Proteomic profiling of wheat and barley grain for cultivar and provenance identification

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

by:

Paul Worden

Department of Molecular Sciences, Macquarie University, Sydney, NSW, 2109, Australia

December 2018

Table of Contents

List of Figures

List of Tables

Declaration

The work presented in this thesis was carried out between June 2014 and October 2017 on a full-time basis. This work represents original research, which has not been submitted for any other degree or to any other University or institution. All work was carried out by the author unless otherwise acknowledged.

Candidates Signature

Paul Worden

Acknowledgements

I would like to thank Professor Robert Willows and Professor Brian Atwell for their help and support over the years of my PhD. I would also like to thank Dr Mehdi Mirzaei, Dr Yunqi Wu, and Dr Samantha Emery for their excellent advice and assistance in the laboratory and other PhD support. Not to be forgotten, I must thank Dr Ante Jerkovic and Dr Artur Sawiki for their tireless efforts and help in editing this thesis.

I would also like to thank the Australian Research Council (ARC) Industrial Transformation Training Centre (ITTC) for funding my scholarship, without which completion of this PhD would have been impossible.

Lastly, I must thank all my friends and family for having the patience to put up with me while completing this PhD project. I would especially like to thank my wife Torgui Worden, my mother Anne and sister Mary, as well as my son Alex Worden and step-son Andres Varela. Through everyone's support I have managed to overcome all the personal and intellectual challenges that this PhD project has thrown at me.

Publications and conference attended

Publication 1. Appendix E

MIRZAEI, M., WU, Y., WORDEN, P., JERKOVIC, A. & ATWELL, B. J. 2016. How Proteomics Contributes to Our Understanding of Drought Tolerance? In: SALEKDEH, G. H. (ed.) Agricultural Proteomics Volume 2. Switzerland: Springer International Publishing.

Publication 2. Appendix F

Willows RD, Worden P, Mirzaei M (2017) Barley Grain Proteomics. Proteomics in Food Science: From Farm to Fork:75-88. doi:10.1016/B978-0-12-804007-2.00005-9

Conference attended and poster presentation

Worden, P., Willows, R., Atwell, B., Merzaei, M. 2016. Using proteomics to assess provenance and quality of wheat and barley grain. Poster session presented at the Advances in Biotechnology for Food and Medical Applications Workshop, Sydney University, Sydney.

Abstract

The proteomes of wheat and barley grain cultivars have evolved through genetic variability and environmental factors. The ability to identify different cultivars, farm origin, and quality of wheat and barley grain are becoming increasingly important to farmers, processors, and food manufacturers. This will help them to diversify into higher value boutique products, or just to add value through security and traceability to bulk grain or flour exports. This project aims to use modern proteomic techniques and transcriptomics to discover proteins that can be used as biomarkers in wheat and barley grain to identify crop cultivars, grain provenance (farm origin) and possibly grain quality. Firstly, several protein extraction methods were assessed for optimal protein yield and diversity. This was important for maximising the discovery of potential protein biomarkers through proteomic analysis. By applying Tandem Mass Tags (TMTs) labelled shotgun proteomics, a subset of grain proteins was detected from wheat and barley that show statistically significant differential expression between different cultivars and different farm locations. Indeed, serpin and chitinase proteins (observed to be involved in stress response) were found to be differentially expressed in the wheat and barley sample proteomes. The results also indicated that the differentially expressed proteins from wheat and barley grain have the potential to be used as biomarkers for probable quality traits. The assigned protein biomarkers between cultivars or a particular cultivar from a different environment (farm location) have almost identical functional summaries (gene ontology [GO] Slims). Investigations into wheat mRNA expression between cultivars showed GO Slims that were analogous to the proteomic results. Further experiments involving proteomics and common traditional quality testing such as, 1000-kernel weight, farinograph, extensograph, baking tests and falling number, are needed to answer this question, and is beyond the scope of this project. Protein-based tests to identify cultivar, farm origin, and grain quality have the potential to address these needs in a manner that would be faster relative to existing quality controls.

Abbreviations

ABA, abscisic acid; APX, Ascorbate peroxidase; bHLH, Basic helix-loop-helix; CAT, Chloramphenicol acetyltransferase; CCCH, Cysteine3Histidine; GAPDH, Glyceraldehyde-3-Phosphate Dehydrogenase; Gene Ontology, GO; GS1-2, Glutamine synthetase cytosolic isozyme 2; iTRAQ, Isobaric tag for relative and absolute quantitation; LC-MS/MS, Liquid Chromatography tandem mass spectrometry; LEA, Late Embryogenesis Protein; MAF1, Matrix attachment region-binding protein-association factor; MFP1, MAR binding filament–like protein 1; MALDI, Matrix-assisted laser desorption/ionization; MALDI-TOF, Matrix-assisted laser desorption/ionization-Time of flight; MFP-1, MAR-binding filament-like protein 1; MS, Mass Spectrometry; nano LC-ESI-MS/MS, Liquid Chromatography-Electrospray ionization-tandem mass spectrometry; NOD26, nucleotide-binding oligomerization domain26; OEC, Oxygen Evolving Complex; OEE1, oxygen evolving enhancer protein 1; PP2C, Protein phosphatase 2C; PPDK, Pyruvate phosphate dikinase; ROS, Reactive oxygen species; RuBisCO, Ribulose-1,5-bisphosphate carboxylase/oxygenase; SCX, Strong cation-exchange chromatography; SNAP, Soluble NSF-attachment proteins; SnRK, SNF-related serine/threonine-protein kinase; SOD, Superoxide dismutase; TCTP, Translationally controlled tumour protein; TPI, Triosephosphate isomerase; V-ATPase, Vacuolar-type H^+ ATPase; WCOR, wheat cold-responsive protein; 2-DE, Two dimensional electrophoresis.

