Grain Proteomics of Sorghum: Improving Digestibility, Nutritional Quality and Starch Conversion to Ethanol

Julia Erin Cremer
BSc (Hon)

A thesis submitted for the degree of Doctor of Philosophy at
The University of Queensland in March, 2014
School of Agricultural and Food Sciences
Abstract

Sorghum is an important food staple in rural Asia, India and Africa and a major feedstock in developed countries. Originating in the semi-arid African savannah, sorghum is well adapted to drought prone environments and produces more biomass per unit water than maize. Extensive variability exists within the sorghum gene pool creating a diversity of plant forms with many commercial and industrial applications. Sorghum, provides an alternative to grain crops with greater irrigation and fertiliser requirements, such as maize, however, the grain is generally less digestible, especially upon cooking, due to extensive disulphide cross-linking among sulphur-rich storage proteins in the protein-starch matrix. This reduces the accessibility of proteolytic enzymes to starch, decreasing palatability and nutritional value, and resulting in the need for increased processing.

The commercial value of cereals is largely determined by the chemistry of the protein-starch matrix. Located on the periphery of storage protein bodies, cysteine-rich β- and γ-kafirins prevent enzymatic access to internally positioned α-kafirins and to starch. The development of mutants is an efficient approach for studying genetic regulation of protein biosynthesis and compartmentalisation and shows how changes in protein profile affect the physical characteristics of the seed. An integrated proteomic approach was employed to characterise 28 sorghum lines with allelic variation at the kafirin loci, including three β-kafirin null mutants, in order to determine the effects of kafirin genetic diversity on the expression of storage proteins.

High performance liquid chromatography (HPLC) and capillary electrophoresis were employed to profile water/salt- and alcohol-soluble protein fractions, revealing a wide range of diversity in protein content across the genotypes. Peak profiles were similar among β-kafirin null lines but different to normal lines, with a significant reduction in alcohol-soluble protein and the disappearance of a major set of protein peaks. The relative content of sequentially extracted soluble, insoluble and residual proteins was measured using RP-HPLC to evaluate the degree of cross-linking among storage proteins in the seed. High levels of insoluble and residual protein were associated with reduced digestibility.

Gel-based separation of seed proteins and subsequent identification with tandem mass spectrometry identified a range of redox-active proteins affecting storage protein biochemistry. Multiple protein fractions were analysed across the β-kafirin null line QL12 and wild-type line 296B. The redox states of endosperm proteins, of which kafirins are a subset, have been shown to impact on quality traits in addition to the expression of proteins. Thioredoxin, active in the processing of storage
proteins at germination, has reported impacts on grain digestibility and was differentially expressed across the genotypes. A range of putative uncharacterised sorghum proteins regulating the folding and disulphide pairing of storage proteins in homologous crop species were detected, as well as a diversity of lysine-rich albumins and globulins, and enzymes involved in starch hydrolysis, such as \( \alpha \)-amylase inhibitor. In addition, a high molecular weight \( \gamma \)-prolamin homolog was identified in sorghum at the sequence level for the first time.

Biochemical factors affecting grain quality were measured across the sample set, including crude protein content, raw and cooked protein digestibility, total starch and flour moisture content. These parameters were evaluated for correlation to the expression of specific classes of proteins, such as the kafirins. Alcohol-soluble RP-HPLC peak distribution profiles showed correlation between a specific peak, likely a \( \gamma \)-kafirin, and digestibility. To test the impact of kafirins on the efficiency of ethanol production from sorghum, a subset of ten diverse genetic lines from the sample population, including three \( \beta \)-kafirin null mutants, were evaluated for ethanol yield and fermentation efficiency. Starch content was positively correlated to total ethanol yield \((R^2=0.74)\), and there was a slight positive correlation between protein digestibility and ethanol yield \((R^2=0.52)\). Increases in free amino nitrogen (FAN) content significantly enhanced fermentation efficiency \((R^2=0.65)\). Multivariate analysis indicated an association between the \( \beta \)-kafirin allele and variation in grain digestibility \((P=0.042)\) and FAN \((P=0.036)\), with subsequent effects on ethanol yield.

Sorghum improvement strategies focussed on enhancing digestibility, nutritional value, and fermentation efficiency will increase the usefulness of the crop as a food, feed and biofuel. The most common technique for enhancing cereal protein quality is seed-specific genetic modification of protein biosynthesis for the introduction of genes encoding proteins rich in deficient amino acids or with favourable biochemical characteristics. To achieve this, a comprehensive understanding of storage protein gene function and expression is required for the stable introduction of foreign genes and the manipulation of protein content and structure in the seed. This work contributes to our understanding of sorghum grain protein composition and how it is impacted by diversity in kafirin genetic background. The acquisition of sequence-based information pertaining to the expression of proteins involved in redox activity in the seed, the accumulation of lysine-rich storage proteins and the activity of starch metabolic enzymes, further augments annotation of the sorghum proteome and lends insight to future improvement strategies.
Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the Copyright Act 1968.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.
Publications during candidature


Publications included in this thesis


<table>
<thead>
<tr>
<th>Name</th>
<th>Contributions</th>
</tr>
</thead>
</table>
| Julia Cremer       | Designed experiments (20%)  
Fermentation efficiency, total starch, RP-HPLC, Lab on Chip (50%)  
Data analysis (60%)  
Wrote the paper (90%) |
| Liman Liu          | Fermentation efficiency experiments (70%) |
| Dr Scott Bean      | Designed Experiments (70%)  
Data analysis and edited paper (30%) |
| Dr Jae B Ohm       | Statistical analysis (10%) |
| Dr Mike Tilley     | Lab on Chip analysis (30%) |
| Dr Jeff Wilson     | Total starch measurements (10%) |
| Rhett Kaufman      | Total starch measurements (40%) |
| T Vu               | Amylose measurements (80%) |
| Dr Edward Gilding  | Edited paper (10%) |
| Prof Ian Godwin    | Designed Experiments and edited paper (20%) |
| Prof Donghai Wang  | Fermentation efficiency experiments (20%)  
Edited paper (20%) |

<table>
<thead>
<tr>
<th>Julia Cremer</th>
<th>Designed experiments (50%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RP-HPLC, Lab on Chip, SDS-PAGE, LC-MS/MS, total starch (70%)</td>
</tr>
<tr>
<td></td>
<td>Data analysis (60%)</td>
</tr>
<tr>
<td></td>
<td>Wrote the paper (80%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dr Scott Bean</th>
<th>Designed Experiments (40%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data analysis and edited paper (30%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dr Jae B Ohm</th>
<th>Statistical analysis (10%)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Brian Ioerger</th>
<th>RP-HPLC (20%)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Dr Mike Tilley</th>
<th>Lab on Chip (50%)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Dr Jeff Wilson</th>
<th>Total starch measurements (20%)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Rhett Kaufman</th>
<th>Total starch measurements (40%)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Dr Edward Gilding</th>
<th>Designed Experiments and edited paper (10%)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Prof Ian Godwin</th>
<th>Designed Experiments and edited paper (40%)</th>
</tr>
</thead>
</table>

**Contributions by others to the thesis**

Cooked digestibility data, figure 2.2 (Bean, S.R.)
Data for Soluble/Insoluble/Residual protein cross-linking study, figure 2.4 (Bean, S.R.)

**Statement of parts of the thesis submitted to qualify for the award of another degree**

None
Acknowledgements

I would like to thank all the people who have helped me over the course of my PhD research and thesis preparation. Thank you to my supervisors Prof Ian Godwin and Dr Edward Gilding for your guidance and assistance throughout the project. Thanks also to Dr Amanda Nouwens at the SCMB proteomics facility for generously giving your time to instruct me in 2D SDS-PAGE and mass spectrometry.

I appreciate the support this project has received from the University of Queensland and through funding from an ARC linkage grant and by Pacific Seeds. I am also grateful for the provision of the Graduate School International Travel Award (GSITA), which provided me with the opportunity to undertake collaborative work in the United States as a component of my PhD.

I extend my sincere gratitude to Dr Scott Bean and his wife Jill, and all the members of the Bean lab at the USDA in Manhattan, Kansas, for enthusiastically supporting my research during visits to the US, and for a very enjoyable travel experience. Thank you for the many lovely meals and conversation and for sharing your tornado shelter with me!

Thank you also to the entire staff at ARS-USDA for providing a warm and welcoming work environment. I am especially thankful to Dr Mike Tilley for teaching me Lab on Chip and for maintaining a great sense of humour even when the chips were down! Thank you to Brian Ioerger and Rhett Kaufman for sharing your extensive technical expertise and for helping me so much in the lab. Thank you also to Peggy Labadie for providing me with very comfortable accommodation and delicious home cooking during my visit, and for trusting me to drive the car on the ‘right’ side of the road.

Thanks to all my colleagues and friends at UQ/SAFS who made my PhD a much more enjoyable experience, especially Louise Shuey, Alice Hayward, Margaret Jewell, Satomi Hayashi, Annaliese Mason, Fiona Filardo, Brett Ferguson and Kerry Condon. Thank you also to Sharon Williams and Kerry Vinall for your support in the lab.

Last but not least, I would like to express my deepest thank you to my family, especially to Dad and Will, for their constant encouragement, understanding, and support over the course of my PhD.
Keywords

Proteomics, seed storage proteins, sorghum, kafirin, mass spectrometry, HPLC, digestibility, ethanol conversion, bio-ethanol, grain quality

Australian and New Zealand Standard Research Classifications (ANZSRC)

060109 Proteomics and Intermolecular Interactions (50%)
070303 Crop and Pasture Biochemistry and Physiology (40%)
070304 Crop and Pasture Biomass and Bioproducts (10%)

Fields of Research (FoR) Classification

0601 Biochemistry and cell biology (50%)
0703 Crop and Pasture Production (50%)
Table of Contents

Abstract ................................................................................................................................. iv
Acknowledgements .............................................................................................................. ix
List of Figures and Tables .................................................................................................. xv
Abbreviations .................................................................................................................. xviii
Chapter 1 ............................................................................................................................ 1

Literature Review ............................................................................................................... 1
  1.1 Introduction .................................................................................................................. 2
  1.2 The diversity and evolution of cereals........................................................................ 2
  1.3 Sorghum production and uses..................................................................................... 3
    1.3.1 Sorghum as a food and animal fodder ................................................................. 4
    1.3.2 Bioethanol........................................................................................................... 4
    1.3.3 Nutrient use efficiency ....................................................................................... 5
  1.4 Sorghum improvement................................................................................................. 5
  1.5 Grain structure and composition .............................................................................. 6
    1.5.1 Grain quality ....................................................................................................... 8
  1.6 Seed storage proteins ................................................................................................. 8
    1.6.1 Protein solubility ............................................................................................... 9
    1.6.2 The protein-starch matrix ................................................................................ 9
    1.6.3 Regulation of protein body structure ............................................................... 13
  1.7 Seed proteomics ......................................................................................................... 13
    1.7.1 SDS PAGE and Tandem Mass Spectrometry ..................................................... 14
    1.7.2 High Performance Liquid chromatography (HPLC) ........................................ 16
    1.7.3 High Performance Capillary Electrophoresis (HPCE) ...................................... 17
  1.8 Grain biochemical analysis ......................................................................................... 18
    1.8.1 Protein digestibility .......................................................................................... 18
    1.8.2 Total starch ....................................................................................................... 18
    1.8.3 Flour moisture content .................................................................................... 19
  1.9 Project Objectives ..................................................................................................... 19

Chapter 2 ........................................................................................................................... 21

Characterisation of sorghum seed proteins through HPLC, capillary electrophoresis and biochemical analysis ................................................................. 21
  2.1 Introduction ................................................................................................................ 22
    2.1.1 Seed storage proteins ..................................................................................... 22
    2.1.2 Impact of protein composition on quality traits .............................................. 22
List of Figures and Tables

Chapter 1:

Figure 1.1  World sorghum production in thousand metric tonnes (TMT). ................................................. 4
Figure 1.2  Sorghum caryopsis showing pericarp, hilium, seed-coat or testa, endosperm and germ. Electron micrograph shows a microscopic view of the protein-starch matrix, where starch granules are embedded in a network of protein bodies ......................... 7
Figure 1.3  Structure of the protein-starch matrix in the endosperm ......................................................... 10
Figure 1.4  Mascot summary for a peptide matches to the β-kafirin protein identified from 2D SDS-PAGE gel spot analysed with LC-MS/MS .......................................................... 15
Figure 1.5  Separation of proteins according to size by size exclusion chromatography (SEC), using columns containing porous particle filters ........................................ 17

Chapter 2:

Table 2.1  Background and origin of sample grain lines ................................................................. 27
Table 2.2  Percent protein, digestibility and total starch content for sample population of sorghum lines with allelic variation in kafirin storage proteins ............................................... 46
Table 2.3  Allelic variant key ........................................................................................................ 47

Figure 2.1  Protein extraction procedure utilised to extract albumin/globulins (AGs) and kafirins (prolamins) ...................................................................................................................... 28
Figure 2.2  RP-HPLC separation of water/salt-soluble (A/G) protein fraction for 28 sorghum grain lines with allelic variation in the kafirins ............................................................................ 32
Figure 2.3  RP-HPLC profiles across β-kafirin null allelic variants QL12, IS17214 and RTx2737 and normal lines .................................................................................................................. 34/35
Figure 2.4  Sequential extraction of soluble, insoluble and residual protein from kafirin allelic variants extracted under non-reducing and reducing conditions ........................................ 36
Figure 2.5  Negative association between soluble (SP) and residual (RP) protein content across sorghum allelic variants evaluated with SEC ................................................................. 37
Figure 2.6  RP-HPLC profiles for alcohol-soluble proteins across a selection of kafirin allelic variants. Tannin-containing lines IS8525 and IS12572C showed binding and precipitation of the γ-kafirin peak, absent at the 4min elution time point .......... 39
Figure 2.7  SEC analysis of the A/G fraction in β-kafirin null mutant QL12 compared to wild-type 296B................................................................................................................................. 40
Figure 2.8  LOC gel image of the kafirin fraction showing allelic variation in β-kafirin across normal and null mutant lines........................................................................................................... 42
Figure 2.9  β-kafirin content of null allelic variant QL12 compared to normal genotypes 296B and KS115................................................................................................................................. 43
Figure 2.10: LOC gel image of sequentially extracted kafirins from QL12 and 296B in 10, 20, 30 and 40% tertiary butanol...................................................................................................................... 45
Figure 2.11  LOC electropherograms of sequentially extracted kafirins from 296B and QL12 in 20 and 30% tert-butanol...................................................................................................................... 45
Figure 2.12  Crude protein content across the kafirin allelic variants...................................................................................................................... 48
Figure 2.13  Dry matter and cooked digestibility of sorghum whole meal flour across sample population...................................................................................................................... 48
Figure 2.14  Feed trial evaluating protein digestibility in standard sorghum feed in the pig gut.. 50
Figure 2.15  Total starch measured with a Megazyme assay across kafirin allelic variants............ 51
Figure 2.16  Absence of correlation between total starch and protein (%).................................. 52

Chapter 3:

Table 3.1  Protein Candidates Identified in Sorghum Grain with Reported Effects on Protein-Starch Matrix Structure, Grain Quality and Stress Responses......................................................... 74

Figure 3.1  RP-HPLC profiles for the A) alcohol-soluble (prolamin) fraction and B) water/salt soluble (albumin/globulin) fraction across β-kafirin null allelic variants QL12, 1517214 and RTx2737, and lines with normal β-kafirin content (B35, 296B and KS115)............................................................................................................................................ 65
Figure 3.2  RP-HPLC analysis shows a significant negative correlation between seed protein digestibility (%) and the alcohol-soluble kafirin fraction peak eluting at 5.5-6 min on the chromatogram........................................................................................................................................... 66
Figure 3.3  Alcohol-soluble protein fraction visualised on Lab on Chip across β-kafirin null lines QL12, 1517214 and RTx2737, and normal β-kafirin lines 296B and M35.............. 67
Figure 3.4  One-dimensional SDS-PAGE separation of water/salt soluble albumin/globulins, alcohol-soluble prolamins, and alkali-soluble glutelins from normal β-kafirin allelic variant 296B, and β-kafirin null QL12..................................................................................................... 69
Figure 3.5  Two-dimensional SDS-PAGE separation of water/salt soluble albumin/globulins.. 70
Figure 3.6  Phylogenetic neighbour joining tree view of the alignment of protein sequences representing prolamin subclasses across grass species using the Grishin protein method. Putative uncharacterised sorghum γ-prolamin homolog groups with 50kD γ-zeins and γ-caneins of maize and sugarcane, separately from smaller Mw γ-prolamins................................................................. 72

Figure 3.7  ClustalW alignment of peptidyl-prolyl cis/trans isomerase (PPIase) peptide sequences across major grass species, illustrating a high level of sequence homology....................................................................................... 77

Chapter 4:

Table 4.1  Sorghum grain lines characterised for allelic variance in the kafirin storage proteins and tested for ethanol production efficiency................................................................. 86

Table 4.2  Kafirin allelic profiles, biochemical measurements and ethanol fermentation characteristics for selection of sorghum grain cultivars, with mean values and standard deviation calculated across replicate sample measurements. Means that do not share a letter are significantly different................................................................. 90

Figure 4.1  Ethanol yield (L/tonne) produced by ten sorghum lines with allelic variation in the kafirin storage proteins.......................................................................................................... 91

Figure 4.2  Relationship between starch content, fermentation efficiency and ethanol yield across ten sorghum genotypes..................................................................................... 92

Figure 4.3  Associations between protein digestibility (%), fermentation efficiency at 72 hours (%), and ethanol yield (L/ton) across ten sorghum genotypes........................................................ 93

Figure 4.4  Fermentation efficiency (%) versus fermentation hour. KS115 is an outlier which exhibits high efficiency in the early stages of the fermentation process relative to other genotypes....................................................................................... 94

Figure 4.5  PCA of parameters affecting fermentation efficiency and ethanol yield................ 95

Figure 4.6  RP-HPLC peak distributions for the alcohol-soluble kafirin protein fraction across ten sorghum genotypes evaluated for ethanol production efficiency............................... 97

Figure 4.7  Lab on Chip size-based separation of alcohol-soluble kafirin fraction in sorghum lines evaluated for ethanol fermentation........................................................................................ 99
Abbreviations

ABC  Ammonium bicarbonate
ACN  Acetonitrile
A/G  Albumin/Globulins
βME  Beta-mercaptoethanol
CHAPS 3-[3-cholamidopropyl]dimethylammonio]-1-propanesulfonate
4VP  4-vinylpyridine
FU   Fluorescent units
KCl  Potassium chloride
kD   Kilodalton
Da   Dalton
DTT  Dithiothreitol
EDTA Ethylenediaminetetraacetic acid
HD   High Digestibility
HPLC High performance liquid chromatography
IAA  Iodoacetamide
IEF  Isoelectric focussing
LC-MS/MS Liquid chromatography-tandem mass spectrometry
LOC  Lab on Chip
MES  2-(N-morpholine)ethanesulfonic acid (buffer)
MeOH Methanol
Mw   Molecular weight
PDI  Protein disulphide isomerase
PPIase Peptidyl-prolyl cis/trans isomerase
QPM  Quality Protein Maize
SDS-PAGE Sodium dodecyl sulphate-polyacrylamide gel electrophoresis
SEC  Size exclusion chromatography
SP/IP/RP Soluble/Insoluble/Residual proteins
TFA  Trifluoroacetic acid
Tris-HCl Tris(hydroxymethyl)aminomethane hydrochloric acid (buffer)
Trx   Thioredoxin
v/v  Volume per volume
w/v  Weight per volume
Chapter 1

Literature Review
1.1 Introduction

Cereals play a critical role in our food chain. Reliance on grain staples with limited amino availability as a main food source can lead to widespread nutritional deficiencies, particularly in developing regions affected by water shortages and/or limited access to alternative food sources. Sorghum, a drought tolerant crop with high photosynthetic efficiency, is suited to cultivation in semi-arid climates and on marginal lands, potentially impacting on food security in impoverished areas. However, the availability of lysine and threonine is limited in sorghum, which contains a highly cross-linked network of proline-rich storage proteins surrounding starch in the grain.

The structure and composition of storage proteins in the grain determines quality traits, such as digestibility and nutritional value. The large-scale study of proteins, or proteomics, provides an integral tool for identifying variation in protein composition across diverse collections of germplasm. The functional characterisation of storage proteins, (ie. linking their proteomic profiles to phenotypic traits), is instrumental to developing molecular markers for enhanced grain quality. The small diploid genome of sorghum makes it an attractive model for functional genomics of the C4 grasses. Identification of elements regulating the synthesis, targeting and degradation of storage proteins will facilitate the manipulation of grain characteristics, such as amino acid content and digestibility (1), in sorghum and other related grain crops. The integration of proteomic analyses with genomic and metabolomic studies carried out in parallel shows how storage protein biosynthesis is regulated at multiple levels throughout grain development (2).

1.2 The diversity and evolution of cereals

Cereals represent a large percentage of our daily food intake and form a relatively new branch of the plant phylogenetic tree. While the angiosperms (flowering plants) are thought to be around 200 million years old, there is evidence that grasses such as maize (Zea mays), rice (Oryza sativa), sorghum (Sorghum bicolor), and wheat (Triticum aestivum) may have diverged from a common ancestor as recently as 50–70 million years ago (3). Domestication of the cereals has occurred in an evidently independent manner, where parallel selection associated with domestication and improvement has transpired in geographically diverse areas (4). Genetic studies indicate that sorghum diverged from Zea mays progenitors around 12 million years ago (5). Early analysis of duplicated maize genes and RFLP mapping across the cereals demonstrated that maize has
undergone a whole genome duplication event, where two of its chromosome sets align with a chromosome from both rice and sorghum (6).

1.3 Sorghum production and uses

Sorghum is a staple food source for millions in rural Africa and Asia and is employed mainly as a feedstock and biofuel in the West. Historically, it is thought that nomadic African communities originally adopted hardy sorghum cultivars suited to the hot dry conditions of the savannah to supplement the food supply provided by herding livestock (7). When semi-nomadic pastoralists mixed with agriculturalists from the river delta regions of Ethiopia and the Nile, robust sorghum varieties were integrated into established agronomic systems, leading to the cultivation of highly productive sorghums. Evidence suggests that the wild sorghum arundinaceums and the guinea sorghums (guineesia) derived from this race evolved across the southern edge of the Sahara. Guineas appear to be the earliest grain producing sorghums and spread throughout Africa and later to India via overland routes or by sea, often as ships provisions (7).

Approximately half of the world production of sorghum is fed to livestock. Demand for sorghum as animal feed has driven increases in global production, particularly in developed and higher income developing countries (8). Sorghum feedstocks contain similar levels of metabolizable energy to maize, and a higher crude protein content. However, due to lower endosperm digestibility, sorghum feed provides around 5% less digestible nutrients to ruminants compared to maize (9). Sorghum is also similar in energy content to wheat, but again with reduced protein availability (10).

Major producers of sorghum include the United States, Nigeria, India and Mexico (Fig. 1.1). Sorghum is commonly grown as a companion crop for wheat rotations in the US and elsewhere in regions that are too hot or dry for maize cultivation (11). The US, Australia and Argentina are the top exporters of sorghum worldwide, with the US contributing around three quarters of global trade (12). The ‘sorghum belt’, composed of large stretches of dryland acreage in the southern half of the Great Plains, supports the majority of sorghum cultivation in the US, predominantly for use as animal feed, industrial, and food products, or for export.
1.3.1 Sorghum as a food and animal feed

Sorghum grain is a valuable source of starch and protein, and sweet sorghum types produce sugars for the production of molasses and fuel (13). Sorghum stems contain high levels of cellulose for fibre-based industries (14-16). Most grain varieties are rich in phosphorus, potassium and iron, as well as providing a good source of minerals and antioxidants (17, 18). The whole grain can be processed into flour for use in various traditional food products including, porridge, unleavened bread, couscous and tortillas. Sorghum grain also provides an alternative source of white flour for production of gluten free foodstuffs. Sorghum malts exhibit similar amylase activities to those of barley, making sorghum a viable alternative in the production of alcoholic and non-alcoholic beverages and in agro-industrial foods (19). Therefore, sorghum is widely employed in industrial scale brewing, with operations including a Nigerian version of Guinness and several gluten-free labels produced in the US. In China, sorghum is used in the production of potent distilled liquors such as Fen and Maotai (20).

1.3.2 Bioethanol

Commercial production of plant-derived fuel has expanded in recent years due to environmental concerns over carbon emissions, the increasing scarcity of fossil-based fuels and the risks associated with their acquisition and transport (21). Sorghum water use efficiency creates a market advantage
for the crop, particularly in drought prone regions. Following processing for ethanol, grain residue can be returned to the feedlot industry as distiller’s dried grain and solubles (DDGS). Sorghum currently represents around 5% of the ethanol market in the US, where maize is the more commonly utilised feedstock (22). However, increased initiatives to develop sorghum as a grain-based fuel could lead to a greater market share for the water-efficient crop as a biofuel (23). Due to ever increasing risks associated with climate change, such as prolonged droughts and heat events, the popularity of sorghum as a feedstock and biofuel is likely to increase in coming years (24, 25).

1.3.3 Nutrient use efficiency

The consequences of water shortages are felt most severely in densely populated areas, where crops with high moisture requirements, such as maize and rice represent a major food source (26). Originating in the semi-arid African savannah, sorghum is robust and well adapted to dry environments, producing more biomass per unit water than maize, and requiring less nitrogen to produce similar yields (27, 28). In semi arid regions, sorghum is grown as a so-called “insurance” crop, because it tends to better survive the severe conditions (29). The development of drought resistant or ‘staygreen’ varieties is leading to further improvements in water use efficiency in sorghum and in other crops (30, 31).

Nitrogen is a major limiting factor of agricultural productivity. Most crops incorporate less than half of the nitrogen applied to the soil, with the remainder lost through leaching or denitrification. Sorghum has similar nitrogen-use requirements to maize, but exhibits a higher efficiency in the uptake and assimilation of nitrogen from the soil, especially under nitrogen limiting conditions (32, 33). This eliminates the need for excessive fertiliser application and prevents the leaching of excess quantities of nitrogen into waterways. Physiological traits underlying enhanced nitrogen uptake in sorghum include a high stem to leaf biomass ratio, efficient nitrogen remobilisation from lower leaves and stem internodes during development and a long duration of vegetative growth (34). It has also been suggested that root architecture plays a role in increased nitrogen uptake (35). Current research is focussed on the evaluation of genotypic factors and management conditions contributing to improved nitrogen-use efficiency in sorghum (36).

1.4 Sorghum improvement

Commercial hybridisation of sorghum was first realised with the detection and utilization of cytoplasmic genic male sterility and fertility restoring genes (37). Crop productivity has been
greatly boosted with the use of male sterile hybrids employed to develop cultivars with improved yield, plant architecture and pest resistance (38). Until recently, commercial cereal breeding programs have been largely geared towards production of improved maize and wheat cultivars. However, since the 1980s, increased resources have been allocated for sorghum initiatives due to the periodic failure of major cash crops as a result of prolonged drought and associated water shortages (39). With the removal of maize subsidies in Africa, water efficient crops such as sorghum have become more competitive, thus allowing for the development of a stable and profitable sorghum seed industry (40).

Contemporary breeding programs utilizing landraces from Africa, India and China have been successful in developing elite lines exhibiting tolerance to various biotic and abiotic stresses, including drought, salinity, lodging, pests and disease (13). Progress has also been made in selecting traits for improved grain quality and yield potential (41, 42). The Sorghum and Millet Improvement Program (SMIP), a joint venture between ICRISAT and the Southern Africa Development Community (SADC), founded in 1983, was designed to support research into conservation of genetic resources, information and technology exchange, natural resource management and food security in poverty stricken areas of sub-saharan Africa, greatly improving the lives of poor subsistence farmers in this region.

Elsewhere, establishment of sorghum breeding programs has lead to increases in crop productivity through the development of improved hybrids (43). In China, large companies are producing hybrid seed, and A2 cytoplasmic-based male sterile hybrids are being commercially cultivated (44). In India, targeted breeding efforts have been ongoing since the development of Coordinated Sorghum Hybrid-1 (CSH-1) in 1964, which lead to a substantial increase in yields (45).

1.5 Grain structure and composition

Grasses produce a caryopsis consisting of a pericarp (outer layer), germ (embryo) and endosperm (storage tissue) (Fig. 1.2). The endosperm is surrounded by the aleurone layer, which is rich in minerals, oils, pigments and produces amylases for starch hydrolysis during germination. Although the basic elements of seed structure remain uniform across the cereals, much variation occurs in grain shape and size, proportion of endosperm, germ and pericarp, amino acid content, pigmentation and oil content (46).
Figure 1.2: Sorghum caryopsis showing pericarp, hilum, seed-coat or testa, endosperm (composed of the sub-aleurone (S. A.) layer, aleurone layer, and corneous, or floury, parts) and germ consisting of the embryonic axis (E. A.) and the scutellum (S) storage tissue. Source: Rooney and Miller (1982). The electron micrograph (left) shows a microscopic view of the protein-starch matrix, where starch granules (S) are embedded in a network of protein bodies (P) (Source: Godwin, 2010).

The major cereal storage proteins, categorised according to solubility, include the albumins, globulins, prolamins and glutelins. The proportion of these proteins in the seed endosperm varies greatly across the cereals. The family Poaceae has defined subdivisions as described by the angiosperm Phylogeny Group (47). The more robust grains, including maize, sorghum and the millets, of the subfamily Panicoideae, contain around 80% prolamin, with non-prolamins such as albumin, globulin and glutelin making up the remaining 20%. Rice and oats, of the subfamily Festucoideae, are rich in albumin and globulin, while barley, wheat and rye, grouped in the tribe Triticeae of the same family, contain relatively high levels of prolamin, but also significant levels of globulin, and are more similar in protein composition to the Panicoideae (48).

Efforts to improve cereal protein quality are predominantly focused on enhancing seed amino acid content to combat nutritional deficiencies. In rice, reduced levels of cysteine and methionine can result in sulphur deficiency. Also, grains containing a high percentage of prolamins, including maize, wheat, barley, sorghum and millet, are deficient in lysine (49, 50). In sorghum, the bioavailability of lysine and threonine is particularly low compared to other amino acids, and needs
to be supplemented in animal feed (51). The nutritional quality of maize has been greatly improved through the development of high lysine varieties with altered starch qualities, referred to as Quality Protein Maize (QPM)(52). High digestibility, high lysine lines have also been developed in sorghum based on the waxy P721Q-derived line (53). In wheat, rye, and barley, a combination of traditional and molecular breeding techniques have been employed to identify and select for improved protein quality at the genetic level (54). QTL mapping of the high grain protein content gene Gpc-B1 has lead to increased protein quality in hexaploid wheat, without yield penalty(55).

1.5.1 Grain quality

The structure and composition of the endosperm protein-starch matrix significantly impacts grain quality traits, including digestibility, nutritional value, grain processing requirements, and fermentation efficiency. Storage proteins make up about 50% of the total protein in mature cereal grains and accumulate surplus nitrogen and sulphur for embryo development (56). In the endosperm, starch is embedded in a network of protein storage bodies and high molecular weight (HMW) polymers linked by disulphide bridges and hydrogen bonds (57). The interconnectivity of these proteins governs the susceptibility of matrix components to proteolysis and therefore affects starch availability. A highly networked structure of endosperm proteins in sorghum limits the digestibility and palatability of the grain, leading to increased processing costs (58). Soft endosperm waxy sorghum, maize and millet varieties are more digestible, but exhibit increased susceptibility to environmental stresses, including moulds and drought (59). High digestibility (HD) sorghum lines, which carry mutations in kafirin or prolamin storage proteins, contain a higher content of non-prolamins and exhibit increased digestibility and nutritional value. These waxy lines have been introgressed with vitreous or hard endosperm traits, therefore contributing to their improved agronomic performance (60).

