliana*	S-adenosylmethionine synthetase	Metabolism
gi|502107091	*C. arietinum*	PREDICTED: inosine-5'-monophosphate dehydrogenase-like	Metabolism
gi|502142165	*C. arietinum*	PREDICTED: bifunctional riboflavin biosynthesis protein RIBA 1, chloroplastic-like	Metabolism
gi|922367746	*M. truncatula*	pfkB family carbohydrate kinase	Metabolism
gi|357508777	*M. truncatula*	cobalamin-independent methionine synthase	Metabolism
gi|1009140994	*Z. jujuba*	PREDICTED: enoyl-CoA delta isomerase 2, peroxisomal-like	Metabolism
gi|595797432	*P. persica*	hypothetical protein PRUPE_ppa008482mg	Metabolism
gi|609557	*P. sativum*	S-adenosylmethionine synthase, partial	Metabolism
gi|147856448	*V. vinifera*	hypothetical protein VITISV_024563	Metabolism
gi|357521193	*M. truncatula*	cytoplasmic phosphoglucomutase	Metabolism
gi|449455026	*C. sativus*	PREDICTED: omega-3 fatty acid desaturase, endoplasmic reticulum	Metabolism
gi|1709006	*A. chinensis*	S-adenosylmethionine synthase 3	Metabolism
gi|225425053	*V. vinifera*	PREDICTED: 6-phosphogluconate dehydrogenase, decarboxylating 3	Metabolism
gi|593219094	*P. vulgaris*	hypothetical protein PHAVU_011G107700g	Metabolism
gi|125539677	*O. sativa*	hypothetical protein OsI_07440	Multi-process regulation
gi|828335547	*C. arietinum*	PREDICTED: annexin-like protein RJ4, partial	Not assigned
gi|4850247	*P. sativum*	14-3-3-like protein	Not assigned
gi|1168189	*V. faba*	14-3-3-like protein A	Not assigned
gi|752855040	*P. sativum*	non-specific lipid transfer protein 3 precursor	Not assigned
gi|22742	*G. max*	pseudo-atpA	Not assigned

Table S 5.2. continued

gi|357452143	*M. truncatula*	seed maturation protein	Not assigned
gi|657950547	*M. domestica*	PREDICTED: annexin-like protein RJ4 isoform X1	Not assigned
gi|565382862	*S. tuberosum*	PREDICTED: starch synthase V precursor isoform X1	Not assigned
gi|119095	*V. faba*	Embryonic abundant protein VF30.1	Not assigned
gi|19658	*M. sativa*	translationally controlled tumor protein	Not assigned
gi|459649445	*G. barbadense*	annexin AnxGb5	Not assigned
gi|976918495	*C. cardunculus*	Nucleotide-binding, alpha-beta plait	Not assigned
gi|449812010	*x. Doritaenopsis*	elongation factor 1-alpha, partial	Not assigned
gi|697184772	*N. tabacum*	PREDICTED: GDSL esterase/lipase At4g10955-like	Not assigned
gi|971508673	*K. nitens*	Putative structural maintenance of chromosome protein	Not assigned
gi|565347586	*S. tuberosum*	PREDICTED: uncharacterized protein LOC102584459 isoform X1	Not assigned
gi|1026758124	*M. polymorpha*	hypothetical protein AXG93_3960s1360	Not assigned
gi|1035939297	*A. comosus*	hypothetical protein ACMD2_06360	Not assigned
gi|1040813396	*D. carota*	PREDICTED: uncharacterized protein LOC108203864	Not assigned
gi|194703160	*Z. mays*	unknown	Not assigned
gi|525313365	*C. arietinum*	18 kDa seed maturation protein-like	Not assigned
gi|242075586	*S. bicolor*	hypothetical protein SORBIDRAFT_06g014743, partial	Not assigned
gi|38636672	*O. sativa*	hypothetical protein	Not assigned
gi|1012263763	*A. duranensis*	PREDICTED: annexin-like protein RJ4	Not assigned
gi|224142475	*P. trichocarpa*	hypothetical protein POPTR_0018s12400g	Not assigned
gi|302828470	*V. carteri*	hypothetical protein VOLCADRAFT_86128	Not assigned
gi|694327669	*Pyrus x bretschneideri*	PREDICTED: multiple C2 and transmembrane domain-containing protein 2-like	Not assigned
gi|743873975	*E. guineensis*	PREDICTED: monothiol glutaredoxin-S7, chloroplastic	Not assigned
gi|920695830	*P. angularis*	hypothetical protein LR48_Vigan03g243700	Not assigned
gi|449442261	*C. sativus*	PREDICTED: dynamin-related protein 4C-like	Not assigned
gi|593509712	*P. vulgaris*	hypothetical protein PHAVU_008G248500g	Not assigned
gi|657959929	*M. domestica*	PREDICTED: GDSL esterase/lipase At1g29670-like	Not assigned
gi|976929520	*C. cardunculus*	CRAL-TRIO domain-containing protein, partial	Not assigned
gi|302754422	*S. moellendorffii*	hypothetical protein SELMODRAFT_403040	Not assigned
gi|566166370	*P. trichocarpa*	PWWP domain-containing family protein	Not assigned
gi|697167626	*N. tomentosiformis*	PREDICTED: uncharacterized protein At4g06744-like	Not assigned
gi|960464152	*B. distachyon*	PREDICTED: uncharacterized protein LOC100834903	Not assigned

Table S 5.2. continued

gi|923702049	*B. oleracea*	PREDICTED: proline-rich receptor-like protein kinase PERK12	Not assigned
gi|470123282	*F. vesca*	PREDICTED: uncharacterized protein LOC101290948	Not assigned
gi|75102461	*P. sativum*	Seed biotin-containing protein SBP65	Not assigned
gi|147822732	*V. vinifera*	hypothetical protein VITISV_040070	Not assigned
gi|226498118	*Z. mays*	uncharacterized protein LOC100284947	Not assigned
gi|728844838	*G. arboreum*	hypothetical protein F383_04038	Not assigned
gi|951027694	*V. radiata*	PREDICTED: U-box domain-containing protein 3	Not assigned
gi|971512911	*K. nitens*	hypothetical protein KFL_002280160	Not assigned
gi|1004148665	*G. pectorale*	hypothetical protein GPECTOR_1g574	Not assigned
gi|1009111613	*Z. jujuba*	PREDICTED: disease resistance protein RPM1-like	Not assigned
gi|552844109	*C. variabilis*	hypothetical protein CHLNCDRAFT_140858	Not assigned
gi|590712660	*T. cacao*	Uncharacterized protein TCM_002467	Not assigned
gi|658005774	*M. domestica*	PREDICTED: pentatricopeptide repeat-containing protein At5g44230	Not assigned
gi|764601987	*F. vesca*	PREDICTED: putative disease resistance protein RGA3 isoform X1	Not assigned
gi|1009120002	*Z. jujuba*	PREDICTED: putative inactive disease susceptibility protein LOV1	Not assigned
gi|388499082	*M. truncatula*	unknown	Not assigned
gi|672167001	*P. dactylifera*	PREDICTED: uncharacterized protein LOC103716998	Not assigned
gi|923850487	*B. napus*	PREDICTED: uncharacterized protein LOC106407489 isoform X1	Not assigned
gi|20139323	*P. sativum*	Defensin-2	Not assigned
gi|674893561	*B. napus*	BnaC06g26960D	Not assigned
gi|303281198	*M. pusilla*	predicted protein	Not assigned
gi|976907055	*C. cardunculus*	Homeodomain-like protein	Not assigned
gi|242066462	*S. bicolor*	hypothetical protein SORBIDRAFT_04g032615, partial	Not assigned
gi|302840630	*V. carteri*	hypothetical protein VOLCADRAFT_92484	Not assigned
gi|308800626	*O. tauri*	Serine proteinase inhibitor (KU family) (ISS), partial	Not assigned
gi|113549328	*O. sativa*	Os03g0668300	Not assigned
gi|147779506	*V. vinifera*	hypothetical protein VITISV_036894	Not assigned
gi|255088179	*M. commoda*	predicted protein	Not assigned
gi|308813309	*O. tauri*	laminarinase (ISS)	Not assigned
gi|971520621	*K. nitens*	hypothetical protein KFL_000010570	Not assigned
gi|303279096	*M. pusilla*	predicted protein	Not assigned
gi|661870294	*C. canephora*	unnamed protein product	Not assigned
gi|674959042	*B. napus*	BnaA04g07860D	Not assigned
gi|11072010	*A. thaliana*	F12A21.11	Not assigned
gi|77550981	*O. sativa*	expressed protein	Not assigned
gi|848900941	*E. guttata*	PREDICTED: uncharacterized protein LOC105970387	Not assigned
gi|565367199	*S. tuberosum*	PREDICTED: putative disease resistance protein RGA3	Not assigned