Chapter 1. Introduction

1.1. The Grasses

1.1.1. Plants and the family Poaceae (grasses)

Plants have been colonising Earth's land masses for at least 475 million years (Wellman, et al., 2003). Over this time, both the number of species and their increased complexity has provided a large variety of plant forms, from simple Bryophytes to the increasingly complex Ferns, Gymnosperms, Angiosperms (flowering plants), are split into Eudicots, also generally known as dicots, and Monocots. These two classes of Angiosperm differ in their seed structure as well as their vascular and pollen structure. Through this evolution, plants have been able to colonise almost all land surfaces including areas of extreme temperatures, and arid conditions with limited nutrients. The most recently evolved group – the Angiosperms – are not only the most successful in terms of species by number (Table 1.1), they are also successful in their ability to inhabit vast areas of land (Figure 1.3). This is especially apparent when looking at the scale of land that is covered by crops, which are themselves no more than monocultures of only a few domesticated plant species from the two major Angiosperm phyla, Dicots and Monocots. In 2014, annual crops (arable land) and permanent crops together made up 12.18% of global land. Other commonly used plants for human consumption make up the major portion at 10.9% according to FAOSTAT (http://www.fao.org/faostat/en/#data/EL). Moreover, in terms of geographic success, it is the family of monocots known as the Poaceae (grasses), that have been the most successful. Especially the small number of domesticated annual crops known as cereals, that have evolved under human selection from the grasses., and through human intervention come to dominate large swathes of land.

As such, it is not difficult to see that in terms of geographic success, for both domesticated and wild species, the most successful family of monocots are the Poaceae (grasses).

Table 1.1. Estimation of plant species numbers

20,000 species of Mosses and liverworts (*Bryophytes)*
13,000 species of Ferns or Fern allies (*Pteridophytes)*
1,000 species of Conifers, Cycads and allies (*Gymnosperms)*
352,000 species of flowering plants or (*angiosperms*)

Data is from (Kew, 2016).

1.1.2. The success story of grasses

Since their relatively recent divergence from their fellow monocots (Figure 1.6), grasses have differentiated into approximately 12,000 different species (Gaut, 2002). These include 12 subfamilies, 51 tribes, and 80 subtribes (Soreng, et al., 2015). A phylogenetic summary of the grasses is shown in

Figure **1.1** which compares approximately 350,000-400,000 vascular plant species from www.theplantlist.org (Kew, 2016), making the grasses - at least in terms of numbers of species - one of the most diverse plant families. As discussed above, the extensive area of land covered by Poaceae make them easily the most successful of all plant families. Not only do they cover a significant portion of the Earth's land mass, they have evolved to survive in almost every terrestrial environment on the planet; from arctic climates, to dense jungle, and hot and dry deserts (Figure 1.2 andFigure 1.3).

In 2005, it was estimated that cultivated and non-cultivated grasslands covered 40.5% of the Earth's land surface – excluding Greenland and Antarctica (FAO, 2005). Due to their importance as a primary food source, as well as an economic commodity, cereal grass crops are the top three harvested crops behind the weight of sugar cane stems (Table 1.2).

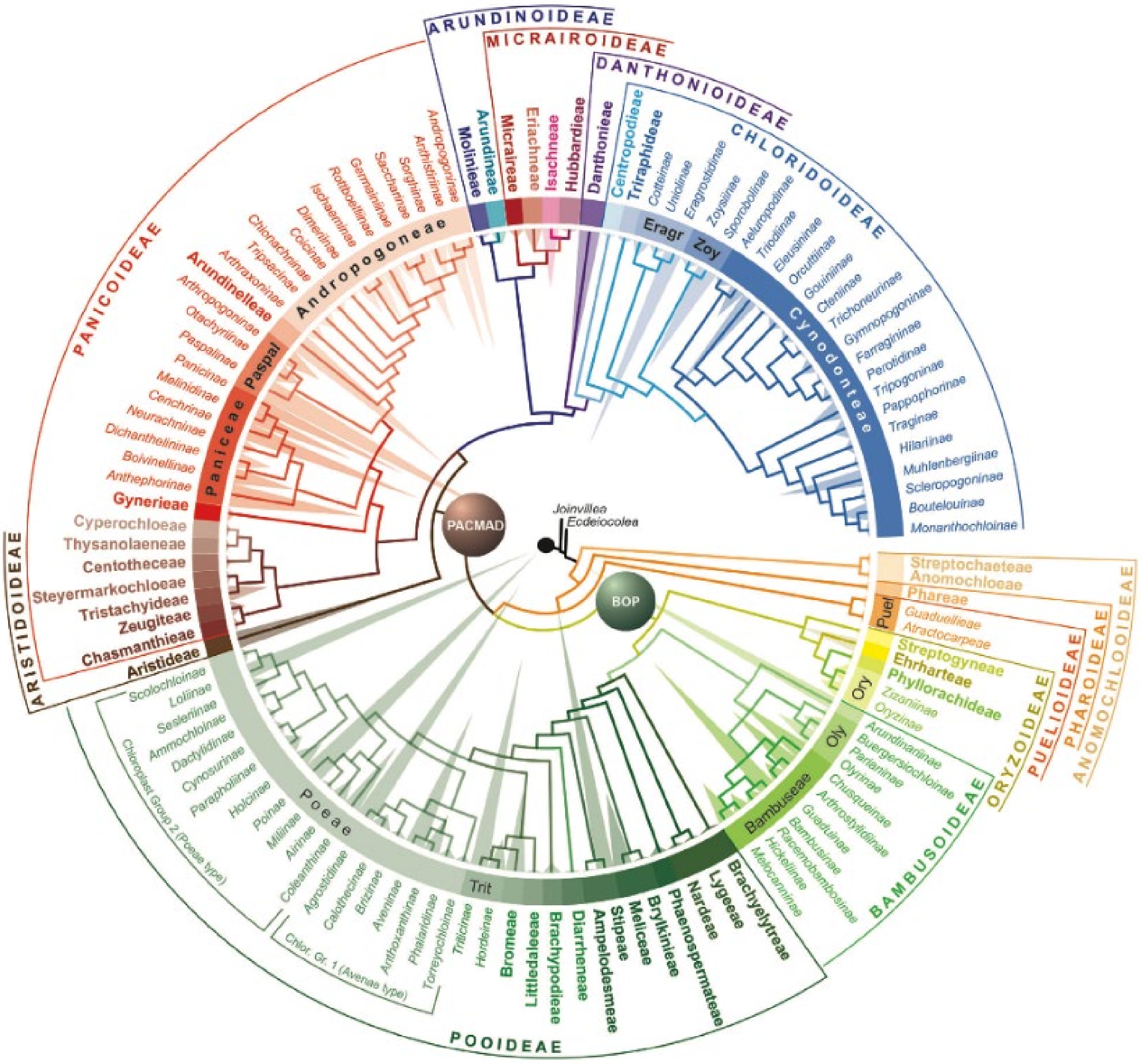


Figure 1.1. Phylogenetic summary of the grasses (Soreng, et al., 2015). PACMAD (previously known as PACC) is one of the two major clades of the Poaceae and is an acronym constructed from the first initials of the included subfamilies Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae, and Danthonioideae. BOP (or BEP) clade is a sister group to the PACMAD clade, and contains the three subfamilies Bambusoideae, Oryzoideae, and Pooideae.