1.6 Seed storage proteins

The tissue-specific biosynthesis of seed storage proteins is regulated at the developmental level. Proteins are assembled in secretory pathways and deposited into discrete protein bodies in the endosperm (61). Kafirins, the most abundant storage proteins in sorghum, are produced on the endoplasmic reticulum and targeted directly to protein bodies depending on their state of folding and internal signalling peptides (56). Albumins originate as precursor proteins, which are later cleaved with the loss of a linker peptide. The 7S and 11S globulins, related to the legumins, are
synthesised on the rough ER and are transported into the lumen, and then to specific protein storage vacuoles via the Golgi apparatus (62).

### 1.6.1 Protein solubility

Functional analysis of the major seed storage proteins has traditionally involved the sequential isolation of each protein fraction according to solubility; ie. water-soluble albumins, saline-soluble globulins, alcohol-soluble prolamins (extracted under reducing conditions), and glutelins, soluble in dilute acid/alkali solutions (63, 64). In maize, sorghum, and millet, prolamins and glutelins make up the bulk of storage proteins deposited to the endosperm (63). Albumins and globulins are found predominantly in the germ, providing a source of lysine during germination, however, they are also predicted to play a role in the encapsulation of starch in the matrix (65). Kafirins are classified into groups according to size and solubility. β-kafirins (18kD), γ-kafirins (20kD), and δ-kafirins (~13kD) are rich in cysteine and sulphur and tend to be highly cross-linked, forming both inter- and intra molecular disulphide bonds (57). The more abundant α-kafirins (23 and 25kD) are rich in non-polar amino acids and do not crosslink as extensively, forming mainly intra-molecular disulphide bonds.

### 1.6.2 The protein-starch matrix

The sorghum protein-starch matrix exists as a network of spherical protein bodies, primarily composed of kafirins, embedded in a glutelin lattice, which surrounds the starch granules (57) (Fig. 1.3). The structure of the protein-starch matrix directly influences the rate of starch digestion, impacting on food quality traits, such as glycemic response and grain nutritional value (66). The proportion of soluble to insoluble proteins in the matrix determines the access of enzymes to starch and oligo- and polysaccharides (67). Non-kafirins provide a readily accessible source of nitrogen and carbon to the embryo during germination and also play an important secondary role in seed storage, providing structure to the protein-starch matrix (68, 69). The successful development and germination of maize null glb mutants suggests that globulins are not strictly required for normal seed function. However, the proportion of albumins and globulins in the grain influences seed biochemistry, having significant impacts on grain nutritional quality and digestibility (56), as in HD and QPM lines.
Glutelin is composed of subunits linked by disulphide bridges, which aggregate through hydrogen bonding to surround protein bodies in the endosperm (70). Disruption of their native conformation requires a combination of reducing agents and chemicals (NaOH, urea and/or detergent) (71). Alkali treatment in conjunction with cooking increases A/G content and significantly reduces glutelins (72). The digestibility of sorghum grain is increased through NaOH treatment, indicating that glutelins play a major role in connecting the protein-starch matrix (73). Moreover, alkaline processing significantly increases the availability of carbohydrates and free amino nitrogen in the flour, and reduces anti-nutritional properties, such as hydrogen cyanide and tannin content (74).

Figure 1.3: Structure of the protein-starch matrix in the endosperm; starch is embedded in a network of disulphide-bound protein bodies (prolam) and encased in a glutelin lattice (Adapted from de Mesa-Stonestreet et al (2010) (57).

1.6.2.1 Protein bodies

Protein bodies are formed at the milky to hard dough stage (at approximately 14-35 dpa) of grain development. This occurs through the accumulation of storage proteins in the lumen of the rough ER (75). Sorghum grain digestibility is limited, compared to other major cereals, by the enzyme-resistant nature of its protein bodies, where cysteine-rich β- and γ-kafirins may prevent enzymatic access to internally positioned α-kafirins (76). The digestibility of sorghum grain is further reduced upon cooking due to increased cross-linking of kafirins in the endosperm (77). Protein body
composition varies according to endosperm type, with vitreous (hard) endosperm types containing a higher proportion of kafirins, while corneous (soft) endosperm (high lysine) varieties contain higher levels of albumins, globulins and glutelins (78, 79). Vitreous endosperm also contains a greater proportion of cross-linked proteins, which are present across a wider size range compared to floury varieties (80).

High digestibility (HD) sorghum lines carry a missense mutation in an α-kafirin signalling peptide, resulting in a reduced kafirin content and increased levels of albumins and globulins (A/Gs) (81). HD lines exhibit an altered protein body structure, where β- and γ-kafirins are re-located from the periphery to the folds of the structure, resulting in an increased exposure of the α-kafirins to proteolytic breakdown (53). Despite increased nutritional value, the commercial success of HD lines has remained limited due to unfavourable agronomic characteristics of the grain associated with the opaque phenotype. However, the introgression of genetic factors controlling vitreousness has increased grain hardness, thus improving agronomic traits and likewise the commercial viability of the HD line. In maize, the negative agronomic factors associated with the opaque phenotype were recently offset through RNAi reduction of α-zeins, producing a high-lysine kernel with a vitreous endosperm (82).

1.6.2.2 Starch

Starch is the most abundant storage polysaccharide in plants. The proportion and distribution of the two starch types, amylose and amylopectin, in the seed has a major impact on quality traits (83, 84). Amylose is a near linear molecule which exists in varying degrees of polymerisation. It consists of many α-(1,4)-linked D-glucopyranosyl units interspersed with a small number of α-(1,6)-glycosidic linkages. Amylopectin is a more highly branched polysaccharide, made up of α-(1,4)-linked D-glucopyranosyl chains interconnected with a greater number of α-(1,6)-glycosidic linkages (85). Starch is deposited to the endosperm as intracellular granules of varying sizes which are organised in a semi-crystalline configuration. The degree of crystallinity is dependent on the structural features of amylopectin and the proportion of amylose to amylopectin. ‘Waxy’ grain varieties contain nearly 100% amylopectin. The activity of starch synthases and starch branching enzymes is important in determining the balance between amylose and amylopectin production (86, 87). The crystallinity of the starch structure governs susceptibility to hydrolysis by amylolytic enzymes, such α-amylase, an endozyme, and glucoamylase, an exozyme (88). At germination starch is catalysed into sugars by amylases to fuel the growth of the developing embryo (89).
1.6.2.3 Tannins

The accumulation of condensed tannins (proanthocyanidins) in the pericarp and testa of some varieties of sorghum has significant impacts on grain quality and nutrition (90, 91). Tannins are polymers of catechin (5-7 flavan-3-ol units) linked by carbon-carbon bonds (92), which contain a large number of hydroxyl or other functional groups and readily form cross-linkages with proteins and other macromolecules (93). The Tan-1 locus, encoding a WD40 protein, controls the biosynthesis of tannins in sorghum. Genotypic variation in grain tannin content occurs as a result of nucleotide polymorphisms in the coding region of the gene, as observed in mutant tan-1a (94). The presence of condensed tannins protects the grain from environmental stress, but can also decrease protein solubility (95). Tannins form complexes with matrix proteins, hydrolytic enzymes, metal ions and polysaccharides in the grain, limiting protein and starch hydrolysis (93). The aggregation of tannins with storage proteins, including kafirins, results in a reduction in extractable alcohol-soluble protein and γ-kafirin is preferentially bound by tannins, due to its high proline content (96).

Sorghum with brown pericarps and pigmented testas have been shown to limit both rat and poultry growth (97). Alkaline dehulling of the grain restores normal growth rates, suggesting that the removal of the pericarp and subsequent breakdown of protein: tannin aggregates in matrix increases digestibility in high tannin sorghums. Alkaline processing, such as soaking grains in sodium hydroxide, has been shown to decrease sorghum tannin by as much as 80% (72). Tannins are difficult to quantify due to their tendency to polymerise. Grain colour is generally an unreliable indicator of polyphenolic content, as white sorghum varieties have also been found to contain anthocyanins in the pericarp (90). The vanillin-HCL assay is the most accurate assay employed for measuring tannins, where commercial catechins are commonly employed as a standard for measurement. Chromatographic techniques, such as size exclusion chromatography, have also been effective in quantifying and characterising tannins in the grain (98).

In sorghum, there is a strong correlation between phenolic content and mould resistance (99). Mould infestation discolours the seed and breaks down the endosperm, adversely affecting grain processing, and various commercial uses, including malting and brewing. Fungal infection is prevented through various mechanisms, such as the binding of tannins to microbial enzymes (100). The high number of hydroxyl groups substituted on their benzene rings also confers antioxidant properties to tannins, which are considered important nutraceuticals (101).
1.6.3 Regulation of protein body structure

Co-suppression of the various prolamin subclasses in sorghum, including α-, β- and γ-kafirins, alters protein body structure and digestibility (102). In previous studies, down-regulation of α-kafirin expression in the seed increased digestibility, while altering the expression of γ-kafirin had no apparent phenotypic effect (103). This was possibly due to functional redundancy between the β- and γ-kafirins (103, 104). In maize, QTL analysis correlates starch digestibility to chromosome regions also linked to the zeins (105). The Opaque2 gene, encoding a leucine zipper element, alters protein composition and vitreousness of the kernel through regulation of α-zein at the genetic level. O2 mutants contain 50% less zein, exhibiting a floury or opaque endosperm with improved in situ starch digestibility and ethanol conversion (106), but with poor agronomic factors, such as brittle grain texture (107).

The activity of redox-active enzymes have a major influence on grain structure. The composition of the protein-starch matrix is largely defined by the degree of protein folding, proportion of α-helical to β-sheet formation, and the dynamic interactions between different classes of storage proteins. The over-expression of a barley thioredoxin in sorghum, which catalyses the reduction of disulphide bonds among seed proteins, has been shown to significantly enhance grain digestibility (108-110).

Transgenic sorghum expressing elevated levels of HMW glutenin subunits also exhibits increased digestibility (103).

1.7 Seed proteomics

The identification and characterisation of proteins impacting on various grain quality parameters, such as nutritional quality, grain hardness, and stress resistance assists breeders in their introgression of genetic factors associated with these traits into established breeding lines.

Purification of cereal proteins is confounded by their poor solubility and tendency to polymerise. Grain proteomic analysis can be simplified by defining potential protein targets and knowing their location within the seed structure. In whole grain extracts, highly abundant proteins may mask the identification of proteins of interest, present in lower abundance. In addition, the presence of seed storage compounds, including starch, lipids and polyphenols can further complicate the extraction of proteins from certain parts of the grain.
The depth of the proteome can be improved by targeting the domains of the seed separately and through sub-fractionation of proteins with distinct functional or physiological properties. An integrated proteomic approach is often required for comprehensive evaluation of proteins with differing biochemical properties and varying abundance in the grain. This may be achieved through the screening of sequentially extracted protein fractions (for example, A/G, prolamin and glutelin) and by using multiple proteomic techniques, including gel-based separation (SDS-PAGE), mass spectrometry (LC-MS/MS), liquid chromatography (HPLC), and capillary electrophoresis (HPCE). Parallel measurement of biochemical parameters for grain quality, including starch content, digestibility and fermentation efficiency allows for the investigation of relationships between quality parameters and protein expression profiles, in order to identify underlying mechanisms controlling commercially important grain quality traits.

1.7.1 SDS PAGE and tandem mass spectrometry

Protein expression is commonly determined using two-dimensional gel electrophoresis (2DE), a technique which separates proteins based on isoelectric point in the first dimension and by molecular weight in the second dimension (111). Following gel separation, molecular mass and sequence information for 2DE protein spots is determined through tandem mass spectrometry and subsequent bioinformatic analysis (Fig. 1.4). Liquid chromatography-tandem mass spectrometry (LC-MS/MS) has been employed to analyse a wide variety of biomolecules with high sensitivity and resolution. For example, the majority of abundant wheat flour proteins have been identified using 2DE and mass spectrometry and mapped back to sequence data to determine the profile of the expressed wheat genome (112). Cultivar specific contigs are used to assist in matching peptides to gene sequences for members of highly similar storage protein families (113). In sorghum, 2DE/MS-MS profiling the proteome has been carried out in the roots and vegetative tissues under drought- and salt-stress (114, 115). Proteomic analysis of sorghum grain has been limited and research in this area will assist in the further annotation, characterisation, and manipulation of the expressed sorghum genome.
Figure 1.4: Mascot summary for a peptide matches to the β-kafirin protein identified from an excised 2D SDS-PAGE gel spot analysed with LC-MS/MS. The protein score is determined by the quality of the spectral data, as illustrated in the window for the first peptide sequence, query 134.

1.7.1.1 Bioinformatic analysis of MS data

Computational analysis of proteomic data involves conversion of three dimensional retention and migration data and intensity readings for trypsin-digested peptides into two-dimensional chromatography electropherograms through the integration of data within a specific range on an axis (m/z axis), referred to as ion extraction or binning \( (116) \). Baseline removal is then conducted to remove background noise and reduce false positives and a mathematical algorithm is applied to identify top peak candidates. The intensities of MS peaks are related to peptide abundance. MS data is converted from binary format to allow for the data to be processed on independent operating systems and software. The major MS proteomic search engines include Mascot (Matrix Science, London, UK), Protein Prospector, and ProFound. Spectral data can then be mapped against
predicted masses for proteins available from publicly available protein sequence databanks (117). The completed genome sequences for sorghum and rice provide valuable references to which sequence data generated from proteomic and transcriptional analysis can be mapped. Information pertaining to the structure and function of protein candidates can be accessed through queries to annotated protein sequence databases, including the Protein Database (PDB) (118), UniProtKB/SWISS PROT (119) and BRENDA (120), which provide comprehensive collections of protein sequence and structure predictions derived from nucleotide sequence translations and previous experimental work. The SWISSPROT database contains thousands of unique protein sequence entries for each of the cereals.

1.7.2 High Performance Liquid Chromatography (HPLC)

The separation of storage proteins by HPLC presents an alternative method of protein analysis and has a number of important applications. These include the evaluation of the main classes of grain proteins, the fingerprinting of these proteins across diverse grain cultivars, and the validating of genetic modifications in the seed (121, 122), among others. Varietal identification ensures use of lines with optimal quality, yield, stress resistance and adaptability, and selection of high quality flours for brewing, baking and food production. Variation in storage protein and amino acid content among grain cultivars is routinely detected with HPLC (123). The most widely utilised chromatographic techniques for analysis of cereal proteins include Reversed-Phase (RP)-HPLC, Ion Exchange HPLC (IEC), and Size Exclusion-HPLC (SEC) (124). The choice of method depends on the chemical features of the analyte and the complexity of the sample mixture. Chromatographic peaks of interest may be isolated and identified using mass spectrometry.

Size exclusion chromatography (SEC), also referred to as gel-filtration or gel-permeation chromatography (GPC), separates proteins according to size using porous particle filters (Fig. 1.5). Proteins are isolated on a column packed with silica particles of varying pore sizes, where smaller molecules enter pores in the column particles, while larger molecules travel through the column unobstructed, with a faster elution time. The retention time of the molecule in the column pore is directly proportional to its size and shape. Starch solubilities and the profiles of major storage proteins are routinely evaluated in sorghum and maize using SEC (125).
**Figure 1.5:** Depiction of the separation of high Mw (green) and small Mw (red) proteins by size exclusion chromatography (SEC). Columns containing porous particle filters separate proteins by size, where smaller proteins enter the pores of silica particles and are retained in the column. Larger proteins are excluded from the pores, and therefore elute more quickly from the column.

HPLC-based techniques include RP-HPLC and IEC, which resolve proteins on the basis of surface hydrophobicity and charge on a non-porous cation-exchange column. Proteins are eluted along an ionic gradient according to the strength of their binding to the column. RP-HPLC has been used to screen for bread-making qualities, including dough extensibility, which was predicted from overall protein composition rather than the detection of individual gliadins (126, 127). Non gel-based methods, including HPLC, have proven particularly effective in separating the highly hydrophobic kafirin subclasses, such as the abundant α-kafirins, with an increased level of resolution (128).

1.7.3 **High Performance Capillary Electrophoresis (HPCE)**

HPCE techniques have been employed to characterise storage proteins across all of the major cereals, including sorghum (129, 130). The system utilises small diameter capillaries for fine separation of complex mixtures of proteins with high speed, resolution and reproducibility. Proteins are fractionated according to size (SDS-CE), charge density (FZCE), or isoelectric point (cIEF) (131). These methods have been useful for distinguishing between cereal genotypes and in assigning functionality to storage proteins (132). FZCE has also been used in the differentiation of wheat cultivars, identification of rye translocations in wheat flour, and to monitor seed maturation (133, 134).
1.7.3.1 Lab on a Chip (LOC)

LOC technology employs a microchip capillary electrophoresis system to provide protein compositional analysis comparable to SDS-PAGE, with increased speed and automation (135, 136). LOC profiling has previously been used to evaluate grain quality, processing properties and in varietal identification of cereals (137). Small sample volumes are loaded onto a chip and processed in a bioanalyser. The technique involves the insertion of electrodes attached to the lid of the analyser into the fluid wells, applying a voltage gradient which moves proteins through a system of tiny channels, employing the principles of capillary electrophoresis for sample separation. LOC has a number of applications at the protein level, such as profiling of variation in protein content in transgenic seed lines and evaluating proteins associated with nutritional value in legumes (135).

1.8 Grain biochemical analysis

1.8.1 Protein digestibility

Direct quantification of the rate of digestion of starch and protein in the small intestine and the fermentation efficiency by enzymatic hydrolysis is important in determining the precise impacts of grain biochemistry parameters such as crude protein content and composition on commercial end uses, such as feed quality and ethanol production efficiency. Protein digestibility can be measured by in vitro methods, where in vivo studies, such as human or animal feed trials are too laborious and expensive or are incapable of detecting small differences among ecotypes. The most common method for chemical estimation of protein digestibility involves pepsin/amylase digestion (138). In vitro methods using enzymes and length of incubations that mimic human digestion can be used with sufficient accuracy to predict digestibility due to their high correlation with in vivo digestibility (139). The pepsin digestion method is widely used to assess cereal digestibility and feed efficiency (140-142).

1.8.2 Total starch

The content and composition of starch is a major factor in determining the functional characteristics of the grain for human uses. Techniques for measuring total starch include acid hydrolysis or enzymic procedures (143, 144). Because acid hydrolysis can only be applied to pure starch samples,
grain starch content is commonly quantified using enzymatic assays, such as the Megazyme© Total Starch assay. The procedure involves sequential hydrolysis of starch into sugars with thermostable α-amylases and amyloglucosidase. The resulting glucose is quantified through colourimetric measurement (**145**). For difficult to gelatinise samples, such as high amylose corn starch, pre-treatment with solvents such as sodium hydroxide or dimethyl sulfoxide (DMSO) dissolve any resistant starch prior to quantification. Procedures for determining the relative quantity of amylase in a cereal starches and flours can be carried out simultaneously by measuring complex formation between the lectin concavalin A and amylpectin (**146**).

**1.8.3 Flour moisture**

Grain moisture can have significant impacts on grain viability, susceptibility to fungi, and various biochemical characteristics (**147**). The digestibility of grain is often higher in samples with lower moisture content because they contain fewer volatiles associated with higher moisture. Also, grain hardness measurements can be affected by moisture, including the time and energy required to grind the seed (**148**). Variation due to moisture content must be corrected for in evaluating the quality of the grain for functional uses, especially malting and fermentation. Moisture is generally calculated as the weight lost during drying of a sample, and can be achieved through an oven drying method using a moisture analyser. Grain quality parameters, such as (%) protein and starch content are usually reported on a dryweight basis, thus accounting for differences across the grain types arising from this variation in grain moisture content.

**1.9 Project Aims**

This research aims to further characterise the sorghum grain proteome in order to facilitate the identification of factors impacting on grain quality traits. This information will support the development of improved sorghum cultivars exhibiting increased nutritional value and energy production potential.

The rapid introgression and selection of desirable grain quality traits will require the linking of the proteome with the genome. Recently developed next generation sequencing (**149**) and genotyping tools (**150**) are being incorporated into sorghum pre-breeding for grain quality traits such as digestibility (**86**). This will facilitate more cost-effective and efficient plant breeding for improved sorghum varieties tailored to specific end-uses.
Project objectives:

- Identify candidate endosperm proteins in sorghum affecting the structure of the protein-starch matrix and/or regulation of protein body synthesis and aggregation

- Evaluate the impact of the β-kafirin mutation on the expression and interaction of storage proteins in the grain endosperm through profiling of genetically diverse sorghum lines, including several null mutant lines

- Further elucidate the role of β-kafirin in the protein-starch matrix, and the effect of the null mutation on grain quality

- Value-add to the genome to phenome data generated for sorghum quality parameters for digestibility, biofuel and biomaterial production
Chapter 2

Characterisation of sorghum seed proteins through HPLC, capillary electrophoresis and biochemical analysis
2.1 Impacts of Protein Composition on Grain Quality

The value of sorghum grain as a food, feedstock and fuel is limited predominantly by grain quality constraints, such as flour quality, susceptibility to mould and disease, and competition with other grains in the market place (59). Proteins associated with starch granules in the grain endosperm play a key role in determining the functional properties of the grain. Substantial work has been carried out into developing methods for analysis of grain storage proteins, such as by high performance liquid chromatography (HPLC) and capillary electrophoresis (HPCE). These techniques are particularly useful for identifying and characterising the main classes of storage proteins in the grain and in fingerprinting these proteins for cultivar identification.

2.1.1 Storage proteins

Proteins involved in seed storage provide a source of nitrogen during germination and are of major importance in nutrition and in determining functional traits, such as flour quality for breadmaking (56). In the seed endosperm, starch is embedded in a network of protein storage bodies and large molecular weight polymers linked by disulphide bridges and hydrogen bonds (57). The interconnectivity of these proteins governs the susceptibility of matrix components to proteolysis and amylolytic breakdown and therefore affects starch availability, nutritional quality and fermentation profile.

2.1.2 Grain protein composition

The amino acid profile of the grain is a good indicator of nutritional quality. Limited amino availability is a major cause of nutritional deficiencies in regions where cereals are consumed as a primary food source. Sorghum, like maize, is deficient in lysine and tryptophan. Increasing the content of lysine-rich storage proteins, such as albumins and globulins, has been shown to enhance nutritional quality (79). High-lysine sorghum varieties, such as the chemically-induced opaque mutant P721, exhibit improved feed efficiency, with half the grain needed to sustain normal growth in livestock (151, 152). Opaque 2 (o2) and floury 2 (fl2) mutations in maize lead to the development of germplasm with improved food value, referred to as Quality Protein Maize (52). O2 and fl2 exhibit a soft, floury endosperm with increased content of lysine and tryptophan. However, the commercial success of these lines was initially limited by negative pleiotropic effects associated
with the soft grain phenotype, such as brittleness and increased susceptibility to pests. Introgression of genetic modifiers for vitreousness or grain hardness has resulted in improved kernel characteristics, while maintaining the enhanced nutritional profile of the grain. QPM hybrids exhibit increased seed density, higher grain yield and twice as much usable protein as other maize varieties (153).

Endosperm texture, or the relative proportion of corneous to vitreous endosperm, dictates the quality of grain for commercial end uses. Vitreous endosperm contains a high degree of cross-linked storage proteins with a greater percentage of disulphide bonds. In both maize and sorghum, floury endosperm contains higher levels of non-kafirins, such as albumins and globulins, while vitreous endosperm contains a higher proportion of β- and γ-kafirin (80, 154). Sorghum high digestibility (HD) lines, which exhibit increased lysine content, retain a vitreous endosperm (60). The basis of the HD mutation occurs in the 22kD α-kafirin signalling peptide and results in faulty targeting of kafirins to the protein bodies (PBs). Interestingly, the mutation is accompanied by an increase in the expression of protein disulphide isomerase (PDI), which occurs presumably due to an elevated unfolded protein response (UPR) (81). Identification of allelic variation in the kafirins has provided germplasm carrying a null mutation in the β-kafirin gene. The precise roles of β- and γ-kafirin in the protein-starch matrix and the possibility of functional redundancy among the two protein types is unclear. Further research is needed to decipher the composition and interconnectivity of these subclasses through mutant analysis and biochemical profiling.

2.1.2.1 Protein cross-linking and polymeric proteins

The chemistry and structure of endosperm proteins has significant impacts on functional uses of the grain. The ratio of soluble to insoluble and residual protein reflects the extent of cross-linking among storage proteins, which impacts on important grain quality traits for food and feed, particularly flour viscosity and protein digestibility. For example, the rate of fermentation depends on the proportion of soluble to insoluble proteins in the matrix, which governs the access of enzymes to starch and polyssacharides (67). Isolation of storage proteins under non-reducing conditions produces groups of ‘soluble’ proteins consisting primarily of monomeric kafirins and metabolic enzymes. The ‘insoluble’ protein fraction, consisting predominantly of larger Mw polymeric protein structures extracted using sonication. The remaining highly cross-linked un-extractable proteins, referred to as ‘residual proteins’, are then solubilised with the addition of a reducing agent, such as β-mercaptoethanol (β-ME). When disulphide bonds are reduced, these polymers (referred to primarily as glutelins) are broken down into smaller polypeptides, which may
be further solubilised in urea, guanidine chloride and sodium-dodecyl sulphate (SDS) for functional analysis (155).

2.1.2.2 Effects of cooking on protein solubility

Swelling in water is restricted by the cross-linking of storage proteins, affecting the digestibility and fermentation profile of the grain (93). In sorghum, kafirin hydrophobicity is increased upon cooking with moisture, altering protein polymerisation and beta-sheet formation (48, 156). Disulphide linkages among cysteine-rich \( \gamma \)-kafirins are thought to be responsible for reduced viscoelasticity in sorghum relative to maize (157). During cooking a change in protein structure from \( \alpha \)-helical to \( \beta \)-sheet occurs, shown in studies using Fourier transform infrared spectroscopy (158). Pepsin-indigestible kafirin residues exist mainly in the \( \beta \)-sheet formation. A significant reduction in kafirin content occurs during wet cooking, accompanied by an increase in the proportion of cross-linked glutelin and non-extractable proteins (57, 72, 77). Cooked sorghum contains greater amounts of \( \beta \)-and \( \gamma \)-prolamin compared to maize, and samples enriched with protein bodies have been found to contain more 45-50kD oligomers than maize, some of which were resistant to digestion upon cooking.

2.1.2.3 Tannin-protein interactions

Condensed tannins protect the grain from environmental stress, but readily form cross-linkages with storage proteins (95). The phenolic hydroxyl groups of tannins form complexes with proteins, reducing dry matter and protein digestibility. Direct quantification of tannin content is difficult and generally involves precipitation, oxidation, colourimetric analysis, and/or UV spectroscopy of tannins (159). Additionally, HPLC is a common technique for quantifying tannins, which has been extensively employed in sorghum and millet (90). The aggregation of tannins and proteins in the grain endosperm reduces levels of alcohol-extractable protein. Tannins have been found to preferentially bind the proline-rich \( \gamma \)-kafirins (96). Profiling the alcohol-soluble kafirin fraction across sorghum grain lines using RP-HPLC has showed that binding of tannins to grain storage proteins results in an altered peak distribution, displaying reduced levels or the complete absence of the \( \gamma \)-kafirin peak in most tannin-containing lines (91). Similar to this, an assay was proposed where the amount of protein precipitated from a solution by the addition of tannins/phenols is taken as a measure of the amount of tannins present (160). The results of the study showed that the relationship between precipitated protein and tannin content was sigmoidal (representing an s-
shaped curve between the two variables), and that the interaction is affected by pH and different proteins interacted differently with the same type of tannin.

2.1.3 Starch impacts on grain digestibility

The structure of starch in the endosperm has additional impacts on grain quality. The production of amylose is controlled by the WAXY (Wx) locus, which encodes the gene for granule bound ADP-glucose-glucosyl transferase or Granule Bound Starch Synthase I (GBSSI) (161). Waxy or glutinous loci have been identified across the major cereals (162). Mutations in GBSSI result in a shift in amylose to amylopectin production in the grain, which results in the production of less highly branched starch structures. Waxy cereal varieties have higher feed efficiency and steam flake more easily, reducing processing inputs (163). Starch from waxy cultivars is valued for its adhesive qualities and used as a commercial thickener and in the production of gelatinous noodles (164). Sorghum grain waxy types exhibit an improved digestibility and fermentation profile (165), but may have poor agronomic characteristics, such as reduced kernel quality and susceptibility to insect attack and moulds, similar to early HD and QPM lines.

2.1.4 Analysis of seed proteins

A variety of methods have been developed for the analysis of cereal proteins. High performance liquid chromatography (HPLC) and capillary electrophoresis (HPCE) are useful for quantitative profiling of grain proteins, particularly where the analysis of large sample sets is required. These techniques may be coupled with mass spectrometry for specific identification of proteins isolated in the analysis. Chromatographic techniques allow for rapid analytical determination of parameters associated with grain quality, such as profiling of the major protein classes, determining the ratio of glutenin to gliadin in breadmaking and quantifying the percentage of unextractable polymeric protein present in cross-linking studies. The degree of cross-linking among storage proteins in vitreous and floury endosperm has been investigated using size exclusion chromatography (SEC), by measuring the relative proportion of unreduced protein present in the sample (80). Reversed phase (RP)-HPLC has also been employed to monitor the effects of wet cooking on protein composition in oats (166).
2.1.5 Study Aims and Objectives

Analysis of grain protein composition across 28 grain sorghum lines previously characterised for kafirin allelic background (Table 2.1) complements genomics initiatives currently underway for the development of improved sorghum cultivars. Separation of water/salt- and alcohol-soluble protein fractions carried out using chromatographic techniques and capillary electrophoresis provides high resolution profiling of specific types of storage proteins across the grain lines. The goal is to determine the impact of variation in the expression of different classes of storage proteins on quality traits, such as protein digestibility. It is expected that a null mutation in the β-kafirin gene may impact on grain protein profiles and biochemical parameters, such as digestibility, as was seen in previous studies measuring flour viscosity across genotypes with kafirin allelic variation (167).