Table S 5.2. continued

gi|460406796	*S. lycopersicum*	PREDICTED: annexin-like protein RJ4	Not assigned
gi|159471932	*C. reinhardtii*	ubiquitin-protein ligase	Not assigned
gi|303289367	*M. pusilla*	predicted protein	Not assigned
gi|137584	*V. faba*	Vicilin	Protein destination and storage
gi|126162	*V. faba*	Legumin type B	Protein destination and storage
gi|126164	*V. faba*	Legumin type B	Protein destination and storage
gi|22008	*V. faba*	legumin A2 primary translation product	Protein destination and storage
gi|12580894	*V. faba*	putative sucrose binding protein	Protein destination and storage
gi|403336	*V. faba*	legumin	Protein destination and storage
gi|2578438	*P. sativum*	legumin (minor small)	Protein destination and storage
gi|164512572	*V. faba*	convicilin	Protein destination and storage
gi|473217	*P. sativum*	PsHSC71.0	Protein destination and storage
gi|3063396	*V. faba*	vcCyP	Protein destination and storage
gi|1297070	*V. narbonensis*	convicilin	Protein destination and storage
gi|403334	*V. faba*	legumin	Protein destination and storage
gi|593555617	*P. vulgaris*	hypothetical protein PHAVU_008G281300g	Protein destination and storage
gi|2827084	*M. sativa*	malate dehydrogenase precursor	Protein destination and storage
gi|29539109	*L. culinaris*	allergen Len c 1.0101	Protein destination and storage
gi|1297072	*V. narbonensis*	vicilin	Protein destination and storage
gi|359807323	*G. max*	uncharacterized protein LOC100814078	Protein destination and storage
gi|159459822	*V. pseudoreticulata*	heat shock protein 90	Protein destination and storage
gi|502129011	*C. arietinum*	PREDICTED: heat shock cognate protein 80	Protein destination and storage
gi|922340417	*M. truncatula*	ATPase, AAA-type, CDC48 protein	Protein destination and storage
gi|326490117	*H. vulgare*	predicted protein	Protein destination and storage
gi|164512548	*L. latifolius*	convicilin	Protein destination and storage
gi|502161581	*C. arietinum*	PREDICTED: subtilisin-like protease SBT1.7	Protein destination and storage
gi|612385867	*B. prasinos*	unknown	Protein destination and storage
gi|922327311	*M. truncatula*	subtilisin-like serine protease	Protein destination and storage
gi|126171	*P. sativum*	Legumin K	Protein destination and storage
gi|399942	*P. sativum*	Stromal 70 kDa heat shock-related protein, chloroplastic	Protein destination and storage
gi|566168859	*P. trichocarpa*	calreticulin family protein	Protein destination and storage

Table S 5.2. continued

gi|461736	*C. maxima*	Chaperonin CPN60-2, mitochondrial	Protein destination and storage
gi|56554815	*S. medusa*	heat shock protein hsp70	Protein destination and storage
gi|802724617	*J. curcas*	PREDICTED: calreticulin-3-like	Protein destination and storage
gi|308813664	*O. tauri*	Karyopherin (importin) beta 3 (ISS)	Protein destination and storage
gi|764562644	*F. vesca*	PREDICTED: uncharacterized protein LOC101298502	Protein destination and storage
gi|164512552	*L. aphaca*	convicilin	Protein destination and storage
gi|502117000	*C. arietinum*	PREDICTED: calreticulin	Protein destination and storage
gi|555891	*G. max*	7S storage protein alpha subunit	Protein destination and storage
gi|470126765	*F. vesca*	PREDICTED: malate dehydrogenase, chloroplastic	Protein destination and storage
gi|922399197	*M. truncatula*	peptidyl-prolyl cis-trans isomerase	Protein destination and storage
gi|387600188	*E. californica*	ABH-like cyclophilin	Protein destination and storage
gi|731318762	*B. vulgaris*	PREDICTED: LOW QUALITY PROTEIN: uncharacterized protein LOC104887018	Protein destination and storage
gi|922345971	*M. truncatula*	legumin J	Protein destination and storage
gi|922357238	*M. truncatula*	E3 ubiquitin-protein ligase COP1	Protein destination and storage
gi|460391363	*S. lycopersicum*	PREDICTED: serine carboxypeptidase-like	Protein destination and storage
gi|118158	*V. mungo*	Vignain	Protein destination and storage
gi|828300518	*C. arietinum*	PREDICTED: legumin J	Protein destination and storage
gi|357453983	*M. truncatula*	serine carboxypeptidase-like protein	Protein destination and storage
gi|357495169	*M. truncatula*	heat shock protein 81-2	Protein destination and storage
gi|593687583	*P. vulgaris*	hypothetical protein PHAVU_007G157200g	Protein destination and storage
gi|819320623	*B. luminifera*	heat shock cognate 70 kDa-like protein, partial	Protein destination and storage
gi|922400239	*M. truncatula*	translation elongation factor EF-2 subunit	Protein synthesis
gi|82408030	*V. faba*	Chain B, Fava Bean Lectin-Glucose Complex	Protein synthesis
gi|13877525	*A. thaliana*	S18.A ribosomal protein	Protein synthesis
gi|3023847	*M. sativa*	Guanine nucleotide-binding protein subunit beta-like protein	Protein synthesis
gi|3122060	*V. faba*	Elongation factor 1-alpha	Protein synthesis
gi|920692533	*P. angularis*	hypothetical protein LR48_Vigan02g190800	Protein synthesis
gi|224094244	*P. trichocarpa*	elongation factor 2 family protein	Protein synthesis
gi|357451779	*M. truncatula*	translation elongation factor EF-2 subunit	Protein synthesis
gi|1031986639	*G. hirsutum*	elongation factor 1-alpha-like	Protein synthesis
gi|1028979514	*G. hirsutum*	PREDICTED: superoxide dismutase	Protein synthesis

Table S 5.2. continued

gi|829282	*N. plumbaginifolia*	eukaryotic initiation factor 5A (1)	Protein synthesis
gi|84468264	*T. pratense*	putative 60S ribosomal protein L1	Protein synthesis
gi|388492036	*L. japonicus*	unknown	Protein synthesis
gi|567874809	*C. clementina*	hypothetical protein CICLE_v10012772mg	Protein synthesis
gi|1005817282	*E. lathyris*	elongation factor 1 alpha	Protein synthesis
gi|729314496	*T. hassleriana*	PREDICTED: polyadenylate-binding protein 4	Protein synthesis
gi|760437487	*A. protothecoides*	U3 small nucleolar ribonucleoprotein MPP10	Protein synthesis
gi|224077760	*P. trichocarpa*	40S ribosomal protein S19	Protein synthesis
gi|564587015	*S. alfredii*	elongation factor 1-alpha	Protein synthesis
gi|108706531	*O. sativa*	40S ribosomal protein SA, putative, expressed	Protein synthesis
gi|255646229	*G. max*	unknown	Protein synthesis
gi|351723425	*G. max*	uncharacterized protein LOC100500179	Protein synthesis
gi|1012220994	*A. hypogaea*	PREDICTED: elongation factor 1-alpha	Protein synthesis
gi|126101	*U. europaeus*	Anti-H(O) lectin 2	Protein synthesis
gi|1009115809	*Z. jujuba*	PREDICTED: 60S ribosomal protein L12-like	Protein synthesis
gi|357512935	*M. truncatula*	elongation factor 1-alpha	Protein synthesis
gi|1021031362	*D. carota*	hypothetical protein DCAR_026222	Protein synthesis
gi|255590183	*R. communis*	PREDICTED: leucine-rich repeat receptor-like protein kinase PXL2	Protein synthesis
gi|527198456	*G. aurea*	hypothetical protein M569_08176, partial	Protein synthesis
gi|950999301	*V. radiata*	PREDICTED: pollen receptor-like kinase 1	Protein synthesis
gi|217075286	*M. truncatula*	unknown	Protein synthesis
gi|388506980	*M. truncatula*	unknown	Protein synthesis
gi|224104009	*P. trichocarpa*	60S ribosomal protein L6	Protein synthesis
gi|502082389	*C. arietinum*	PREDICTED: serine/threonine-protein kinase EDR1 isoform X1	Protein synthesis
gi|126115	*L. cicera*	Lectin alpha-1 chain	Protein synthesis
gi|13359453	*P. sativum*	putative senescence-associated protein	Protein synthesis
gi|217070962	*M. truncatula*	unknown	Protein synthesis
gi|604298231	*E. guttata*	hypothetical protein MIMGU_mgv1a007381mg	Protein synthesis
gi|674251037	*A. alpina*	hypothetical protein AALP_AA1G174600	Protein synthesis
gi|720020966	*N. nucifera*	PREDICTED: L-type lectin-domain containing receptor kinase IV.1-like	Protein synthesis
gi|126123	*V. sativa*	Mitogenic lectin alpha chain	Protein synthesis
gi|357492613	*M. truncatula*	60S acidic ribosomal protein P0-1	Protein synthesis
gi|25809056	*P. sativum*	DEAD box RNA helicase	Protein synthesis
gi|729468664	*T. hassleriana*	PREDICTED: elongation factor 1-delta 1	Protein synthesis
gi|238625281	*J. regia*	flavanone 3-hydroxylase	Secondary metabolism
gi|12229615	*C. sinensis*	Chalcone synthase 1	Secondary metabolism
gi|168042500	*P. patens*	predicted protein	Secondary metabolism
gi|1705840	*P. sativum*	Chalcone synthase 1A	Secondary metabolism