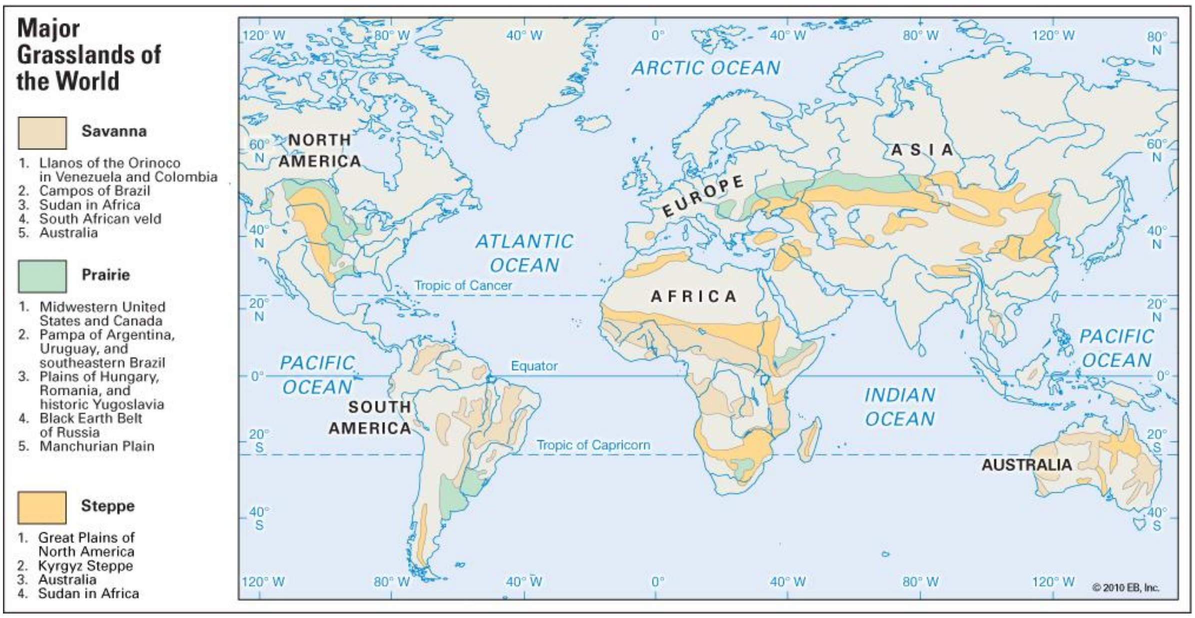


Figure 1.2. Types and extent of grasslands. Figure taken from the encyclopaedia Britannica website - https://www.britannica.com/science/grassland).

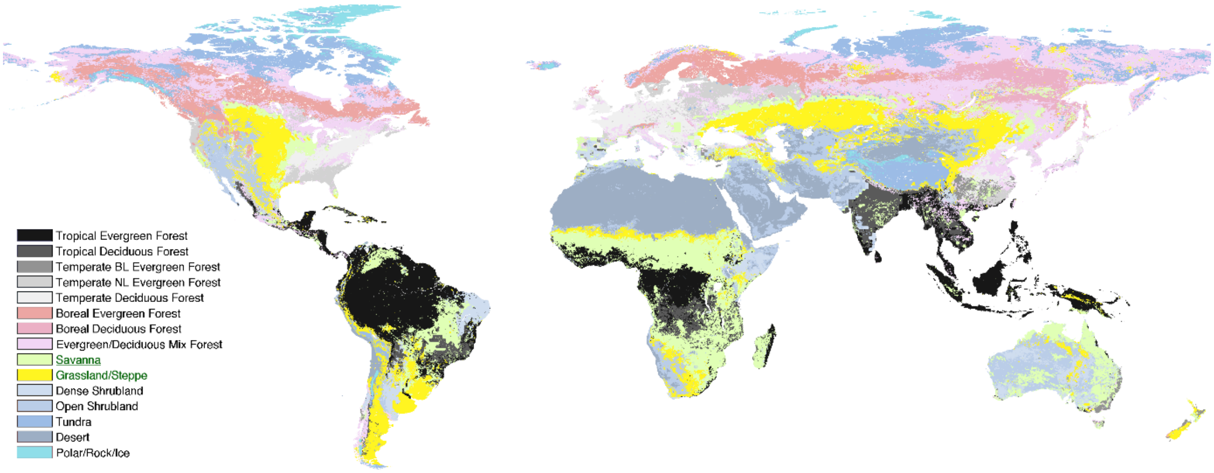

Tropical Evergreen Forest
Tropical Deciduous Forest
Temperate BL Evergreen Forest
Temperate NL Evergreen Forest
Temperate Deciduous Forest
Boreal Evergreen Forest
Boreal Deciduous Forest
Evergreen/Deciduous Mix Forest
Savanna
Grassland/Steppe
Dense Shrubland
Open Shrubland
Tundra
Desert
Polar/Rock/Ice

Figure 1.3. Types of vegetation and their geographical extent. Figure was taken from "http://www.earthstat.org/" and Ramankutty and Foley (1999), and re-coloured to emphasise both savanah and grassland.

Table 1.2. Top ten harvested foods in the world for 2014 by weight (tonnes).

Item	World Total (tonnes)
Sugar cane	1,884,246,253
Maize	1,037,791,518
Rice, paddy	741,477,711
Wheat	729,012,175
Potatoes	381,682,144
Soybeans	306,519,256
Vegetables	289,788,862
Oil, palm fruit	274,618,164
Sugar beet	269,714,066
Cassava	268,277,743

Table 1.3. Top ten harvested foods in the world for 2014 by area (hectare).

Item	World Total (ha)
Wheat	220,417,745
Maize	184,800,969
Rice, paddy	162,716,862
Soybeans	117,549,053
Barley	49,426,652
Sorghum	44,958,726
Rapeseed	36,117,722
Seed cotton	34,747,265
Millet	31,432,088
Beans, dry	30,612,842

Data is from FAOSTAT (http://www.fao.org/faostat/en/#data/QC), and the grey shading represents crops from the family Poaceae.