2.2 Materials and Methods

2.2.1 Plant Material

Mature grain from 28 inbred sorghum lines, with allelic variation in the β-, γ-, and δ-kafirin storage proteins were evaluated in the study (Table 2.1). The primary focus of the analysis was on comparisons between the β-kafirin null lines QL12, IS17214, and RTx2737, and lines with normal expression of the β-kafirin protein. Plants were cultivated under field conditions at the University of Queensland, Gatton campus, QLD, Australia, over the 2011/2012 summer growing season. Whole grain was milled through a UDY sample mill (UDYCorp, Fort Collins, CO.) fitted with a 0.5mm mesh screen for all analytical procedures.
Table 2.1: Background and origin of sample grain lines

<table>
<thead>
<tr>
<th>Genotype/Line</th>
<th>Origin</th>
<th>Background/Traits</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTx7000/FF-RTx7000</td>
<td>USA</td>
<td>Milo-Kaura, Senescent background</td>
</tr>
<tr>
<td>IS22457C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS8525</td>
<td>Ethiopia</td>
<td>Parent of mapping population for ergot resistance, highly resistant to sugary disease</td>
</tr>
<tr>
<td>QL12</td>
<td>Australia</td>
<td>Stay green, yellow endosperm</td>
</tr>
<tr>
<td>IS12572C/FF-6214E</td>
<td></td>
<td>Caudatum-nigricans-conspicuum, photoperiod insensitive, midge resistance</td>
</tr>
<tr>
<td>Ai4</td>
<td>China</td>
<td>Dwarf, photoperiod insensitive, possible cold tolerance</td>
</tr>
<tr>
<td>BTx3197</td>
<td>USA</td>
<td>Inbred line, Kafir, stay green</td>
</tr>
<tr>
<td>Hegari/ Early Hegari</td>
<td>Sudan</td>
<td>Caudatum, white, chalky seeds, highly palatable</td>
</tr>
<tr>
<td>ISCV745</td>
<td>India</td>
<td>Zerazera-caudatum, Parent of mapping population</td>
</tr>
<tr>
<td>B923296</td>
<td>Australia</td>
<td>Elite line, stay green parent</td>
</tr>
<tr>
<td>ISCV400</td>
<td>Mali</td>
<td>2-dwarf, Zerazera caudatum bred by ICRISAT as a food and malting, white grain</td>
</tr>
<tr>
<td>QL39</td>
<td>Australia</td>
<td>Stay green, midge resistance</td>
</tr>
<tr>
<td>Karper 669</td>
<td>USA</td>
<td>Diverse yellow endosperm germplasm line</td>
</tr>
<tr>
<td>QL41</td>
<td>Australia</td>
<td>Staygreen</td>
</tr>
<tr>
<td>R931945-2-2</td>
<td>Australia</td>
<td>Elite stay green parent</td>
</tr>
<tr>
<td>R89562/ R890562-1-1</td>
<td>Australia</td>
<td>Elite moderately senescent parent line</td>
</tr>
<tr>
<td>M35-1/FF-BM351</td>
<td>India</td>
<td>Durra landrace selection, drought resistance, cooking quality</td>
</tr>
<tr>
<td>KS115</td>
<td>USA</td>
<td>Breeding line, large seed, yellow endosperm</td>
</tr>
<tr>
<td>IS17214</td>
<td>Nigeria</td>
<td>Landrace</td>
</tr>
<tr>
<td>SC1270-6-8/FF-SC1270-6-8</td>
<td>Ethiopia</td>
<td>Caudatum line ex Ethiopia, high yielding genotype and parent of mapping population.</td>
</tr>
<tr>
<td>LR91918</td>
<td>China</td>
<td>Breeding line, 2-dwarf, male of a good hybrid in China</td>
</tr>
<tr>
<td>IS12611C-F-F_SC11114E</td>
<td>Ethiopia</td>
<td>Zera-zera, breeding line, parent in cross for ICSV400</td>
</tr>
<tr>
<td>R9733</td>
<td>USA</td>
<td>Breeding line from Texas A&amp;M breeding program</td>
</tr>
<tr>
<td>BOK11/FF-BPK11</td>
<td>USA</td>
<td>Inbred breeding line: Dwarf Hydro x Rice, Kafir</td>
</tr>
<tr>
<td>BTx623/FF-BTx623</td>
<td>USA</td>
<td>Zerazera-caudatum, elite inbred line, stay green</td>
</tr>
<tr>
<td>B355/FF-B35</td>
<td>Ethiopia</td>
<td>Partially converted durra landrace IS 1255, highly stay green</td>
</tr>
<tr>
<td>296B/FF-296B</td>
<td>India</td>
<td>Elite inbred line, dwarf, high yield, resistant to foliar diseases</td>
</tr>
<tr>
<td>RTx2373/FF-BTx2373</td>
<td>USA</td>
<td>Commercial hybrid breeding line, yellow endosperm, stay green, midge resistance</td>
</tr>
</tbody>
</table>

2.2.2 Protein Extraction methods

2.2.2.1 Albumin/globulins and prolamins

*Water/salt-soluble (albumin/globulin)* proteins were pre-extracted using a modified version of previous methods (168). Briefly, flour samples (100mg) were soaked in 1ml extraction solvent (50mM Tris-HCl ph 7.8, 100mM KCl and 5mM EDTA), vortexed (5 min), centrifuged (10k rpm for 4min), and 500µl supernatant was retained. The process was repeated and an additional 500µl aliquot was removed. Pooled aliquots were lyophilised and re-dissolved in 500µl of 50% ethylene glycol for analysis.

*Alcohol-soluble (prolamin)* proteins were extracted sequentially from the A/G pellet above using previously described methods (169). The pellet was washed with 1 ml dH2O, centrifuged and then dissolved in 1ml solvent (60% tertiary butanol, 0.5% sodium acetate and 2% β-mercaptoethanol).
The sample was then vortexed (5 min), centrifuged (10k rpm for 4min) and 500µl supernatant was retained as above. The procedure was repeated and a second 500µl aliquot of supernatant was removed. Pooled extracts were alkylated by the addition of 33.3µl of 100% 4-vinylpyridine with vortexing (10min) to prevent the re-formation of disulphide bonds. Protein extraction methods for A/Gs and prolamins are illustrated in a flow chart (Fig 2.1). Protein was extracted using the above methods from standard sorghum feed and the ileal tract and fecal matter of swine fed on these rations (section 2.3.4.2). Samples were treated with phyto-phosphatase prior to extraction to control for the digestion of phytate, which does not occur in monogastric animals.

![Protein extraction procedure utilised to extract albumin/globulins (AGs) and kafirins (prolamins).](image)

**Figure 2.1:** Protein extraction procedure utilised to extract albumin/globulins (AGs) and kafirins (prolamins).

### 2.2.2.2 Extraction of polymeric proteins

Analysis of polymeric proteins involved a sequential extraction procedure based on previous methods (80), which was modified to accommodate for increased sample volumes. Soluble protein (SP) was extracted from 100mg flour, which was soaked for 30min in 1mL 12.5 mM sodium borate pH 10 plus 2% SDS, with continual vortexing. Samples were centrifuged and supernatant removed. Insoluble protein (IP) was extracted from the SP pellet for 30min with sodium borate/SDS buffer pH 10 using sonication (10W for 30s). Samples were then centrifuged and supernatant removed. Finally, residual protein (RP) was extracted from the IP pellet for 30min under reducing conditions in sodium borate/SDS buffer plus 2% βME. The sample was centrifuged and supernatant was removed. Aliquots of each extract were analysed with size exclusion chromatography (SEC) and the relative proportion of each protein type was calculated by summing the total peak area across all
2.2.3 High Performance Liquid Chromatography (HPLC)

2.2.3.1 Reversed-Phase HPLC (RP-HPLC)

Protein samples were injected into an Agilent 1100 HPLC system (Agilent, Foster City, CA), fitted with a Poroshell column of varying specifications, as indicated below for the various protein fractions below. The system employs a binary gradient with a constant flow rate of 0.7ml/min and a column temperature maintained at 55°C. Solvent gradients included water (A) and acetonitrile (B), both containing 0.1% trifluoroacetic acid (v/v) with gradient flow specifications as follows: 0-18min, 45%-60% B; 18-19min, decreased to 45% B; followed by a 7min post run. Protein peaks were visualised with a UV detector measuring at 206nm. Sample processing was as follows:

Albumin/globulins (A/G): 250μl A/G aliquots were freeze dried in a speed vac. Samples were resuspended in 100μl 50% ethylene glycol to stabilise proteins and 5μl injections run on the Poroshell30058 C-3 column, 2.1 x 75 (S/N USYV001432).

Prolamins: For alkylation of proteins, 33μl 4-vinylpyridine added to each prolamin sample and vortexed 10 min. Protein samples were then run in 5μl injections on the Poroshell 30058 C-18 column, 2.1 x 75 (S/N USU5003087).

2.2.3.2 Size exclusion chromatography (SEC)

SEC was carried out using BioSep SEC 3000 column with 50mM sodium phosphate buffer, pH 7 containing 1% SDS. Flow rate was 1 mL/min and column temp 40 C. Water/salt-soluble A/G fractions were analysed using SEC. Samples were extracted as above, resuspended in 50% ethylene glycol and run in 50μl injections.

2.2.4 Lab on Chip (LOC)

The Lab on Chip procedure was carried out on the Agilent 2100 Bioanalyser. Samples were prepared using the Protein 80 assay kit (Agilent Technologies, Palo Alto, CA). A 1:1 mixture of protein to denaturing buffer was used. Sample tubes and an additional tube containing ladder were
heated to 95°C for 5min, cooled and centrifuged. 84μl dH₂O was added to each tube and samples were vortexed and spun briefly. Protein samples (6μl), ladder (6μl) and gel dye (12μl) were loaded into the appropriate well on the chip. The chip was inserted into the Bioanalyser and analysed per manufacturer’s instruction.

2.2.5 Crude protein digestibility

Protein digestibility was determined in duplicate using the previously described method for measuring *in vitro* pepsin digestibility (140). Briefly, 200mg milled sorghum flour was mixed in 35ml pepsin solution (1.5mg/ml in 0.1 phosphate buffer (KH₂PO₄ + H₃PO₄) pH 2) and incubated at 37°C for 2hrs. After incubation, 2ml 2M NaOH was added, the sample vortexed and centrifuged (3220xg, 15min). The supernatant was then discarded. Residue was washed in 10ml 0.1 phosphate buffer pH2, centrifuged (3220xg, 15min) and supernatant discarded. Washing steps were repeated and after second wash and centrifugation the samples were placed in a -80°C freezer (Romulus Holding Company, New York, NY). Prior to nitrogen analysis, samples were lyophilised (Labconco, Kansas City, MO), and total protein content was measured using nitrogen combustion Leco nitrogen measurement (LECO Corporation, St. Joseph, MI). Nitrogen content was measured and converted to protein using the conversion factor 6.25.

2.2.6 Starch Analysis

Total starch in 200mg milled sorghum samples was determined in duplicate through a colorimetric technique outlined in AACCI method 76-13.01 (170), utilizing a DMSO pre-treatment for resistant starch (K-TSTA Megazyme International Ireland Ltd., Co. Wicklow, Ireland). A corn starch standard was included with each set of measurements. Sample absorbance was measured at 510 nm against a reagent blank and converted to % starch (of flour weight) using a d-glucose standard.

2.2.7 Flour moisture content

Moisture readings were taken from 1g milled sorghum flour per sample in duplicate using a MX-50 moisture analyser (A&D, Tokyo, Japan). Average moisture content was calculated as the mean across duplicates.
2.2.8 Statistical Analysis

Significance of correlation among factors affecting grain quality, including total protein, total starch, and protein digestibility (%), was illustrated using Microsoft Excel. Multivariate analysis carried out in Minitab generated p values reflecting the significance of correlations among grain quality parameters using an ANOVA general linear model with pairwise comparisons according to Tukey methods with 95% confidence. Absorbance values from RP-HPLC of alcohol-soluble proteins were interpolated by a spline method and absorbance area values were calculated to 0.01-min intervals using MATLAB (The MathWorks, R2012a, Natick, MA). Linear correlation coefficients were calculated between individual absorbance area and digestibility and shown as a continuous spectrum over RP-HPLC retention time.

2.3 Results and Discussion

2.3.1 RP-HPLC analysis of sorghum seed proteins

HPLC and capillary electrophoresis were employed to evaluate protein composition across the population of kafirin allelic variants and to further characterise the β-kafirin null mutation. Separation of water/salt-soluble (A/G) (Fig. 2.2) and alcohol-soluble (kafirin-containing) (Fig. 2.3) protein fractions by RP-HPLC revealed significant variation in protein accumulation across the grain lines.

2.3.1.1 Albumin/Globulins (A/G)

Chromatographic techniques have been widely employed for the characterisation and fingerprinting of water/salt-soluble proteins in legumes and grains, especially oats and rice (166, 171). Seed proteins are traditionally classified according to their biological function as storage proteins (prolamins and glutelins), and metabolic and structural proteins (albumins and globulins). However, certain A/Gs, particularly HMW proteins, have been found to play a role in starch storage in the grain. Here, profiling of the water/salt-soluble fraction reveals the composition of A/Gs in sorghum across a diverse selection of grain genotypes (Fig. 2.2)
Figure 2.2: RP-HPLC separation of water/salt-soluble (A/G) protein fractions for sorghum lines with allelic variation in the kafirins. Protein samples were freeze-dried, resuspended in 50% ethylene glycol to stabilise proteins and run in 5μl injections on a C3 column.

Major peaks eluting at 5.5 and 6.5 min likely represent highly abundant proteins also identified on SDS PAGE gels (Chapter 3), such as HMW globulins and cupin-like proteins. Several lines contain a high level of A/Gs, including Karper 669, KS115, R9733, BOK11, QL12 and RTx7000. This was shown through the presence of wide protein peak areas and a large number of peaks distributed across the chromatogram. These grain lines may be rich in essential amino acids, such as lysine and methionine, due to their high A/G content. Genotypes with a relatively low A/G content included R890562, R91918, IS12611C, BTx3197, Hegari and QL39, possibly representing lysine-poor lines. Extensive variability was observed in a peak eluting at 9 min, however the peak did not appear to be correlated to grain protein digestibility according to statistical analysis performed against biochemical traits measured in the grain (data not shown).
2.3.1.2 Kafirins

HPLC analysis of the alcohol-soluble protein fraction in sorghum was carried out to evaluate the impact of prolamin diversity on grain quality parameters. Peak distribution profiles across the lines were highly variable. Two relatively large peaks were visible in the prolamin fraction across the whole sample set, eluting at approximately 11 and 12 min. These peaks likely represent the major sets of α-kafirins (22 and 25kD). An additional large peak eluting at 12.5 min was visible across three genotypes, KS115, R733 and B923296, and appears to be differentially expressed across the lines.

High peak areas across the chromatogram were observed in KS115, QL41, BTx623, RTx7000, BTx3197, ISCV745 and ICSV400, indicating that these lines are rich in prolamins, particularly in α-kafirin. β-kafirin null variants QL12, IS17214, RTx2737, along with grain lines 296B, SC177068, R91918 and B923296, exhibited the lowest prolamin contents in terms of the heights of the major peaks. Reduction in the height of protein peaks in the prolamin fraction tended to accompany an increase in the diversity or number of peaks in the sample. This indicates possible variation in levels of cross-linking and the formation of complexes among storage proteins across the samples.
Figure 2.3 (Continued below)
Figure 2.3: RP-HPLC profiling across kafirin allelic variants (including β-kafirin nulls QL12, IS17214 and RTx2737). The alcohol-soluble protein fraction was extracted in 60% tertiary butanol. Proteins were alkylated with 4-vinylpyrididine (4-VP) and run in 5μl injections on a C18 column. β-kafirin null lines were missing a major set of peaks eluting at 10-11min. Peak profiles are highlighted for possible γ-kafirin (blue), β-kafirin (purple), and α-kafirin (red).
2.3.1.3 Soluble, insoluble and residual proteins

The degree of cross-linking among storage proteins in the endosperm matrix was evaluated through measurement of the relative content of soluble (SP), insoluble (IP) and residual (RP) protein (Fig. 2.4). A sequential extraction scheme was employed, which was modified from previous methods (103), where monomeric and low Mw (LMW) disulphide-bonded polymeric protein structures were removed under non-reducing conditions as the ‘soluble’ component of the grain. Insoluble proteins were then extracted with sonication, which isolated larger, more highly linked polymeric protein structures. Finally, residual proteins were removed under reducing conditions (β-ME). Each respective protein fraction was analysed using SEC and the resulting peak areas were calculated to quantify relative amounts of each type of protein.

![Figure 2.4: Sequential extraction of soluble, insoluble and residual protein from kafirin allelic variants extracted under non-reducing and reducing conditions.](image)

Increased SP levels tended to be associated with decreased levels of RP across the grain lines (Fig. 2.5). This weak correlation ($R^2=0.337$) ($p=0.001$) across a relatively large sample set indicates that further investigation into the extent of protein cross-linking could be warranted in experiments.
repeated under similar conditions. A tightly bound protein network reduces the solubility of LMW monomeric protein bound by disulphide linkages in the matrix. The extractability of the soluble protein fraction may be limited by the extent of cross-linking among proteins in the endosperm. Therefore, a shift in the distribution of proteins from SP to IP/RP may indicate increased cross-linking among endosperm proteins (80).

**Figure 2.5**: A negative association exists between soluble (SP) and residual (RP) protein content across sorghum allelic variants evaluated with SEC. SP was extracted under non-reducing conditions, while RP was extracted in the presence of 2% βME.

Beta-kafirin null mutant QL12 and normal genotype 296B exhibit similar profiles for SP, IP and RP content, indicating that the degree of cross-linking among endosperm proteins is consistent across the two lines. Protein digestibility is significantly higher in 296B than QL12, despite the β-kafirin null mutation in QL12, which should confer a lower kafirin content, and therefore higher digestibility. The basis for the high digestibility of 296B is unclear and could be due to inherently low levels of β-kafirin or could also be attributed to other factors, such as the activity of thioredoxin (Chapter 3) and/or starch branching enzymes (not measured).
Genotype QL39 contains a high level of residual protein and has a corresponding low digestibility. IS125726, IS17214, BTx3197 and BTx623 contain higher proportions of residual protein relative to insoluble protein, and also contain low soluble protein. This protein profile tended to be accompanied by low digestibility, except in the β-kafirin null line IS17214, which had the highest cooked digestibility (33%) across the sample population. This digestibility profile may be attributed to the mutation in the β-kafirin gene in this line. R9733 exhibited the highest soluble protein content across the sample set, with the lowest cooked digestibility and an average raw digestibility. This illustrates that high levels of total protein are generally negatively associated with digestibility, especially during wet cooking, even if levels of insoluble and residual protein are relatively low.

2.3.1.4 RP-HPLC screening for the presence of tannins

Tannins bind preferentially to cysteine-rich prolamins, particularly γ-kafirin. The RP-HPLC profile of alcohol-soluble proteins in sorghum line IS8525, a line reported to contain tannins (98), indicates that interactions between protein and tannins may occur in this line. Protein content was significantly reduced in IS8525 and a set of protein peaks thought to include γ-kafirin, eluting at 4-6 min, was absent, compared to the other lines. IS12572C exhibited a similar change in RP-HPLC profile for the kafirins, with missing peaks at 4-6 min and lower overall levels of alcohol-soluble protein observed across the chromatogram (Fig. 2.6). Significant reductions in raw and cooked digestibility were also observed in both lines (IS8525 and IS12572C), as shown through subsequent biochemical analysis of the grain (pepsin digest). Tannin content was not directly measured in the grain, however, RP-HPLC may present a useful method for screening protein composition and tannin content in parallel, where the biochemical effects of tannins binding protein in the seed endosperm can be evaluated in terms of their effects on grain quality and commercial end-uses, such as digestibility.
**Figure 2.6**: RP-HPLC profiles for alcohol-soluble proteins across kafirin allelic variant sample population. Protein extracts were alkylated and loaded in 5μl injections onto a C18 column for separation. Chromatograms show the absence of the γ-kafirin peak, eluting at 4 min, in characterised tannin-containing line IS8525 and in an additional line IS12572C, which may also contain tannins. The γ-kafirin peak was visible across the entire sample population, except in these lines (IS8525 and IS12572C), indicating the probable binding and precipitation of the protein by tannins in these samples.
2.3.2 Size Exclusion Chromatography (SEC)

SEC profiling proteins in β-kafirin null QL12 and wild-type line 296B (Fig. 2.7) was carried out to further investigate the effects of the β-kafirin mutation on the expression of water-soluble proteins in the grain. The differential expression of a small Mw thioredoxin (Trx) in QL12 (absent) compared to 296B (present) was observed using 2D SDS-PAGE/LC-MS/MS analysis, outlined in Chapter 3. Changes in the expression of this enzyme may be linked to the β-kafirin mutation and may also account for the increased digestibility of 296B relative to QL12, given the documented effects of increased Trx on protein solubility (172).

Figure 2.7: SEC analysis of the A/G fraction in β-kafirin null mutant QL12 compared to wild-type 296B. Water/salt-soluble (A/G) protein fractions were freeze dried and resuspended in 50% ethylene glycol prior to analysis on BioSep SEC 3000.
SEC profiles were generally similar across the two genotypes, however, a HMW protein aggregate, which eluted from the column at 10min in 296B, was absent in QL12. Additionally, the height of a protein peak eluting at 22min was approximately doubled in QL12, compared to 296B. Because elution time is inversely proportional to the size of the protein and HMW proteins are eluted first, the peak eluting at 10min is unlikely to represent a LMW Trx protein. However, the enzyme may be masked in the peak eluting at 22min or could be represented in a smaller peak eluting at ~32min, which is visible in 296B and absent in QL12.

SEC chromatograms indicate that, across the entire distribution of protein peaks, QL12 generally contains higher levels of water/salt-solubles overall compared to 296B, in agreement with RP-HPLC profiles for water-solubles across these lines (Fig. 2.2). However, the presence of the large peak in 296B, which may also represent a protein aggregate, or contaminant, indicates that this line may contain high levels of a specific type of protein or polymer, which could influence the protein digestibility and biochemical characteristics of this line relative to QL12. Conversely, in QL12, the most notable differentially expressed protein entity in the A/G fraction is represented by the approximate doubling of the peak height of a protein or set of similarly sized proteins eluting at 22min.

Grain types with increased levels of A/Gs should be more digestible than those with a higher prolamin content, as previously reported (65, 173). As discussed above, 296B exhibits a higher raw and cooked digestibility than QL12. Therefore, although QL12 exhibits above average digestibility relative to the other lines, possibly due to a high A/G levels and low β-kafirin content, 296B exhibits an even higher digestibility, due to low total protein and possibly to the differential expression of Trx observed LC-MS/MS and the increased presence of this high Mw protein in the water-soluble fraction shown here with SEC.

2.3.3 Lab on Chip

2.3.3.1 β-kafirin expression in null allelic variants

Size-based separation of alcohol-soluble proteins using LOC provides further verification of the differential expression of β-kafirin across the sample population. This was apparent through changes to the profiles of peaks eluting within the 17-20 kD size range (Figure 2.8) and through a
general reduction in the height of the 19kD peak in mutant compared with wild-type lines. The expression of β-kafirin appears to be decreased in β-kafirin null lines QL12 and IS17214, and to a lesser extent in RTx2737. Interestingly, genotype 296B also appears to contain relatively low levels of β-kafirin compared to the other normal lines, although it carries a functioning allele for the gene. The height of the 19kD β-kafirin peak is only slightly higher in 296B relative to QL12 (Fig. 2.9), which may explain, in part, the high digestibility profile of 296B.

Corresponding to the computerised gel image generated in LOC, electropherograms showed the reduced expression of β-kafirin with greater resolution (Fig. 2.9). Despite the mutation in β-kafirin, a small peak is visible in the mutants at 19kD. This protein may represent a different protein of a similar size, such as PPIase, identified in the same protein band/spot as β-kafirin with LC-MS/MS (Chapter 3). In any case, there appears to be a similar change in the distribution of proteins in the 17-20 kD size range across the mutants, where the major peak visible in wild-type genotypes is reduced in the mutants and a second peak representing a slightly larger (~20 kD) protein appears more prominent (Figure 2.8).

**Figure 2.8:** LOC gel image of the kafirin fraction showing allelic variation in β-kafirin across normal and null mutant lines.
Figure 2.9: Alcohol-soluble protein content of sorghum grain in the β-kafirin null QL12 compared to normal genotypes 296B and KS115. Peak intensity was measured in fluorescent units (FU) and shows the height of the 19kD β-kafirin-containing peak to be variable across grain lines, i.e. peak height is significantly reduced in null mutant QL12.
2.3.3.2 High Mw (HMW) prolamin

A small protein peak observed on LOC electropherograms at 46kD (Fig. 2.9), which is also visible as a weak band in the corresponding size range on the computerised gel image (Fig. 2.8), represents a HMW kafirin or kafirin dimer. The visibility of this protein in samples treated with βME indicates that it is not likely to represent a polymer composed of LMW kafirins bound by disulphide linkages, which can be broken down by treatment with a reducing agent. Instead, it is possible that this protein could potentially be classified as a HMW protein entity, which may be linked by hydrogen bonds. In previous studies, this protein has been identified as a HMW kafirin due to its resistance to breakdown by reducing agents (132). In the current study, a HMW protein was isolated within a similar size range (40-50kD) on 2D SDS-PAGE and identified as a homolog to 50 kD γ-prolamins from other grass species, such as sugarcane and maize, through LC-MS/MS (Chapter 3). The protein did not exhibit any association with protein digestibility across this sample set when LOC peaks were plotted against digestibility values (data not shown).

2.3.3.3 Sequential extraction of alcohol-soluble proteins

Separation of the kafirins into fractions of unique polypeptide composition can be achieved by isolating proteins at varying alcohol concentrations. In maize, differences in solubility across the zeins have been exploited to separate α-zeins, soluble in higher concentrations of alcohol, from other zein constituents (174). In the current study, proteins were solubilised in 10, 20, 30 and 40% butanol. LOC analysis was then employed to investigate differences in the expression of various subclasses of kafirins, soluble at varying concentrations of alcohol, in QL12 compared to 296B (Fig. 2.10). Peak distributions for prolamin extracts solublized in 30 and 40% butanol are visible with increased resolution on corresponding electropherograms generated in LOC (Fig. 2.11). Higher levels of kafirin were extracted in lower concentrations of alcohol in QL12, indicating that this genotype contains a higher relative prolamin content. As expected, β-kafirin was visible in 296B in the 19kD size range of the fraction extracted with 40% butanol. However, due to the null mutation, the peak was not visible in QL12. The complete absence of a peak at 19kD in QL12 indicates that the small peak present in samples extracted previously with 60% tert-butanol (Fig. 2.8/2.9) is only visible at higher alcohol concentrations.
Figure 2.10: LOC gel image of sequentially extracted kafirins from QL12 and 296B in 10, 20, 30 and 40% tertiary butanol.

Figure 2.11: LOC electropherograms of sequentially extracted kafirins from 296B and QL12 in 30 and 40% tert-butanol.
### 2.3.4 Biochemical analyses

The biochemical properties of storage proteins, such as the increased hydrophobicity of the kafirins, impact on the chemical composition of the seed, with subsequent effects on grain end uses. Non-kafirins, including albumins, globulins, and glutelins, are also thought to play a role in starch storage, forming a shell around protein bodies, thus providing structure in the protein-starch matrix (68, 69). These proteins, in addition to the kafirins, influence grain quality, particularly nutritional value, and their further characterisation and association with grain biochemistry is warranted as a means to mitigate grain quality issues. The proximate biochemical composition of the grain across the sample population was determined through quantification of crude protein content, raw and cooked protein digestibility, total starch, and moisture content (Table 2.2/2.3). The data was analysed for correlations to the expression of specific classes of kafirins profiled with HPLC and LOC (Chapter 3, figure 3.2).

<table>
<thead>
<tr>
<th>Genotype/Line</th>
<th>β</th>
<th>γ</th>
<th>δ</th>
<th>Total Protein (%) uncooked</th>
<th>Dryweight protein (%)</th>
<th>% Digestibility uncooked</th>
<th>Total Protein (%) Cooked</th>
<th>% Digestibility Cooked</th>
<th>% Starch</th>
<th>Dryweight Starch (%)</th>
<th>% Moisture</th>
<th>Dryweight</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTx2737/FF-BTx2737</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>12.8</td>
<td>11.5</td>
<td>51.5</td>
<td>14.3</td>
<td>21.2</td>
<td>70.1</td>
<td>64.4</td>
<td>8.1</td>
<td>91.9</td>
</tr>
<tr>
<td>IS22457C</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>11.9</td>
<td>11.0</td>
<td>60.2</td>
<td>13.4</td>
<td>24.0</td>
<td>54.9</td>
<td>50.9</td>
<td>7.3</td>
<td>92.7</td>
</tr>
<tr>
<td>IS12572C-F--6214E</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>10.6</td>
<td>9.5</td>
<td>30.8</td>
<td>14.0</td>
<td>26.0</td>
<td>62.9</td>
<td>56.4</td>
<td>10.4</td>
<td>89.6</td>
</tr>
<tr>
<td>IS12611C-F-F</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>12.8</td>
<td>11.5</td>
<td>40.5</td>
<td>15.4</td>
<td>25.4</td>
<td>66.5</td>
<td>59.7</td>
<td>10.3</td>
<td>89.7</td>
</tr>
<tr>
<td>A14</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>12.7</td>
<td>11.4</td>
<td>45.1</td>
<td>13.8</td>
<td>21.8</td>
<td>62.5</td>
<td>56.0</td>
<td>10.5</td>
<td>89.5</td>
</tr>
<tr>
<td>B923296</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>11.9</td>
<td>11.0</td>
<td>62.1</td>
<td>13.3</td>
<td>22.3</td>
<td>64.0</td>
<td>59.1</td>
<td>7.6</td>
<td>92.4</td>
</tr>
<tr>
<td>C5115</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>11.9</td>
<td>11.0</td>
<td>66.8</td>
<td>13.6</td>
<td>29.6</td>
<td>68.1</td>
<td>62.9</td>
<td>7.6</td>
<td>92.4</td>
</tr>
<tr>
<td>QL39</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>11.0</td>
<td>9.9</td>
<td>40.1</td>
<td>13.3</td>
<td>17.0</td>
<td>67.3</td>
<td>60.9</td>
<td>9.5</td>
<td>90.5</td>
</tr>
<tr>
<td>Karper 669</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12.6</td>
<td>11.4</td>
<td>55.3</td>
<td>14.0</td>
<td>25.6</td>
<td>67.3</td>
<td>60.7</td>
<td>9.7</td>
<td>90.3</td>
</tr>
<tr>
<td>QL41</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>10.7</td>
<td>9.6</td>
<td>51.6</td>
<td>12.5</td>
<td>22.7</td>
<td>68.0</td>
<td>61.4</td>
<td>9.7</td>
<td>90.3</td>
</tr>
<tr>
<td>R931945-2-2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11.6</td>
<td>10.8</td>
<td>62.3</td>
<td>12.9</td>
<td>17.3</td>
<td>66.8</td>
<td>62.7</td>
<td>6.1</td>
<td>93.9</td>
</tr>
<tr>
<td>B98562/ R88562/1-1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>13.2</td>
<td>12.4</td>
<td>64.9</td>
<td>14.9</td>
<td>18.7</td>
<td>67.3</td>
<td>62.9</td>
<td>6.6</td>
<td>93.5</td>
</tr>
<tr>
<td>ISCV746</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11.8</td>
<td>11.0</td>
<td>62.9</td>
<td>13.6</td>
<td>29.6</td>
<td>68.1</td>
<td>62.9</td>
<td>7.6</td>
<td>92.4</td>
</tr>
<tr>
<td>QL12</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>16.5</td>
<td>14.9</td>
<td>40.8</td>
<td>18.1</td>
<td>16.8</td>
<td>64.1</td>
<td>57.9</td>
<td>9.6</td>
<td>90.4</td>
</tr>
<tr>
<td>Hegari/ Early Hegari</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>11.8</td>
<td>10.9</td>
<td>54.4</td>
<td>15.2</td>
<td>29.0</td>
<td>71.0</td>
<td>65.7</td>
<td>7.5</td>
<td>92.5</td>
</tr>
</tbody>
</table>

**Table 2.2:** Percentage total protein, protein digestibility (cooked and uncooked), starch, and moisture content for the sample population of sorghum lines with allelic variation in the kafirins. Values are presented ‘as is’ and on a dryweight basis and represent the mean of duplicate measurements.
<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>γ</th>
<th>δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GU732401</td>
<td>M73688</td>
<td>AY834250 (with 2 possible SNPs)</td>
</tr>
<tr>
<td>2</td>
<td>AJ717660</td>
<td>X62480</td>
<td>AY043223</td>
</tr>
<tr>
<td>3</td>
<td>GU732403 - NULL MUTANT</td>
<td>GU732407</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>GU732404</td>
<td>GU732408</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 2.3: Allelic variant key

2.3.4.1 Protein content and digestibility

Total crude protein content of raw flour ranged between 10-16.5% across the genotypes, and was between 12.1-16.8% in cooked flour (Fig. 2.12). Raw digestibility ranged between 40.8- 68.2% and cooked digestibility ranged between 14.7-33.8% (Fig. 2.13). The data shows an approximate decrease of up to 50% in the digestibility of cooked samples compared to raw flour. Genotype 296B exhibited the lowest crude protein content and the highest raw flour digestibility across the lines, accompanied by a relatively high cooked digestibility. ISCV745 exhibited low raw and cooked protein content and was also highly digestible. Lines with high crude protein content and corresponding low digestibility included KS115 and IS8525. The digestibility of IS8525 was most likely decreased due to the binding of storage proteins by tannins (Fig. 2.6). R9733 and QL12 contain high levels of crude protein, but exhibited above average digestibility, possibly due to a high A/G content (Fig. 2.2), and a high proportion of soluble to residual protein (Fig. 2.4). However, R9733 exhibited a below average cooked digestibility, which may be due to extensive cross-linking due to its high crude protein content.
Figure 2.12: Crude protein content across the kafirin allelic variants

Figure 2.13: Dry matter and cooked digestibility of sorghum whole meal flour across sample population
Protein digestibility was above average across β-kafirin null lines, relative to other lines. Interestingly, the β-kafirin null variant IS17214 exhibited the highest cooked digestibility across the population of allelic variants. QL12 was also highly digestible, but less so than 296B, possibly due to a higher crude protein content. Relatively high A/G content in this line appears to be responsible for the high crude protein measurement. According to RP-HPLC data, QL12 contains low levels of alcohol-soluble protein, which could subsequently account for its relatively high digestibility. Although QL12 exhibits similar crude protein levels to KS115, which has a markedly lower digestibility, the high proportion of A/G to prolamin content, coupled with a concomitant reduction in β-kafirin, resulting from the null mutation, likely accounts for the relatively high raw and cooked digestibility of QL12. Alternately, the high levels of alcohol-soluble protein isolated from KS115, shown with RP-HPLC (Fig. 2.3), likely account for the low digestibility of this line. QL39 exhibits a low crude protein content and low digestibility due to a relatively dense composition of insoluble and residual proteins (Fig. 2.4), which is also reflected in the RP-HPLC prolamin distribution profile for this line (Fig. 2.3). The low digestibility of QL39, similar to KS115, may be due to the increased abundance of insoluble and residual proteins in this line relative to soluble protein content.