Table S 5.2. continued

gi|756179851	*V. faba*	anthocyanidin reductase	Secondary metabolism
gi|685340759	*B. campestris*	nitrile-specifier protein 4-like	Secondary metabolism
gi|71534989	*M. sativa*	polygalacturonase inhibitor protein	Signal transduction
gi|37051109	*P. sativum*	polygalacturonase inhibiting protein	Signal transduction
gi|470142813	*F. vesca*	PREDICTED: E3 ubiquitin-protein ligase KEG isoform X1	Signal transduction
gi|527200750	*G. aurea*	hypothetical protein M569_06682	Transcription
gi|593699727	*P. vulgaris*	hypothetical protein PHAVU_005G142000g	Transcription
gi|657991039	*M. baccata*	PREDICTED: histone H2A-like	Transcription
gi|502149239	*C. arietinum*	PREDICTED: glycine-rich RNA-binding, abscisic acid-inducible protein-like	Transcription
gi|474166006	*T. urartu*	Histone H4	Transcription
gi|6273331	*M. sativa*	glycine-rich RNA binding protein, partial	Transcription
gi|629097846	*E. grandis*	hypothetical protein EUGRSUZ_G01238	Transcription
gi|357473273	*M. truncatula*	RNA-binding (RRM/RBD/RNP motif) family protein	Transcription
gi|502149245	*C. arietinum*	glycine-rich RNA-binding, abscisic acid-inducible protein-like	Transcription
gi|920701048	*P. angularis*	hypothetical protein LR48_Vigan05g187800	Transcription
gi|802541061	*J. curcas*	protein argonaute 1	Transcription
gi|527196987	*G. aurea*	hypothetical protein M569_09119, partial	Transcription
gi|307136393	*C. melo*	aquarius	Transcription
gi|1028941064	*G. hirsutum*	probable histone H2A.5	Transcription
gi|19611	*M. sativa*	histone H3 (AA 1-123)	Transcription
gi|552848723	*C. variabilis*	hypothetical protein CHLNCDRAFT_48410	Transcription
gi|302828570	*V. carteri*	histone H2A	Transcription
gi|1021037380	*D. carota*	hypothetical protein DCAR_018405	Transcription
gi|470116864	*F. vesca*	histone H2B	Transcription
gi|149213011	*A. megacarpus*	maturase K, partial (chloroplast)	Transcription
gi|674893103	*B. napus*	BnaA04g13850D	Transcription
gi|33326203	*Z. muricata*	RNA polymerase II largest subunit	Transcription
gi|1009106743	*Z. jujuba*	transcription factor GTE12	Transcription
gi|302839302	*V. carteri*	aurora kinase	Transcription
gi|108773317	*O. viridis*	beta" subunit of RNA polymerase	Transcription
gi|1035953488	*A. comosus*	Mediator of RNA polymerase II transcription subunit 12	Transcription
gi|1012251148	*A. duranensis*	uncharacterized protein LOC107468799 isoform X1	Transcription
gi|4469288	*M. sativa*	ferritin	Transporters
gi|734387082	*G. soja*	Putative copper-transporting ATPase 3	Transporters
gi|21027	*P. vulgaris*	ferritin	Transporters
gi|302854600	*V. carteri*	hypothetical protein VOLCADRAFT_100128	Transporters
gi|658010594	*M. domestica*	sodium/hydrogen exchanger 8 isoform X1	Transporters
gi|1029071605	*G. hirsutum*	piezo-type mechanosensitive ion channel homolog isoform X3	Transporters

Table S 5.3. Details of seed protein band labels in Figure 5.2 that were previously identified by mass spectrometer (Warsame et al. 2020).

Abbreviated name	Full names
Lox 107 kDa	Lipoxygenase 107 kDa
Lox 96 kDa	Lipoxygenase 96 kDa
Convc 79 kDa	Convicilin 79 kDa
HSP 75 kDa	HSP 75 kDa
Convc 65 kDa	Convicilin 65 kDa
Convc 54 kDa	Convicilin 54 kDa
Vc 48-50 kDa	Vicilin 50 kDa
SBP 46 kDa	SBP 46 kDa
Leg 43 kDa	Legumin 43 kDa
Leg A 39 kDa	α Legumin A 39 kDa
Leg B 37 kDa	α Legumin B 37 kDa
Vc 30 kDa	Vicilin 30 kDa
Leg 26 kDa	Legumin 26 kDa
β Leg A&B 22-23 kDa	β Legumin A&B 22-23 kDa
Alb 12 kDa	Albumin 12 kDa

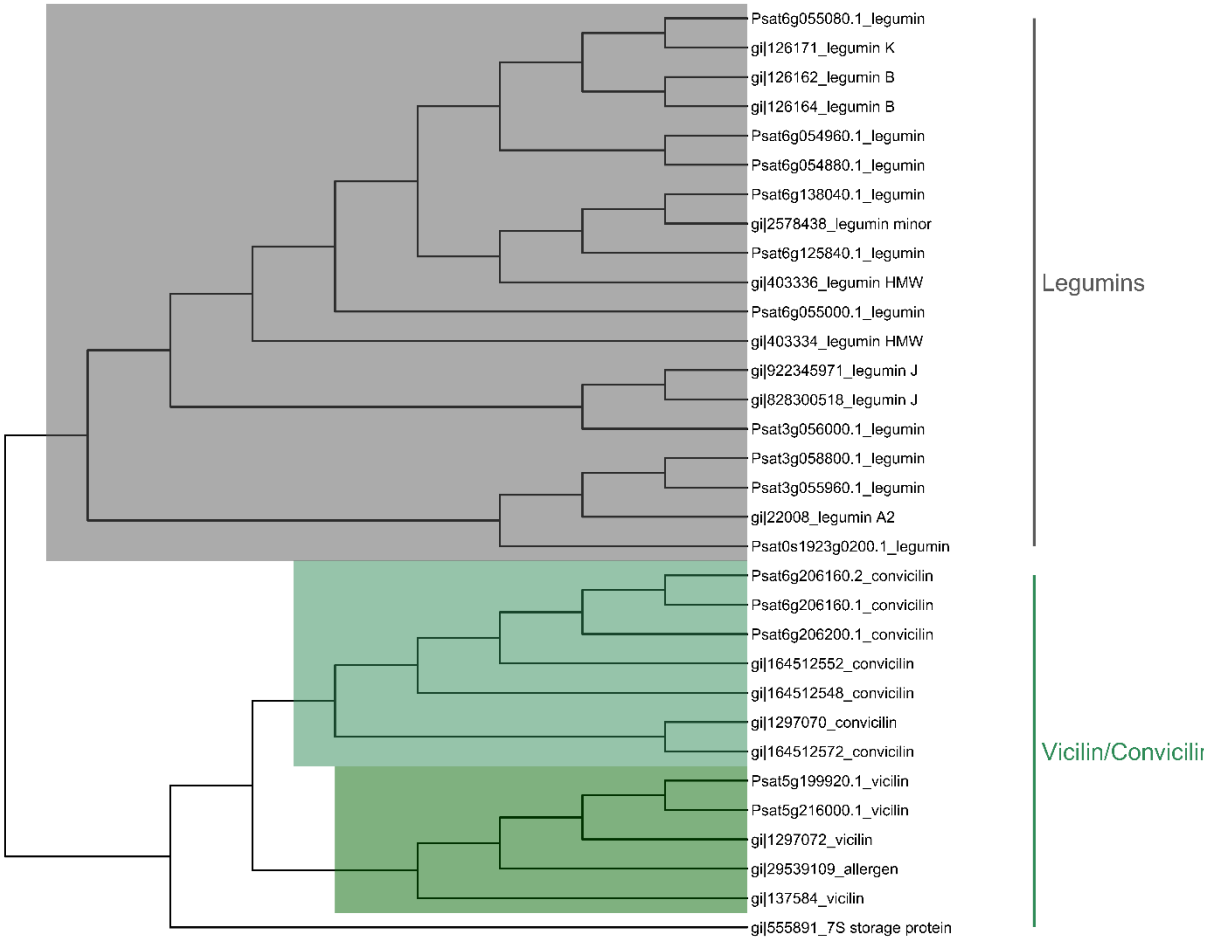

Figure S 5.1. Phylogenetic tree showing evolutionary relationships among globulins identified in this study and those recently reported in pea (Kreplak et al. 2019). The tree was constructed using MEGA X 10 with UPGMA method and 5000 replications.

Chapter 6 Fine-mapping of *hc* locus controlling seed hilum colour in faba bean (*Vicia faba*, L.)

6.1 Abstract

The genetic basis of many important agronomic and quality traits in *Vicia faba* are not yet known. In this work, seed hilum colour, which is an important quality trait in beans destined for export to Middle East markets, was used to explore the potential of high-density SNP array and bulk segregant analysis (BSA) in achieving a rapid and cheap way of trait fine-mapping in *Vicia faba*. A single DNA bulk constituted from 84 F_3 lines carrying the recessive allele phenotype of pale hilum (*hc*) and their parental lines was genotyped with a high-density single nucleotide polymorphism (SNP) genotyping array. Scoring homozygosity of previously mapped SNPs in the bulk sample showed a distinct segment of chromosome I where SNP genotypes of the bulk sample were enriched for the homozygous parental line carrying the pale hilum. A subsequent search for candidate genes in the identified region showed that the syntenic region in *Medicago truncatula* contained three tandem copies of the gene dihydroflavonol 4-reductase-like (DFR), which plays a critical role in anthocyanin biosynthesis in plants. The candidacy of DFR was further supported by linkage mapping in a separate biparental population which narrowed the interval containing both *hc* and DFR to 0.7 cM. Cloning of PCR amplicons from the candidate gene revealed the presence of two near identical copies of DFR gene in *Vicia faba* and, therefore, the full complexity of the candidate gene cluster could not be resolved. In conclusion, though the identified gene needs further characterization, the approach employed in this work illustrated the potential of this novel application of high-density genotyping array in BSA to accurately and efficiently map simply inherited traits in *Vicia faba*.

Key words: *Vicia faba*; hilum colour, dihydroflavonol 4-reductase-like

6.2 Introduction

Seed hilum colour is one of the quality attributes in *Vicia faba* (hereafter *Vf*) seeds destined for human consumption in Middle East markets where pale colour in the seed coat and hilum fetch premium prices (PGRO, 2017). This preference is probably related to the link between hilum colour and vicine and convicine (vc) content. As shown by genetic analysis, low *vc* is in-phase linkage with pale hilum in the original donor low vicine genotypes (Duc *et al.*, 2004; Khamassi *et al.*, 2013). In addition, hilum colour has been used as phenotypic marker for varietal purity during seed certification (Bould and Crofton, 1987).