1.1.3. Wheat and barley

The coverage of the grasses was expanded by a growing human need for a stable source of food, hence both pasture and crops and have come to dominate vast areas of land over largely differing environmental conditions. A tiny subset of grass species are major crops (the cereals) and in 2014, they were reported to cover an estimated 693,753,042 hectares of land (Table 1.3). The two crops that are the focus of this project, wheat and barley, account for a total of 873,502,171 tonnes of grain harvested (Table 1.4), and 269,844,397 hectares in crop area (Table 1.3). The vastness and varied environments of these cereal crops are evident in the maps shown in Figures 4 and 5.

Table 1.4. Top eight Cereals in terms of worldwide tonnes harvested in 2014.

Number	Cereal	Tonnes of grain
1	Maize (*Zea mays*)	1,037,791,518
2	Rice (*Oryza sativa)*	741,477,711
3	Bread Wheat and Durum Wheat (*Triticum aestivum and Triticum durum)*	729,012,175
4	Barley (*Hordeum vulgare)*	144,489,996
5	Sorghum (*Sorghum bicolor)*	68,938,587
6	Foxtail Millet and Pearl Millet (*Setaria italic* and *Pennisetum glaucum)*	28,384,668
7	Oat (*Avena sativa)*	22,721,702
8	Triticale [Wheat and Rye hybrid] (× *Triticosecale)*	16,953,565

Data is from FAOSTAT (http://www.fao.org/faostat/en/#data/QC).

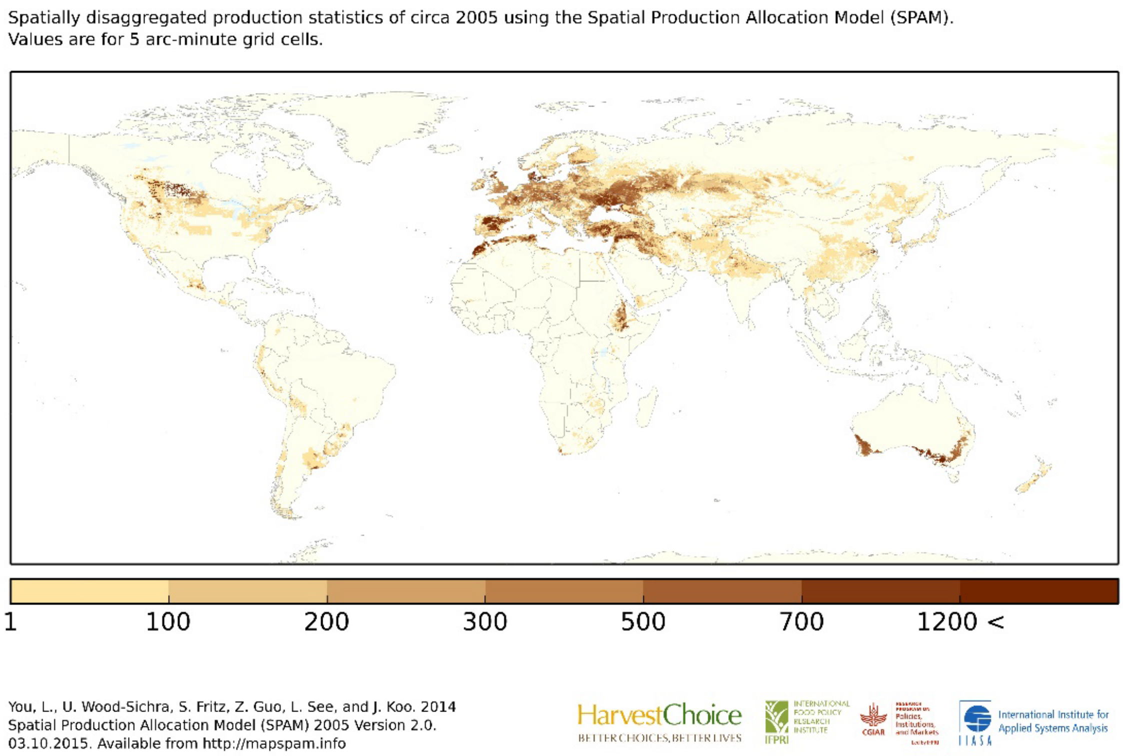

Spatially disaggregated production statistics of circa 2005 using the Spatial Production Allocation Model (SPAM).
Values are for 5 arc-minute grid cells.
1
100
200
300
500
700
1200 <
You, L., U. Wood-Sichra, S. Fritz, Z. Guo, L. See, and J. Koo. 2014
Spatial Production Allocation Model (SPAM) 2005 Version 2.0.
03.10.2015. Available from http://mapspam.info
HarvestChoice
BETTER CHOICES, BETTER LIVES
INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI
CGIAR
Policies, Institutions, and Markets
International Institute for Applied Systems Analysis
IIASA

Figure 1.4. Area of wheat under harvest in hectares (You and Wood, 2005).