A significant correlation was expected, but not observed, between total protein content and raw digestibility ($R^2=0.140$) (p=0.135) across the sample population. However, there did appear to be a weak correlation between total protein and cooked digestibility ($R^2=0.148$) (p=0.035). Outliers contributing to the non-association between protein content and raw digestibility included QL12, which exhibited a high protein content (14.6%) and relatively high raw digestibility (58.4%). As discussed above, the high protein content of QL12 may be due to high levels of A/G proteins, which are more readily digested than kafirins. In addition, QL12 contains reduced levels of β-kafirin due to the null mutation, which would also increase digestibility of the grain. Similar to QL12, outlier R9733 contained high levels of total protein (14.6%), with a relatively high raw digestibility (59.6%). Interestingly, this line had a very low cooked digestibility (14.7%), only slightly higher than that of the tannin-containing line IS8525. The absence of a correlation between total protein content and raw digestibility could also be due in part to the impact of additional factors, such as phytic acid content, which is negatively correlated with digestibility, as well as the activity of several inhibitors, which could also play a role in influencing digestibility (Bean pers. comm).

The probable presence of tannins in IS8525 is associated with both raw and cooked digestibility, while possible binding of tannins in IS12572C may only reduce raw digestibility, with cooked
digestibility at levels comparable to the other genotypes. This data reflects the likelihood that IS8525 may contain higher tannin content than IS12572C. The crude protein content of IS8525 was, however, relatively high compared to the other lines, indicating that the binding of tannins may not affect protein quantification with the pepsin digest.

2.3.4.2 Digestibility of kafirins in sorghum feed

The digestibility of standard sorghum feed was evaluated through extraction of alcohol-soluble proteins from the grain, as well as from ileal tract and fecal matter from swine fed on the grain (Fig. 2.14). Alcohol-soluble proteins separated on 1D SDS-PAGE (methods chapter 3) show the digestion profile of major prolamin storage proteins in stock feed. Substantial digestion of the kafirins occurred in the digestive tract across the trial, made apparent through the large reduction in kafirin isolated from ileal and fecal samples compared to standard feed. There was, however, a significant amount of α-kafirin (22-25kD) remaining in fecal samples following digestion in the gut, possibly due to the presence of substantial amounts of insoluble and residual protein in standard sorghum feed or, alternately, because intestinal tract proteases may exercise substrate preference, as reported in malting studies (132).

![Figure 2.14](image.png)

**Figure 2.14**: Feed trial evaluating protein digestibility in standard sorghum feed in the pig gut. Kafirins were extracted in duplicate from standard sorghum feed, ileal tract and fecal samples, with or without phyto/phosphatase treatment. Phyto-phosphatase treatment controls for the digestion of phytate, which does not occur in monogastric animals. Protein extracts were run on the gel in duplicate (1= Ladder, 2=control BSA, 3/4=Standard feed, 5/6=Control ileal, 7/8=phyt/phos treated ileal, 9/10=control fecal, 11/12=phyt/phos treated fecal).
2.3.4.3 Starch

Significant variation was observed in total starch content (approximately ~15%) across the sample population (Fig. 2.15). Starch levels were relatively low in some lines, such as in genotype IS22457C. Hegari and ISCV745 exhibited the highest total starch content across the sample population. The absence of waxy or floury lines in the study limited the extent of variability in starch composition across the samples and the potential effects on grain quality traits and end-uses, indicating that variability in digestibility were probably not due to variation in starch structure.

Figure 2.15: Total starch measured with a Megazyme assay across kafirin allelic variants

Changes in starch levels did not appear to significantly affect (%) protein content in the grain ($R^2=0.0029$) ($p=0.658$) in this sample population (Fig. 2.16). In addition, total starch content was not correlated to raw ($R^2=0.1405$) ($p=0.433$) or cooked ($R^2=0.0002$) ($p=0.799$) digestibility. This indicates that the majority of differences in biochemical characteristics across the lines could largely be attributed to allelic variation in protein structure and composition, independent of variation in starch.
Figure 2.16: Absence of correlation between total starch and total protein (%) across sorghum lines.

2.4 Conclusion

Profiling of storage proteins across a collection of diverse sorghum lines increases our understanding of grain endosperm development and the structure and composition of the protein-starch matrix, which has significant impacts on starch and amino acid availability. Extensive variation exists in the sorghum gene pool, which can be exploited in breeding programs for improved grain quality. The ability to differentiate cultivars on a DNA and protein basis is important in breeding, marketing and research. The development of chromatographic techniques and capillary electrophoresis for the profiling of grain storage proteins is central to these efforts.

The evaluation of protein content across the sample set was carried out to determine the effects of genetic background on grain quality traits and commercial end uses, with emphasis on protein composition in the β-kafrin null mutants. HPLC and HPCE were effective techniques for screening variation in protein composition at a high level of resolution. Limited profiling of the water-soluble A/G fraction has been carried out previously, therefore optimisation of HPLC and LOC for the analysis of A/Gs was achieved in the course of the study. Improved methods for evaluating A/G composition in cereal grains, such as sorghum, will enhance efforts to develop grain varieties with increased nutritional value, as lines exhibiting an elevated A/G content may contain higher levels of lysine and other essential amino acids.

The impact of storage proteins on grain quality is well documented. However, the precise mechanism controlling their expression and deposition into protein bodies and the roles of specific
protein subclasses, such as β-kafirin, in maintaining the connectivity of the protein starch matrix is still unclear. RP-HPLC profiling revealed a correlation between digestibility and a reduced kafirin content. Evaluation of β-kafirin expression using an integrated proteomic approach has provided functional information about the protein at multiple levels, including the effects of the mutation on chromatographic protein peak profiles, and the identification of potential direct or indirect impacts of the differential expression of β-kafirin on other storage proteins and grain quality traits in general. According to this preliminary chromatographic data, it appears that the null mutation may be associated with reduced levels of prolamins and increased levels of A/Gs in the grain. In this respect, lines expressing low levels of β-kafirin may be appropriate candidates for development of grain varieties with increased lysine content and higher protein digestibility. However, as this trial was carried out in a single growing season, further evaluation of the effects of the mutation on proteomic profiles and commercial traits is required across multiple growing seasons, locations and conditions. The β-kafirin mutation has been further characterised using gel-based techniques and mass spectrometry in the next chapter.

Molecular breeding and biotechnology initiatives based on the discovery of novel proteins and their functional traits will contribute to improving the commercial value of sorghum, particularly by increasing the digestibility and nutritional quality of the grain. Further functional characterisation of storage proteins in sorghum and their associations with grain quality traits will contribute to a greater understanding of grain development and the diversity of storage proteins within breeding populations. The impact of this diversity on commercial end uses, such as bioethanol production, is outlined in chapter 4, where fermentation efficiencies are evaluated across a subset of the sample population of allelic variants.
Chapter 3

Manuscript: Journal of Agricultural and Food Chemistry

Grain sorghum proteomics: Integrated approach towards characterisation of seed storage proteins in kafirin allelic variants

Cremer J.E.¹, Bean S.R.², Tilley M², Ioerger B.P.³, Ohm J.B. ³, Innes D.⁴, Gilding E.K. ⁵, Godwin I.D.¹

¹School of Agriculture and Food Sciences, The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia
²USDA-ARS Centre for Grain and Animal Health Research, Manhattan, KS 66502, USA
³USDA-ARS Cereal Crops Research Unit, Fargo, ND 58102, USA
⁴Department of Agriculture, Fisheries and Forestry, St. Lucia, Brisbane, QLD 4072, Australia
⁵Institute for Molecular Bioscience, The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia

Names are necessary to report factually on available data; however, the U.S. Department of Agriculture neither guarantees nor warrants the standard of the product, and use of the name by the U.S. Department of Agriculture implies no approval of the product to the exclusion of others that may also be suitable. USDA is an equal opportunity provider and employer.
3.1 Abstract

Grain protein composition determines quality traits, such as value for food, feedstock and biomaterials uses. The major storage proteins in sorghum are the prolamins, known as kafirins. Located primarily on the periphery of the protein bodies surrounding starch, cysteine-rich β- and γ-kafirins may limit enzymatic access to internally positioned α-kafirins and starch. An integrated approach was used to characterise sorghum with allelic variation at the kafirin loci to determine the effects of this genetic diversity on protein expression. RP-HPLC and Lab-on-a-Chip analysis showed reductions in alcohol-soluble protein in β-kafirin null lines. Gel-based separation and LC-MS/MS identified a range of redox active proteins affecting storage protein biochemistry. Thioredoxin, involved in the processing of proteins at germination has reported impacts on grain digestibility and was differentially expressed across genotypes. Thus, redox states of endosperm proteins, of which kafirins are a subset, could affect quality traits in addition to the expression of proteins.
3.2 Introduction

Improving agricultural productivity is imperative to feeding an expanding global population set to peak at 9 billion by the year 2050 (175). Diminishing soil quality and increasing demand for limited water reserves, coupled with unpredictable weather patterns associated with climate change are placing mounting pressure on farming systems. Growers are being prompted to adopt more nutrient efficient crop alternatives (176) and to invest in improving drought resistance through conventional and genomics assisted breeding (177). Sorghum produces grain with greater water and nutrient use efficiency, yielding up to 33% more biomass per unit water than maize, a close crop relative (27, 178). However, grain digestibility and nutritional value is less optimal in sorghum compared to maize, due in part to the highly cross-linked nature of cysteine–rich storage proteins in the endosperm (179).

3.2.1 Grain protein biochemistry

The commercial value of cereals is largely determined by the physio-chemistry of the endosperm, and the composition and interactions of protein and starch present there. Soft endosperm sorghum, such as waxy or high lysine varieties, are more digestible in terms of starch and protein (180), but may exhibit increased susceptibility to environmental stresses, including moulds and drought (181). Considerable variation occurs in endosperm protein content across the cereals. Maize, sorghum and millet, of the subfamily Panicoideae, contain around 80% prolamin, whereas rice and oats, of the subfamily Festucoideae, contain greater proportions of albumin and globulin. Barley, wheat and rye, classified in the same subfamily as rice and oats, but grouped in the tribe Triticeae, exhibit greater similarity in protein composition to the Panicoideae, but with a higher albumin/globulin (A/G) content (48).

The kafirins are classified into groups according to various properties, including molecular weight, solubility, structure, and amino acid composition. The β-kafirins (18kD), δ-kafirins (13kD) and γ-kafirins (28kD) are rich in cysteine and methionine. Cysteine enrichment contributes to the intra- and inter- connectivity of the protein matrix surrounding starch (57). The α-kafirins (22 -26kDa) (182) are rich in non-polar amino acids and do not crosslink as extensively, forming mainly intramolecular disulphide bonds. There are 23 members of the α-kafirin family. In maize, 42 members of the homologous 19-22kDa α-zein family have been identified, including the wild-type allele for the floury2 mutation (183, 184). Although extensive sequence homology and functional similarity
exists between sorghum and maize storage proteins, sorghum protein bodies are more highly cross-linked, leading to increased levels of covalent bonding and protein polymerisation in the matrix (48, 57). This increased level of cross-linking reduces overall protein digestibility, which compounds nutritional deficiencies of sorghum, such as low inherent levels of lysine and threonine, and often needs to be supplemented in food and feed (185).

3.2.2 Proteomic analysis

Identification and characterisation of endosperm proteins and the elements regulating their synthesis, localisation and degradation facilitates enhancement of amino acid content and starch accessibility for improved grain nutritional value (1). This study involves the evaluation of grain protein composition in sorghum lines with allelic variation in kafirin storage proteins using a range of proteomic tools, including reversed-phase HPLC (RP-HPLC), Lab on a Chip (LOC), SDS-PAGE, and liquid chromatography-tandem mass spectrometry (LC-MS/MS). The focus of the work is to identify candidate proteins, which may impact on quality traits, particularly those which are differentially expressed across the genotypes.

Protein structure and composition is influenced by a diverse set of properties, including charge, size and hydrophobicity. Therefore, a combination of biochemical techniques is required for complete proteome analysis (186). Multi-dimensional gel electrophoresis (2DE), coupled with gel-free methods such as liquid chromatography, allow for fine separation of proteins according to differences in molecular weight, charge density, and surface hydrophobicity. Structural and biochemical information describing isolated proteins is generated through tandem mass spectrometry and subsequent bioinformatic analysis, where mass and sequence data is mapped to annotated protein databases (119).

3.3 Methods

3.3.1 Plant Material

Twenty eight inbred sorghum lines, with previously determined allelic variation in the β-, γ-, and δ-kafirin storage proteins were included in the study (167). The primary focus of the analysis was on comparisons between the β-kafirin null line QL12 and in 296B, an important food line of Indian
origin. Plants were cultivated under field conditions at the University of Queensland, Gatton campus, QLD, Australia, over the 2008/2009 and 2011/2012 summer growing seasons. Genotypes 296B and QL12 were sampled and analysed across both growing seasons, while data for the entire sample population was generated from grain grown in the second season only.

3.3.2 Protein Extractions

3.3.2.1 Albumin/globulins (A/G)

Wholegrain flour (100mg) prepared from kernels ground in liquid nitrogen using a mortar and pestle was mixed in 1mL extraction solvent (50mM Tris-HCl pH 7.8, 100mM KCl and 5mM EDTA). The sample was vortexed and centrifuged and a 500µL aliquot of supernatant was removed to a new tube (169). The process was repeated by adding an additional 1mL extraction solvent to the pellet, repeating the extraction process and removing an additional 500µL supernatant to the same tube.

3.3.2.2 Prolamins

The A/G pellet from the above extraction procedure was retained, washed with 1 mL dH2O, and 1mL solvent was added (60% tertiary butanol, 0.5% sodium acetate w/v and 2% β-mercaptoethanol v/v). The pellet was then vortexed and centrifuged and 500µL supernatant was transferred to a new tube. The procedure was repeated once and a further 500µL supernatant was removed to same tube.

3.3.2.3 Glutelin/Residual Proteins

The pellet from the prolamin extraction was washed with 1mL dH2O and 1ml sodium borate buffer (125mM pH 10.0 containing 1%SDS w/v and 1% βME v/v) was added to the pellet. The suspension was vortexed and centrifuged and 500μl supernatant was transferred to a new tube. The procedure was repeated once and a further 500μl supernatant was removed to same tube.
3.3.3 Reversed-Phase HPLC (RP-HPLC)

Protein samples were analysed using an Agilent 1100 HPLC system (Agilent, Foster City, CA), fitted with a Poroshell column of varying stationary phases (as indicated below for the various protein fractions). The system employed a binary gradient with a constant flow rate of 0.7mL/min and a column temperature was maintained at 55°C. Solvents for RP-HPLC included water containing 0.1% trifluoroacetic Acid (TFA) w/v (A), and acetonitrile (ACN) containing 0.07% TFA w/v (B), with gradient flow specifications as follows: 0-18min, 45%-60% B; 18-19min, decreased to 45% B; followed by a 7min post run. Absorbance was measured with a UV detector at 214nm.

3.3.3.1 Albumin/globulins

After extraction, 250μL A/G aliquots were freeze dried in a speed vac overnight. Prior to RP-HPLC analysis, the aliquots were resuspended in 100μL 50% ethylene glycol to stabilise proteins (168). Samples (5μL) were analysed using a 2.1x75mm Poroshell 300 SB C3 column.

3.3.3.2 Prolamins

After extraction, protein samples (1 mL) were alkylated by adding 33μL 4-vinylpyridine and vortexed for 10 min. Samples (5μL) were then injected directly for RP-HPLC analysis using 2.1x75mm Poroshell 300 SB C18 column (169). Proteins were detected by UV at 214 nm.

3.3.4 Lab on Chip (LOC)

The Lab on a Chip procedure was carried out on the Agilent 2100 Bioanalyser using the Protein 80 assay kit (Agilent Technologies, Palo Alto, CA). A 1:1 mixture of protein to denaturing buffer was used.
3.3.5 Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS PAGE)

3.3.5.1 Protein sample cleanup

Samples were cleaned and concentrated with a 2D Cleanup kit (GE Healthcare Ltd, Little Chalfont, Buckinghamshire, UK) according to manufacturer’s instructions. Proteins were precipitated to remove detergent, salts, lipids, phenolics, nucleic acids, and other contaminants.

3.3.5.2 Protein Quantification

Protein sample concentrations were obtained using a 2D quant kit (GE Healthcare Ltd, Little Chalfont, Buckinghamshire, UK) according to manufacturer’s instructions. The method involves the specific binding of copper ions to precipitated protein, where unbound copper is measured with a colourimetric agent. Absorbance measured at 480 nm is inversely proportional to protein concentration.

3.3.5.3 One dimensional SDS-PAGE

A 10µL sample was resuspended in 5µL 3X loading buffer (100mM Tris, 1% SDS (w/v), 0.01% Bromophenol blue, 15% glycerol, 0.05% β-ME), and heated to 95 °C, then chilled on ice and loaded onto a small format precast anyKd™ (10-250kD range) gradient gel (BioRad, Hercules, CA), with 15µL ColorPlus prestained marker (broad range 7-175kDa) added to the first and last wells. Gels were run in SDS buffer (25mM Tris, 200mM glycine, 0.1% SDS w/v) at 200V for 40min until the loading dye front reached the bottom of the gel. Gels were fixed in 50% MeOH, 7% acetic acid for 1hr with shaking, then stained in ~50mL Coomassie stain overnight. Gels were destained in 1% acetic acid for 30min, then rinsed with dH₂O and scanned.

3.3.5.4 Isoelectric focusing (IEF)

A rehydration solution containing 100µg protein was prepared to a total volume of 125µL (20µL protein [conc. 5 µg/ µL], 100µL rehydration buffer ([7M urea, 2M thiourea, 2% CHAPS, 0.5% (v/v) IPG Buffer, 0.0002% Bromophenol blue], 4µL fresh 1M dithiothreitol (DTT), 1.25µL carrier ampholytes [pH 3-11]). The sample mixture was loaded onto 7cm IPG strip (3-11 NL small strips),
placed in a carrier coffin and 0.8mL dry strip cover fluid was applied to minimise evaporation and urea crystallisation. IEF was run for a total of ~16000 Volt Hours at 20 °C on the IPGphor with the following program settings: 30V for 14hrs (rehydration), 100V for 2hrs (step n hold), 500V for 1.5hrs (step n hold), 1000V for 1.5hrs (step n hold), 5000V for 1hr (gradient), 5000V for 2 hrs (step n hold) and 50V (hold). It was ensured that the voltage reached at least 5000V to achieve a standard protein gradient. Strips were removed from coffin and either used immediately for 2D SDS-PAGE gel analysis or stored at -20 °C until required.

### 3.3.5.5 Two-dimensional SDS-PAGE

IPG strips were saturated with MES running buffer [50mM MES, 50mM Tris base, 0.01% SDS w/v, 1mM EDTA], 6M urea, 30% glycerol, 2% SDS w/v and 0.001% bromophenol blue) by first equilibrating in MES buffer plus 50mg/mL (DTT) for 15 min with shaking to reduce disulphide bonds, then soaking strips in MES buffer plus 125mg/mL iodoacetamide (IAA) for 15 min with shaking to alkylate the free cysteine residues. IPG strips were fitted into Bis/Tris small format precast gels (15%) (Invitrogen, Carlsbad, CA). The loading well was then sealed with ~0.8mL agarose sealing solution (0.5% agarose, 0.001% bromophenol blue). The gel was run in 1X MES running buffer at low voltage (25-50V) for 30 min, then voltage increased to 125V and run until loading dye front disappeared off the bottom of the gel (approx 1.5-2hrs). Gels were fixed in 50% MeOH, 7% acetic acid for 1hr with shaking, then transferred to a Sypro Ruby silver stain solution overnight with shaking and destained for 30 min (10% MeOH, 7% acetic acid) prior to scanning using a Typhoon 9400 (GE Healthcare). The gel was then stained in coomassie overnight with shaking to visualise spots for excision from gel and subsequent mass spectrometric analysis. Coomassie stained gels were destained in 1% acetic acid and scanned with Odyssey at 700nm.

### 3.3.6 Mass fingerprinting (LC-MS/MS)

#### 3.3.6.1 Reduction/alkylation of gel pieces

Protein spots were manually excised from gels and destained in 50mM ammonium bicarbonate (ABC)/50% acetonitrile (ACN), 2 x 200µL with shaking for 3-4 hrs or overnight. For 1D gel pieces, buffer/destain was removed from gel pieces with pipetting and the sample was soaked in 40µl 10mM DTT to reduce cysteine residues. Gel pieces were incubated at 60 °C for 30 min, DTT solution was removed and 40µL of 55mM fresh IAA was added. Samples were then incubated at
room temperature for 30 min in the dark. IAA was discarded and (for 1D and 2D gel pieces), 100µl of 50mM ammonium bicarbonate (ABC) was added, with vortex/shake 1-2 min, removal of solution and repeated wash. Gel pieces were then dehydrated in 50µl 100% ACN for 5 min, with vortexing.

3.3.6.2 Enzymatic digestion and peptide extraction

Gel pieces were rehydrated in a 50mM ABC solution containing 8µL trypsin (10ng/µL) (Sequencing-grade modified trypsin, Promega, Madison, WI) for 10-20 min at 4 °C. An additional volume (6-16µL) 50mM ABC buffer was then added, depending on size of gel piece and samples were incubated overnight at 37 °C. Then, 50µL 50% ACN/0.1% TFA was added to gel pieces and samples sonicated in a water bath for 10 min, centrifuged briefly and supernatant transferred to a new tube. An additional 50 µL aliquot of 50% ACN/0.1% TFA was added, sonication repeated and supernatant combined with the first extract. Supernatant was lyophilised in a speed vac at 45 °C and peptides were resuspended in 10µL 5% ACN/0.1% TFA. A Ziptip® cleanup was carried out according to manufacturer’s instructions (Merck Millipore, Darmstadt, Germany) for removal of acrylamide contamination prior to MS analysis.

3.3.6.3 Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS)

Digested peptides were run on LC-MS/MS using an ESI-QTOF instrument. MS parameters were similar to that in Kappler and Nouwens (2013) (187), with the following modifications: Samples were desalted for 5 min on an Agilent C18 trap (0.3 x 5mm, 5 um), followed by separation on a Vydc Everest C18 column (300A, 5 um, 150 mm x 150 um) at a flow rate of 1 ul/min, using a gradient of 10 – 60% buffer B over 30 min, where buffer A = 1% ACN/0.1% formic acid and buffer B = 80% ACN/0.1% formic acid. Eluted peptides were directly analysed on a TripleTof 5600 mass spectrometer (ABSciex) using a Nanospray III interface. Gas and voltage were set as required. MS TOF scan across m/z 350 – 1800 was performed for 0.5 s, followed by data-dependent acquisition of 20 peptides with intensity above 100 counts across m/z 40-1800 (0.05 s per spectrum) with rolling collision energy. MS data was converted to mascot generic format and submitted to MASCOT.
3.3.6.4 MS Data Analysis

The automated Mascot search engine (Matrix Science, London, UK) was used to identify best matched protein sequences to the peptides detected by MS. Trypsin was specified as the proteolytic enzyme and carbamidomethyl (C) of cysteine and oxidation (M) of methionine residues was taken into account. Charged states of 2+, 3+ and 4+ were considered for parent ions. Using a Ludwig-based search of plant species, a profile of best matched protein(s) was generated employing an algorithm to rank the proteins identified based on their peptide mass fingerprints. Individual ion scores were calculated as \(-10 \times 10^g(P)\), where P is the probability that the observed match is a random event. Peptides identified with an ions score > 40 were considered to indicate identity or extensive similarity. Each of the peptide identifications was manually inspected and verified to ensure that the spectra was composed of a wide series of intense fragments, which could be designated to major fragments (b or y) of the proposed peptide. Each protein match also generated a protein score in the MS/MS search report, which is the sum of the highest ions scores for each distinct sequence. The exponentially Modified Protein Abundance Index (emPAI) provided an approximate, relative quantification of the proteins present in the sample. Functionally characterised homologs to putative uncharacterised sorghum proteins identified with MS were identified through BLAST searches based on FASTA sequences derived through queries to the Uniprot database.

3.3.7 Statistical Analysis

Absorbance values from RP-HPLC of alcohol soluble proteins were interpolated by a spline method and absorbance area values were calculated to 0.01-min intervals using MATLAB (The MathWorks, R2012a, Natick, MA). Linear correlation coefficients were calculated between individual absorbance area and digestibility and shown as a continuous spectrum over RP-HPLC retention time.
3.4 Results and Discussion

3.4.1 RP-HPLC profiling of grain proteins

Profiling of alcohol-soluble prolamin (Fig. 3.1a and Supporting Information, fig. S1) and water/salt-soluble albumin/globulin (A/G) (Fig. 3.1b and Supporting Information, fig. S2) protein fractions across 28 sorghum lines using RP-HPLC revealed a range of diversity, particularly in the alcohol-soluble kafirin-containing fraction. Variation observed across wild-type and mutant lines could contribute to phenotypic differences, such as increased flour viscosity and fermentation efficiency (188, 189). Among the β-kafirin null lines QL12, IS17214 and RTx2737, a high degree of similarity was observed in the alcohol-soluble peak distributions compared to lines with functional β-kafirin alleles. An approximately 50% reduction in the height of a peak eluting at 9 min was observed in β-kafirin null lines compared to wild-type (Fig. 3.1a, highlighted in orange), indicating a likely elution profile for β-kafirin (under the analysis conditions described here) and presenting a potential marker for the protein. Alternately, the β-kafirin peak could have been masked by a series of larger peaks eluting from the column at 10-11 min, which were absent in null lines (Figure 3.1a, highlighted in blue).

Further work to positively identify the β-kafirin in RP-HPLC separations under these exact conditions is needed. RP-HPLC analysis of the water/salt-soluble fraction showed less similarity among peak profiles for the β-kafirin null mutants relative to other lines. However, substantial quantitative differences in peak heights were observed among the genotypes and there was a high level of variability across the sample set in a peak eluting at 9 min, which was unrelated to fluctuations in total sample protein concentration.
Figure 3.1: RP-HPLC profiles for the A) alcohol-soluble (prolamin) fraction and B) water/salt-soluble (albumin/globulin) fraction across β-kafirin null allelic variants QL12, 1517214 and RTx2737, and lines with normal β-kafirin content (including B35, 296B and KS115). In the prolamin fraction β-kafirin null lines showed a reduction in the size of a peak eluted at 9 minutes (outlined in orange). Peak distribution profiles at 10-11 minute elution times were similar across lines with normal β-kafirin content, while a significant proportion of protein in this area of the chromatogram was missing in the β-kafirin null mutants (outlined in blue).

Interestingly, a statistically significant negative correlation was identified between a set of protein peaks eluted in the kafirin fraction and protein digestibility across sorghum grain lines, providing evidence for links between seed protein composition and end-use traits (Figure 3.2). Further research is needed to identify the protein(s) present in this peak, however, given the sequential extraction procedure utilized in this work, it is likely that this is a member of the kafirin family.
Figure 3.2: RP-HPLC analysis showing a significant negative correlation (outlined in red) between seed protein digestibility (%) and the alcohol-soluble kafirin fraction peak eluting at 5.5-6 min on the chromatogram. Analysis included 26 diverse sorghum lines (Table 2.1), excepting two possible tannin-containing lines (IS8525 and IS12572C).

3.4.2 Lab on Chip (LOC) analysis of storage proteins

LOC analysis, employing microfluidic chip-capillary electrophoresis, generated size-separated protein profiles for alcohol-soluble (prolamin) (Fig. 3.3) and water/salt-soluble (A/G) fractions (Supplementary Information, fig. S3). The prolamin fraction showed a large set of peaks present in the 22-26kDa size range, representing the previously characterised α- and γ-kafirins. Small peaks were also visible at 11 and 19kDa, with the 19kDa representing the β-kafirins and the 11kDa potentially the δ-kafirins. The β-kafirin peak was diminished in size in β-kafirin nulls, but, interestingly, a peak or set of peaks was visible at 19-20kDa across all lines. This indicates that an additional protein of a similar size to β-kafirin may be present in the peak. This protein could represent a different protein, such as peptidyl-prolyl cis/trans isomerase (PPIase), identified in the same band/spot as β-kafirin through SDS-PAGE and LC-MS/MS, as outlined in the next section. Chip profiles indicate that several cultivars in addition to the β-kafirin null mutants contain relatively low levels of β-kafirin, including 296B, a highly digestible line (189).
Figure 3.3: Alcohol-soluble protein fraction visualised on Lab on Chip across β-kafirin null lines QL12, 1517214 and RTx2737, and normal β-kafirin lines 296B and M35. A 19kD β-kafirin peak (highlighted in orange) precedes a large 22-25kD α- and γ-kafirin peak and a high Mw prolamin is just visible across the genotypes at ~50kD (highlighted in blue).

LOC analysis also revealed the presence of a small peak visible at 46kDa in alcohol-soluble samples, which could represent a high Mw kafirin or kafirin dimer. In subsequent analysis, a protein spot extracted from the corresponding size range on 2D gels was identified through LC-MS/MS as a ~36kDa γ-prolamin homolog (Supporting Information, fig. S4 and table 1), thus providing sequence information for this kafirin entity at the protein level. As observed with RP-HPLC, the LOC profiling of the water/salt-soluble fraction shows that QL12 contains a relatively high content and diversity of A/Gs compared to other genotypes (Supporting Information, fig. S3). High-lysine sorghum varieties, such as P721, exhibit an elevated A/G content, with increased nutritional value (79). This indicates that QL12, among other lines, may contain increased levels of lysine and other essential amino acids due to a high A/G content.
3.4.3 SDS PAGE and LC-MS/MS protein profiling

Profiling of water/salt-soluble (A/G), alcohol-soluble (prolamin), and alkali-soluble (glutelin) protein fractions across the β-kafirin null mutant QL12 and wild-type line 296B using SDS-PAGE gel separation methods, coupled with mass spectrometry (LC-MS/MS) (Fig. 3.4 and Supporting Information, fig. S4/table 1) identified a range of proteins involved in storage protein biochemistry and starch metabolism. One-dimensional gel separation of each of the protein fractions by size (Fig. 3.4), showed that the A/G fraction contains a highly concentrated, complex array of proteins present across a broad size range, similar to profiles generated in previous studies (190). The prolamin fraction was less complex and more concentrated within the 23-26kDa size range, which contains the more abundant α-kafirin storage proteins. The alkali-soluble glutelin fraction exhibited some similarities to the prolamin fraction, likely due to incomplete extraction of the prolamin fraction, but contained a wider diversity of protein bands visible across a greater size range, which was also verified on 2D gels, and has been reported in previous studies (191, 192). It is hypothesised that the glutelin proteins are more diverse because they play both a role in connecting the protein:starch matrix as well as providing a source of hydrolytic enzymes for the breakdown of starch and protein during germination (193).