Erith (1930) was the first to report that hilum colour is controlled by a single locus in which black is dominant over pale colour. AlsoMore recently, the hilum colour locus (*hc*) has been mapped to the telomeric region of chromosome I, and its relatively close linkage to the *vc* locus was confirmed (Khazaei *et al.*, 2015) but no candidate genes or diagnostic markers have been reported so far. In related species like soybean, hilum colour has been reported to be under control of *I* and *R* loci on two different chromosomes (Sonah *et al.*, 2015) in which brown colour (rather than black wild type state) results from a loss of function mutation in a R2R3 MYB transcription factor gene, which in turn leads to decreased expression of flavonoid 3-O-glucosyltransferase (UF3GT) gene required for the final step in anthocyanin biosynthesis (Gillman *et al.*, 2011).

Bulked segregant analysis (BSA) was developed as a fast and cost-effective method for identifying markers linked to traits of interest by bulking and genotyping a DNA sample(s) from individuals with extreme phenotypes of the trait (Michelmore *et al.*, 1991). With the advent of Next Generation Sequencing technology, the method underwent several modifications including sequencing of DNA or RNA bulks from two or single phenotype extremes (Zou *et al.*, 2016). For instance, Schneeberger *et al.* (2009) proposed mapping-by-sequencing method in which deep sequencing of a single DNA pool containing individuals with the mutant allele phenotype

allowed not only the identification of the candidate gene but also the allelic mutation causing the phenotype. The method has even been successfully applied in polyploid species (Gardiner *et al.*, 2016). In *Vf*, since genome sequencing is not yet feasible due to its large genome (~13 Gb), high-throughput SNP genotyping platforms and dense genetic maps are expected to provide an alternative route towards rapid and cost-effective fine-mapping of genomic regions associated with agronomic and nutritional quality traits.

In this study, high density SNP genotyping array and a single DNA bulk was used to fine-map the genomic region containing the *hc* locus in *Vf*. The candidate locus was further confirmed by linkage mapping in biparental population. Based on the synteny between *Vf* and *Medicago truncatula,* dihydroflavonol 4-reductase-like (DFR) was identified as the candidate gene for *hc* in *Vf*.

6.3 Materials and Methods

6.3.1 Plant materials:

A DNA bulk was constructed from 84 pale hilum-coloured lines from an F_3 segregating population (**Figure 6.1**) that was developed from a cross between the black hilum line NV639, an inbred line of the German cultivar Hedin, and the pale hilum inbred line NV866 derived from the French low vicine cultivar 'Disco'. For linkage mapping of the *hc* locus, an F6 recombinant inbred line (RIL) population consisting of 76 RILs from cross between NV153 (black) and NV644 (pale) was used.

6.3.2 DNA extraction and Genotyping:

DNA of all members of the F3 population described above had been previously extracted from young plant leaves using the CTAB method and stored at -80°C. Based on the phenotypic information, DNA samples from pale coloured individuals were selected and its quality was assessed using Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher Scientific, UK) following

manufacturer's guidelines. Then, a single DNA bulk was created by mixing ~2 µg DNA from each of the pale hilum individuals. Finally, about 100 µl DNA sample (~20 ng/µl concentration) of the bulk sample and both parental lines were genotyped with the *Vf* Axiom 58K SNP genotyping array using Affymetrix GeneTitan® system. SNP calling was performed in Axiom Analysis Suite (version 4.0, Thermofisher) following the recommended 'Best Practices Workflow' with default settings.

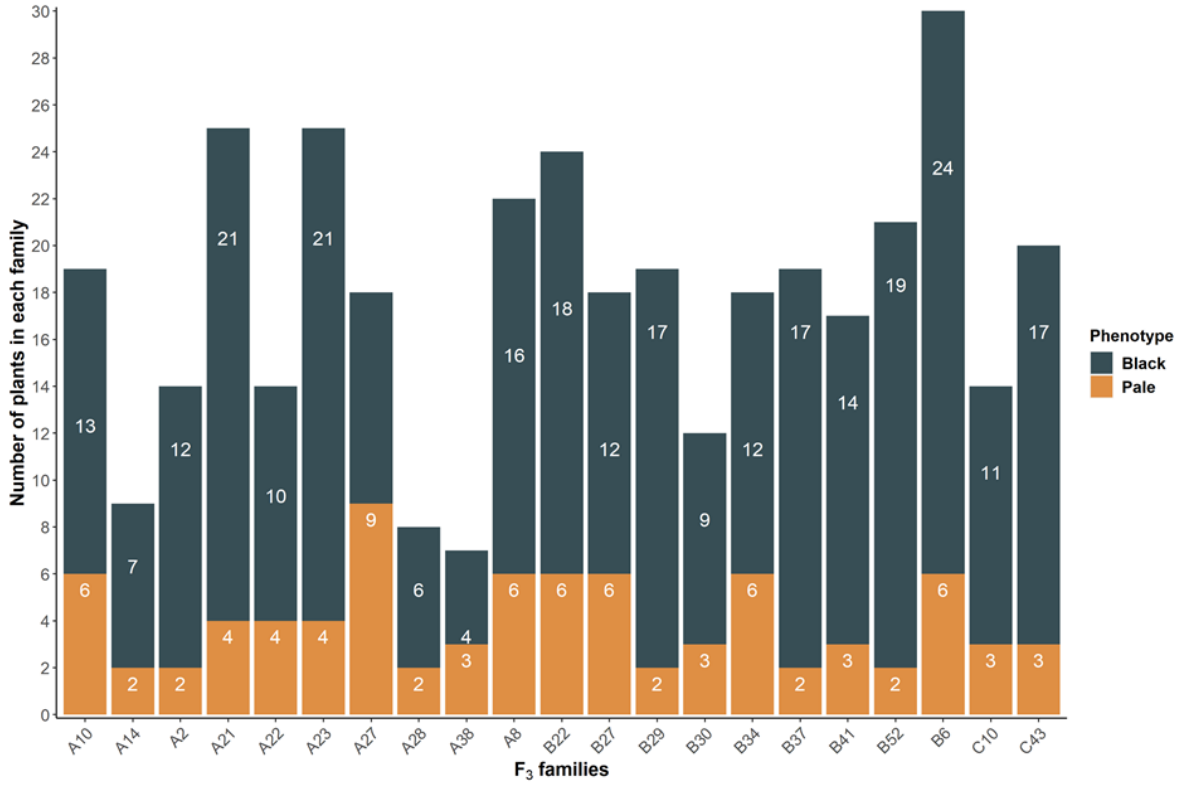

Figure 6.1. Segregation of hilum colour among NV866-1×NV639-2 F_3 families. The overall ration of black to pale was 253:84, which perfectly fits the 3:1segregation ratio expected if the *hc* was segregating as F_2 (see the results).

6.3.3 Mapping the *hc* locus

For homozygosity mapping, SNP call data was processed in MS Excel where genotypic data was cleaned by removing SNPs that were monomorphic, heterozygous or missing in the parental lines to confine the comparison between bulk and parental genotypes to only segregating markers. Then, SNP genotypes of the bulked sample were compared to parental genotypes at

each locus and SNPs were scored as “3”, “2” or “1” if heterozygous, missing or equal to pale parent, respectively.

For mapping in the RIL population, a linkage mapping was performed using Rqtl package (Broman *et al.*, 2003) with hilum colour coded as “0” for pale and “1” for black hilum and mapped as a Mendelian trait.

6.3.4 Sequencing the candidate region

A ~2.4 kb contig of the candidate region was first assembled from the public genomic data of *Vf* and then used to design PCR primers (**Table S 6.1**) to amplify the candidate gene from the parental lines, NV639 and NV866. PCR was conducted using PCRBIO HiFi Polymerase kit (PCR Biosystems Ltd, UK) following manufacturer’s protocol. The PCR product was assessed on 1.5% agarose gel and purified using GeneJET Gel Extraction and DNA Cleanup Micro Kit (ThermoFisher, UK) following the protocol of the manufacturer. In order to distinguish gene copies, the cleaned PCR product was cloned using Zero Blunt™ TOPO™ PCR Cloning Kit (ThermoFisher, UK) and the plasmid DNA was isolated from positive cultures using GeneJET Plasmid Miniprep Kit (ThermoFisher, UK). Finally, the plasmid was sequenced using M13 F and M13 R primers located at both ends of the insert. All PCR samples were sequenced at Source Bioscience, UK.

6.4 Results and discussion

6.4.1 Mapping the *hc* loci

The segregating F_3 population used for bulk segregant analysis originated from the selfing of F_2 plants that were heterozygous for SNP markers in the *vc* and *hc* region and, therefore, it can be considered as F_2 at the *hc* locus. This was confirmed by the overall observed 3:1 segregation ratio for hilum colour in the population (**Figure 6.1**). It is expected that within a bulk of DNA samples from such F_3 individuals carrying the recessive pale hilum phenotype, all SNPs

at/near the causative locus will be homozygous for the pale hilum parent (NV866) genotype, and of random genotype (therefore making the bulk appear heterozygous) at unlinked loci across the rest of the genome. After data cleaning, there were 6,450 polymorphic SNPs between the two parental lines of which 593 mapped to chromosome I.