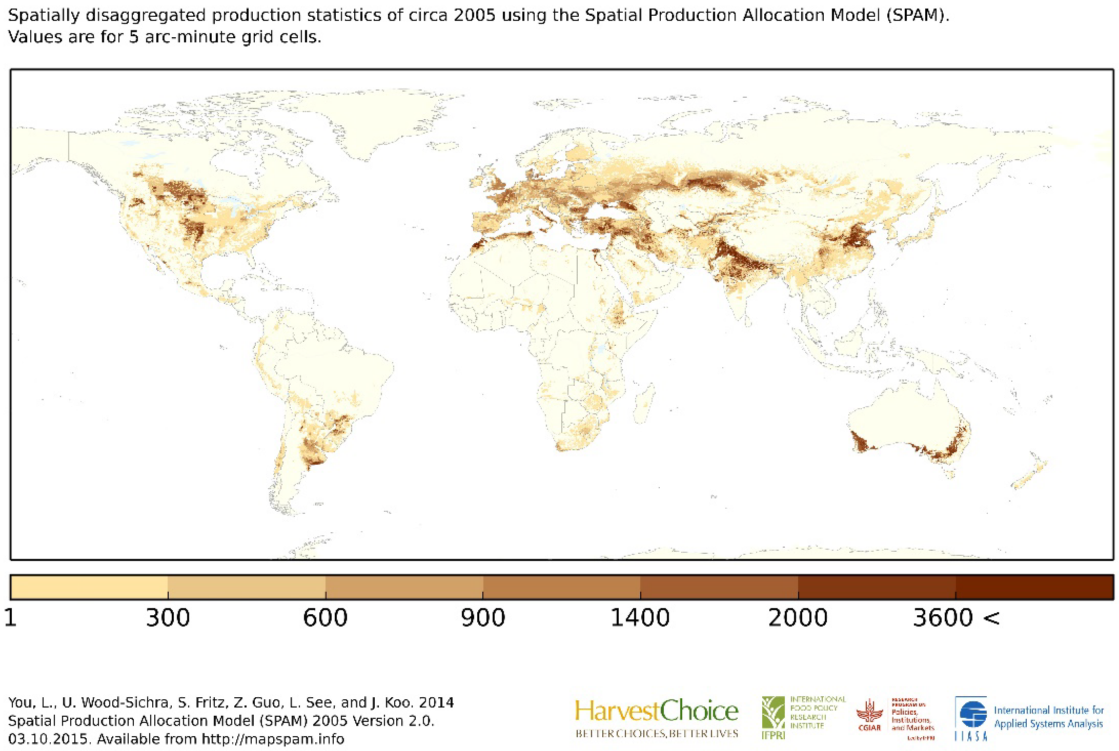

Spatially disaggregated production statistics of circa 2005 using the Spatial Production Allocation Model (SPAM).
Values are for 5 arc-minute grid cells.
1
300
600
900
1400
2000
3600 <
You, L., U. Wood-Sichra, S. Fritz, Z. Guo, L. See, and J. Koo. 2014
Spatial Production Allocation Model (SPAM) 2005 Version 2.0.
03.10.2015. Available from http://mapspam.info
HarvestChoice
BETTER CHOICES, BETTER LIVES
INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI
CGIAR
Policies, Institutions, and Markets
International Institute for Applied Systems Analysis
IIASA

Figure 1.5. Area of barley under harvest in hectares (You and Wood, 2005).

1.1.4. Origins of the cereals

The initial divergence of Poaceae was around 75-50 million years ago and further divergence into two clades occurred around 50 million years ago, called the PACMAD (previously known as PACC) and the BOP (or BEP) clade (Figure 6). One of the sub-families to emerge from the PACMAD clade at around 28 million years ago was the Panicoideae. This sub-family gave rise to the ancestors of foxtail millet (*Setaria*) and pearl millet (*Pennisetum*) at around 14 million years ago, and the ancestors of sorghum (*Sorghum*) and maize (*Zea*) at around 9 million years ago. The BEP clade, from which wheat and barley evolved, the subfamily *Oryzoideae* (rice) was the first to branch off at around 46 million years ago. This was followed by the sub-family Pooideae at around 25 million years ago, which then gave rise to, *Avena* (oats), followed by *Hordeum* (barley) and *Triticum* (wheat) diverging at around 13 million years ago (Figure 1.6).

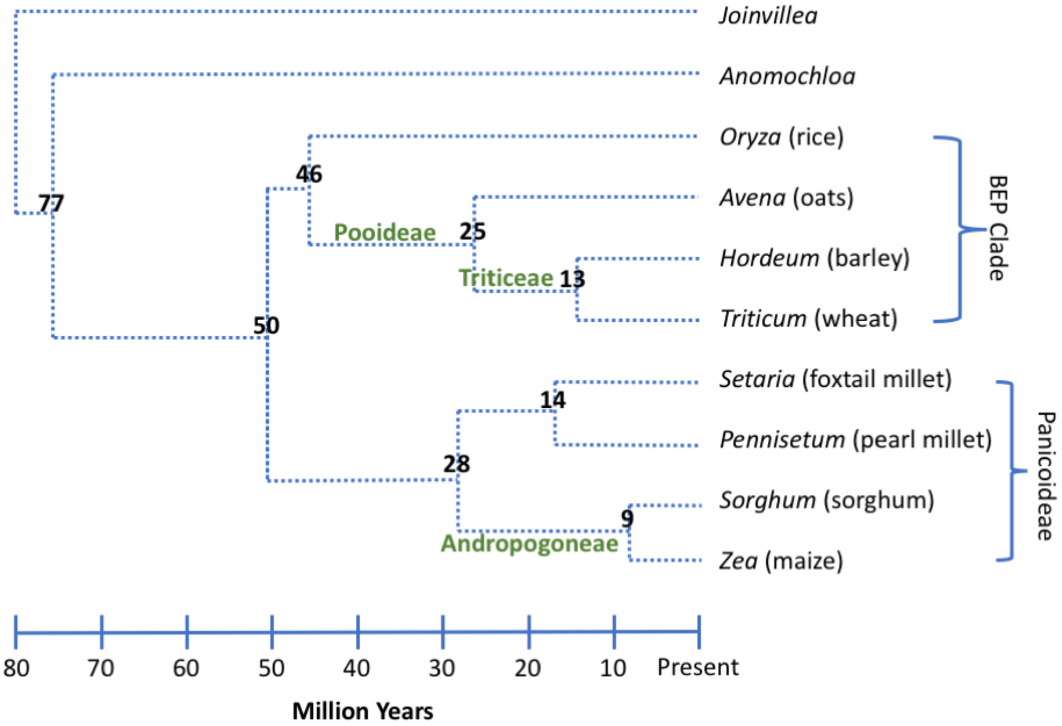


Figure 1.6. Evolution of cereals, adapted from Gaut (2002) and Ji, et al. (2013).

1.1.5. The domestication of wheat and barley

The domestication of wheat and barley resulted from the development of human agriculture, both in terms of creating pastures for livestock which was mostly dominated by grasses as well as the various cereal crops such as maize, rice, sorghum, millet, oats, and rye. Modern wheat and barley are examples of how grasses and their genetic plasticity (Figure 1.7) have been cultivated through selection of favoured traits to dominate large swathes of the earth enabling them to be geographically dominant.