3.4.4 Characterisation of kafirin subclasses and the β-kafirin null mutation

Each kafirin subclass in the alcohol-soluble fraction was isolated using SDS-PAGE and identified through LC-MS/MS. α- and γ-kafirin were visualised within the 22-25kDa size range on the 1D SDS-PAGE gel (Fig. 3.4, highlighted in white). Using 2D SDS-PAGE separation of proteins by both size and charge, the δ-kafirin protein was isolated in the prolamin fraction for genotype 296B (spot 11) (Supporting Information, fig. S4/table 1). Matches to α- and γ-kafirins were also generated on 2D gels in prolamin gel spots 8 and 11 for genotype QL12, and with spot 9 returning α-kafirin as a top hit. Differential expression of the β-kafirin protein was detected in the alcohol-soluble fraction across mutant and wild-type germplasm. This provided a positive control for the MS technique and allowed for further characterisation of the mutant at the protein level. 1D separation of the prolamin fraction revealed altered expression of β-kafirin in a 19kD protein band, present in 296B and absent in QL12 (Fig. 3.4, highlighted in white), as predicted based on previous molecular characterisation (167). On 2D gels differential expression of β-kafirin was identified in 296B prolamin spot 11 and glutelin spot 4 (Supporting Information, fig. S4/table 1). Identification of β-kafirin in the glutelin fraction (296B spot 4) may reflect the incomplete sequential extraction of
alcohol-soluble proteins in the previous step (194). β-kafirin was completely absent in QL12, with the exception of a weak hit in the 2D prolamin spot 1, which had a low Mascot score achieved across two peptide matches and was discarded due to poor quality spectral data. This match may represent a truncated form of the protein, as it was identified in a smaller size range on the gel.

Figure 3.4: One-dimensional SDS-PAGE separation of water/salt soluble albumin/globulins, alcohol-soluble prolamins, and alkali-soluble glutelins from normal β-kafirin allelic variant 296B, and β-kafirin null QL12. Excised protein bands (outlined in white) were digested with trypsin and analysed using MALDI TOF MS (listed in Supplementary Information, table 1). Ladder: P7709V ColourPlus Prestained Protein Marker, Broad Range (7-175kDa) mW ladder.

Although protein expression was generally similar across the genotypes within the range of spots analysed on 2D gels, differential expression of a low Mw thioredoxin (Trx) enzyme was identified in the water/salt-soluble fraction, where spot 4 was present in 296B and absent in QL12. This was the only Trx identified across the protein fractions analysed in the study, with the finding replicated in triplicate (Fig. 3.5 and Supporting Information, table 1). Trx catalyse the conversion of seed proteins from the oxidised to the reduced state during germination, with significant impacts on protein digestibility and grain nutrition.
Figure 3.5: Two-dimensional SDS-PAGE analysis of the water/salt-soluble protein fraction in β-kafirin null allelic variant compared with wild-type 296B. Protein samples were loaded onto 7cm IPG strips (3-11 NL) and run on IPGphor machine for isoelectric focussing. SDS-PAGE gels (4-12% Bis/Tris small format precast) were utilised for separation of proteins by size (Mw). Protein spots (circled in yellow) were excised across a range of size and pI, digested with trypsin and identified with LC-MS/MS. Spot 4 was identified as a differentially expressed thioredoxin, present in 296B and absent in QL12.

3.4.5 Sorghum 50kD γ-prolamin homolog

High Mw (HMW) γ-prolamins (~50 kDa) have been identified and characterised across a range of plant species, including maize and wheat (195, 196). Previous studies on sorghum have identified a protein band at ~45kDa in the alcohol-soluble fraction using 1D SDS-PAGE. However, this protein has not yet been characterised at the genetic level (167). The HMW protein band was previously reported as a kafirin dimer because it diminishes in intensity upon treatment with increasing concentrations of reducing agent, indicating that it can be broken down into smaller peptides (48, 49).
However, because the original 45kDa band is still intact on reduced gels, it has been suggested that the protein may be linked by bonds that are not broken by reducing agents (132).

In the present study, this HMW protein was visible in reduced alcohol-soluble protein samples from both genotypes QL12 and 296B, on 1D and 2D gels. LC-MS/MS data indicates that this protein represents a homolog to HMW γ-prolamin identified in closely related species (Supporting Information, fig. S4/table 1). Across both genotypes, 2D prolamin spot 4 returned a match for a putative uncharacterised sorghum protein (accession C5XDK9), exhibiting homology to 50kDa γ-zein, γ-canein and γ-coixin (Fig. 3.6). The sorghum γ-prolamin homolog is larger than any previously characterised kafirin, with a calculated Mw of 36,615 daltons. LOC analysis of the prolamin fraction revealed the presence of two small peaks at ~44 and 46kD across the majority of the sample collection (Fig. 3.3, highlighted in blue), which may correspond to two genetic variants of the HMW γ-prolamin (48, 195). In a previous study using LOC, similar peak profiles were also observed for the HMW prolamin peak in reduced alcohol-soluble protein samples extracted from three grain genotypes (QL12, QL41, and 296B) from the kafirin allelic variant sample set, which had been harvested in a previous growing season (198).
Figure 3.6: Phylogenetic neighbour joining tree view of the alignment of protein sequences representing prolamin subclasses across grass species using the Grishin protein method. Putative uncharacterised sorghum γ-prolamin homolog groups with 50kD γ-zeins and γ-caneins of maize and sugarcane, separately from smaller Mw γ-prolamins.
The general structure and distribution of prolams in the protein bodies has been shown to be uniform across maize and sorghum. In maize, 50kD γ-zeins localise to the periphery of the protein bodies, similar to other γ-zeins (195). Immunocytochemistry shows that γ-kafirins also localise to the periphery of protein bodies, and may prevent proteolysis of internally positioned α-kafirins (76). Sorghum lines with increased digestibility exhibit a change in the protein body structure from spherical to invaginated, with the γ-kafirins located from the periphery to the folds of the structure, resulting in an increased exposure of the α-kafirins to proteolytic breakdown (53).

In maize, a clear physical distribution of zein classes is observed within the seed. Protein bodies in the sub-aleurone layer are smaller and contain mainly β- and γ-zeins, while those encapsulating starch in the inner endosperm are larger and contain continuous central regions of α-zeins with β- and γ-zein located on the periphery (199). Construction of mutant γ-zein proteins has pinpointed the proline-rich N-terminal domain as being critical in wild-type protein body development (200). Furthermore, deletion of a cysteine-rich γ-zein domain results in abnormally structured protein bodies.

The different kafirin classes exhibit variable solubilities according to their degree of polymerisation and cross-linking (197). Interactions between β- and γ- kafirin on the periphery of protein bodies may limit enzyme accessibility to α-kafirin, impeding digestion of protein and starch encased in the matrix. This study presents evidence for the seed-specific expression of a high Mw γ-prolamin homolog in sorghum. Characterisation of this protein in sorghum could enhance efforts to further distinguish the roles of the different classes of kafirins in maintaining endosperm connectivity and access to starch.

### 3.4.6 Identification of proteins with potential impacts on grain quality

Profiling of water/salt-, alcohol- and alkali-soluble proteins across the sample population of kafirin allelic variants using mass spectrometry identified proteins with potential various impacts on the structure of the protein-starch matrix, such as enzymes functioning in protein cross-linking and the mobilisation of starch during germination (108, 201) (Table 3.2, Figs. 3.5, 3.6 & 3.7 and Supporting Information, fig. S4 and table 1). Identification and localisation of these proteins in the mature grain indicates they may have an impact on endosperm development and/or germination. Deciphering the precise activities of these proteins in sorghum will involve additional analysis of the transcriptome and proteome throughout grain development.
<table>
<thead>
<tr>
<th>Accession</th>
<th>Protein</th>
<th>296B spot</th>
<th>QL12 spot</th>
<th>Protein Fraction</th>
<th>Homologs</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5XB72/ C5YGL1 (sorghum)</td>
<td>Thioredoxin</td>
<td>4</td>
<td>absent</td>
<td>A/G</td>
<td>Maize B6SX54</td>
<td>Converts seed proteins from the oxidised to the reduced state during germination</td>
</tr>
<tr>
<td>C5YBX1 (sorghum)</td>
<td>Glutaredoxin Grx_C2.2 – glutaredoxin subgroup 1</td>
<td>9 &amp; 10</td>
<td>5 &amp; 6</td>
<td>A/G</td>
<td>Maize B6THA1</td>
<td>Converts proteins with reduced sulphide groups to those with oxidised disulphide bonds</td>
</tr>
<tr>
<td>F2DDK (Barley)</td>
<td>Periredoxin</td>
<td>5</td>
<td>5</td>
<td>Prolamin</td>
<td>Maize HB6T2Y1 &amp; rice PR2E1</td>
<td>Interacts with glutaredoxins, thioredoxins and cyclophilin as both reductants and non-dithio-disulphide exchange proteins</td>
</tr>
<tr>
<td>C5XT06, B3GQV9 and C5Z9C6 (sorghum)</td>
<td>Peptidyl-prolyl-cis/trans isomerase (PPIase)</td>
<td>9, 10 &amp; 11</td>
<td>9 &amp; 10</td>
<td>Prolamin/ Glutelin</td>
<td>Maize B4FZZ2 &amp;B4FY3T, wheat AZLM55 Sugar Cane C7E3V7</td>
<td>Facilitates protein folding through slow isomerisation of peptide bonds in oligopeptides and through the amino acid proline in cellular proteins</td>
</tr>
<tr>
<td>A5A5E7 (maize)</td>
<td>Protein disulphide Isomerase (PDI)</td>
<td>2</td>
<td>2</td>
<td>Glutelin</td>
<td>Maize C0PLF0 &amp; A5A5E7</td>
<td>Promotes the correct disulphide pairing in proteins</td>
</tr>
<tr>
<td>P81368 (sorghum)</td>
<td>α-amylase inhibitor</td>
<td>1, 2, 8 &amp;11</td>
<td>1 &amp; 2</td>
<td>Prolamin</td>
<td>Sorghum IAA5_SORI</td>
<td>Slows the conversion of starch to sugars (and ethanol)</td>
</tr>
</tbody>
</table>

Table 3.1: Protein Candidates Identified in Sorghum Grain with Reported Effects on Protein-Starch Matrix Structure, Grain Quality and Stress Responses

### 3.4.6.1 Thioredoxins (Trx) and Glutaredoxin (Grx)

Trx are small proteins containing a site with a redox-active disulphide, which functions in the reversible oxidation of protein SH-groups to a disulphide bridge (202) (Table 3.1). Differential expression of a sorghum Trx protein accession C5XB72, homologous to maize Trx B6SX54, was observed in the A/G fraction, with 2D protein spot 4 present in 296B and absent in QL12 (Figure 3.5). Matches to sorghum Trx had a high emPAI protein quantification score, and a pI of 5.79 and Mw 13061 Da. Trx labelling studies have shown *in vitro* and *in vivo* that the enzyme catalyses the reduction of seed proteins during germination (203). Therefore, Trx has been linked to enhanced grain digestibility in wheat and sorghum (109, 110). Alternate expression of Trx in this study may
be associated with the β-kafirin null mutation and could have potential downstream impacts on digestibility, in addition to effects of the mutation.

The sorghum homolog to a maize glutaredoxin Grx_C2.2 – glutaredoxin subgroup 1 was identified in the A/G fraction in 296B spots 9 and 10, and in QL12 spots 5 and 6, which localised to the same size range (Mw ~13kDa) (Figure 3.5). Grxs, including the maize Grx_C2.2 identified in this study, catalyse the enzymatic reaction where proteins with reduced sulphide groups are converted to those with oxidised disulphide bonds (Table 3.2). Grxs have been implicated in the oxidative stress response through regeneration of enzymes involved in peroxide and methionine sulfoxide reduction (204). These proteins therefore have dual roles in protein aggregation and stress responses, creating opportunities for the development of streamlined grain improvement strategies affecting multiple traits.

3.4.6.2 Protein Disulphide Isomerases (PDI)

PDIs function as molecular chaperones in disulphide-mediated protein folding. They contain two Trx domains with a redox site (205). Matches to HMW PDI from maize (C0PLF0 and A5A5E7) were generated in glutelin spot 2 across both genotypes (Supporting Information, fig. S5/table 2). These proteins appear to migrate to the same location on A/G and prolamin gels, but the spots were not analysed in this fraction using LC-MS/MS. However, it is possible that they share the same identity as the PDIs identified in the glutelin fraction because their distribution on the gel is similar. Again, the presence of these spots across multiple fractions may be a result of incomplete sequential extraction of the proteins.

Previous studies indicate that PDI mutations can result in irregular starch granule formation and chalky grain phenotypes (206). The rice mutant esp2 lacks protein disulfide isomerase-like (PDIL)1:1, but shows enhanced expression of the thiol disulphide oxidoreductase OsEro1 (207). The grain exhibits altered seed storage protein compartmentalisation through inhibition of disulfide bond formation. It has been proposed that the formation of native disulfide bonds in proglutelins also depends on an electron transfer pathway involving the OsEro1 and the PDI-like OsPDIL (208). In the flouncy2 mutant, PDIL is up-regulated, where a mutation in a signalling peptide results in the abnormal processing and accumulation of small precursor α-zeins and high levels of luminal binding protein (BiP) in irregularly shaped protein bodies (209). Mutations in sorghum protein isomerases, such as PDI, may have a similar effect on protein body formation.
3.4.6.3 Peptidyl-prolyl cis/trans isomerases (PPIase)

PPIases catalyse protein folding through isomerisation of peptide bonds in oligopeptides and on proline residues in cellular proteins (205, 210, 211). PPIases were identified in the 19kD band co-localised with β-kafirin on the 296B 1D prolamin gel, as well as in the 26kD band, co-localised with α-kafirin (Fig. 3.4). On 2D gels, PPIases were detected in the 296B prolamin fraction gel spots 9, 10 and 11, and in QL12 gel spots 9 and 10, as well as in the QL12 glutelin fraction spot 4 (Supporting Information, fig. S4/table 1). Sorghum PPIases identified in the study (accessions C5XT06, C5Z9C6 and B3GQV9), consistently localised to the same size region (~18kD) on 1D and 2D PAGE gels, and within the same pl range pH 8-9. PPIases co-localised with β- and δ- kafirin in the 296B 2D prolamin gel spot 11 and with α- kafirin in QL12 spot, as well as in the prolamin fraction on the 1D SDS-PAGE gel, indicating that they are hydrophobic enzymes, which may interact with the kafirins and/or be present in the 19kD β-kafirin peak visualised with RP-HPLC and LOC.

The PPIase family is encoded by multiple genes, exhibiting some redundancy in plants (212). An alignment of the sorghum PPIases identified in this study, with orthologs from closely related species, reveals a high level of sequence homology at the protein level (Fig. 3.7). The sorghum PPIases exhibited close homology to their counterparts in Zea mays (B4FZZ2 and B4FY3T), Triticum aestivum (AZLM55), Saccharum officinarum (C7E3V7), Citrus sinensis (D0ELH5) and Gerbera ABCYN7. PIN1-type PPIases are encoded by multiple genes and are present in a wide variety of plant species. A four amino acid insertion site is situated next to the phosphor-specific recognition site of the active site, regulating the activity of the enzyme (213). The PPIase activity of cyclophilins is regulated by thioredoxin (214), also identified in this study and discussed above. PIN1At in Arabidopsis encodes a PPIase, which regulates flowering time through phosphorylation-dependent prolyl cis/trans isomerisation of key regulatory pathway specific transcription factors (215). Alterations to protein folding status catalysed by PPIases with the the cis/trans isomerisation of proline imidic peptide bonds could also affect protein body composition in sorghum and other grain crops (158).
Figure 3.7: ClustalW alignment of peptidyl-prolyl cis/trans isomerase (PPIase) peptide sequences across major grass species, including *zea mays*, *oryza sativa* and *brachypodium*, illustrates a high level of sequence homology. PPIases identified in sorghum grain prolamin and glutelin protein fractions (figure S3/supplementary table 1) are included. Dark purple indicates exact homology, light purple indicates highly conserved homology.
3.4.6.4 Heat shock protein (HSP)/ Luminal Binding Protein (BIP)

HSP/BIP chaperones are key regulators of protein body biogenesis, with significant impacts on grain quality traits and composition (216) (Table 3.2). HSP/BIPs were isolated predominantly from the water/salt- (A/G) and alkali-soluble (glutelin) protein fractions across both genotypes (Supplementary Information, fig. S4/S5 and tables 1 & 2). A/G spot 2 produced a strong match to sorghum protein (accession C5XY25), which exhibits homology to the *Oryza sativa* 19 kDa class II HSP. The protein has a calculated Mw 19982 Da and pI 5.71, appropriate to its position on the gel. A sorghum homolog to a maize 16.9 kDa class I heat shock protein was also identified (C5XQR9) in a lower Mw region of the gel within a more basic pl range, relative to spot 2 discussed above. The HSP homolog C5XQR9 localised to an appropriate position on the gel, corresponding to Mw 17121 Da and pl 6.18. Matches to C5XQR9 were observed in 296B spots B, 7, 8, 9, and 10, and in QL12 spots B and 7. QL12 spot 8 contained an additional sorghum homolog (accession C5XML7) to a maize 17.5 kDa class II HSP.

In addition, several HMW (50-75 kDa) putative uncharacterised proteins with homology to HSP/chaperones or luminal binding proteins (BIP) were identified in the alkali-soluble glutelin fraction across both 296B and QL12 genotypes (accessions C5YU58, C5WNX8 and C5XPN2). Sorghum HSP C5YU58 has also been identified in grain proteomic analyses carried out by other researchers (217). In maize, elevated HSP/BiP levels during development are linked to mutations in α- and γ-zein resulting in an opaque or floury phenotype (216, 218, 219). BiP has been found to associate with the surface of rice protein bodies, assisting in prolamin deposition through disulphide bonding at specific cysteine residues (220). In wheat, these chaperones are localised to the interior, rather than on the surface of protein bodies, indicating diverse roles for HSP/BiP in protein body aggregation across different plant species (221).

3.4.6.4 α-amylase inhibitors

Sorghum α-amylase inhibitors were isolated from the prolamin fraction in both genotypes (Table 3.2). The enzyme was isolated from a low Mw (10-15 kDa) size range, within pI 7-8. The inhibitor was identified in 296B gel spots 1, 2, 8 and 11, and in QL12 spots 1 and 2, which was observed across duplicate gels. The activity of α-amylase inhibitors in the grain affects starch digestion into glucose, slowing the conversion of starch to ethanol during the fermentation process (222). Lines expressing low levels of α-amylase inhibitor may be appropriate candidates for the biofuels.
industry. α-amylase inhibitors also play a role in plant defence, where they deter crop destruction by snails and birds in tannin-containing sorghum and have been shown to enhance insect resistance in wheat (223-225).

### 3.4.7 Non-prolamins in the protein-starch matrix

Non-prolamin proteins, such as albumins and globulins, are rich in essential amino acids, including lysine and tryptophan, and provide the embryo with additional readily accessible nitrogen reserves during germination. Sorghum homologs to the major non-kafirin storage proteins were identified across all protein solubilities (Supporting Information, fig. S4/table 1). In the water/salt-soluble fractions sorghum homologs to globulin and cupin-like proteins in Zea mays were abundant, including accessions C5WY16 and C5WQD2. Globulins are saline-soluble secondary storage proteins, many belonging to the cupin superfamily. 7S globulins, for example, function exclusively as storage proteins, but are not required for normal seed function (226). Globulin S-1 (63kDa) and globulin S-2 (45kDa) collectively represent ~20% of seed protein content in maize and share amino acid sequence similarity with the 7S seed proteins of wheat and legumes (227). These proteins have significant impacts on the nutritional quality of the grain.

Proteins identified in the alkali-soluble glutelin fraction included a range of vicilin- and legumin-like storage proteins. The sorghum homolog to uncleaved maize legumin, accession C5YY38, and the globulin/vicilin-like sorghum homolog C5WUN6, were identified in this fraction. Immuno-localisation studies in Medicago trunculata using anti-vicilin antibodies show preferential targeting of vicilins to the periphery of the protein bodies (228). Sorghum glutelins have not yet been extensively characterised, but it is hypothesised that they may form complex highly-linked protein networks encasing protein bodies and providing additional structure to the protein-starch matrix (68, 69). Protein members of the legumin superfamily are most abundant in legumes, oats and rice. Wheat legumin-like protein, or triticin, accumulates in globulin inclusion bodies at the periphery of prolamin bodies. Overexpression of pea legumin in wheat forms crystalline patterns contributing to the altered structure of the protein-starch matrix (229, 230).

A range of proteins were identified in the glutelin fraction in addition to legumin and vicillin, which included HSP/BIPs, cell division cycle proteins, homologs to maize caleosin, glutathione-s-transferases, RuBisCo large subunit binding protein and wheat Mother of FT and TFL1 (MFT), which regulates seed dormancy and the onset of germination(231). Significant amounts of α-, β- and δ- kafirins were also identified in this fraction, although quantification scores indicate that their
concentration was considerably less than in the prolamin fraction, and their presence may be due to incomplete sequential extraction of the alcohol-soluble fraction as previously noted.

Across the study, protein spots excised from HMW areas of 2D gels generally produced matches specifically to HMW proteins, while spots excised from LMW areas tended to produce matches to both high and low Mw proteins. This indicates that spots in LMW areas of the gel may contain a mix of intact proteins and protein subunits, or cleaved products of larger proteins. For example, sorghum homologs to HMW globulin S-1 and 2 (C5WY16 and C5WQD2), cupin family proteins (C5WUN and C5X0T3) and uncleaved legumin (C5YY38) were isolated from every protein fraction, from both high and low Mw areas of the gel, whereas LMW proteins (~13kDa), such as thioredoxin, glutaredoxin, and α-amylase inhibitors were isolated exclusively from the lower Mw area of the gel. Highly abundant sorghum homologs to glyceraldehyde-3-phosphate dehydrogenase as well as various HSPs were detected in the glutelin fraction. These proteins were generally found in the area of their calculated pIs, but localised across a broader size range. Because the presence of these proteins was so widespread, the matches were not listed in results tables (Supporting Information, table 1), unless they represented the only quality match returned for a protein spot.

3.5 Conclusion

Our proteomic analysis compiles sequence and biochemical information describing a range of proteins affecting sorghum endosperm structure and composition. The derived dataset will further augment annotation of the sorghum proteome and facilitate identification of potential targets for improved grain quality. Furthermore the data provides a basis for comparative analysis with other major grain crops. Possible avenues for utilising protein sequence data include the development of high-lysine varieties with increased A/G content for enhanced nutritional quality, as well as the modification of enzyme-regulated protein aggregation in the endosperm for increased starch availability.

Differential regulation of thioredoxin in the β-kafirin null mutant QL12 indicates that expression of the enzyme may be either directly or indirectly linked to the β-kafirin mutation. Evaluation of Trx expression at the RNA/protein level across additional β-kafirin mutants and at varying stages of development could provide further insights into this relationship. Future research on previously uncharacterised proteins identified in this study may reveal elements related to grain quality parameters, such as digestibility and flour pasting properties. In particular, the HMW γ-prolamin
homolog identified provides an interesting candidate for further functional analysis as it has been shown to localise to the periphery of the protein bodies in maize, possibly limiting enzymatic access to internally located α-zeins and starch. Identification of proteins with desirable amino acid content and enhanced digestibility will also benefit future research initiatives to improve sorghum grain nutritional value and palatability.

The commercial success of grain crops with altered protein composition, such as high lysine and increased digestibility lines, can be limited by negative pleiotropic effects accompanying changes in the amino acid profile. For example, maize endosperm protein mutants, o2 (opaque2) and fl-2 (floury-2) have an improved amino acid content, but initially exhibited a number of undesirable traits such as reduced grain yield and increased susceptibility to diseases and pests, which needed to be improved. This work supports the development of high throughput screening methods for biomarkers associated with grain quality and endeavours for the genetic improvement and biofortification of sorghum through molecular breeding and transformation. The identification and characterisation of proteins impacting on various grain quality parameters in sorghum, such as nutritional quality, grain hardness, and stress resistance will assist breeders in their introgression of genetic factors associated with these traits into established breeding lines. The rapid introgression and selection of desirable grain quality traits will require the linking of the proteome with the genome. Next generation genome re-sequencing (149) and genotyping by selection (150) tools are now publically available and are being incorporated into sorghum pre-breeding for other grain quality traits such as digestibility (86). This will ultimately lead to more cost-effective and efficient plant breeding for improved sorghum varieties tailored to specific end-uses.

3.6 Acknowledgements

The work is supported by funding from an ARC linkage grant with Pacific Seeds and by the USDA-ARS, Manhattan, KS, USA. Extensive support in proteomic techniques was kindly provided by Dr Amanda Nouwens at the SCMB protein facility (University of Queensland). Identification of MALDI TOF MS proteins was facilitated through open access queries to the Mascot server.
Chapter 4

Impacts of kafirin allelic diversity, starch content and protein digestibility on ethanol conversion efficiency in grain sorghum

Cremer J.E.,⁷, Liu L., Bean S.R.,²,³ Ohm J.B.,⁴ Tilley, M.,², Wilson J. D.,³, Kaufman R.C.,⁴,⁵ Vu T.H.,⁵ Gilding E.K.,⁶ Godwin I.D.,¹, Wang D.²

¹School of Agriculture and Food Sciences, The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia
²Department of Biological and Agricultural Engineering, Kansas State University, Manhattan, KS 66506, USA
³USDA-ARS Centre for Grain and Animal Health Research, Manhattan, KS 66502, USA
⁴USDA-ARS Cereal Crops Research Unit, Fargo, ND 58102, USA
⁵Department of Grain Science and Industry, Kansas State University, Manhattan, KS 66506, USA
⁶Institute for Molecular Bioscience, The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia
⁷Corresponding author j.cremer@uq.edu.au

Names are necessary to report factually on available data; however, the U.S. Department of Agriculture neither guarantees nor warrants the standard of the product, and use of the name by the U.S. Department of Agriculture implies no approval of the product to the exclusion of others that may also be suitable.
4.1 Abstract

Seed protein and starch composition determine the efficiency of the fermentation process in the production of grain-based ethanol. Sorghum, a highly water- and nutrient-efficient plant, provides an alternative to fuel crops with greater irrigation and fertiliser requirements, such as maize. However, sorghum grain is generally less digestible due to extensive disulfide cross-linking among sulfur-rich storage proteins in the protein/starch matrix. Thus, the fine structure and composition of the seed endosperm directly impacts grain end use, including fermentation performance. To test the hypothesis that kafirin (prolamin) seed storage proteins specifically influence the efficiency of ethanol production from sorghum, ten diverse genetic lines with allelic variation in the β-, γ- and δ-kafrins, including three β-kafrin null mutants, were tested for ethanol yield and fermentation efficiency. Our selected lines showed wide variation in grain biochemical features, including total protein (9.96-16.47%), starch (65.52-74.29%), and free amino acid (FAN) (32.84-73.51 mg/L). Total ethanol yield (ranging 384-486 L/tonne), showed a significant positive correlation to starch content ($R^2=0.74$), and there was a small significant positive correlation between protein digestibility and ethanol yield ($R^2=0.52$). Increases in FAN content enhanced fermentation efficiency ($R^2=0.65$). The highest ethanol producer was elite staygreen breeding line B923296, and the line with the highest fermentation efficiency at the 72 hour time point was inbred BTx623. A large seeded genotype, KS115, carrying a novel γ-kafrin allele, was rich in FAN and exhibited excellent short term fermentation efficiency at 85.68% at the 20 hour time-point. However, the overall ethanol yield from this line was comparatively low at 384 L/tonne, due to insufficient starch, low digestibility and high crude protein. Multivariate analysis indicated an association between the β-kafrin allele and variation in grain digestibility (P=0.042) and FAN (P=0.036), with subsequent effects on ethanol yield. RP-HPLC profiling of the alcohol-soluble kafrin protein fraction revealed diversity in protein content and composition across the lines, with similarities in peak distribution profiles among β-kafrin null mutants compared to normal lines.
4.2 Introduction

Commercial production of plant-derived fuels presents an effective strategy for reducing our reliance on fossil fuels and increasing energy security (232). Grain ethanol production involves the conversion of starch to ethanol using the enzyme α-amylase to aid gelatinisation and liquefaction and gluco-amylase for production of fermentable sugars or saccharification. Bioethanol can be employed as a gasoline extender and fuel oxygenate, where distribution is aided by existing infrastructure (233). The fermentation process generates valuable by-products, including distillers dried grain and solubles (DDGS), which are marketed as high-quality feed products. Under the Renewable Fuels Standard (RFS)-Energy Independence and Security Act (22), the recommendation has been for an increase in the production of bioethanol to 136 billion litres by 2022, with government legislation in the European Union, the United States, Brazil, Australia and others, mandating increased fuel ethanol components of up to 30% by 2025 (234). The largest producers of bioethanol are the United States and Brazil, accounting for almost 90% of global production (235). Maize is the major feedstock for the bioethanol industry in the US, representing 95% of the total 56 billion litres produced there in 2011 (236). In Brazil, nearly half the nation’s vehicles run on fuel ethanol produced from sugarcane. However, in recent years negative impacts have been associated with the production of some forms of biofuels. For example, maize cropping requires relatively high irrigation and fertiliser inputs, causing drought susceptibility and nitrification of waterways (237, 238). Additionally, it has been reported that burning cane trash for fuel production generates carbon emissions similar to those associated with fossil fuels (239). Thus, sustainable alternatives to these bioethanol feedstocks are being sought, particularly in regions of low water availability. One viable solution is biofuel from sorghum grain. An efficient C4 assimilator with less fertiliser inputs required to achieve optimal yields compared to other crops (178), sorghum tolerates a range of soil conditions and exhibits a high level of drought tolerance, requiring half the water needed to produce equivalent yields compared to corn and a quarter the water required for sugar cane (240). Ethanol produced from grain sorghum at facilities using ‘biogas digesters’ and specifically combined heat and power technology qualify as advanced biofuels, meeting greenhouse gas emissions reduction thresholds (241).