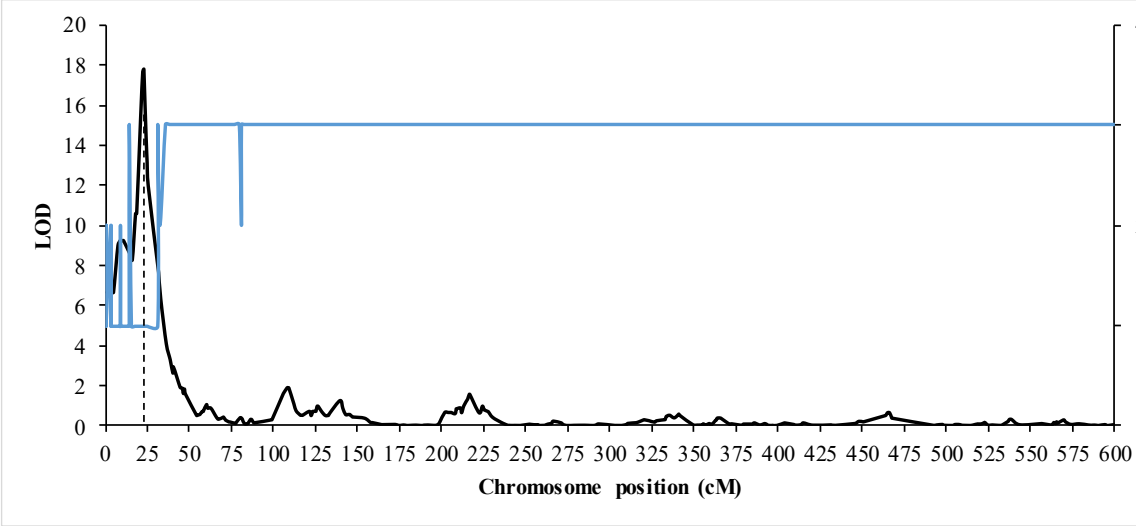

Figure 6.2. *Vf* chromosome I showing the region containing the candidate locus for hilum colour. The blue line is the genotype scores SNPs in the DNA bulk compared to the parental lines. The black line is LOD scores from linkage mapping which shows a strong peak overlapping with the candidate region identified by homozygosity mapping. The vertical dashed line indicates the QTL peak, where the putative causative SNP/candidate gene lies.

Plotting SNP scores of the bulk on chromosome I showed a segment between 16-31 cM with considerably enriched homozygosity for all SNPs (**Figure 6.2**). The SNP calling algorithm effectively coverts a continuous gradient of allele frequencies found in the DNA bulk into 3 crude 'bins', hence unlike the QTL scan, which quantitatively reflects the number of recombinations degrading the correlation between marker and phenotype, the homozygosity scan reports a score of 1 (homozygous NV866 allele) until the level of flanking recombination reaches a threshold that forces it to assign a score of 2 (NoCall) or 3 (Het) and thus most of the genetic resolution that could be captured in the bulk is lost. The resolution of the homozygosity scan could be improved by developing a bespoke scoring system that directly plotted the parent A:B signal ratios rather than the output of an algorithm designed to recognise only those discrete

genotypic states possible in a diploid individual, and the power of this approach would be magnified of course by increasing the number of bulk components which, unlike in the alternative QTL mapping approach, contribute to the genetic resolution without additional genotyping cost.

Table 6.1. *Vicia faba* chromosome I segment showing SNP names and positions, graphical genotype of RILs showing recombination close to *hc* and functional annotations of genes in the syntenic region of *Medicago truncatula*.

The interval in *V. faba* chromosome I containing the candidate region				Hilum colour of genotypes with genetic recombinations near the candidate region											*M. truncatula* syntenic region
SNP ID	**Chr**	**Pos.(cM)**	**LOD**	**Black**				**Pale**							**Annotation of the homologous genes**
AX-181147929_Medtr2g009270.1	1	15.37	8.26	B	A	A	A	A	B	B	A	A	A	A	3,4-dihydroxy-2-butanone 4-phosphate synthase
AX-181172107_Medtr2g008820.1	1	16.01	8.27	B	A	A	A	A	B	B	A	A	A	A	potassium transporter-like protein
AX-181473397_Medtr2g012630.1	1	17.31	10.57	B	A	B	B	A	B	B	A	A	A	A	katanin p80 WD40 repeat subunit B1-like protein
AX-181461804_Medtr2g012670.1	1	17.96	10.58	B	A	B	B	A	B	B	H	A	A	A	strubbelig receptor family 3 protein
AX-181455153_Medtr2g012630.1	1	18.61	10.59	B	A	B	B	A	B	B	A	A	A	A	katanin p80 WD40 repeat subunit B1-like protein
AX-181471930_Medtr2g012990.1	1	20.93	15.31	B	A	B	B	A	A	A	A	A	A	A	TIR class disease resistance protein
AX-181162128_Medtr2g013350.1	1	22.23	17.67	B	B	B	B	A	A	A	A	A	A	A	splicing factor 3B subunit 2
AX-181491099_Medtr2g013230.1	**1**	**22.56**	**17.73**	B	B	B	B	A	A	A	A	A	A	A	**dihydroflavonol-4-reductase-like protein**
AX-181154349_Medtr2g013450.1	1	22.89	17.79	B	B	B	B	A	A	A	A	A	A	A	PPR containing plant protein
AX-181154974_Medtr2g013690.1	1	24.19	14.45	B	B	B	B	A	A	A	A	B	A	B	DUF616 family protein
AX-181497306_Medtr2g014020.1	1	24.83	13.13	B	B	B	B	A	A	A	B	B	A	B	WD domain, G-beta repeat protein
AX-181189945_Medtr2g014370.1	1	25.47	11.97	B	B	B	B	A	A	A	B	B	B	B	hypothetical protein
AX-181175571_Medtr2g014480.1	1	25.79	12.00	B	B	B	B	A	A	A	B	B	B	B	DnaJ heat shock family protein
AX-181189944_Medtr2g014370.1	1	26.11	12.03	B	B	B	B	A	A	A	B	B	B	B	hypothetical protein
AX-181197124_Medtr2g014560.1	1	26.75	12.09	B	B	B	B	A	A	A	B	B	B	B	LRR receptor-like kinase family protein
AX-181151052_Medtr2g015040.1	1	29.07	10.76	A	B	B	B	B	A	A	B	B	B	B	ethylene response factor
AX-181180898_Medtr2g015320.1	1	30.04	10.21	A	B	B	B	B	A	A	B	B	B	B	ATP-dependent RNA helicase DHX35
AX-181168731_Medtr2g015310.1	1	30.36	9.87	A	B	B	B	B	A	A	B	B	B	B	glutamate receptor 2.7

The location of *hc* was fully consistent with QTL mapping in the RIL population which showed peak LOD score at 22.56 cM (**Figure 6.2**). One of the markers at the QTL peak and close to the centre of the bulk homozygous segment was a SNP from a gene whose homologue in *M. truncatula* is annotated as dihydroflavonol-4-reductase-like protein in Mt.4 CDS (Tang *et al.*, 2014). This enzyme is known to play a critical role in the final steps of anthocyanin production where dihydroflavonols are reduced to leucoanthocyanins before they are converted to anthocyanidins and finally to anthocyanins (Verdier *et al.*, 2012; Hossain *et al.*, 2018) and could therefore be considered a biological candidate for a pigment-related phenotype. In addition, by examining the phenotypic and genotypic data of RIL genotypes with genetic recombination near the candidate region, there was a 0.7 cM window which contained three mapped markers with peak LOD score whose genotype perfectly correlated with hilum colour,

including the candidate gene SNP (**Table 6.1**). This region of *Vf* chromosome I has a considerable collinearity with chromosome II of *M. truncatula* where three tandem copies of a predicted dihydroflavonol 4-reductase (DFR) gene are located (**Table 6.1**).

It has been shown that loss of function mutations in DFR abolished anthocyanin pigments from different plant tissues in various species. For instance, silencing of DFR gene in *Ipomoea batatas* resulted in loss of anthocyanin pigments on the leaves, stems and storage roots (Wang *et al.*, 2013). In related species, *Ipomoea nil*, CRISPR/Cas9 mediated mutation in DFR completely abolished stem and flower pigments (Watanabe *et al*., 2017). Moreover, in *Arabidopsis thaliana*, induced mutations in this gene was associated with anthocyanin reduction on seedlings and seed testa (Bharti and Khurana, 2003; Appelhagen *et al*., 2014). It was therefore considered a worthy hypothesis that a mutation in a DFR is potentially responsible for seed pale hilum colour in *Vf*.

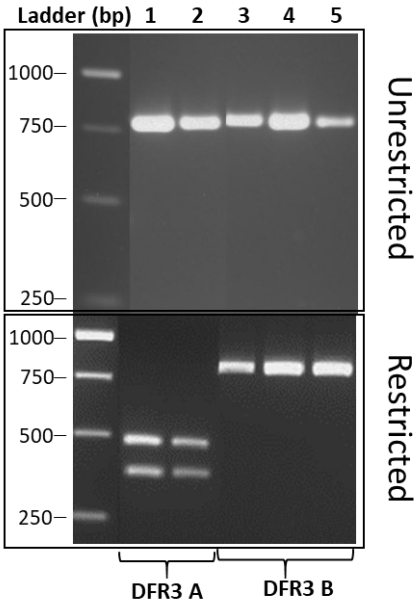

Figure 6.3. Agarose gel showing selected lanes (1-5) containing the cloned gene copies. The two gene copies are distinguished by *PstI* restriction enzyme.