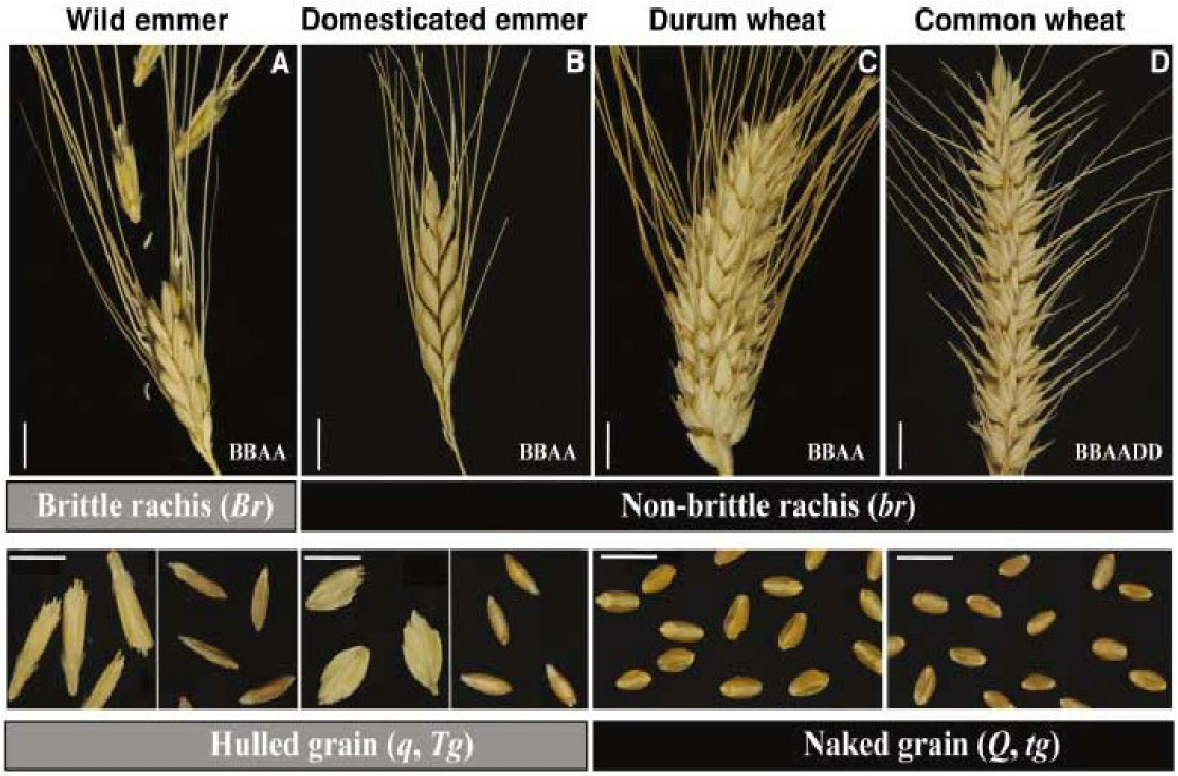


Figure 1.7. Phenotypic changes to wheat spikes and grain with domestication. The figure taken from Dubcovsky and Dvorak (2007) showing the loss of a brittle rachis and grain hull with increasing domestication.

1.1.6. Origins of wheat and barley

There is a large body of evidence that modern durum and bread wheat, as well as barley and a number of other crops had their origin within an area located around southeast Turkey and northern Syria (Figure 1.8), known as the fertile-crescent (Abbo, et al., 2006, Lev-Yadun, et al., 2000, Salamini, et al., 2002). This area is home to a number of wild cereals, one of which is the tetraploid wild emmer wheat (*Triticum dicoccoides*). This species evolved from a natural hybridization that occurred 500,000 years ago between a diploid wild wheat *Triticum urartu* (AA) and an unidentified diploid goat grass (genus: *Aegilops* - BB). Millennia later, Emmer wheat had a seemingly important role in the evolution of modern durum and bread wheat. The diploid wild einkorn wheat (*Triticum boeoticum*) with a somewhat overlapping but more extensive range than emmer wheat as well as wild barley (*H. spontaneum*), had a major but different role in the development of agriculture within the fertile-crescent.