The efficiency with which grain is converted into fuel is largely dependent on the structural features and chemical composition of the seed endosperm. Factors directly affecting fermentation efficiency, yield and DDGS quality include starch content and composition, proportion of amylose to amylopectin, flour viscosity, crude protein content, digestibility, and condensed tannins in tannin...
containing sorghum (93, 242-244). These traits are genetically controlled but are also strongly influenced by agronomic factors such as water and nutrient availability. Sorghum digestibility is reduced by extensive cross-linking among proteins in the grain endosperm (245). This may limit enzymatic accessibility to starch in grain sorghum relative to maize, impacting on ethanol conversion efficiency. Storage proteins are synthesised on the endoplasmic reticulum and deposited as vacuolar protein aggregates, which develop into protein bodies (PBs), encasing starch in the seed (57). Sorghum and maize endosperm contain relatively high proportions of proline-rich ‘prolamins’, which are hydrophobic in nature and develop highly folded, complex tertiary structures, reducing in vitro endosperm solubility (246, 247). Sorghum and maize prolamins, referred to as kafirins or zeins, respectively, exhibit extensive sequence homology and a similar relative distribution in the PBs (248). However, variation in the functional characteristics and degree of polymerisation of the proteins accounts for differences in grain quality traits like digestibility and ethanol conversion across maize and sorghum (65, 67). Immunolocalisation studies indicate that cysteine-rich β- and γ-kafirins are located on the periphery of sorghum PBs, while the α-kafirins fill the interior of the structure along with small amounts of δ-kafirin (53). A line of high digestibility (HD) sorghum mutants were found to exhibit an altered PB structure in the grain, with the more hydrophobic β- and γ-kafirins re-located from the periphery of the bodies into folds in the structure, increasing centrally located α-kafirin exposure to protease activity (53, 249). Improvements in endosperm digestibility in these lines was directly correlated to the HD mutant allele dosage and translated into higher ethanol production efficiency in fermentation studies (250). Sequencing and molecular analysis of the β-, γ- and δ-kafirin genes across a variety of sorghum commercial hybrids and wild relatives has revealed a wide range of allelic diversity (167, 251). A mutation in the β-kafirin gene has been identified in several sorghum varieties, where a single cytosine insertion results in a frameshift and early termination codon. Functional analysis shows that β-kafirin null line QL12 exhibits altered flour viscosity, presumably due to changes in β-kafirin levels in endosperm PBs. Genetic characterisation of the kafirins across diverse sorghum lines has facilitated investigation of the effects of variation in seed storage proteins on ethanol production in sorghum grain. Thus, the aim of this study is to evaluate the impact of kafirin allelic diversity on ethanol conversion efficiency across a selection of sorghum lines, characterised for kafirin genetic background, seed biochemistry and composition. The study will carry forward research into sorghum-based ethanol production and subsequently allow for identification of key factors affecting ethanol bioconversion, illustrating how they are influenced by the composition of the protein-starch matrix.
4.3 Methods

4.3.1 Plant genotypes, seed weight and sample preparation

Mature grain from ten commercial grain sorghum hybrid parent inbreds (Table 4.1) with varying genetic background for the kafirin seed storage proteins was harvested at the University of Queensland Gatton campus during the 2011-2012 summer cropping season. Sorghum varieties were selected from a sample population previously characterised for allelic variation in β-, γ- and δ-kafirins. Lines with sequenced genomes and additional reported resistance to environmental stress were included in the panel. Seed weight was measured in grams per hundred grains and then averaged to mg per grain. Whole grain was milled through a UDY sample mill (UDYCorp, Fort Collins, CO.) fitted with a 0.5mm mesh screen for all analytical procedures.

<table>
<thead>
<tr>
<th>Lines/Genotypes</th>
<th>β-kafirin allele</th>
<th>γ-kafirin allele</th>
<th>δ-kafirin allele</th>
<th>Origin</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL12</td>
<td>GU732403 null</td>
<td>M73688</td>
<td>AY834250</td>
<td>Australia</td>
<td>Staygreen male parental line, yellow endosperm</td>
</tr>
<tr>
<td>B923296</td>
<td>GU732401</td>
<td>M73688</td>
<td>AY834250</td>
<td>Australia</td>
<td>Elite staygreen female parental line</td>
</tr>
<tr>
<td>ICSV400</td>
<td>GU732404</td>
<td>M73688</td>
<td>AY834250</td>
<td>Mali</td>
<td>Breeding line for malting and grain yield</td>
</tr>
<tr>
<td>M35-1</td>
<td>GU732401</td>
<td>GU732407</td>
<td>AY834250</td>
<td>India</td>
<td>Drought resistant, cooking quality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Breeding line, large seed, yellow</td>
</tr>
<tr>
<td>KS115</td>
<td>GU732401</td>
<td>GU732408</td>
<td>AY834250</td>
<td>USA</td>
<td>endosperm</td>
</tr>
<tr>
<td>1S17214</td>
<td>GU732403 null</td>
<td>M73688</td>
<td>AY043223</td>
<td>Nigeria</td>
<td>Landrace</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inbred breeding line: Dwarf Hydro x Rice, Kafir</td>
</tr>
<tr>
<td>BOK11</td>
<td>GU732401</td>
<td>M73688</td>
<td>AY043223</td>
<td>USA</td>
<td></td>
</tr>
<tr>
<td>BTx623</td>
<td>AJ717660</td>
<td>M73688</td>
<td>AY834250</td>
<td>USA</td>
<td>Inbred breeding line, zerazera-caudatum</td>
</tr>
<tr>
<td>296B</td>
<td>GU732401</td>
<td>GU732407</td>
<td>AY834250</td>
<td>India</td>
<td>Inbred female parental line: Dwarf, high yielding</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Commercial hybrid breeding line, staygreen</td>
</tr>
<tr>
<td>RTx2737</td>
<td>GU732403 null</td>
<td>M73688</td>
<td>AY834250</td>
<td>USA</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Sorghum grain lines characterised for allelic variance in the kafirin storage proteins and tested for ethanol production efficiency
4.3.2 Starch Analysis

Total starch in 200mg milled sorghum samples was determined in duplicate using a colorimetric technique according to AACCI method 76-13.01 (170) utilizing a DMSO pretreatment for resistant starch (K-TSTA Megazyme International Ireland Ltd., Co. Wicklow, Ireland). Starch was isolated from the milled samples by the sonication method of Park et al (252). Amylose content was determined on the isolated starch samples by the Concanavalin A precipitation assay (K-AMYL, Megazyme International Ireland Ltd., Co. Wicklow, Ireland).

4.3.3 Crude protein digestibility

Protein digestibility was determined in duplicate using the previously described method for measuring \textit{in vitro} pepsin digestibility (140). Briefly, 200mg milled sorghum flour per sample was mixed in 35ml pepsin solution [1.5mg/ml pepsin from porcine gastric mucosa (#P6887 Sigma Aldrich, St Louis MO) (enzyme activity: one unit produces ΔA$_{280}$ of 0.001/min at pH 2, at 37° C) pepsin solution containing 0.1 phosphate buffer containing KH$_2$PO$_4$ and H$_3$PO$_4$, pH 2], incubated at 37°C for 2hrs. After incubation, 2ml 2M NaOH was added, the sample vortexed and centrifuged (3220 x g, 15min). The supernatant was then discarded. Residue was washed in 10ml 0.1 phosphate buffer pH2, centrifuged (3220 x g, 15min) and supernatant discarded. Washing steps were repeated and after second wash and centrifugation the samples were placed in a -80°C freezer (Romulus Holding Company, New York, NY). Prior to nitrogen analysis, samples were lyophilised (Labconco, Kansas City, MO) and total protein content was measured using nitrogen combustion Leco nitrogen measurement (LECO Corporation, St. Joseph, MI) and a protein conversion factor of 6.25.

4.3.4 Flour moisture content

Moisture readings were taken from 1g milled sorghum flour per sample in duplicate using a MX-50 moisture analyser (A&D, Tokyo, Japan). Average moisture content was calculated as the mean across the duplicates.
4.3.5 Ethanol production and fermentation efficiency

Analysis of total ethanol production and fermentation efficiency was carried out as previously described (Wu et al, 2008). Ground sample (30g dry mass) was combined with 100ml heated (~60-70°C) enzyme solution (0.1g KH₂PO₄ and 20μl Liquozyme per liter) (Novozymes North America, Inc., Franklinton, NC) in an Erlenmeyer flask to form a uniform slurry with shaking (180rpm) at 70°C. Liquefaction of the slurry was initiated by increasing temperature from 70°C to 90°C for 30 min, holding at 90°C for 5min, then reducing the temperature to a constant 86°C for a further 60min. Material on the sides of flasks was pushed back into slurry and rinsed with 3-5 mL dH₂O. The mashes were cooled to room temperature and adjusted to pH 4.2 with 2N HCl. Dry ethanol yeast (Ethanol Red, Lesaffre Yeast Co. Milwaukee, WI) was activated with 1 mL pre-culture broth (20g glucose, 5g peptone, 3g yeast extracts, 1g KH₂PO₄, and 5g MgSO₄• 7H₂O per liter) and incubated at 38°C for 25-30min with shaking at 200rpm. Simultaneous saccharification and fermentation was initiated with 1ml activated yeast culture, 100μl Spirizyme (Novozymes North America, Inc., Franklinton, NC) and 0.3g yeast extract. Flasks were sealed with an S-airlock filled with mineral oil. Fermentation was carried out at 30°C for 72hr with shaking at 1500rpm. Fermentation efficiency was calculated by recording weight loss of the mash through CO₂ diffusion during the fermentation process. Ethanol concentration following distillation (conducted as described in Yan et al. 2010) was quantified by HPLC with a Rezex RCM-monosaccharide column (300mmx7.8mm) and a reflective index detector (Shimadzu RID-10A, Columbia, MD). The mobile phase through the column was 0.6ml/min of dH₂O at a constant temperature of 80°C. Fermentation efficiency was calculated according to theoretical yield of 56.72g from 100g dry starch.

4.3.6 Free Amino Acid (FAN) analysis

Free amino nitrogen (FAN) is a measure of the concentration of individual amino acids and small peptides that can be utilised to fuel yeast growth and proliferation during the fermentation process. FAN was determined according to previously described methods (253) with modification. Sorghum flour (150 mg) was mixed with 1.5 mL of deionized distilled water in a 2.5 mL microcentrifuge tube and vortexed five times in 10 min, then centrifuged at 12,000 rpm (20,000 x g) for 20 min. A 1.0 mL aliquot of supernatant was diluted with 9.0 mL distilled water and then analysed for FAN using the ninhydrin colourimetric method.
4.3.7 Reversed Phase HPLC (RP-HPLC)

Alcohol soluble proteins (prolamins) were isolated from milled flour samples as described in Bean et al (2010). Briefly, flour was dissolved in 1ml extraction solvent (60% tertiary butanol, 0.5% sodium acetate and 2% β-mercaptoethanol). The pellet was then vortexed 5 min, centrifuged 10k rpm (14,000 x g) for 4 min and 500µl supernatant was transferred to a new tube. The procedure was repeated once and a further 500µl supernatant was collected and pooled 1:1. Finally, 33µl 4-vinylpyridine was added to each prolamin sample and vortexed 10 min for alkylation of proteins. Protein samples (5µl injections) were analysed on an Agilent 1100 HPLC system (Agilent, Foster City, CA) fitted with a Poroshell C-18 column, 2.1mm x 75mm (Agilent, Foster City, CA) using a previously described gradient (169). Detection was by UV at 214nm.

4.3.8 Lab-on-a-Chip

The Lab on a Chip (LOC) procedure was carried out on the Agilent 2100 Bioanalyser with alcohol-soluble protein samples extracted as for RP-HPLC methods above and processed using the Protein 80 assay kit (Agilent Technologies, Palo Alto, CA). A 4µl aliquot of each protein sample was combined with 2µl denaturing buffer, containing β-mercaptoethanol, in a 0.5ml micro centrifuge tube. Sample tubes and an additional tube containing 6µl of ladder were heated to 95°C for 5min, cooled and centrifuged. 84µl dH₂O was added to each tube and samples were vortexed and spun briefly. Protein samples (6µl), ladder (6µl) and gel dye (12µl) were loaded into the appropriate well on the chip. The chip was inserted into the bioanalyser and analysed per manufacturer’s instruction.

4.3.9 Statistical Analysis

Correlation among parameters affecting ethanol conversion (% protein, crude starch, digestibility etc.) was illustrated using Microsoft Excel. Letters indicating significance of differences (table 2) were generated in Minitab using an ANOVA general linear model with pairwise comparisons according to Tukey methods with 95% confidence. Principle Component Analysis (PCA) was carried out in Minitab for grain quality parameters (table 2) using multivariate analysis within a correlation matrix. In addition, Generalised Linear Mixed Models (GLMM), were applied in the ‘R’ platform (fitted with a Poisson distribution) for multivariate analysis of the relationship between seed biochemical features, kafirin alleles and ethanol production efficiency.
4.4 Results and Discussion

4.4.1 Genotypic associations with fermentation efficiency and ethanol yield

The performance of a selection of sorghum genotypes (Table 4.1) was evaluated for fermentation efficiency and total ethanol yield to determine how variation in endosperm protein-starch matrix impacts on ethanol conversion (Table 4.2). Total ethanol yields across the lines varied by 10.8% and fermentation efficiency varied by 26% at 20 hours, and by 5.4% at 72 hours (Table 4.2). The grain types produced between 384 and 426 L/tonne ethanol (Fig. 4.1), with fermentation efficiencies ranging between 68-85% at 20 hours and 87-92% at 72 hours, similar to values obtained in comparable studies (93, 250). Total starch ranged between 65.52% -74.29%, similar to other studies (254, 255).

<table>
<thead>
<tr>
<th>Cultivar</th>
<th>Seed weight (mg)</th>
<th>% Crude protein</th>
<th>% Protein Digestibility</th>
<th>Starch % dry base</th>
<th>Amylose %</th>
<th>FAN mg/L</th>
<th>Fermentation efficiency % @ 20 hr</th>
<th>Fermentation efficiency % @ 72 hr</th>
<th>Ethanol yield (liters/tonne)</th>
<th>Ethanol (ml per kg starch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL12</td>
<td>28.6</td>
<td>14.57 b</td>
<td>58.42</td>
<td>70.11 b</td>
<td>21.76 b</td>
<td>33.0 ± 3</td>
<td>70.44 ± 0.40 ±</td>
<td>88.40 ± 1.39 ±</td>
<td>410.69 ± 1.40 ±</td>
<td>585.79 ± 1.39 ±</td>
</tr>
<tr>
<td>B923296</td>
<td>28.0</td>
<td>11.9 ± d</td>
<td>61.16</td>
<td>71.0 ± b</td>
<td>23.58 ±</td>
<td>32.8 ± 3</td>
<td>73.63 ± 0.40 ±</td>
<td>90.42 ± 1.40 ±</td>
<td>426.91 ± 1.40 ±</td>
<td>601.41 ± 1.39 ±</td>
</tr>
<tr>
<td>ICSV400</td>
<td>44.6</td>
<td>11.95 d</td>
<td>56.93</td>
<td>71.50 b</td>
<td>21.93 b</td>
<td>39.67 b</td>
<td>69.97 ± 0.40 ±</td>
<td>88.57 ± 1.39 ±</td>
<td>420.94 ± 1.39 ±</td>
<td>588.71 ± 1.39 ±</td>
</tr>
<tr>
<td>M35-1</td>
<td>30.4</td>
<td>12.93 c</td>
<td>48.65</td>
<td>71.06 b</td>
<td>21.53 ±</td>
<td>36.54 b</td>
<td>69.69 ± 0.40 ±</td>
<td>89.55 ± 1.39 ±</td>
<td>415.80 ± 1.40 ±</td>
<td>585.13 ± 1.39 ±</td>
</tr>
<tr>
<td>KS115</td>
<td>72.5</td>
<td>16.47 a</td>
<td>40.83</td>
<td>65.52 c</td>
<td>23.19 a</td>
<td>73.51 ±</td>
<td>85.68 ± 0.40 ±</td>
<td>90.21 ± 1.39 ±</td>
<td>384.37 ± 1.39 ±</td>
<td>586.61 ± 1.39 ±</td>
</tr>
<tr>
<td>IS17214</td>
<td>43.0</td>
<td>11.43 d</td>
<td>59.22</td>
<td>74.29 ±</td>
<td>23.31 ±</td>
<td>37.61 b</td>
<td>67.98 ± 0.40 ±</td>
<td>88.14 ± 1.39 ±</td>
<td>420.30 ± 1.40 ±</td>
<td>565.78 ± 1.39 ±</td>
</tr>
<tr>
<td>BOK11</td>
<td>29.1</td>
<td>11.67 d</td>
<td>65.03</td>
<td>70.34 b</td>
<td>22.60 b</td>
<td>51.33 b</td>
<td>72.55 ± 0.40 ±</td>
<td>88.19 ± 1.40 ±</td>
<td>414.11 ± 1.40 ±</td>
<td>588.74 ± 1.39 ±</td>
</tr>
<tr>
<td>BTx623</td>
<td>32.2</td>
<td>12.93 c</td>
<td>50.8</td>
<td>66.29 ±</td>
<td>22.47 b</td>
<td>48.13 c</td>
<td>75.13 ± 0.40 ±</td>
<td>92.05 ± 1.40 ±</td>
<td>395.66 ± 1.40 ±</td>
<td>596.83 ± 1.39 ±</td>
</tr>
<tr>
<td>296B</td>
<td>31.1</td>
<td>9.96 ± e</td>
<td>68.21</td>
<td>73.85 ±</td>
<td>19.74 e</td>
<td>64.27 b</td>
<td>74.24 ± 0.40 ±</td>
<td>87.67 ± 1.40 ±</td>
<td>415.67 ± 1.40 ±</td>
<td>562.88 ± 1.39 ±</td>
</tr>
<tr>
<td>RTx2737</td>
<td>31.1</td>
<td>13.3 ± e</td>
<td>58.12</td>
<td>71.4 ±</td>
<td>19.74 f</td>
<td>40.45 ± 3</td>
<td>73.3 ± 0.40 ±</td>
<td>87.36 ± 1.40 ±</td>
<td>413.45 ± 1.40 ±</td>
<td>579.00 ± 1.39 ±</td>
</tr>
<tr>
<td>mean</td>
<td>37.1</td>
<td>12.7</td>
<td>56.7</td>
<td>70.5 ± 0.35</td>
<td>22.0 ± 0.38</td>
<td>1.452</td>
<td>na</td>
<td>na</td>
<td>1.551 ± 0.339</td>
<td>3.959 ± 0.339</td>
</tr>
</tbody>
</table>

Table 4.2: Biochemical measurements (seed weight, protein, digestibility, starch, amylose, and FAN) and ethanol fermentation characteristics (yield and fermentation efficiency) for selection of sorghum grain cultivars. Amylose was determined as a percentage of starch content. Mean values and standard deviation are calculated across replicate sample measurements. Means that do not share a letter are significantly different.

Commercial line B923296 was competitive with maize in terms of ethanol yield (256), producing the highest level of ethanol across the lines at 426 L/tonne (2.86 gallons per bushel). This genotype displayed high fermentation efficiency at the 20 and 72 hour time points, high starch and good digestibility, traits which likely combined to enhance yield (Table 4.2). Genotypes IS17214 and ICSV400 also yielded well (420 L/tonne each), both with a high starch content and moderately large seed size, although fermentation efficiency was relatively low in these lines. KS115 and...
BTx623, each with a high crude protein content, low digestibility, and low starch produced the lowest overall ethanol yields at 384 and 395 L/tonne, respectively. However, the proportion of ethanol produced per gram of starch (Table 4.2) was comparable to most other genotypes, signifying that low starch content is the major factor restricting yield in these lines. Fermentation efficiency in KS115 of 85.68% at 20 hours, exceeded efficiencies of other genotypes by more than 10% and indicating that this line could perform well in a short term fermentation system, particularly if starch content were improved.

Figure 4.1: Ethanol yield (L/tonne) (data from table 4.2 depicted) across ten sorghum lines with allelic variation in the kafirin storage proteins. β-kafirin null mutants QL12, IS17214 and RTx2737 (highlighted in light brown) produced above average ethanol yields.

4.4.2 Endosperm starch and protein effects on fermentation

The structure and degree of cross-linking among protein bodies in the starch matrix has significant effects on digestibility, starch accessibility and ethanol conversion (93, 245). Crude protein and starch levels across the lines in the current study were similar to those attained in previous analyses (67, 165, 257). Starch content was shown to have a major influence on grain ethanol production, where a strong positive linear correlation was observed between total starch and ethanol yield ($R^2=0.736$) (Fig. 4.2), as observed in past studies (23, 165). Of the genotypes tested, the β-kafirin
null line IS17214 contained the highest amount of total starch and produced the second highest ethanol yield, with above average protein digestibility and medium-level fermentation efficiency. Beta-kafrin null allelic variants QL12 and RTx2737 contained lower starch compared to IS17214, which is likely to be the main factor accounting for their reduced ethanol yield, comparatively, as protein digestibility was above average in these lines.

![Figure 4.2: Relationship between starch content, fermentation efficiency and ethanol yield across ten sorghum genotypes. A significant positive correlation ($R^2=0.74$) is observed between total starch (% dry base) and ethanol yield (L/ton). There was also a weak, although significant negative correlation ($R^2=0.54$) between starch content and fermentation efficiency at 72 hours.]

Regarding starch composition, all the genotypes tested in the study were “normal” in terms of the amount of amylose present (Table 4.2), i.e. none of the samples were waxy genotypes. Only minor differences in amylose content (~19-24%) were found across all the genotypes. There was no correlation between amylose content and ethanol yield (data not shown) suggesting that within this “normal” range, amylose:amylopectin ratio did not play a role in determining ethanol fermentation properties of the sorghum genotypes used in this study. It has been demonstrated that more subtle changes in starch structure contribute to enhanced digestibility (86). These changes were not
measured in this study, however, it is conceivable that starch structural differences associated with higher digestibility may, at least in part, also contribute to ethanol yield differences.

Protein digestibility and ethanol yield were positively correlated across the genotypes (P=0.019) (Fig. 4.3), as previously observed (67, 93, 258). Additionally, β-kafirin had significant, although relatively small impacts on digestibility as observed using a multi-variant analysis (P=0.0416), where digestibility was slightly higher than average in null β-kafirin mutant lines. Previously, Zhao et al (2008) have shown that increases in ethanol yield and conversion efficiency occur as the amount of extractable proteins from sonication of mashed samples increases, indicating that endosperm digestibility impacts directly on ethanol production. Conformational changes in endosperm protein structure following cooking further influence the accessibility of amylolytic enzymes to starch. Zhan et al (2006) demonstrated this with supercritical fluid extrusion cooking of sorghum grain prior to fermentation, showing that alterations to the protein matrix which enhance starch accessibility also increase ethanol yield.

Figure 4.3: Associations between protein digestibility (%), fermentation efficiency at 72 hours (%), and ethanol yield (L/ton) across ten sorghum genotypes. A positive linear relationship was observed between digestibility and ethanol yield ($R^2=0.5194$), indicating a low, yet significant correlation, also shown with multivariant analysis (P=0.019). A trend towards an association between digestibility and fermentation efficiency was also observed ($R^2=0.3483$), although this was not a significant correlation (P=0.072) within this data set.
4.4.3 Free Amino Nitrogen (FAN) impact on fermentation efficiency

Fermentation requires an adequate supply of nitrogenous compounds to fuel yeast growth and proliferation. Release of FAN during enzymatic breakdown of endosperm proteins provides a grain-specific source of amino nitrogen. Nitrogen deficiencies have been reported as a major cause for a slow fermentation (259, 260). Low FAN levels can be supplemented in the industrial fermentation process with a mix of amino acids and ammonium sulphate at the exponential phase of yeast growth, enhancing fermentation rate and ethanol conversion efficiency, but increasing commercial costs. A strong positive association was observed in the current study between FAN and fermentation efficiency at the 20th hour timepoint ($R^2=0.647$), in agreement with past work (243, 244, 250). This relationship was exemplified by KS115, which displayed relatively high levels of FAN and performed well in the early stages of the fermentation process (Fig. 4.4).

![Fermentation efficiency vs hour](image)

**Figure 4.4:** Fermentation efficiency (%) versus fermentation hour. KS115 is an outlier which exhibits high efficiency in the early stages of the fermentation process relative to other genotypes.
An antagonistic relationship appears to occur among factors affecting ethanol yield and fermentation efficiency, as illustrated through PCA (Fig. 4.5), where protein digestibility and total starch are positively associated with ethanol yield, while FAN content and total protein are associated with increased fermentation efficiency. Moreover, a significant correlation was identified between the β-kafirin allele and FAN levels (P= 0.0357), indicating that diversity at this locus will have a potential effect on fermentation efficiency. In a previous analysis of normal versus waxy sorghums, genotypes with high crude protein and FAN content exhibited high fermentation efficiencies, but generally with reduced ethanol yields compared with low protein, high digestibility varieties (165). KS115, a large seeded grain type, with a high proportion of FAN-rich embryo, exhibits remarkably high fermentation efficiency in the early stages of the fermentation process (Fig. 4.4), but performed poorly in terms of overall ethanol yield.

![Figure 4.5: PCA of parameters affecting fermentation efficiency and ethanol yield. Starch and protein digestibility are positively correlated to yield, and crude protein content and FAN levels are positively correlated to fermentation efficiency. The β-kafirin allele is more strongly related to digestibility and ethanol yield, whereas γ-kafirin shows a positive association with fermentation efficiency and FAN.](image-url)
4.4.4 Kafirin allelic effects on ethanol conversion efficiency

The interconnectivity of seed storage proteins governs susceptibility of matrix components to proteolysis, impacting on starch availability and fermentation profile (245). Previous work has shown that grain kafirin content is significantly linked to variation in seed biochemistry, including fat, protein and starch content, as well as seed weight (130). In the current investigation, the β-kafirin null lines QL12, IS17214 and RTx2737 produced higher than average ethanol yields across the study (Table 4.2, fig. 4.1). Digestibility, starch and FAN profiles were similar for lines carrying a β-kafirin null allele, indicating that an analogous kafirin profile may produce corresponding similarities in grain characteristics. Among the genotypes, KS115 exhibits a significantly higher seed weight. Seed weight is controlled by a number of factors, including various embryo- and endosperm-specific regulators (261). However, the composition of endosperm storage proteins, such as the kafirins, is also linked to the large-seeded trait (130). KS115 carries a novel γ-kafirin allele and exhibits high fermentation efficiency in the early stages of the conversion process, but with low ethanol yields produced overall.

Alterations in the positioning of the kafirins located on the periphery of the protein bodies has been shown to increase grain digestibility and ethanol production efficiency, indicating that regulatory changes to β- and/or γ-kafirin expression may instigate these changes, similar to the 22 kD α-kafirin mutation in HD lines (81). Lines M35 and B923296 each displayed higher digestibility, starch content and lower FAN compared to KS115. Variation in γ-kafirin allelic background, where β- and δ-kafirin alleles are identical between the poor ethanol producer KS115 and mid to high yielding M35 and B923296 indicates a possible link between kafirin genetic background and ethanol conversion. The various γ-kafirin alleles encode for identical protein sequences. However, kafirin expression may be differentially regulated at the transcriptional level, resulting in variable grain kafirin content. In any case, significant diversity was observed in the kafirin RP-HPLC peak distribution profile for KS115 compared to other genotypes, with a greater number of peaks eluting at 10-12 min in KS115, and an additional peak appearing at 12.5 min, which was not observed in other lines (Fig. 4.6). This raises the question as to whether there is a specific kafirin allelic combination linked to increased protein digestibility and ethanol conversion efficiency. Through this study it appears that the content of β-kafirin in the seed significantly impacts on ethanol production and that γ-kafirin could also play a distinct role, as observed in KS115.
Figure 4.6: RP-HPLC peak distributions for the alcohol-soluble kafirin protein fraction across ten sorghum genotypes evaluated for ethanol production efficiency. Peak profiles for β-kafirin null allelic variants QL12, IS17214 and RTx2737 are similar in size and distribution, indicating similarities in protein composition. Chromatograms for low digestibility lines, such as KS115 and BTx623, show large peak heights and a relatively diverse peak distribution profile.

4.4.5 Regulation of kafirin seed storage proteins

Co-suppressing the synthesis of various prolamin subclasses in sorghum, including α-, β- and γ-kafirins, has been shown to increase grain protein digestibility (102). Kumar et al (2012) report that down-regulation of α-kafirin in the sorghum endosperm results in increased endosperm digestibility, however altering γ-kafirin expression in isolation had no apparent phenotypic effects on protein body morphology or cooked flour digestibility (103, 104). In maize, QTL analysis correlates starch digestibility to chromosome regions already linked to the zeins (105). Mutations in the maize Opaque2 gene result in a 50% reduction in zein, exhibiting a floury or opaque endosperm, with improved in situ starch digestibility and ethanol conversion (106). Variation in the β-kafirin allele across the collection of grain types analysed in the present study is a result of altered expression of the gene, such as in QL12, where production of a truncated protein causes distinct
changes to seed biochemistry, including digestibility. The relatively small, but significant increase in digestibility and ethanol yield in the β-kafirin null lines compared to normal lines indicates that some degree of functional redundancy may exist between β- and γ-kafirin. The relationship between seed kafirin content and grain quality parameters, such as digestibility and ethanol conversion efficiency, observed in this study justifies further investigation into interactions among the kafirins and the mechanisms regulating their targeting to the endosperm.

4.4.6 Effects of seed size on ethanol conversion

Large-seeded hybrids often contain higher levels of crude protein and fat and less starch than small-seeded lines and hybrids (262). KS115, with a high seed weight, rich in protein and fat and low in starch (263), produced the lowest ethanol yield across the lines. Grain containing low levels of amylase, for example waxy or heterozygous waxy types, are known to perform better in the fermentation process (23, 242, 250). Ethanol conversion may have been impeded in this line by low total starch content or by relatively high levels of amylase-lipid complex in the grain endosperm. The KS115 seed contains a higher proportion of protein-rich embryo relative to starchy endosperm (264), providing less starch for conversion to ethanol, despite the large grain size. It is unknown whether the large seeded structure is related to changes in storage protein composition, such as increased kafirin content, or to changes in the regulation of protein aggregation. In rice, the polycomb complex OSFIE2 has been linked with certain aspects of grain filling and seed size, including regulation of the starch synthesis rate limiting step and multiple storage compounds (265). Homologs to the polycomb complex gene family were identified in other major cereals, such as maize, indicating that large seeded sorghum may carry an alternate allele for an OSFIE2 ortholog, causing specific changes to seed morphology.

Reversed-phase HPLC analysis of protein content and composition across the genotypes shows that levels of alcohol-soluble protein were higher in KS115 grain than in the more digestible, higher yielding lines, with the KS115 chromatogram exhibiting larger peak areas for the kafirin-containing fraction and a more diverse peak distribution profile (Fig. 4.6). A greater proportion of kafirin in the grain endosperm is likely to contribute to the low digestibility and poor ethanol yield through reduced enzyme accessibility to starch. Lab on Chip size-based separation of alcohol-soluble proteins across the genotypes provides additional evidence of the high content of insoluble protein in KS115, relative to other lines (Fig. 4.7). KS115 grain has been recommended as a valuable component in animal feed, due to a high fat content, which provides a major energy source for
livestock (263). However, with a low starch content and high susceptibility to grain moulds, breeding strategies for large seeded cultivars for fuel ethanol would likely involve introgression of genes for large seededness and embryo size into backgrounds with high starch content and protein digestibility coupled with improved stress resistance.

Figure 4.7: Lab on Chip size-based separation of alcohol-soluble kafirin fraction in sorghum lines evaluated for ethanol fermentation. β-kafirin null mutants QL12 and IS17214 have low kafirin content, similar to highly digestible line 296B. β-kafirin null RTx2737 shows higher levels of kafirin protein comparatively, as was observed with RP-HPLC, where peaks areas were larger and distribution profile was more diverse. KS115 and M35 exhibit relatively high kafirin content, contributing to the reduced digestibility of these lines.

Variation in grain prolamin profile has been reported to have major impacts on grain quality, accounting for differences in digestibility and ethanol conversion rates across different genotypes (179, 242). RP-HPLC peak distribution profiles for alcohol-soluble protein from β-kafirin null allelic variants (QL12, IS17214, and RTx2737), compared to genotypes expressing a functional
form of β-kafirin, reveal that the mutation appears to be associated with the disappearance of a major set of protein peaks eluted around 10.5-11 min (Fig. 4.6). Further investigation beyond the scope of this paper into the precise identity of the missing alcohol-soluble proteins in β-kafirin null lines and their impact on grain quality is warranted. QL12 crude protein levels are relatively high (Table 4.2), compared to other genotypes, which is not reflected in RP-HPLC and LOC analysis of the alcohol-solubles. This may, therefore, be attributed to a high content of albumins, globulins or glutelins as observed with subsequent analysis of water/salt-soluble protein for this line (Supplementary Information, figure S3 and S4). The reduced levels of kafirins, and relatively high levels of A/Gs, in the null mutant results in an increased grain digestibility, which appears to enhance ethanol production efficiency. However, below average FAN levels in the mutants may have had corresponding negative effects on fermentation efficiency, and was perhaps the reason why fermentation efficiency of the β-kafirin nulls (as well as overall ethanol yield) were only slightly above average, despite a high protein digestibility.