6.4.2 Cloning the candidate gene

To identify possible causative mutations responsible for pale colour, a 2.4 kb transcript contig spanning the whole Open Reading Frame (ORF) was used for designing PCR primers (**Table S 6.1**) to amplify and sequence the entire candidate gene coding sequence from the parental lines, NV866 and NV639. However, even after using multiple primer combinations

under different PCR conditions, the whole target region could not be amplified as a single fragment. It was, however, established that the gene of interest had five exons and that intron 2 was responsible for the failure in amplifying the whole target sequence. Therefore, two sets of primers that amplified 5′ and 3′ ends of the gene on either side of the second intron were used in PCR reactions.

Sanger sequencing of the amplified gene fragments revealed the presence of multiple heterozygous SNPs and, in the case of NV639, exons from 3 to 5 resulted in mixed sequencing data indicating that the primers amplified more than one target sequence. In order to resolve the constituent sequences which were co-amplifying, the fragment spanning exons 3 to 5 was cloned and sequenced. As shown in **Figure 6.3**, there were two distinct copies of the target gene, denoted here as *Vf* DFR3 A and B (**Figure 6.3**). The sequenced segments of these copies showed high similarity explaining why they could not be specifically amplified by the primer sets used. The existence of multiple, closely related DFR genes in *Vf* is not surprising considering that dihydroflavonol-4-reductase enzyme belongs to a gene family containing more than 12 copies dispersed across the *M. truncatula* genome, Mt4.0 (Tang *et al.*, 2014). Of these, three tandem copies are found in the *M. truncatula* region syntenic to the mapped *hc* interval of *Vf.* Blast results showed that the sequenced two *Vf* gene copies were indeed orthologues of the syntenic *M. truncatula* DFR copies with highest matches between DFR3 A and Medtr2g013230 (82.74%), and DFR3 B and Medtr2g013250 (82.72%). This confirms that the two gene copies are likely to belong to the *hc* region. Two other DFR cDNAs (DFR1 and DFR2) had been previously reported (Ray *et al.*, 2015) which, based on the *M. truncatula* synteny information, can be hypothesized to reside on chromosome III of *Vf.*

Despite the difficult in sequencing the whole gene, the coding sequence was assembled from the amplified segments of the candidate gene, which contained some features discriminating between the gene copies and the parental lines, including multiple synonymous

and nonsynonymous SNPs, and a *Pstl* restriction site in copy A in NV639 (**Figure S 6.1**). However, it was not possible to draw any conclusions on the effects of these SNPs on hilum colour due to the lack of full sequence of the gene copies and the possibility that the causative mutation leading to change in hilum colour may reside in the upstream promoter region or regulatory regions around the candidate gene, as suspected in the case of the *TTG1-a* white flower colour mutation in *Vf* (Webb *et al.*, 2016). Therefore, pinpointing the mutation leading to pale hilum colour requires deeper characterization of the candidate region which was beyond the scope of this study.

In conclusion, pale hilum is a consumer preference quality trait but has only been previously given a very approximate map position on *Vf* chromosome I. In this work, the interval containing *hc* was narrowed to just 0.7 cM using high density genotyping of a RIL population making it feasible to think about identifying the gene itself. In a further technical innovation, it was shown that categorical traits such as *hc* can actually be mapped by genotyping a single recessive bulk DNA sample, although not with the same resolution as would be possible when characterizing the segregating population line by line. Since the 0.7cM *hc* interval contained a convincing biological candidate gene – a dihydroflavonol-3-reductase - putatively involved in a key step of the anthocyanin biosynthesis pathway, efforts were made to re-sequence DFR from pale and dark hilum parents. Although some partial DFR sequences and a set of putative polymorphisms between the pale and dark hilum parents were found, two problems prevented assembly of full copies of the dark and pale hilum alleles of DFR: the existence of a gene family with an unknown number of highly similar gene copies co-amplifying and problems amplifying intron 2. However, the expected release of a polished genome assembly of the reference line Hedin/2 (dark hilum) could make it easier to determine the total number of DFR copies in the *Vf* genome, the extent of non-coding elements including introns and promoters and their location, so that allele re-sequencing initiated here can be brought to a full conclusion.

6.5 References

Appelhagen, I., Thiedig, K., Nordholt, N., Schmidt, N., Huep, G., Sagasser, M. & Weisshaar, B. (2014). Update on transparent testa mutants from *Arabidopsis thaliana*: characterisation of new alleles from an isogenic collection. *Planta,* **240** (5), 955-970.

Bharti, A. K. & Khurana, J. P. (2003). Molecular characterization of transparent testa (tt) mutants of *Arabidopsis thaliana* (ecotype Estland) impaired in flavonoid biosynthetic pathway. *Plant Science,* **165** (6), 1321-1332.

Bould, A. & Crofton, G. R. A. (1987). Variability in the expression of hilum colour in field bean varieties in relation to seed certification standards. *Seed Science and Technology,* **15**, 651-662.

Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics,* **19** (7), 889-890.

Duc, G., Marget, P., Page, D. & Domoney, C. (2004). Facile breeding markers to lower contents of vicine and convicine in faba bean seeds and trypsin inhibitors in pea seeds. *In:* Muzquiz, M., Hill, G. D., Cuadrado, C., Pedrosa, M. M. & Burbano, C., eds. Recent Advances of Research in Antinutritional Factors in Legume Seeds and Oilseeds, 2004 Toledo, Spain. Wageningen: Wageningen Academic Publishers, 281–285.

Erith, A. G. (1930). The inheritance of colour, size and form of seeds, and of flower colour in *Vicia Faba* L. *Genetica* **12**, 477–510.

Gardiner, L.-J., Bansept-Basler, P., Olohan, L., Joynson, R., Brenchley, R., Hall, N., O'Sullivan, D. M. & Hall, A. (2016). Mapping-by-sequencing in complex polyploid genomes using genic sequence capture: a case study to map yellow rust resistance in hexaploid wheat. *The Plant Journal,* **87** (4), 403-419.

Gillman, J. D., Tetlow, A., Lee, J.-D., Shannon, J. G. & Bilyeu, K. (2011). Loss-of-function mutations affecting a specific *Glycine max* R2R3 MYB transcription factor result in brown hilum and brown seed coats. *BMC Plant Biology,* **11**, 155-155.

Hossain, M., Kim, H.-T., Shanmugam, A., Nath, U., Goswami, G., Song, J.-Y., Park, J.-I. & Nou, I.-S. (2018). Expression profiling of regulatory and biosynthetic genes in contrastingly anthocyanin rich strawberry (Fragaria × ananassa) cultivars reveals key genetic determinants of fruit color. *International Journal of Molecular Sciences,* **19** (3), 656.

Khamassi, K., Ben Jeddi, F., Hobbs, D., Irigoyen, J., Stoddard, F., O'Sullivan, D. M. & Jones, H. (2013). A baseline study of vicine–convicine levels in faba bean (*Vicia faba* L.) germplasm. *Plant Genetic Resources,* **11** (3), 250-257.

Khazaei, H., M. O'Sullivan, D., Jones, H., Pitts, N., Sillanpää, M., Pärssinen, P., Manninen, O. & Stoddard, F. (2015). Flanking SNP markers for vicine–convicine concentration in faba bean (*Vicia faba* L.). *Molecular Breeding,* **35** (38).

Michelmore, R. W., Paran, I. & Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences,* **88** (21), 9828-9832.

PGRO (2017). PGRO Pulse Agronomy Guide 2017. Peterborough, UK. : Processors and Growers Research Organisation.

Ray, H., Bock, C. & Georges, F. (2015). Faba bean: Transcriptome analysis from etiolated seedling and developing seed coat of key cultivars for synthesis of proanthocyanidins, phytate, raffinose family oligosaccharides, vicine, and convicine. *The Plant Genome,* **8** (1). doi: 10.3835/plantgenome2014.07.0028.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jorgensen, J. E., Weigel, D. & Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods,* **6** (8), 550-551.

Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I. & Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnology Journal,* **13** (2), 211-221.

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., Gentzbittel, L., Childs, K. L., Yandell, M., Gundlach, H., Mayer, K. F., Schwartz, D. C. & Town, C. D. (2014). An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics,* **15** (1), 312.

Verdier, J., Zhao, J., Torres-Jerez, I., Ge, S., Liu, C., He, X., Mysore, K. S., Dixon, R. A. & Udvardi, M. K. (2012). MtPAR MYB transcription factor acts as an on switch for proanthocyanidin biosynthesis in *Medicago truncatula*. *Proceedings of the National Academy of Sciences of the United States of America,* **109** (5), 1766-1771.

Wang, H., Fan, W., Li, H., Yang, J., Huang, J. & Zhang, P. (2013). Functional characterization of dihydroflavonol-4-reductase in anthocyanin biosynthesis of purple sweet potato underlies the direct evidence of anthocyanins function against abiotic stresses. *PLOS ONE,* **8** (11), e78484.

Watanabe, K., Kobayashi, A., Endo, M., Sage-Ono, K., Toki, S. & Ono, M. (2017). CRISPR/Cas9-mediated mutagenesis of the *dihydroflavonol-4-reductase-B (DFR-B)* locus in the Japanese morning glory *Ipomoea (Pharbitis) nil*. *Scientific Reports,* **7** (1), 10028.

Webb, A., Cottage, A., Wood, T., Khamassi, K., Hobbs, D., Gostkiewicz, K., White, M., Khazaei, H., Ali, M., Street, D., Duc, G., Stoddard, F. L., Maalouf, F., Ogbonnaya, F. C., Link, W., Thomas, J. & O'Sullivan, D. M. (2016). A SNP-based consensus genetic map for synteny-based trait targeting in faba bean (*Vicia faba* L.). *Plant Biotechnology Journal,* **14** (1), 177-185.