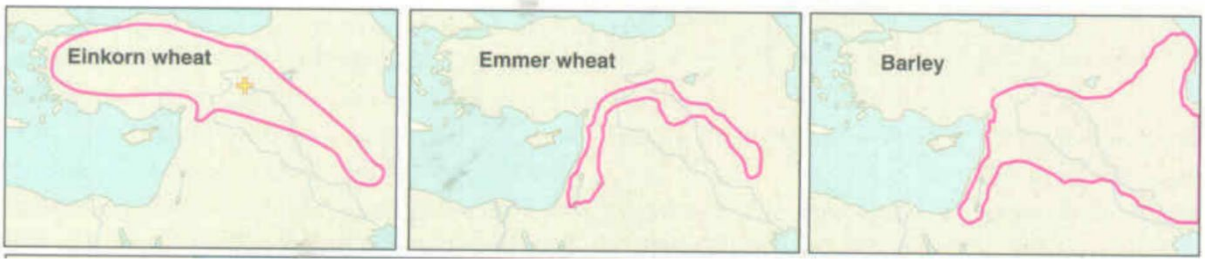


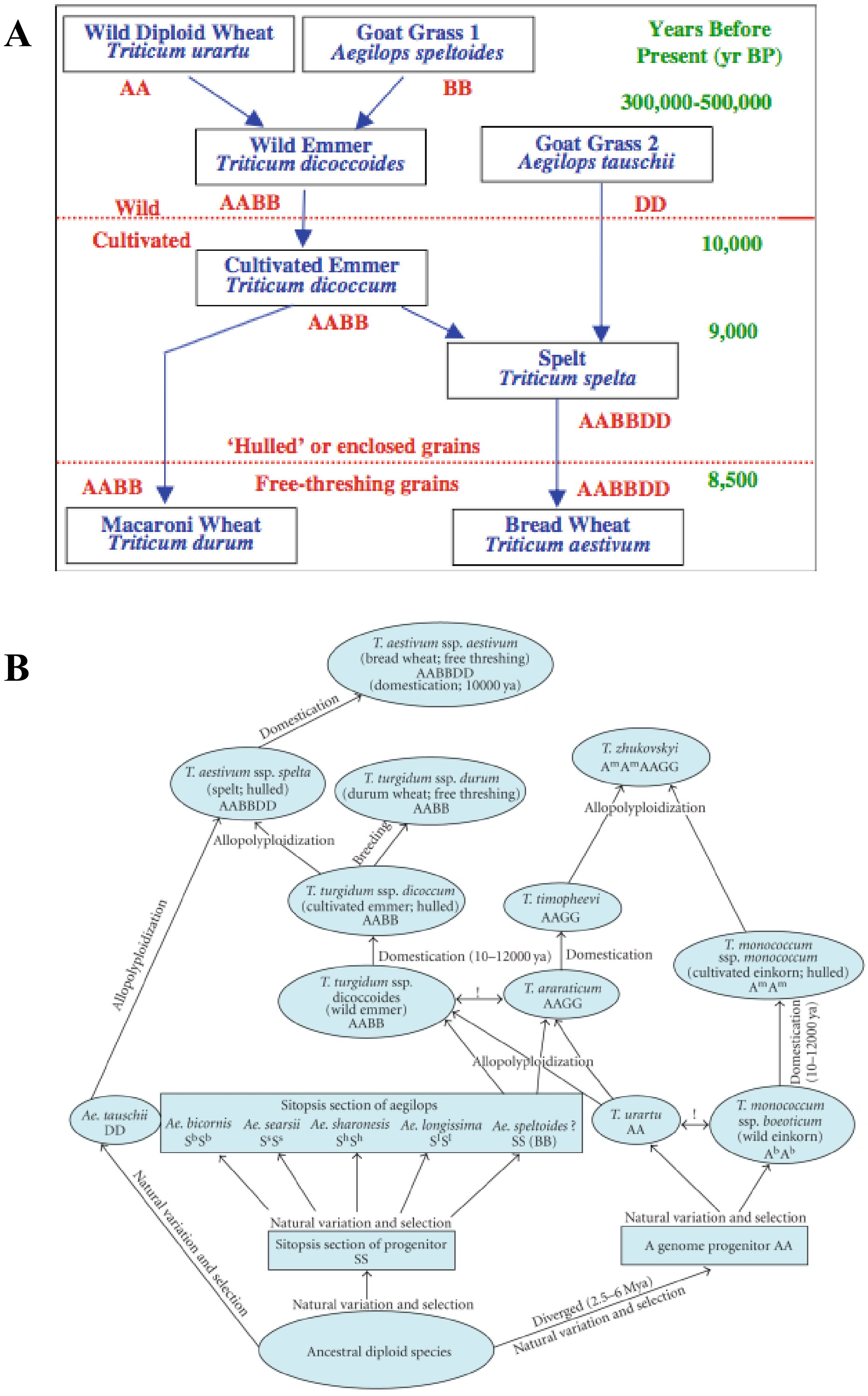
Figure 1.8. Extent of wild cereal coverage from (Lev-Yadun, et al., 2000).

1.1.7. Early cereal cultivation

The earliest gathering of wild cereals has been dated to approximately 19,000 years ago (Piperno, et al., 2004). It appears that 10,000 years ago, just after the Younger Dryas period (12,900 to 11,700 years ago), temperatures and rainfall diminished (Nesbitt, 2001). This created a shift in human food gathering from foraging to farming (Nesbitt, 2002). There is evidence of new sets of tools being invented around this time (Lev-Yadun, et al., 2000), as well as evidence of crop cultivation, through charred plant remains. This resulted in: The harvesting of cereal grains and lentils outside their natural habitats; barley and various wheats appearing at different times; weeds that only grow in cultivated fields; a decrease in the amount of smaller grass seeds; barley grains becoming larger; and large amounts of rodent droppings indicating large scale storage (Willcox and Willcox, 2008).

1.1.8. Domestication of wheat

The first evidence of domesticated emmer wheat (*Triticum dicoccum*) appeared around 10,600 to 9,900 years ago, with domesticated einkorn (*Triticum monococum*) also appearing around this time (Weiss and Zohary, 2011). Full domestication was a slow process, selecting for traits such as seed size and weight, flowering time, grain yield, plant height, number of spikes and kernels, weight of spikes, as well as the lack of a brittle rachis and a weak glume (Peng and Peng, 2011). The latter two traits are particularly important in terms of improved efficiency of grain harvest and grain processing (Peng, 2003). Wild wheat has a brittle rachis to allow seed dispersal and tough glumes for seed protection. Early domesticated relatives also have tough glumes but have evolved a strengthened rachis so that seed spikes do not easily break before or during harvesting, allowing efficient collection under human selection (Hillman and Davies, 1990, Nesbitt, 2001). Similarly, over thousands of years, the modern wheat and barley cultivars had evolved a fully toughened rachis, as well as weak glumes that result in free threshing, or naked grain (Feldman and Kislev, 2007, Nesbitt, 2001). Quantitative trait loci (QTL) analysis demonstrated that while some traits seemed mostly independent of one another, many traits seemed to be under the control of multiple loci (gene interactions), complicating and potentially slowing the pace of selection (Peleg, 2011, Peng, 2003). This was possibly a result of the difficulty (or rarity) of selection, reduced survival fitness of domesticates, and/or farming practices that slowed trait selection (i.e. harvesting before maturity and not selecting for a toughened rachis). As such, it appears that wild cereals were grown alongside domesticates for almost a thousand years (Kislev, 1984). During this reportedly slow domestication, agriculture expanded and brought domesticated emmer (AABB) into the habitat of another goat grass *Aegilops tauschii* (DD), which prompted a hybridization between the two, resulting in a hexaploid wheat (AABBDD) called Spelt (Giles and Brown, 2006, Kerber, 1964, Kislev, 1980, Matsuoka and Nasuda, 2004, Nesbitt and Samuel, 1996, Salamini, et al., 2002). With the continued cultivation of wild and domesticated forms of wheat occurring side by side, it appears that domestication arose multiple times according to a polycentric origin of the cereals (Balter, 2007, Tanno and Willcox, 2006, Willcox and Willcox, 2008). Sometime around 8,500 years ago, the selection of spelt and emmer traits which are desired by the farmer societies resulted in more free-threshing types (Peng and Peng, 2011) with evidence that these traits were sourced from genomes of wild emmer (Dvorak et al. 2006). Over the next few millennia, these domesticated wheat genomes were not only adaptable enough for continued selection to enable the emergence of durum and bread wheat from emmer and spelt respectively, they were also diverse enough to allow the expansion of wheat agriculture into the multiple and varied terrestrial environments that we see today. A summary of the domestication of bread wheat and durum is shown in Figure 1.9.