4.4.7 Candidate traits for sorghum grain biofuels breeding program

A number of quality traits contribute to the efficiency with which starch is converted to ethanol in the production of grain-based fuels. This investigation reveals that differential expression of the β-kafirin gene has significant impacts on ethanol production through changes to protein digestibility and FAN content. The important role of starch in determining the suitability of a grain crop for conversion to bioethanol has been further verified, where a strong positive correlation was observed between total starch content and ethanol yield within this dataset. Lines BTx623 and KS115 had the lowest starch contents and, subsequently, the lowest ethanol yields. Elite Australian line B923296 produced high ethanol yields with an efficient fermentation profile and was the third most digestible grain among the genotypes tested, indicating its value as a potential bioethanol feedstock, among other high yielding lines in the study. B923296 was previously incorporated into breeding programs for staygreen traits and midge resistance. The β-kafirin null line QL12 and the Indian line M35 were also incorporated into breeding programs as staygreen parental varieties, and are efficient ethanol producers, although they exhibit slightly lower starch and digestibility relative to B923296. Large seeded types, exemplified by KS115, ferment efficiently in the initial stages of the conversion process, but produce very low yields due to insufficient starch. However, the line does appear to produce ethanol efficiently per kilogram starch available, compared with other lines (Table 4.2). These lines represent a valuable source of germplasm for use in research and development and could be exploited to contribute useful ethanol-related traits to breeding programs aimed
particularly at increasing seed size and improving early stage ethanol production efficiency in sorghum.

4.5 Conclusion

Endosperm protein structure and composition play an important role, in addition to starch, in determining the suitability of a grain crop for bioethanol production. Here it is illustrated that fermentation yield and efficiency is determined by key quality parameters, such as starch and protein content. These traits are strongly influenced by the specific expression and interaction of endosperm storage proteins. Genetic variation in β-kafirin alters digestibility and FAN content, with subsequent effects on ethanol conversion. This work recommends sorghum with high starch content, high digestibility and low levels of β-kafirin, for further development in the grain-based ethanol industry. Large seeded varieties produce ethanol more efficiently in the short term due to a high FAN content and possible variation in the regulation of γ-kafirin, providing a valuable source of germplasm for breeding initiatives aimed at improving fermentation rate. Specific investigation of correlations between the γ-kafirin allele and the large seeded phenotype will be useful in deciphering the effects of variation in kafirin genetic background on end use traits. Furthermore, transcriptional profiling of the kafirin genes across these genotypes will contribute to our understanding of the impact of this genetic diversity on seed biochemical traits and ethanol production. Sorghum currently represents ~5% of the grain ethanol market in the US (22), but is ideally positioned for expansion in the industry. Genotypes exhibiting optimal endosperm composition and storage protein profile for converting ethanol offer commercially competitive alternatives to fuel crops with greater environmental impacts, such as maize and sugarcane.

4.6 Acknowledgements

JEC was supported by a Graduate School International Travel Award (GSITA) provided by the University of Queensland and an Australian Postgraduate Award Industry Scholarship, as supported by the Australian Research Council and Pacific Seeds (Linkage project LP0883808). Research conducted at the USDA-ARS in Manhattan, KS was partially supported by a NC-213 Andersons Research Grant. Additional research funding was provided by Kansas State University, Manhattan, KS. Assistance with statistical analysis was kindly provided by Mrs Delma Greenway and Dr Margaret Jewell at the University of Queensland.
Chapter 5

General Discussion
5.1 Introduction

Cereals are a major source of nutrition for humans and livestock. Grain protein composition has significant impacts on quality traits, including digestibility and nutritional value. Prolamins, which represent the major seed storage proteins in the hardier grains, including maize, sorghum and the millets, are poor in the essential amino acids lysine, threonine and tryptophan. In sorghum, amino availability and digestibility is further reduced, particularly upon cooking, by a high degree of cross-linking among the cysteine-rich prolamins or kafirins. Proteomic profiling of the main classes of grain proteins across sorghum lines with allelic variation in the kafirins increases our understanding of the impact of changes in the expression of storage proteins on the structure of the protein starch-matrix and on grain quality traits. This study has generated a range of biochemical information pertaining to the accumulation of proteins in the grain endosperm, which will contribute to the improvement of sorghum for human and animal consumption, as well as for various industrial applications, such as ethanol production and fermentation efficiency. The acquisition of sequence data for previously uncharacterised sorghum seed proteins through mass spectrometric techniques augments the annotation of the sorghum proteome and facilitates the identification of potential targets in sorghum for the development of improved grain cultivars.

5.2 Impact of kafirin background mutations

Kafirins are encoded by single or low copy gene loci, with the exception of the α-kafirins, of which 19 are reported expressed (183). Mutations in α- and γ-prolamins in both sorghum and maize result in altered endosperm storage body composition. Thus far, investigations into the effects of allelic variation in β-kafirin on protein body structure have been limited. Three sorghum lines, QL12, IS17214 and RTx2737, recently characterised at the genetic level, carry an alternate β-kafirin allele with a single cytosine insertion, resulting in a frameshift and early termination codon (167). Profiles of alcohol-soluble proteins in QL12 show that this line does not produce β-kafirin.

In this work, β-kafirin null lines were further characterised at the protein level using multiple proteomic and biochemical techniques, including measurement of protein digestibility and fermentation efficiency. Across the null mutants, similarities in proteomic profile were observed compared to wild-type lines. RP-HPLC analysis showed that the mutants contain less alcohol-soluble protein and are missing a major set of peaks compared to most other lines. Based on this analysis, a likely candidate peak for the β-kafirin protein has been proposed. Despite the similarities
in profile, there was also some evidence, provided through LOC analysis, that β-kafirin expression profiles are not entirely uniform across the mutant lines. For example, peaks present in the 17-20kD area of the LOC electropherogram, representing the size range for β-kafirin, were larger than expected in RTx2737. However, there appeared to be similar changes to peak heights for this size range across the mutants, where the first protein peak, present at ~19kD, was smaller than the second, eluting at ~20kD, whereas in normal genotypes the first peak was generally larger than the second. Interestingly, RP-HPLC and LOC data show that the β-kafirin mutation does not result in the complete absence of the protein peak at 19kD. This indicates the likely presence of a similarly sized protein or set of proteins, which co-localise with β-kafirin within this size range. LC-MS/MS analysis of the β-kafirin band/spot excised from SDS-PAGE gels indicated that this protein may represent a peptidyl prolyl cis/trans isomerase (PPIase), involved in protein folding.

Previous work has shown that β- and γ-kafirins localise to the periphery of protein bodies, and may block enzymatic access to internally located α- and δ-kafirins, impacting on protein digestibility and starch solubility (53). Raw and cooked protein digestibility was above average in the mutants, and IS17214 displayed the highest cooked digestibility across the entire sample population of 28 lines. This result indicates that the loss of β-kafirin has significant effects on protein composition in the grain, and on cooking qualities, which could be further evaluated in future research.

The development of mutants with altered expression of prolamins or the regulatory elements controlling their synthesis and aggregation into protein bodies, has generated germplasm with altered grain quality traits. Mutations in the 22kD α-kafirin result in a non-vitreous grain phenotype. In these mutants, modified protein body shape results from changes to the spatial distribution of storage proteins and the kinetics of their deposition in the protein bodies (218, 266). An unfolded protein response (UPR) accompanies the mutation, where impaired targeting of endosperm proteins is linked to the increased expression of stress-related factors, including PPIase, PDI and various luminal binding protein BiP/heat shock proteins (HSP70) (219, 267).

Significantly more is known about the zeins in maize than the kafirins in sorghum. Information describing the regulation of zeins and their targeting to protein bodies in maize may provide clues as to how kafirin biosynthesis is regulated in sorghum. Point mutations have been identified in the 16kD γ-zein and in signalling peptides governing the targeting of α-zeins to the endosperm. In maize, the γ-zein mutant o15 produces fewer protein bodies (268), whereas increased levels of γ-zein enhances protein body number, resulting in the formation of more vitreous endosperm (269). Further characterisation of kafirin regulatory elements in sorghum, particularly β-kafirin and the
HMW γ-prolamin homolog, identified at the sequence level here, could lead to further insights as to how changes to the expression profiles of the prolamins impact on the makeup of the protein-starch matrix in sorghum.

Additional mutations causing changes in zein synthesis, which result in the sequestering of irregularly shaped protein bodies include O2, DeB30, Mc and Fl2 (266, 270). O2 encodes a transcription factor that regulates the transcription of 22kD α-zein, among other proteins (271, 272). Fl2 corresponds to a point mutation in the signalling peptide of a 22kd α-zein. The failure to correctly process this signal peptide appears to cause the protein to become anchored to the ER membrane, leading to the accumulation of α-zeins at the surfaces of protein bodies (273, 274). This mutant also displays a dramatic increase in the synthesis of binding protein and other ER-resident chaperones, indicating that an unfolded protein response is occurring in endosperm cells (275). In Mc and DeB30, the expression of ER-resident chaperones is also dramatically induced (216).

The effects of the related stress-responses on protein stability, post-translational modification, protein-protein interactions and the folding and sorting of proteins by the various chaperones and protein folding enzymes brings about major changes to the structure of the protein-starch matrix with resulting effects on grain quality (276). Molecular chaperones impacting on protein body structure and composition identified in the current study through LC-MS/MS, include PDI, PPIase, BiP and thioredoxin. Subsequent functional characterisation of these proteins will allow for the improved profiling of the UPR response in sorghum, which may exhibit modified protein body structure as a result of this response.

5.3 Development of improved grain cultivars

The floury mutation enhances the lysine content of the grain through increased A/G content, but the soft texture of the grain is unsuitable for commercial production. Suppression of zein production is accompanied by an increase in the expression of non-zein proteins, indicating a shift in the flux of metabolic intermediates through biosynthetic pathways producing storage proteins in the grain (277). Transformation of rice with maize O2 transcription factors results in activation of glutelin and globulin promoters, revealing a mechanism for this heightened production of non-prolamins. Increased expression of globulin and glutelin, coupled with a reduction in prolamins, enhances the lysine content of transgenic rice plants (278). Opaque endosperm and floury mutants, o2 and fl-2 exhibit improved amino acid content, but harbour a number of undesirable agronomic traits such as
reduced grain yield and increased susceptibility to diseases and pests. Profiling of the water-soluble fraction of storage proteins in sorghum, carried out through this work, contributes towards the further optimisation of methods for the direct quantification of these proteins in the grain using high resolution techniques, such as tandem mass spectrometry and HPLC.

5.4 Biological significance of proteins identified by proteomic analysis

Proteomic analysis using gel-based techniques and mass spectrometry (LC-MS/MS) identified a range of proteins impacting on seed biochemistry. In addition to their effects on protein body formation and digestibility, several of these proteins also participate in stress-related pathways. Thioredoxin (Trx) alters the redox state of storage proteins in the protein-starch matrix, mobilising starch and nitrogen reserves during germination (109). The absence of trx in β-kafirin null line QL12 may be directly or indirectly linked to the kafirin mutation and could also contribute to the lower digestibility of this line relative to 296B.

Trx overexpression in barley results in an increase in the relative proportion of soluble to insoluble proteins (172). This is due to changes in the sulphhydril disulphide (SH_SS) status of the protein, where the solubility of reduced sulphhydril proteins is higher than that of oxidised proteins. Trx activity has also been associated with increased expression of pullulanase (starch debranching enzyme) and alpha-amylase through up-regulation of gibberellin synthesis, resulting in enhanced solubility of starch in the endosperm matrix (172, 201). Thus, alterations to the biochemistry of storage proteins through trx result in increased starch and nitrogen availability to the growing seedling during germination. In addition, trx plays a central role in the oxidative stress response by providing reducing power to reductases or in scavenging oxidised proteins (279). Therefore, the activity of trx alters the redox state of endosperm proteins, increasing protein solubility, with additional effects on stress resistance.

The role of phosphorylation-specific isomerases in protein folding has reported effects on the structure and interactions of seed storage proteins (280, 281). Proteins with phosphorylation activity identified with LC-MS/MS, included peptidyl-prolyl cis/trans isomerase (PPIase) and maize protein disulphide isomerase (PDI). Through structural processing of proteins at the tertiary level, PPIases regulate protein-protein interactions and the folding status of endosperm storage proteins. These proteins may have effects on protein body aggregation and endosperm solubility in sorghum. Their expression is also regulated by stress signalling pathways. The enhancement of PPIase activity in response to stress is associated with the induction of cyclophilins. Differential expression of PPIase
has been observed in cultivars displaying varying levels of drought tolerance in sorghum and rice (282, 283).

Similar to PPIase, the molecular chaperone PDI was identified in sorghum grain tissues using mass spectrometry. PDI is involved in oxidative protein folding and regulates the accumulation of storage proteins in the endosperm. Mutations in PDI confer a chalky phenotype, with irregularly shaped protein bodies and starch granules (281). The reduced expression of PDI in rice is associated with the decreased activity of plastidial phosphorylase and pullulanase, as well as increased activity of soluble starch synthase 1 and ADP-glucose pyrophosphatase (267). Maize floury2 mutants contain elevated levels of PDI and produce abnormally processed α-zein and high amounts of binding protein (BiP). Increased levels of PDI and binding protein in the mutant infers that PDI functions as a molecular chaperone, in conjunction with BiP, for assembly of zeins into protein bodies (209). In addition to regulating the composition of protein bodies in the grain, PDI has also been implicated in responses to temperature stress during seedling development (284).

Luminal binding protein (BiP) and/or heat shock proteins (HSP), identified here in sorghum grain, are stress-responsive proteins with auxiliary impacts on grain quality. Changes in the abundance of small HSPs in the endosperm of wheat, for example, significantly affect the dough forming potential of the flour (285). In maize, the differential expression of HSPs was also linked to grain hardness and bread-making qualities (286). The activity of HSPs is central to the oxidative stress response. The identification of proteins with dual impacts on grain quality and stress resistance allows for an integrated approach towards introgression of multiple traits in the grain, accelerating breeding programs and contributing substantially towards the development of improved grain lines.

Plants produce proteinaceous amylase inhibitors to combat against amylase enzymes from infesting insects. The impact of α-amylase inhibitors on insect growth and proliferation through interference with carbohydrate absorption has been well documented (287). The identification of α-amylase inhibitors in this study, allows for their screening and quantification in sorghum. Large-scale production of sorghum α-amylase inhibitor could have valuable industrial applications. Plant α-amylase inhibitors are employed as an organic insecticides and their potential value to the pharmaceutical industry as novel plant-based therapeutics for hyperglycemia, diabetes and obesity is currently being evaluated.
5.5 Improving the nutritional value of sorghum

In recent years, traditional agronomic approaches have been combined with modern molecular techniques to improve the nutritional quality of grain crops. The development of high-lysine waxy lines with improved amino availability is central to this effort (52, 81, 104, 154, 191). Increased expression of albumins and globulins leads to a higher content of essential amino acids. In addition, 2S albumins are known to inhibit the growth of pathogenic fungi (288), suggesting a dual role for these proteins in seed storage and plant defence.

Recently, grain quality has been enhanced in other major crops through RNAi silencing of expression and/or catabolism of specific storage proteins, which impact on nutrition and digestibility. In such cases, knowledge of the mechanisms facilitating in vivo biosynthesis and degradation of the protein target was necessary. Identification and profiling of lysine-rich non-prolamin storage proteins and the enzymes regulating their biosynthesis in sorghum facilitates the screening and manipulation of these elements in the grain.

In the current study, previously uncharacterised sorghum seed proteins with homology to lysine-rich storage proteins from related crop species were identified through LC-MS/MS. These included proteins showing homology to globulin-1 S, globulin 2, vicillin-like storage proteins, cupin family proteins and legumin-1/glutelin B. Identification and localisation of these elements in the grain confirms that their gene products are actively expressed in these tissues and facilitates biofortification efforts for improved nutritional value in sorghum. In order to improve methods for screening of non-prolamins in sorghum grain across large sample sets, established chromatographic methods for analysis of water/salt-soluble proteins were further developed in the study (168). Size-based separation of A/Gs using LOC was also optimised within the sample set. Through this work, a collection of diverse sorghum grain lines could be differentiated according to both A/G and prolamin content.

The evaluation of differential expression of kafirins and the various metabolic enzymes affecting their structure and function was linked to variability in alcohol-soluble protein profiles, with variation in grain quality parameters across the sample population, such as protein digestibility and fermentation. Correlations between specific prolamin peaks on RP-HPLC chromatograms, and protein digestibility was also observed. A/G levels varied significantly across the sample population suggesting that the content of essential amino acids, including lysine, was also variable. This has implications in evaluating breeding lines for nutritional quality and protein digestibility.
Additionally, the isolation and identification of redox-active proteins, such as thioredoxin, could facilitate subsequent purification, screening, and further manipulation of these enzymes in sorghum and other crops.

Modification of protein expression at the DNA level through transgenics has significantly enhanced grain nutrition (1). Amino deficiencies have been addressed in wheat and maize through the introduction of genes encoding, for example, lysine-rich albumins from the highly nutritious Amaranthus hypochondriatus (289, 290). Additional biofortification measures have included the expression of pea legumins in wheat under the control of a glutenin promoter (230), and the overexpression of methionine-rich storage proteins, which was achieved by altering the post-transcriptional regulation of the Dzs10 gene. Protein supplementation is costly, and may be unfeasible in developing countries that rely on cereals as a main staple. Increasing in vivo production of methionine has eliminated the need for the addition of synthetic forms to animal feed (291). With a greater understanding of the structure and composition of proteins contributing to amino availability in the grain, this strategy could be applied in sorghum. Improvements in nutritional value have been achieved more recently by blocking lysine catabolism through RNAi, which resulted in an up to 55% increase in grain lysine content in maize (292). Here, the identification of previously uncharacterised lysine-rich storage proteins in sorghum opens avenues for similar approaches to be carried out using the sequence information compiled in this study.

5.6 Identification of polyphenolic anti-nutritional factors

The content of polyphenolics in sorghum grain correlates highly with antioxidant activity and has significant impacts on stress responses (293). However, tannins generally also represent antinutritional factors in the grain, due to their affinity for binding storage proteins, which consequently reduces enzymatic access to starch. For a range of commercial applications, low polyphenolic content in the grain is desirable. Our sample set of grain lines included a previously characterised tannin-containing line IS8525 (98). RP-HPLC profiling indicates that the specific binding of kafirins by tannins may have occurred in this line, because the peak thought to represent γ-kafirin (169) was absent on this chromatogram. An additional line, IS12572C, displayed a similar chromatographic profile, and both lines exhibited abnormally low digestibility compared to other lines, indicative of the presence of tannins in the grain. Chromatographic evaluation of interactions between tannins and storage proteins may complement efforts for screening of anti-nutritional elements in the grain and deciphering their effects on quality traits, including protein digestibility.
However, direct measurement of tannins across the lines included in the study (using the vanillin-HCL assay for example) is required in order to verify the relationship between tannin content and changes to HPLC profiles in the alcohol-soluble prolamin fraction in the grain across these lines.

5.7 Impacts of protein composition on commercial end-uses

The composition and biochemistry of storage proteins has significant effects on grain processing and milling, and on the suitability of the grain for specialty uses, like brewing, baking and ethanol production for biofuels. High starch content is usually associated with increased digestibility, whereas high protein content tends to result in decreased digestibility. The present study goes some way towards determining the impact these factors on grain solubility. Biochemical analysis showed that digestibility was more strongly correlated to protein content (p=0.035 cooked) than to starch (p=0.799 cooked) across the sample population of allelic variants, indicating that much of the variation in protein digestibility could be attributed to protein compositional effects, rather than interactions with starch. The activity of various starch enzymes and inhibitors was not quantified, however, which may have also had significant impacts on digestibility.

The evaluation of fermentation efficiency and ethanol yield was carried out across a selection of ten grain lines from the sample population, including the β-kafirin null mutants. This analysis revealed that endosperm protein structure and composition play an important role, in addition to starch, in determining the suitability of a grain crop for commercial end uses, such as bioethanol production. Here, it was shown that fermentation efficiency and overall ethanol yield is determined by a combination of key quality parameters, such as protein digestibility, as well as starch content. Allelic variation in β-kafirin altered digestibility and FAN content, with subsequent effects on ethanol conversion. Therefore, sorghum with high starch content, high digestibility and low levels of β-kafirin, was recommended for further development in the grain-based ethanol industry. It is likely that decreased FAN levels in QL12 result from a low prolamin content as shown with RP-HPLC. This may be directly or indirectly related to the null mutation in β-kafirin. It was found that the large seeded variety, KS115, exhibited a high short term fermentation efficiencies due to a high FAN content and possibly due to variation in the expression of γ-kafirin. Correspondingly, KS115 contains remarkably higher levels of alcohol-soluble protein as observed in LOC and RP-HPLC analysis.
5.8 Conclusions and future directions

Strategies focussed on coordinated enhancement of digestibility, nutritional quality, fermentation efficiency and stress resistance will improve the commercial value of sorghum as the industry expands. The development of cultivars with combined traits for increased grain nutrition and resistance to environmental stress will have significant effects on alleviating malnutrition and preventing food shortages in marginal grain producing regions, like Africa and India. Cross-talk between protein biosynthetic pathways and defence responses is mediated by the activity of enzymes which interact with both storage proteins and stress signalling elements. Protein sequence data from this study may be employed to modify protein aggregation in the endosperm, such as though enhancement of thioredoxin expression for increased digestibility, which may have concomitant impacts on stress resistance. Further evaluation of Trx expression at the mRNA/protein level in QL12 and across additional β-kafirin null mutants could provide additional insights into the activity of this enzyme in the sorghum grain and its impacts on seed biochemistry. The identification of previously uncharacterised sorghum proteins with roles in the processing of storage proteins and in stress responses, such as BiP/HSP, provides further possibilities for the evaluation of the activity of these proteins and enzymes under varied environmental conditions. Future work could involve, for example, the monitoring of protein body aggregation in heat shock mutants under temperature stress.

Soluble seed storage proteins isolated across multiple fractions in the study, including the sorghum homolog to maize globulin-S-1, may play an integral role in connecting the protein-starch matrix through interactions with the kafirins. Increases in the content of non-kafirins, particularly high Mw (HMW) albumins and globulins, occur in high-lysine grain lines. Further characterisation of these proteins in sorghum will complement efforts to develop highly digestible lines with improved nutritional value. The expression and activity of the grain proteins and regulatory enzymes identified here will need to be measured across multiple growing seasons and in varied environmental and agronomic conditions in order to more fully determine a more precise compositional profile. Experimental examination of protein-protein interactions may also be useful in deciphering the effects of differential expression of specific storage proteins, such as β- and γ-kafirin, on protein-starch matrix connectivity.

To date breeding efforts have delivered numerous cereal varieties with altered starch and protein profiles, however, there is still room for further development, especially in sorghum. High digestibility mutants tend to have low resistance to mould. It would therefore be helpful to link
improved starch content, protein digestibility and ethanol production efficiency with increased mould resistance, particularly in lines already characterised for ergot and midge resistance, and staygreen. Many of the lines employed in this study originate from breeding programs aimed at improving these traits (Table 2.1).

Expression profiling of the kafirins at the transcriptional level across the sample population evaluated in this study may further contribute to our understanding of the impact of genetic variation in storage proteins on grain quality. Identification of a HMW γ-prolamin homolog in sorghum provides a basis for the functional analysis of this protein subclass with regards to protein body structure. The localisation of 50kD γ-zeins to the periphery of maize protein bodies, where they potentially obstruct enzymatic access to internally located α-prolamins and starch, suggests that this protein may play a similar role in sorghum. Integration of the current profiling strategies for storage proteins employed here, with subsequent epigenetic studies investigating the transcriptional and translational stability of these proteins under varied environmental conditions will give us a broader understanding of how these proteins are regulated at the genetic level. This will also provide a basis for comparative studies on grain development in other commercially important crops, including maize, wheat and rice.
Literature Cited


39. Dar, W., and Mahmoud, S. (2012) CGIAR research program on dryland cereals: A global alliance for improving food security, nutrition and economic growth for the world’s most vulnerable poor, ICRISAT and ICARDA.


Zhang, F., and Boston, R. (1992) Increases in binding protein (BiP) accompany changes in protein body morphology in three high-lysine mutants of maize, *Protoplasma* 171, 142-152.

Buckner, R. J. (1997) Developmental and crosslinking factors related to low protein digestibility of sorghum grain (Dissertation), Purdue University.


241. U. S. Environmental Protection Agency. (2012) EPA issues notice of data availability concerning renewable fuels produced from grain sorghum under the RFS program.


## Appendices

### Supplementary Information Table 1

### LC-MS/MS peptide matches

**Water/Salt-Soluble (Albumin/Globulin)**

<table>
<thead>
<tr>
<th>296B</th>
<th>Spot</th>
<th>Protein Match</th>
<th>Acc #</th>
<th>Score</th>
<th>Mw</th>
<th>pI</th>
<th>emPAI</th>
<th>Homology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Triosephosphate isomerase [Sorghum]</td>
<td>C5XQ07</td>
<td>2018</td>
<td>27177</td>
<td>5</td>
<td>7.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IN2-1 protein [Zea mays]</td>
<td>B6TB10</td>
<td>1172</td>
<td>27482</td>
<td>5</td>
<td>5.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Triosephosphate isomerase [Sorghum]</td>
<td>C5XQ07</td>
<td>4256</td>
<td>27177</td>
<td>5</td>
<td>17.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Proteasome subunit alpha type [Sorghum]</td>
<td>C5WVT0</td>
<td>437</td>
<td>27512</td>
<td>6</td>
<td>1.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Thioredoxin [Sorghum ] (PRESENT 296B/ABSENT QL12)</td>
<td>C5XB72</td>
<td>1400</td>
<td>13061</td>
<td>6</td>
<td>31.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Putative uncharacterized protein Sb10g022780 [Sorghum]</td>
<td>C5Z557</td>
<td>1292</td>
<td>25785</td>
<td>6</td>
<td>5.23</td>
<td>Glutathione S-transferase GST 19 (Zea mays) Q9FQC0_MAIZE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb05g022950 [Sorghum]</td>
<td>C5Y5D5</td>
<td>259</td>
<td>16405</td>
<td>8</td>
<td>0.75</td>
<td>Win1 (Zea mays) B6SH12_MAIZE</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Putative uncharacterized protein Sb03g006870 [Sorghum]</td>
<td>C5XQR9</td>
<td>2668</td>
<td>17121</td>
<td>6</td>
<td>6.28</td>
<td>16.9 kDa class I heat shock protein 1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Putative uncharacterized protein Sb03g006870 [Sorghum]</td>
<td>C5XQR9</td>
<td>485</td>
<td>17121</td>
<td>6</td>
<td>0.43</td>
<td>16.9 kDa class I heat shock protein 1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Putative uncharacterized protein Sb03g006870 [Sorghum]</td>
<td>C5XQR9</td>
<td>753</td>
<td>17121</td>
<td>6</td>
<td>1.06</td>
<td>16.9 kDa class I heat shock protein 1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Putative uncharacterized protein Sb06g022060 [Sorghum]</td>
<td>C5YBX1</td>
<td>548</td>
<td>13563</td>
<td>7</td>
<td>1.46</td>
<td>Grx_C2.2-glutaredoxin subgroup I (Zea mays) B6THA1_MAIZE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb05g006870 [Sorghum]</td>
<td>C5Y5D5</td>
<td>105</td>
<td>17121</td>
<td>6</td>
<td>0.43</td>
<td>16.9 kDa class I heat shock protein 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb06g022060 [Sorghum]</td>
<td>C5YBX1</td>
<td>91</td>
<td>13563</td>
<td>7</td>
<td>0.25</td>
<td>Grx_C2.2-glutaredoxin subgroup I (Zea mays) B6THA1_MAIZE</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Putative uncharacterized protein Sb01g039380 [Sorghum]</td>
<td>C5WPE0</td>
<td>2118</td>
<td>15336</td>
<td>5</td>
<td>6.41</td>
<td>glyoxalase family protein superfamily B6SGF3_MAIZE</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Putative uncharacterized protein Sb01g001170 [Sorghum]</td>
<td>C5XKA6</td>
<td>1397</td>
<td>15980</td>
<td>5</td>
<td>4.65</td>
<td>Late embryogenesis protein Lea14-A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb01g012840 [Sorghum]</td>
<td>C5WQF6</td>
<td>555</td>
<td>15167</td>
<td>6</td>
<td>3.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pathogenesis-related protein 10d [Sorghum]</td>
<td>Q4VQB3</td>
<td>788</td>
<td>17059</td>
<td>5</td>
<td>9.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb01g037940 [Sorghum]</td>
<td>C5WMM0</td>
<td>745</td>
<td>17165</td>
<td>5</td>
<td>7.61</td>
<td>Pathogenesis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pathogenesis-related protein 10b Tax_Id=4558 [Sorghum]</td>
<td>Q4VQB2</td>
<td>671</td>
<td>17128</td>
<td>5</td>
<td>7.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb01g037950 [Sorghum]</td>
<td>C5WMM1</td>
<td>206</td>
<td>17173</td>
<td>5</td>
<td>2.51</td>
<td>Pathogenesis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Superoxide dismutase (Cu-Zn) [Ananas comosus]</td>
<td>Q9SQL5</td>
<td>330</td>
<td>15280</td>
<td>5</td>
<td>2.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polyubiquitin [Saccharum officinarum]</td>
<td>Q9ZSV9</td>
<td>1985</td>
<td>42642</td>
<td>8</td>
<td>1.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nucleoside diphosphate kinase [Zea mays]</td>
<td>B4FK49</td>
<td>246</td>
<td>16531</td>
<td>6</td>
<td>1.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nucleoside diphosphate kinase [Cyanothecae sp.]</td>
<td>B1WQB7</td>
<td>104</td>
<td>16533</td>
<td>5</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Soluble (Prolamin)

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Accession</th>
<th>MSp1</th>
<th>MSp2</th>
<th>MSp4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triosephosphate isomerase</td>
<td>C5XQ07</td>
<td>2079</td>
<td>27177</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb02g027930</td>
<td>C5X3S7</td>
<td>1366</td>
<td>25825</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb02g030620</td>
<td>C5X6S5</td>
<td>1659</td>
<td>35273</td>
<td>8</td>
</tr>
<tr>
<td>Triosephosphate isomerase</td>
<td>C5XQ07</td>
<td>4557</td>
<td>27177</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb04g007600</td>
<td>C5XY27</td>
<td>1010</td>
<td>19982</td>
<td>6</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb03g010730</td>
<td>C5XGM0</td>
<td>440</td>
<td>26462</td>
<td>7</td>
</tr>
<tr>
<td>Proteasome subunit alpha type</td>
<td>C5WVT0</td>
<td>423</td>
<td>27512</td>
<td>6</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb03g006870</td>
<td>C5XQR9</td>
<td>44281</td>
<td>17121</td>
<td>6</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb06g022060</td>
<td>C5YBX1</td>
<td>568</td>
<td>13563</td>
<td>7</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb06g022060</td>
<td>C5YBX1</td>
<td>317</td>
<td>13563</td>
<td>7</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb03g006870</td>
<td>C5XQR9</td>
<td>2033</td>
<td>17121</td>
<td>6</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb03g003530</td>
<td>C5XML7</td>
<td>223</td>
<td>17844</td>
<td>6</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb06g022060</td>
<td>C5YBX1</td>
<td>512</td>
<td>13563</td>
<td>7</td>
</tr>
<tr>
<td>5-hydroxyisourea hydrolase</td>
<td>B4FS54</td>
<td>407</td>
<td>24634</td>
<td>6</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb06g022060</td>
<td>C5YBX1</td>
<td>102</td>
<td>13563</td>
<td>7</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb01g039380</td>
<td>C5WPE0</td>
<td>2238</td>
<td>15336</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb03g001170</td>
<td>C5XKA6</td>
<td>212</td>
<td>15980</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb09g009070</td>
<td>C5YYE5</td>
<td>165</td>
<td>14824</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb01g0446490</td>
<td>C5WX5R9</td>
<td>515</td>
<td>27937</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb01g039380</td>
<td>C5WPE0</td>
<td>445</td>
<td>15336</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb01g039380</td>
<td>C5WMM0</td>
<td>355</td>
<td>17165</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb01g039380</td>
<td>C5WMM0</td>
<td>213</td>
<td>17173</td>
<td>5</td>
</tr>
<tr>
<td>Putative uncharacterized protein Sb01g0200010</td>
<td>C5WZ52</td>
<td>1528</td>
<td>61322</td>
<td>6</td>
</tr>
<tr>
<td>Enolase 1</td>
<td>P26301</td>
<td>857</td>
<td>48262</td>
<td>5</td>
</tr>
<tr>
<td>Polyubiquitin</td>
<td>P93135</td>
<td>1725</td>
<td>42656</td>
<td>7</td>
</tr>
<tr>
<td>Nucleoside diphosphate kinase</td>
<td>B4FK49</td>
<td>119</td>
<td>16531</td>
<td>6</td>
</tr>
</tbody>
</table>