Zou, C., Wang, P. & Xu, Y. (2016). Bulked sample analysis in genetics, genomics and crop improvement. *Plant Biotechnology Journal,* **14** (10), 1941-1955.

6.6 Supplementary

Table S 6.1. List of DNA primers used for sequencing the hilum colour candidate gene from the parental lines NV639 and NV866.

Primer ID	Name	Sequence	Tm (°C)
HA12565128	DFR_ex1_F	ATGGAAAGGAGTTGCAAGGT	62.4
HA12565131	DFR_ex2_R	GGTGTTGAAAGGCAGTAGCA	62.9
HA12565132	DFR_ex3_F	GCA CAGTTTAAGAGCATTGAAGAA	63
HA12565133	DFR_ex3_R	AGAGCGGTGAGAGGAAGATG	63.5
HA12565135	DFR_ex4_R	GGGCTTCATTGTCTTTCACC	63.4
HA12565136	DFR_ex5_R1	TGTTTGGCCCTTCCAAATAC	63.5
HA12565138	DFR_ex5_R3	CCCCATACTAAGATCACTTATTC	57.3

Figure S 6.1. Alignment of the coding sequences of the candidate gene which shows multiple SNPs distinguishing between the two gene copies and the two parental lines. The two sequences of NV639 (A & B) contain the cloned segments of the two gene copies (base pairs 246-903) while ambiguous nucleotide symbols indicate polymorphic SNPs between the two gene copies. The blue arrows denote nonsynonymous SNP polymorphisms between gene copies or between the parental lines.

Chapter 7 General discussion and outlook

Seed quality attributes including protein content and composition are among the most important traits for the utilization of *Vicia faba* (*Vf)*. Yet, very little effort has been made to understand the genetic control of these traits. In addition, the available literature on some basic aspects of seed proteins such as their identities and physico-chemical properties is relatively old and incomplete. Therefore, this thesis was aimed to contribute towards genetic improvement of *Vf* quality traits by conducting a systematic investigation into protein subunit composition and diversity, the genetic loci underlying crude protein content and protein subunit abundance, and accumulation dynamics of different proteins during seed development.

Past efforts in *Vf* seed proteins have mainly focused on identifying structural genes coding for major classes of legumin (Horstmann *et al.*, 1993; De Pace *et al.*, 1991; Schlesier *et al.*, 1990; Baumlein *et al.*, 1986) and vicilins (Fuchs *et al.*, 1994; Jiri *et al.*, 1993; Weschke *et al.*, 1988). However, in the context of genetic improvement of seed protein composition, the complex genetic coding and the resulting structural heterogeneity within the major storage proteins (Müntz *et al.*, 1999; Tucci *et al.*, 1991) pose a major methodological challenge to unambiguously identify and quantify certain seed proteins for phenotypic screening of large number of germplasm required in genetic mapping studies. Therefore, as a first step of this thesis, 25 protein bands were accurately identified on 1D SDS-PAGE using three *Vf* genotypes with contrasting protein pattens. Despite its limited separation power for highly complex protein samples like seed protein extract, one-dimensional sodium dodecyl sulphate–polyacrylamide gel electrophoresis (1D SDS-PAGE) has been widely used in genetic analysis of protein composition in various crops (Panthee *et al.*, 2004; Tzitzikas *et al.*, 2006; Le Signor *et al.*, 2017; Boehm *et al.*, 2017; Schatzki *et al.*, 2014; Kerfal *et al.*, 2010) and, therefore, this well-annotated 1D SDS-

PAGE, which captures most protein bands occurring in our panel of 35 diverse inbred genotypes, is expected be a useful reference for seed composition analysis within the *Vf* research community. Regarding the genetic diversity in *Vf* seed protein composition among genotypes, SDS-PAGE analysis highlighted two genotypes which carried rare variants of major legumin subunits–LG Cartouche and NV657 (INRA 29H), which are ideal genetic materials to uncover the genetic basis for such protein subunit variation. In soybean, genotypes lacking some or all subunits of certain storage protein have been developed and exploited in understanding the relationship between protein physicochemical properties and subunit composition (Poysa *et al.*, 2006) or mapping loci associated with protein subunit variants (Boehm *et al.*, 2017).

Accurate quantification of the different protein constituents was another methodological challenge on top of the identification issue; to tackle this aspect, the potential of SE-HPLC in determining the proportions of legumin, vicilin and convicilin was explored. The analysis of the SE-HPLC fractions on 1D SDS-PAGE showed two major peaks belonging to legumin and vicilin/convicilin aggregates that could be quantified. Although the SE-HPLC method in this study could be used reliably to quantify the proportions of legumin and vicilin/convicilin in seed protein mixtures, its suitability of for genetic analysis of protein composition may be limited by the inability to adequately separate vicilin and convicilin and also other abundant seed proteins including lipoxygenases, heat shock proteins, sucrose binding proteins and albumin. Thus, to realize the full potential of this method, further optimization of the mobile phase and column length and pore size is required.

To investigate the underlying genomic loci of protein content and composition, inbred lines from a multi-parent population developed by cross-pollination with bees, and a high-density SNP array were used for genome-wide association mapping of QTL linked to crude seed protein content and the abundance of seed protein subunits. These traits showed significant genetic variation among genotypes and a clear environmental effect. For the total protein content, three

QTL on chromosomes 1, 4, and 6, which individually explained 5.6%, 3% and 11%, respectively, of the phenotypic variation, were identified. Protein content is known to be a genetically complex quantitative trait involving many loci, as reported in other legumes like soybean (Li *et al.*, 2018; Kim *et al.*, 2016; Sonah *et al.*, 2015), and therefore, these QTL could be the larger effect ones while others have been masked by some confounding factors including a large environmental effect, relatively small population size and also the part-inbred status of the population. These results are the first report of *Vf* protein content QTL and further work is needed to confirm the identified loci and to uncover more QTL for this trait. As for protein subunit composition, GWAS analysis detected 59 significant marker-trait associations for 18 protein composition traits. These included loci associated with the abundance of major storage proteins like legumins, vicilins, convicilins. Synteny between *Vf* and *M. truncatula* which allowed examination of predicted gene content of regions surrounding MTAs, revealed some candidate structural and regulatory genes within the significant QTL regions. Generally, protein bands with extreme phenotypes such as presence or absence pattern are expected to be under simple Mendelian genetic control (Boehm *et al.*, 2017; Tucci *et al.*, 1991), while the relative abundance of quantitatively inherited subunits is not only affected by genotype and environment but also modulated by competition from other subunits as reflected by the significant negative correlation between several pairs of major protein bands. Such negative correlation can be potentially useful in selecting for increased abundance of certain proteins like legumins for their higher content of sulphur-containing amino acids (S-AA), as higher legumin content should result in a lower content of the S-AA poor vicilins.

From a methodological point of view, the fact that independently quantified bands belonging to the same protein class, e.g. convicilin, had a common genetic locus was an indication of a good accuracy and reproducibility of SDS-PAGE based protein quantification. However, this was achieved at the expense of the number of the quantified protein bands where, due to the trade-off between capturing all protein bands on the same gel and the resolution of

protein separation, protein bands of molecular weights less than ~30 kDa were allowed run out of the gels. Yet, the advantage of SDS-PAGE is that it provides a direct visual observation of the genetic variation in protein composition rather than sole dependence on spectral intensity in the case of liquid chromatography-based quantification. For this reason, it will probably remain as a key tool for protein subunit composition analysis, but its resolution and separation power of the complex *Vf* seed proteins needs to be enhanced. This may include testing different combinations of polyacrylamide concentrations and sodium dodecyl-sulphate concentration in the sample and running buffers.

Seed development phase can be considered the most important growth stage in a plant's lifecycle as it is the time during which nutritional compounds like proteins and carbohydrates are synthesized and accumulated. Several studies in legumes have reported temporal differences in the expression of seed proteins or seed protein genes during seed development (Gallardo *et al.*, 2003; Verdier *et al.*, 2008; Kreplak *et al.*, 2019; De Pace *et al.*, 1991). Although there are many ways by which environment can modulate protein composition, one main hypothesis is that the temporal differences in the accumulation could be a key deriver of protein composition variation across environments depending on the genotypic performance and prevailing environmental conditions at the seed-filling stages. Thus, to establish the accumulation patterns of the major *Vf* seed proteins, the proteomic profile of 12 seed developmental stages of the reference *Vf* inbred line Hedin/2 was assessed by liquid chromatography mass spectroscopy (LC-MS)**.** As expected, seed developmental stages were characterized by marked morphological and proteomic changes. The relative abundance of 340 proteins was quantified including 17 globulins which showed a diverse accumulation patterns. This diversity in the timing of accumulation potentially helps to understand the interplay between protein composition and environmental conditions by revealing which time periods during grain fill are most critical for a particular protein class. On the other hand, though it could be of limited applications in large scale screening of study materials due to cost and time considerations, the LC-MS method has

been used for the absolute quantification of different soybean proteins like β-Conglycinin subunits (Ippoushi *et al.*, 2019) and, by virtue of its combination of specificity and sensitivity, this method could in the future be used directly to map genetic factors controlling an even greater repertoire of specific proteins.