**Alcohol-Soluble (Prolamin)**
296B

1 Alpha-amylase inhibitor 5 [Sorghum]  P81368  690  13394  9  8.64
   Putative uncharacterized protein Sb02g006470 [Sorghum]  C5XDC3  594  16082  7  2.81
   IAAS_SORBI Alpha-amylase inhibitor 5 (Sorghum bicolor)
2 Putative uncharacterized protein Sb02g006470 [Sorghum]  C5XDC3  1897  16082  7  5.76
   Alpha-amylase inhibitor 5 [Sorghum]  P81368  1620  13394  9  6.69
   IAAS_SORBI Alpha-amylase inhibitor 5 (Sorghum bicolor)
3 Putative uncharacterized protein Sb09g000830 [Sorghum]  C5YY38  2477  53637  6  1.05
   Homology to Q948J8_MAIZE Uncleaved legumin-1 (Zea mays)
4 Putative uncharacterized protein Sb02g025490 [Sorghum]  C5XDK9  956  36615  7  1.18
   C0L964_COILA Gamma-coxin (Coix lachryma-jobi), C0P381_MAIZE Putative uncharacterized protein, Q946W1_MAIZE 50kD gamma zein, A9XEB9_SACOF 50kD gamma canein (Saccharum officinarum)

5 Beta-kafirin [Sorghum]  C5XYX4  441  21381  8  0.55
   Predicted protein (Fragment) [Hordeum vulgare]  F2DDK1  95  24204  9  0.14
   HB6T2Y1_MAIZE Peroxiredoxin-5
5a DNA topoisomerase (Fragment) [Selaginella]  D8SSL3  56  89376  9  0.04
6 Proteasome subunit alpha type [Sorghum]  C5WVT0  384  27512  6  1.5
   Beta-kafirin [Sorghum]  C5XYX4  329  21381  8  0.55
   B4FFZ9_MAIZE Lipoprotein (Zea mays)
6 Putative uncharacterized protein Sb04g031810 [Sorghum]  C5Y0Y5  209  26471  6  0.81
   B6UH67_MAIZE Late embryogenesis abundant protein D-34
7 Beta-kafirin [Sorghum]  C5XYX4  606  21381  8  0.55
   Putative uncharacterized protein Sb02g040650 [Sorghum]  C5X3B9  230  24283  6  1.17
8 Putative uncharacterized protein Sb05g022950 [Sorghum]  C5Y5D5  1203  16405  8  4.4
   Putative uncharacterized protein Sb02g002500 [Sorghum]  C5X8V7  480  16123  7  3.61
   Win1 B6SH12_MAIZE, vacuolar defense protein Q1BER7_WHEAT
9 Peptidyl-prolyl cis-trans isomerase [Sorghum]  C5XT06  907  18503  9  6.43
10 Peptidyl-prolyl cis-trans isomerase [Sorghum]  C5XT06  101  18503  9  0.18
   Peptidyl-prolyl cis-trans isomerase [Sorghum]  C5Z9C6  67  23166  10  0.14
11 Beta-kafirin [Sorghum]  C5XYX4  2587  21381  8  3.96
   Delta kafirin-2 [Sorghum]  C5WPV0  212  21674  8  0.16
   Peptidyl-prolyl cis-trans isomerase [Citrus sinensis]  D0ELH5  62  18310  9  0.18
12 Putative uncharacterized protein Sb06g019110 [Sorghum]  C5Y9W6  693  30730  7  2.09
   Prohibitin Q9M588_MAIZE
   Beta-kafirin [Sorghum]  C5XYX4  363  21381  8  0.79

Observed in rep3 with Sorghum G3FMQ0

QL12

1 Putative uncharacterized protein Sb02g006470 [Sorghum]  C5XDC3  280  16082  7  2.15
   Alpha-amylase inhibitor 5 [Sorghum]  P81368  239  13394  9  3.89
   IAA5_SORBI Alpha-amylase inhibitor 5 (Sorghum bicolor)
2 Putative uncharacterized protein Sb02g006470 [Sorghum]  C5XDC3  1386  16082  7  4.59
   Alpha-amylase inhibitor 5 [Sorghum]  P81368  835  13394  9  3.89
3 Enolase Tax_Id=4577 [Zea mays]  B8A0W7  3306  48334  5  4.95
Putative uncharacterized protein Sb03g046410 [Sorghum]  C5XIK1  743  49578  7  0.47  S-locus-specific glycoprotein B9SIR0_RICCO or Serine-threonine protein kinase B9SIR1_RICCO (Ricinus communis)

4  Enolase Tax_Id=39947 [Oryza sativa subsp. japonica]  QSVNT9  590  48136  5  0.59  C0L964_COILA gamma-coxin (Coxl lachryma-jobs), Q946W1_MAIZE 50kD gamma-zein, A9XEB9_SACOF 50kD gamma-canein (Saccharum)

5a (r2)  Peptidyl-prolyl cis-trans isomerase [Sorghum]  B3GQV9  123  18570  8  0.18  Ribonuclease 3 B4FBD6_MAIZE

5b  Peptidyl-prolyl cis-trans isomerase [Zea mays]  B4FZZ2  90  18578  9  0.18  19.0 kDa class II heat shock protein (Oryza sativa)

5  Putative uncharacterized protein Sb07g021330 [Sorghum]  C5YL90  2710  25112  5  2.49  B6TDA9_MAIZE 2-cys peroxiredoxin BAS1

5  Putative uncharacterized protein [Zea mays]  B4FM07  1490  28318  6  2.8  Glutathione S-transferase GST 19 B4JVJ9_MAIZE and B6TQ2_MAIZE Homology to Chitinase B in Q94EL3_SORHL (Sorghum halepense)

6  Putative uncharacterized protein Sb04g031810 [Sorghum]  C5YOY5  462  26471  6  1.59  B4FFE9_MAIZE Lipoprotein (Zea mays)

6  Triosephosphate isomerase Tax [Zea mays]  C0PE24  192  27248  6  0.41  190 kDa Class I heat shock protein (Oryza sativa)

6  Putative uncharacterized protein Sb04g007585 [Sorghum]  C5XY25  154  22103  6  0.76  19.0 kDa class II heat shock protein (Oryza sativa)

7  Putative uncharacterized protein Sb04g031810 [Sorghum]  C5YOY5  306  26471  6  1.04  B4FFE9_MAIZE Lipoprotein (Zea mays)

7  Putative uncharacterized protein Sb02g040650 [Sorghum]  C5X3B9  270  24283  6  1.46  Glutathione S-transferase GST 19 B4JVJ9_MAIZE and B6TQ2_MAIZE Homology to Chitinase-B in Q94EL3_SORHL (Sorghum halepense)

7  Putative uncharacterized protein Sb10g022780 [Sorghum]  C5ZS57  177  25785  6  0.13  19.0 kDa class II heat shock protein (Oryza sativa)

8  Putative uncharacterized protein Sb06g021230 [Sorghum]  C5YBE8  416  29178  9  0.91  190 kDa Class I heat shock protein (Os sativa)

8  23 kDa alpha-kafirin (Fragment) [Sorghum]  B3VT9  276  25977  9  0.27  Adenylate kinase KAD2_ORYSJ (Oryza sativa)

8  Gamma-kafirin [Sorghum]  C5XDL2  187  22900  8  0.15  GTP-binding nuclear protein Ran-A1 (B4G1P4_MAIZE) C5YW22_SORBI

9  23 kDa alpha-kafirin (Fragment) [Sorghum]  B3VT9  351  25977  9  0.27  Heat shock/chaperone

9  Peptidyl-prolyl cis-trans isomerase [Sorghum]  B3GQV9  134  18570  8  0.4  GTP-binding nuclear protein Ran-A1 (B4G1P4_MAIZE) C5YW22_SORBI

9  Peptidyl-prolyl cis-trans isomerase [Sorghum]  C5XT06  108  18503  9  0.65  GTP-binding nuclear protein Ran-A1 (B4G1P4_MAIZE)

9  Peptidyl-prolyl cis-trans isomerase [Sorghum]  C5XT06  1222  18503  9  7.78  GTP-binding nuclear protein Ran-A1 (B4G1P4_MAIZE)

10  Peptidyl-prolyl cis-trans isomerase [Sorghum]  C5Z9C6  106  23166  10  0.71  GTP-binding nuclear protein Ran-A1 (B4G1P4_MAIZE)

11  Putative uncharacterized protein Sb08g008400 [Sorghum]  C5YU11  872  26638  8  3.64  Adenylate kinase KAD2_ORYSJ (Oryza sativa)

11  Putative uncharacterized protein [Zea mays]  C0HHCl  251  20184  9  1.16  GTP-binding nuclear protein Ran-A1 (B4G1P4_MAIZE) C5YW22_SORBI

11  Gamma-kafirin [Sorghum]  C5XDL2  187  22900  8  0.15  GTP-binding nuclear protein Ran-A1 (B4G1P4_MAIZE)

12  23 kDa alpha-kafirin (Fragment) [Sorghum]  B3VT9  129  25977  9  0.27  GTP-binding nuclear protein Ran-A1 (B4G1P4_MAIZE)

12  Putative uncharacterized protein [Zea mays]  90  20184  9  0.36  GTP-binding nuclear protein Ran-A1 (B4G1P4_MAIZE)

Alkali-soluble (Glutelin)

296B

1  Putative uncharacterized protein Sb08g009580 [Sorghum]  C5YU58  5825  74643  5  5.65  Heat shock/chaperone

C4-specific pyruvate orthophosphate dikinase [Miscanthus]  Q7XJK1  393  103084  6  0.28  RuBisCO large subunit-binding protein?

2  Protein disulfide isomerase [Zea mays]  A5ASE7  1879  56921  5  2.08  RuBisCO large subunit-binding protein?

2  Putative uncharacterized protein Sb01g000380 [Sorghum]  C5WRV5  579  60974  5  0.78  RuBisCO large subunit-binding protein?

C4-specific pyruvate orthophosphate dikinase [Miscanthus]  Q7XJK1  385  103084  6  0.32  RuBisCO large subunit-binding protein?

3  Putative uncharacterized protein Sb08g009580 [Sorghum]  C5YU58  4426  74643  5  3.92  RuBisCO large subunit-binding protein?

C4-specific pyruvate orthophosphate dikinase [Miscanthus]  Q7XJK1  186  103084  6  0.13  RuBisCO large subunit-binding protein?
<table>
<thead>
<tr>
<th>Entry</th>
<th>Description</th>
<th>Accession</th>
<th>Score</th>
<th>E-value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Heat shock/chaperone</td>
<td>C5YU58</td>
<td>4887</td>
<td>74643</td>
<td>6.25</td>
</tr>
<tr>
<td>2</td>
<td>Heat shock/chaperone</td>
<td>C5WNX8</td>
<td>2007</td>
<td>52635</td>
<td>4.82</td>
</tr>
<tr>
<td>3</td>
<td>Luminal-binding protein 2</td>
<td>P24067</td>
<td>4796</td>
<td>73211</td>
<td>2.72</td>
</tr>
<tr>
<td>4</td>
<td>Putative uncharacterized protein Sb03g039360</td>
<td>C5XPN2</td>
<td>986</td>
<td>71261</td>
<td>0.88 Luminal-binding protein or heat shock/chaperone</td>
</tr>
<tr>
<td>5</td>
<td>C4-specific pyruvate orthophosphate dikinase [Miscanthus]</td>
<td>Q7XJK1</td>
<td>1683</td>
<td>103084</td>
<td>1.05</td>
</tr>
<tr>
<td>6</td>
<td>C4-specific pyruvate orthophosphate dikinase [Sorghum]</td>
<td>Q7XJK1</td>
<td>8216</td>
<td>103084</td>
<td>4.58</td>
</tr>
<tr>
<td>7</td>
<td>Pyruvate phosphate dikinase [Sorghum]</td>
<td>Q84N32</td>
<td>7802</td>
<td>103021</td>
<td>5.14</td>
</tr>
<tr>
<td>8</td>
<td>Putative uncharacterized protein Sb01g020910 [Sorghum]</td>
<td>C5X0G5</td>
<td>716</td>
<td>90698</td>
<td>0.83 Cell division cycle protein 48 [Chlorella ellipsoidea]</td>
</tr>
<tr>
<td>9</td>
<td>Putative uncharacterized protein Sb01g046840 [Sorghum]</td>
<td>C5WXV4</td>
<td>676</td>
<td>90631</td>
<td>0.83 Cell division cycle protein 48</td>
</tr>
<tr>
<td>10</td>
<td>Putative uncharacterized protein Sb04g025560 [Sorghum]</td>
<td>C5XX05</td>
<td>188</td>
<td>53173</td>
<td>0.52 Acetolactate synthase/ amino acid binding protein</td>
</tr>
<tr>
<td>11</td>
<td>Predicted protein [Hordeum vulgare]</td>
<td>F2D6B8</td>
<td>168</td>
<td>36874</td>
<td>0.29 Glyceraldehyde 3-phosphate Dehydrogenase</td>
</tr>
<tr>
<td>12</td>
<td>Acetolactate synthase/ amino acid binding protein</td>
<td>C5XX05</td>
<td>513</td>
<td>53173</td>
<td>1.78</td>
</tr>
<tr>
<td>13</td>
<td>Putative uncharacterized protein Sb04g025120 [Sorghum]</td>
<td>C5XX52</td>
<td>356</td>
<td>36437</td>
<td>0.69 Glyceraldehyde 3-phosphate dehydrogenase</td>
</tr>
<tr>
<td>14</td>
<td>Putative uncharacterized protein Sb06g017850 [Sorghum]</td>
<td>C5Y979</td>
<td>312</td>
<td>25085</td>
<td>0.87 Heat shock</td>
</tr>
<tr>
<td>15</td>
<td>Putative uncharacterized protein Sb09g001700 [Sorghum]</td>
<td>C5YYX3</td>
<td>310</td>
<td>23562</td>
<td>0.94 Chloride intracellular channel/ Dehydroascorbate reductase</td>
</tr>
<tr>
<td>16</td>
<td>Putative uncharacterized protein Sb04g033080 [Sorghum]</td>
<td>C5XRZ8</td>
<td>256</td>
<td>25975</td>
<td>0.44 3-isopropylmalate dehydratase small subunit</td>
</tr>
<tr>
<td>17</td>
<td>Superoxide dismutase [Sorghum]</td>
<td>C5YVR0</td>
<td>205</td>
<td>25341</td>
<td>0.64</td>
</tr>
<tr>
<td>18</td>
<td>Putative uncharacterized protein Sb01g042490 [Sorghum]</td>
<td>C5WSN7</td>
<td>347</td>
<td>30714</td>
<td>0.85 Caleosin (Zea mays) B6U6A9_MAIZE</td>
</tr>
<tr>
<td>19</td>
<td>Proteasome subunit alpha type [Sorghum]</td>
<td>C5WVT0</td>
<td>333</td>
<td>27512</td>
<td>1.23</td>
</tr>
<tr>
<td>20</td>
<td>Putative uncharacterized protein Sb07g025220 [Sorghum]</td>
<td>C5YH68</td>
<td>175</td>
<td>40130</td>
<td>0.08 Sorbitol dehydrogenase B6TEC1_MAIZE</td>
</tr>
<tr>
<td>21</td>
<td>23 kDa alpha-kafrin (Fragment) [Sorghum]</td>
<td>C5WSN7</td>
<td>459</td>
<td>30714</td>
<td>1.27 Caleosin (Zea mays) B6U6A9_MAIZE</td>
</tr>
<tr>
<td>22</td>
<td>Putative uncharacterized protein Sb07g025220 [Sorghum]</td>
<td>C5YH68</td>
<td>144</td>
<td>40130</td>
<td>0.08 Sorbitol dehydrogenase B6TEC1_MAIZE</td>
</tr>
<tr>
<td>23</td>
<td>Putative uncharacterized protein Sb01g042490 [Sorghum]</td>
<td>C5WSN7</td>
<td>165</td>
<td>30714</td>
<td>0.85 Caleosin (Zea mays) B6U6A9_MAIZE</td>
</tr>
<tr>
<td>24</td>
<td>Putative uncharacterized protein Sb06g033080 [Sorghum]</td>
<td>C5YAK8</td>
<td>147</td>
<td>28216</td>
<td>0.25 WAL17 (Triticum aestivum)</td>
</tr>
<tr>
<td>25</td>
<td>Putative uncharacterized protein Sb07g025220 [Sorghum]</td>
<td>C5YH68</td>
<td>147</td>
<td>40130</td>
<td>0.08 Sorbitol dehydrogenase B6TEC1_MAIZE</td>
</tr>
<tr>
<td>26</td>
<td>23 kDa alpha-kafrin (Fragment) [Sorghum]</td>
<td>C5Z557</td>
<td>105</td>
<td>25785</td>
<td>0.13 Glutathione S-transferase GST 19 (Q9FQC0_MAIZE)</td>
</tr>
<tr>
<td>27</td>
<td>Putative uncharacterized protein Sb10g013070 [Sorghum]</td>
<td>C5Z2A2</td>
<td>64</td>
<td>19876</td>
<td>0.37 MFT (Triticum aestivum) G1UE17_WHEAT</td>
</tr>
<tr>
<td>28</td>
<td>Putative uncharacterized protein Sb10g013070 [Sorghum]</td>
<td>C5Z2A2</td>
<td>80</td>
<td>19876</td>
<td>0.17 MFT (Triticum aestivum) G1UE17_WHEAT</td>
</tr>
<tr>
<td>29</td>
<td>Beta-kafrin (Fragment) [Sorghum]</td>
<td>G3FMQ0</td>
<td>570</td>
<td>21289</td>
<td>0.55</td>
</tr>
</tbody>
</table>

QL12
<table>
<thead>
<tr>
<th>Entry</th>
<th>Description</th>
<th>Accession</th>
<th>M2</th>
<th>M3</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Putative uncharacterized protein [Zea mays]</td>
<td>C0PLF0</td>
<td>1265</td>
<td>56804</td>
<td>2.08</td>
</tr>
<tr>
<td></td>
<td>Protein disulfide isomerase [Zea mays]</td>
<td>A5A5E7</td>
<td>1095</td>
<td>56921</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb01g000380 [Sorghum]</td>
<td>C5WRV5</td>
<td>343</td>
<td>60974</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Glyceraldehyde-3-phosphate dehydrogenase (Fragment) [Nicotiana] A1BQW0</td>
<td>A1BQW0</td>
<td>138</td>
<td>21799</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Fructose-bisphosphate aldolase [Sorghum]</td>
<td>C5XFH6</td>
<td>110</td>
<td>38990</td>
<td>0.18</td>
</tr>
<tr>
<td>3</td>
<td>Putative uncharacterized protein Sb01g012640 [Sorghum]</td>
<td>C5WQD2</td>
<td>3734</td>
<td>56804</td>
<td>5.26</td>
</tr>
<tr>
<td></td>
<td>Protein disulfide isomerase [Zea mays]</td>
<td>A5A5E7</td>
<td>1095</td>
<td>56921</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>RuBisCO large subunit-binding protein B6SXW8_MAIZE</td>
<td>B4G1D2_MAIZE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb01g012640 [Sorghum]</td>
<td>C5WQD2</td>
<td>3734</td>
<td>69817</td>
<td>2.46</td>
</tr>
<tr>
<td></td>
<td>Glyceraldehyde-3-phosphate dehydrogenase (Fragment) [Nicotiana] A1BQW0</td>
<td>A1BQW0</td>
<td>138</td>
<td>21799</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Fructose-bisphosphate aldolase [Sorghum]</td>
<td>C5XFH6</td>
<td>110</td>
<td>38990</td>
<td>0.18</td>
</tr>
<tr>
<td>4</td>
<td>Peptidyl-prolyl cis-trans isomerase [Sorghum]</td>
<td>C5XT06</td>
<td>518</td>
<td>18503</td>
<td>2.81</td>
</tr>
<tr>
<td></td>
<td>Peptidyl-prolyl cis-trans isomerase [Sorghum]</td>
<td>B3GQV9</td>
<td>283</td>
<td>18570</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb09g024570 [Sorghum]</td>
<td>C5Z0E2</td>
<td>104</td>
<td>21843</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>23 kDa alpha-kafirin (Fragment) Tax_Id=4558 [Sorghum]</td>
<td>B3VT9A</td>
<td>87</td>
<td>25977</td>
<td>0.27</td>
</tr>
<tr>
<td>5</td>
<td>Pyruvate phosphate dkinase [Sorghum]</td>
<td>Q64N32</td>
<td>7923</td>
<td>103021</td>
<td>6.88</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb01g046840 [Sorghum]</td>
<td>C5WXV4</td>
<td>812</td>
<td>90631</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb01g020910 [Sorghum]</td>
<td>C5X0G5</td>
<td>787</td>
<td>90698</td>
<td>0.89</td>
</tr>
<tr>
<td>6</td>
<td>Putative uncharacterized protein Sb03g046410 [Sorghum]</td>
<td>C5XIK1</td>
<td>395</td>
<td>49578</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>S-locus-specific glycoprotein, putative B9SIR0_RICCO or Ser-</td>
<td>Q6DIB5</td>
<td>87</td>
<td>36665</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb04g025560 [Sorghum]</td>
<td>C5XX95</td>
<td>85</td>
<td>53173</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb04g025560 [Sorghum]</td>
<td>C5XX95</td>
<td>531</td>
<td>53173</td>
<td>1.62</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb03g046410 [Sorghum]</td>
<td>C5XIK1</td>
<td>243</td>
<td>49578</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>S-locus-specific glycoprotein, putative B9SIR0_RICCO or Ser-</td>
<td>Q6DIB5</td>
<td>87</td>
<td>36665</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb04g025560 [Sorghum]</td>
<td>C5XX95</td>
<td>85</td>
<td>53173</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb04g025560 [Sorghum]</td>
<td>C5XX95</td>
<td>531</td>
<td>53173</td>
<td>1.62</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb03g046410 [Sorghum]</td>
<td>C5XIK1</td>
<td>243</td>
<td>49578</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>S-locus-spec glycoprotein or Ser-threonine kinase (Ricinus)</td>
<td>G5DAC5</td>
<td>163</td>
<td>36586</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Fructose-bisphosphate aldolase [Sorghum]</td>
<td>C5XFH6</td>
<td>132</td>
<td>38990</td>
<td>0.18</td>
</tr>
<tr>
<td>8</td>
<td>Proteasome subunit alpha type [Zea mays]</td>
<td>B4FFX7</td>
<td>353</td>
<td>25848</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb01g038760 [Sorghum]</td>
<td>C5WNL8</td>
<td>315</td>
<td>27314</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb09g001700 [Sorghum]</td>
<td>C5YYX3</td>
<td>296</td>
<td>23562</td>
<td>1.54</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb04g033080 [Sorghum]</td>
<td>C5XRZ8</td>
<td>238</td>
<td>25975</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb04g033080 [Sorghum]</td>
<td>C5XRZ8</td>
<td>238</td>
<td>25975</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>3-isopropylmalate dehydratase small subunit?</td>
<td>P29036</td>
<td>230</td>
<td>28178</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>Ferritin-1, chloroplastic [Zea mays]</td>
<td>P29036</td>
<td>230</td>
<td>28178</td>
<td>1.19</td>
</tr>
<tr>
<td>9</td>
<td>Triosephosphate isomerase [Sorghum]</td>
<td>C5XQ07</td>
<td>204</td>
<td>27177</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb04g007010 [Sorghum]</td>
<td>C5XXW5</td>
<td>577</td>
<td>47318</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Delta kafirin-2 [Sorghum]</td>
<td>C5WPV0</td>
<td>294</td>
<td>21674</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Mitochondrial protein target At5g10860 [Saccharum]</td>
<td>C7IVU7</td>
<td>289</td>
<td>22476</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Mitochondrial protein target At5g10860 [Saccharum]</td>
<td>C7IVU7</td>
<td>289</td>
<td>22476</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>ADP-glucose brittle-1 transporter</td>
<td>C5WPV0</td>
<td>294</td>
<td>21674</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>CBS domain protein (Zea mays)</td>
<td>C5WPV0</td>
<td>294</td>
<td>21674</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>CBS domain protein (Zea mays)</td>
<td>C5WPV0</td>
<td>294</td>
<td>21674</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Ferritin-1, chloroplastic [Zea mays]</td>
<td>P29036</td>
<td>230</td>
<td>28178</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>Triosephosphate isomerase [Sorghum]</td>
<td>C5XQ07</td>
<td>204</td>
<td>27177</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Putative uncharacterized protein Sb04g007010 [Sorghum]</td>
<td>C5XXW5</td>
<td>577</td>
<td>47318</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Delta kafirin-2 [Sorghum]</td>
<td>C5WPV0</td>
<td>294</td>
<td>21674</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Mitochondrial protein target At5g10860 [Saccharum]</td>
<td>C7IVU7</td>
<td>289</td>
<td>22476</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Mitochondrial protein target At5g10860 [Saccharum]</td>
<td>C7IVU7</td>
<td>289</td>
<td>22476</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Supplementary Figure S1

RP-HPLC profiles for the alcohol-soluble (prolamin) fraction across the sorghum sample population with allelic variation in kafirin storage proteins (data also shown in figure 2.3).
Supplementary Figure S2

RP-HPLC peak distribution profiles for the water/salt soluble protein fraction across 26 grain sorghum lines with allelic variation in the kafirin seed storage proteins.
Supplementary Figure S3

LOC analysis of the water/salt-soluble A/G fraction across β-kafirin null mutants QL12, IS17214, and RTx2737, and wild-type line 296B.
**Supplementary Figure S4**

Two-Dimensional SDS-PAGE separation of water/salt-soluble (albumin/globulin), alcohol-soluble (prolamin) and alkali-soluble (glutelin) protein fractions across wild-type line 296B, and β-kafirin null mutant QL12. Protein samples were loaded onto 7 cm IPG strips (3-11 NL) and run on IPGphor machine for isoelectric focusing. SDS-PAGE gel (4-12% Bis/Tris small format precast) were utilised for protein separation by size (Mw). Protein spots (circled in yellow) were excised across a range of sizes and pl values, digested with trypsin and identified using LC-MS/MS. Proteins identified from each of the fractions through 2D SDS PAGE/ LC-MS/MS are listed in supplementary table 1. Technical replicates of the gels were run for the A/G and prolamin fractions and spot profiles analysed using LC-MS/MS. Spot 4 from the A/G fraction is a differentially expressed thioredoxin, present in 296B and absent in QL12.

Albumin/Globulin (A/G) (rep 1)
A/G (rep 2)

A/G (rep 3)

296B

QL12
Prolamin (rep 1)

296B

QL12

Prolamin (rep 2)

296B

QL12
Glutelins

296B

QL12

[Image of gel electrophoresis patterns labeled with numbers 1 to 11]
Supplementary Figure S5

Two-dimensional SDS-PAGE analysis of the glutelin fraction across genotypes 296B and QL12. Circled spots were excised from the gel, digested with trypsin and analysed with LC-MS/MS.

Supplementary Information Table 2

HMW alkali-soluble proteins isolated from the glutelin fraction on 2D SDS PAGE gels for sorghum lines 296B and QL12, and identified using LC-MS/MS. Proteins identified include PDI, C4-specific pyruvate orthophosphate dikinase, and a HMW heat shock protein, also identified by others during previous proteomic work on sorghum (Buckner, 1997).

<table>
<thead>
<tr>
<th>Glutelin fraction</th>
<th>Protein Accession</th>
<th>Homology Score</th>
<th>Mw</th>
<th>pl</th>
<th>Conc</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>296B</strong> spot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Putative uncharacterised protein</td>
<td>C5YU58</td>
<td>Heat shock/chaperone</td>
<td>5825</td>
<td>74643</td>
<td>5.12</td>
</tr>
<tr>
<td>C4-specific pyruvate orthophosphate dikinase (Miscanthus)</td>
<td>Q7XJK1</td>
<td></td>
<td>393</td>
<td>103084</td>
<td>5.5</td>
</tr>
<tr>
<td>2 Protein disulphide isomerase</td>
<td>A5A5E7</td>
<td></td>
<td>1879</td>
<td>56921</td>
<td>5.01</td>
</tr>
<tr>
<td>Putative uncharacterised protein</td>
<td>C5WRV5</td>
<td></td>
<td>579</td>
<td>60974</td>
<td>5.2</td>
</tr>
<tr>
<td>C4-specific pyruvate orthophosphate dikinase (Miscanthus)</td>
<td>Q7XJK1</td>
<td></td>
<td>385</td>
<td>103084</td>
<td>5.5</td>
</tr>
<tr>
<td>3 Putative uncharacterised protein</td>
<td>C5YU58</td>
<td>Heat shock/chaperone</td>
<td>4426</td>
<td>74643</td>
<td>5.12</td>
</tr>
<tr>
<td>C4-specific pyruvate orthophosphate dikinase (Miscanthus)</td>
<td>Q7XJK1</td>
<td></td>
<td>186</td>
<td>103084</td>
<td>5.5</td>
</tr>
<tr>
<td>4 Laminal-binding protein (maize)</td>
<td>P24067</td>
<td></td>
<td>4796</td>
<td>73211</td>
<td>5.1</td>
</tr>
<tr>
<td>Putative uncharacterised protein</td>
<td>C5XPN2</td>
<td></td>
<td>986</td>
<td>71261</td>
<td>5.1</td>
</tr>
<tr>
<td>C4-specific pyruvate orthophosphate dikinase (Miscanthus)</td>
<td>Q7XJK1</td>
<td></td>
<td>1683</td>
<td>103084</td>
<td>5.5</td>
</tr>
<tr>
<td><strong>QL12</strong> spot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Putative uncharacterised protein</td>
<td>C5YU58</td>
<td>Heat shock/chaperone</td>
<td>4887</td>
<td>74643</td>
<td>5.12</td>
</tr>
<tr>
<td>Putative uncharacterised protein</td>
<td>C5WNX8</td>
<td>Heat shock/chaperone</td>
<td>2007</td>
<td>52635</td>
<td>5.46</td>
</tr>
<tr>
<td>2 Putative uncharacterised protein (maize)</td>
<td>COPLF0</td>
<td>Protein disulphide isomerase</td>
<td>1265</td>
<td>56804</td>
<td>4.94</td>
</tr>
<tr>
<td>Protein disulphide isomerase (maize)</td>
<td>A5A5E7</td>
<td></td>
<td>1095</td>
<td>56921</td>
<td>5.01</td>
</tr>
<tr>
<td>Putative uncharacterised protein</td>
<td>C5WRV5</td>
<td></td>
<td>343</td>
<td>60974</td>
<td>5.2</td>
</tr>
</tbody>
</table>