Finally, from the end-user point of view, seed quality is a multi-facetted topic which also encompasses the sensory quality attributes. Therefore, the aim of Chapter 6 was to fine-map the seed pale hilum colour which is a desirable trait in *Vf* traded for direct human consumption in the Middle East. A loci on chromosome 1 containing dihydroflavonol 4-reductase-like (DFR) was identified as the likely candidate gene harbouring the pale hilum colour mutation in *Vf*. However, cloning the candidate gene and pinpointing the mutation leading to pale hilum colour was hindered by the presence of tandem gene copies in the candidate loci. It is hoped that the upcoming *Vf* genome sequence will help in resolving the gene content of this locus. In general, it can be speculated that hilum colour in *Vf* could be genetically more complex than previously thought. This is supported by the detection of a second hilum colour loci by GWAS analysis in Chapter 4 and the observation of an intermediate (not black neither pale) hilum colour in some genotypes. Also, hilum colour appears to be genetically related to seed coat colour where seeds with coloured coats tend to have coloured hila. In soybean, hilum and seed coat colour are controlled by the epistatic interaction between four independent loci which also affect the colours of pubescence and flowers (Gillman *et al.*, 2011).

Overall, the systematic approach of this thesis has provided not only new information on *Vf* seed quality attributes but also a toolkit to further expand these findings and explore more traits. The seed storage proteins and their structural genes reported here could be exploited in annotating *Vf* genome and, depending on the advances in *Vf* regeneration techniques, introducing more sulphur-containing amino acids through gene editing technologies like CRISPR. On the other hand, multi-parent mapping populations are powerful tools in genetic mapping of traits due

to their higher genetic diversity and minimal population structure (Scott *et al.*, 2020) which makes the population developed in this study good genetic material for studying other nutritional quality traits such as the content of undesirable proteins like lipoxygenase and trypsin inhibitors. This population can also be used to map agronomic traits including yield components, drought tolerance and disease resistance in *Vf*.

References

Baumlein, H., Wobus, U., Pustell, J. & Kafatos, F. C. (1986). The legumin gene family: structure of a B type gene of *Vicia faba* and a possible legumin gene specific regulatory element. *Nucleic Acids Research,* **14** (6), 2707-2720.

Boehm, J. D., Nguyen, V., Tashiro, R. M., Anderson, D., Shi, C., Wu, X., Woodrow, L., Yu, K., Cui, Y. & Li, Z. (2018). Genetic mapping and validation of the loci controlling 7S α′ and 11S A-type storage protein subunits in soybean [*Glycine max* (L.) Merr.]. *Theoretical and Applied Genetics*, **131**, 659-671.

Cernay, C., Pelzer, E. & Makowski, D. (2016). A global experimental dataset for assessing grain legume production. *Scientific Data,* **3**, 160084.

De Pace, C., Delre, V., Mugnozza, G. T. S., Maggini, E., Cremonini, R., Frediani, M. & Cionini, P. G. (1991). Legumin of *Vicia faba* major: accumulation in developing cotyledons, purification, mRNA characterization and chromosomal location of coding genes. *Theoretical and Applied Genetics,* **83**, 17-23.

Fuchs, J., Joos, S., Licheter, P. & Schubert, I. (1994). Localization of vicilin genes on field bean chromosome II by fluorescent in situ hybridization. *Journal of Heredity,* **85** (6), 487-488.

Gallardo, K., Le Signor, C., Vandekerckhove, J., Thompson, R. D. & Burstin, J. (2003). Proteomics of *Medicago truncatula* seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. *Plant Physiology,* **133** (2), 664-82.

Gillman, J. D., Tetlow, A., Lee, J.-D., Shannon, J. G. & Bilyeu, K. (2011). Loss-of-function mutations affecting a specific Glycine max R2R3 MYB transcription factor result in brown hilum and brown seed coats. *BMC Plant Biology,* **11**, 155-155.

Horstmann, C., Schlesier, B., Otto, A., Kostka, S. & Muntz, K. (1993). Polymorphism of legumin subunits from field bean (*Vicia faba* L. var. minor) and its relation to the corresponding multigene family. *Theoretical and Applied Genetics,* **86** (7), 867-874.

Ippoushi, K., Wakagi, M., Hashimoto, N. & Takano-Ishikawa, Y. (2019). Absolute quantification of the α, α', and β subunits of β-conglycinin from soybeans by liquid chromatography/tandem mass spectrometry using stable isotope-labelled peptides. *Food Research International,* **116**, 1223-1228.

Jiri, M., Winfriede, W., Helmut, B., Uta, P., Andreas, H., Ulrich, W. & Ingo, S. (1993). Localization of vicilin genes via polymerase chain reaction on microisolated field bean chromosomes. *The Plant Journal,* **3** (6), 883-886.

Kerfal, S., Giraldo, P., Rodríguez-Quijano, M., Vázquez, J. F., Adams, K., Lukow, O. M., Röder, M. S., Somers, D. J. & Carrillo, J. M. (2010). Mapping quantitative trait loci (QTLs) associated with dough quality in a soft×hard bread wheat progeny. *Journal of Cereal Science,* **52** (1), 46-52.

Kim, M., Schultz, S., Nelson, R. L. & Diers, B. W. (2016). Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. *Crop Science,* **56** (1), 219-225.

Kreplak, J., Madoui, M.-A., Cápal, P., Novák, P., Labadie, K., Aubert, G., Bayer, P. E., Gali, K. K., Syme, R. A., Main, D., Klein, A., Bérard, A., Vrbová, I., Fournier, C., D'Agata, L., Belser, C., Berrabah, W., Toegelová, H., Milec, Z., Vrána, J., Lee, H., Kougbeadjo, A., Térézol, M., Huneau, C., Turo, C. J., Mohellibi, N., Neumann, P., Falque, M., Gallardo, K., McGee, R., Tar'An, B., Bendahmane, A., Aury, J.-M., Batley, J., Le Paslier, M.-C., Ellis, N., Warkentin, T. D., Coyne, C. J., Salse, J., Edwards, D., Lichtenzveig, J., Macas, J., Doležel, J., Wincker, P. & Burstin, J. (2019). A reference genome for pea provides insight into legume genome evolution. *Nature Genetics,* **51** (9), 1411-1422.

Le Signor, C., Aimé, D., Bordat, A., Belghazi, M., Labas, V., Gouzy, J., Young, N. D., Prosperi, J.-M., Leprince, O., Thompson, R. D., Buitink, J., Burstin, J. & Gallardo, K. (2017).

Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. *New Phytologist,* **214** (4), 1597-1613.

Li, Y.-h., Reif, J. C., Hong, H.-l., Li, H.-h., Liu, Z.-x., Ma, Y.-s., Li, J., Tian, Y., Li, Y.-f., Li, W.-b. & Qiu, L.-j. (2018). Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. *Plant Science,* **266** (Supplement C), 95-101.

Müntz, K., Horstmann, C. & Schlesier, B. (1999). Vicia globulins. *In:* Shewry, P. R. & Casey, R. (eds.) *Seed Proteins.* Dordrecht: Springer Netherlands, pp. 259-284.

Panthee, D. R., Kwanyuen, P., Sams, C. E., West, D. R., Saxton, A. M. & Pantalone, V. R. (2004). Quantitative trait loci for β-conglycinin (7S) and glycinin (11S) fractions of soybean storage protein. *Journal of the American Oil Chemists' Society,* **81** (11), 1005-1012.

Poysa, V., Woodrow, L. & Yu, K. (2006). Effect of soy protein subunit composition on tofu quality. *Food Research International,* **39** (3), 309-317.

Schatzki, J., Ecke, W., Becker, H. C. & Möllers, C. (2014). Mapping of QTL for the seed storage proteins cruciferin and napin in a winter oilseed rape doubled haploid population and their inheritance in relation to other seed traits. *Theoretical and Applied Genetics,* **127** (5), 1213-1222.

Schlesier, B., Bassüner, R., Van Hai, N. & Müntz, K. (1990). The cDNA derived primary structure of two distinct legumin A subunit precursors from field bean (*Vicia faba* L.). *Nucleic Acids Research,* **18** (23), 7146-7146.

Scott, M. F., Ladejobi, O., Amer, S., Bentley, A. R., Biernaskie, J., Boden, S. A., Clark, M., Dell'Acqua, M., Dixon, L. E., Filippi, C. V., Fradgley, N., Gardner, K. A., Mackay, I. J., O'Sullivan, D., Percival-Alwyn, L., Roorkiwal, M., Singh, R. K., Thudi, M., Varshney, R. K., Venturini, L., Whan, A., Cockram, J. & Mott, R. (2020). Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity,* **125** (6), 396-416.

Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I. & Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnology Journal,* **13** (2), 211-221.

Tucci, M., Capparelli, R., Costa, A. & Rao, R. (1991). Molecular heterogeneity and genetics of Vicia faba seed storage proteins. *Theoretical and Applied Genetics,* **81** (1), 50-58.

Tzitzikas, E. N., Vincken, J.-P., de Groot, J., Gruppen, H. & Visser, R. G. F. (2006). Genetic variation in pea seed globulin composition. *Journal of Agricultural and Food Chemistry,* **54** (2), 425-433.

Verdier, J., Kakar, K., Gallardo, K., Le Signor, C., Aubert, G., Schlereth, A., Town, C. D., Udvardi, M. K. & Thompson, R. D. (2008). Gene expression profiling of *M. truncatula* transcription factors identifies putative regulators of grain legume seed filling. *Plant Molecular Biology,* **67** (6), 567-580.

Weschke, W., Bassüner, R., Van Hai, N., Czihal, A., Baümlein, H. & Wobus, U. (1988). The structure of a *Vicia faba* vicilin gene. *Biochemie und Physiologie der Pflanzen,* **183** (2-3), 233-242.

WHO/FAO/UNU (2007). Protein and amino acid requirements in human nutrition. *WHO Technical Report Series,* (935), 1-265.