! Morphologically almost indistinguishable but not interfertile

Figure 1.9. Summary of: A) The domestication of Durum and Bread Wheat, and B) The domestication of bread wheat. Figures taken from (http://www.newhallmill.org.uk/wht-evol.htm; (Gupta, et al., 2008) respectively. The repeated capital letters represent chromosome sets of the species described (eg. for figure A above; AA, BB, CC and their combinations). Note the sitopsis section of Aegilops comprises: *Ae. bicornis*, *Ae. longissima*, *Ae. sharonensis*, *Ae. searsii* and *Ae. speltoides*.

1.1.9. The domestication of barley

The domestication of barley had a similar, although slightly less complex story. McCorriston (2000) reported that the evidence lies in favour of barley domestication occurring a few hundred years after the first wheat domesticates arose, possibly due to beer brewing (Katz and Voigt, 1986). Domestication of barley appears to have first taken place in Turkey between 10,600 to 9,900 years ago (Weiss and Zohary, 2011) within parts of the fertile crescent and areas to the north and east of this location and may have been concomitant with wheat (Figure 1.8). Like wheat, wild barley (*Hordeum spontaneum*) was an important food for people within its growing range for thousands of years before a domesticated form of barley (*Hordeum vulgare*) was produced (Nadel, et al., 2015). As with wheat, domesticated and wild barley seem to have been cultivated side-by-side for hundreds of years (Hopf, 1983). Genetic flow between the wild and domestic barley populations, in concert with active selection, eventually led to the emergence of the six-row and higher protein content barley around 8800 years ago (Helbaek, 1959). Finally, over thousands of years of continued cultivation and selection – including under irrigation on the steppes and deserts of the near east – led to the evolution of the two-and six-row barley cultivars that are grown today. Botanical studies observed that the period of ripening for barley evolved into a narrower time-frame with the establishment of a single-season seed dormancy (McCorriston, 2000).

In summary, grasses have taken advantage of genomic plasticity, together with a number of other advantageous traits such as fast growth and regeneration time, and a general tolerance to abiotic and biotic stress. These traits have allowed the grasses to evolve and adapt to a myriad of different environments all over the planet. Additionally, the growth and development of agriculture has been particularly important in the success of domesticated cereal crops. Whether the need for expanding pasturelands or crops, grasses have been the major beneficiaries of the synergy that exists between humans and plants. In terms of wheat and barley alone, the large area and number of differing environments that they cover is remarkable. Clearly the genomes of these cereals have been plastic enough to introduce self-induced changes to their genes and gene control mechanisms, as well as exchange genes with wild relatives (close and distant). Thus, with human intervention, this has resulted in the evolution of the modern commercial forms (*Triticum aestivum*, *Triticum durum*, and *Hordeum vulgare* L.*),* and allowed for the breeding of the hundreds of cultivars currently used.

1.2. Early wheat grain protein extraction methods

Proteins in wheat were first investigated in 1745 (Aliscioni, et al., 2012), when Bartolomeo Beccari used water to wash away all soluble wheat flour proteins to investigate the insoluble gluten fraction. Later, Taddei (1819) used aqueous alcohol to separate gluten into the gliadin (alcohol soluble) and glutenin (alcohol insoluble) fractions. These extraction methods were refined further by Osborne and Voorhees (1893) and Aliscioni, et al. (2012) resulting in the four major protein fractions, known as the "Osborne" fractions: 1) Albumins, which are soluble in water; 2) globulins, which are soluble in salt water; 3) gliadins, which are soluble in aqueous alcohol; and 4) glutenins, which are insoluble in aqueous alcohol (Osborne, 1907).

1.2.1. Modern extraction methods for wheat grain proteins

Improvements in water, salt and solvent-based wheat grain protein extraction methods were developed by further separation and purification by gel filtration (size-exclusion chromatography). Several different buffers and detergents, such as urea (Pomeranz, 1965), SDS (Bottomley, et al., 1982, Danno, et al., 1974, Graveland, et al., 1979), urea/SDS (Gao and Bushuk, 1992), acetic acid/urea/cetyltrimethyl ammonium bromide (Kurowska and Bushuk, 1988, Meredith and Wren, 1966), and Triton X (Blochet, 1991) have been used to solubilise proteins in order to prepare them for gel filtration. As a result, the major wheat grain proteins may be separated into four main fractions of decreasing size; high-molecular-weight glutenin subunits (MHW-GS), ‘low-molecular weight glutenin subunits (LMW-GS), gliadins and albumin/globulins (Bottomley, et al., 1982, Huebner and Wall, 1976).

Gel filtration is time-consuming, unfortunately, and the protein analysis data obtained were difficult to reproduce. The introduction of size exclusion-high performance liquid chromatography (SE-HPLC) allowed for a faster and more quantitative approach that could also be automated (Batey, 1991, Dachkevitch and Autran, 1989). Further improvements in wheat grain protein analysis included the use of an ultrasonic-probe in the extraction process (Singh, et al., 1990, Singh and MacRitchie, 1989).

1.3. Protein extraction methodology

Modern proteomics relies on the assumption that protein extracts are accurate snapshots of the sample proteomes at a given point in time. This can only be achieved if extraction protocols satisfy three conditions: 1) samples are free from contamination of non-target tissue, 2) samples are ground into a fine powder or paste to enable efficient protein solubilisation by buffers and/or detergents, solvents, and 3) all proteins are extracted while removing any contaminants that would affect downstream processing.

Given that wheat grain is relatively dry and that an optimal sample homogenization method for grain had been previously optimised in our laboratory (Jerkovic, 2011), this study focuses on further optimising the protein extraction method. The development of a method to extract protein that is representative of the sample proteome has held the greatest challenge and has a long evolution in the history of proteomic analysis.

With the introduction of 2-dimensional electrophoresis (2-DE) in the mid 1970’s by O'Farrell (1975), the use of two independent properties – molecular weight and isoelectric point –

allowed researchers to resolve proteins into discrete protein spots on a gel (Figure 10). The protein spots could then be further analysed for relative abundance and identification using gel image analysis software and mass spectrometry. However, 2-DE and recent proteomic techniques require protein samples to be free of interfering contaminants as they can modify proteins and/or cause charge heterogeneity and result in the generation of artefacts (Hari, 1981) that may lead to poor protein spot resolution and/or streaking on 2-DE gels. Plant samples are particularly recalcitrant for 2-DE as they have varying levels of interfering polyphenolics, organic acids, lipids, pigments, terpenes, polysaccharides, secondary metabolites, salts, proteases and oxidative enzymes (Wang, et al., 2003).

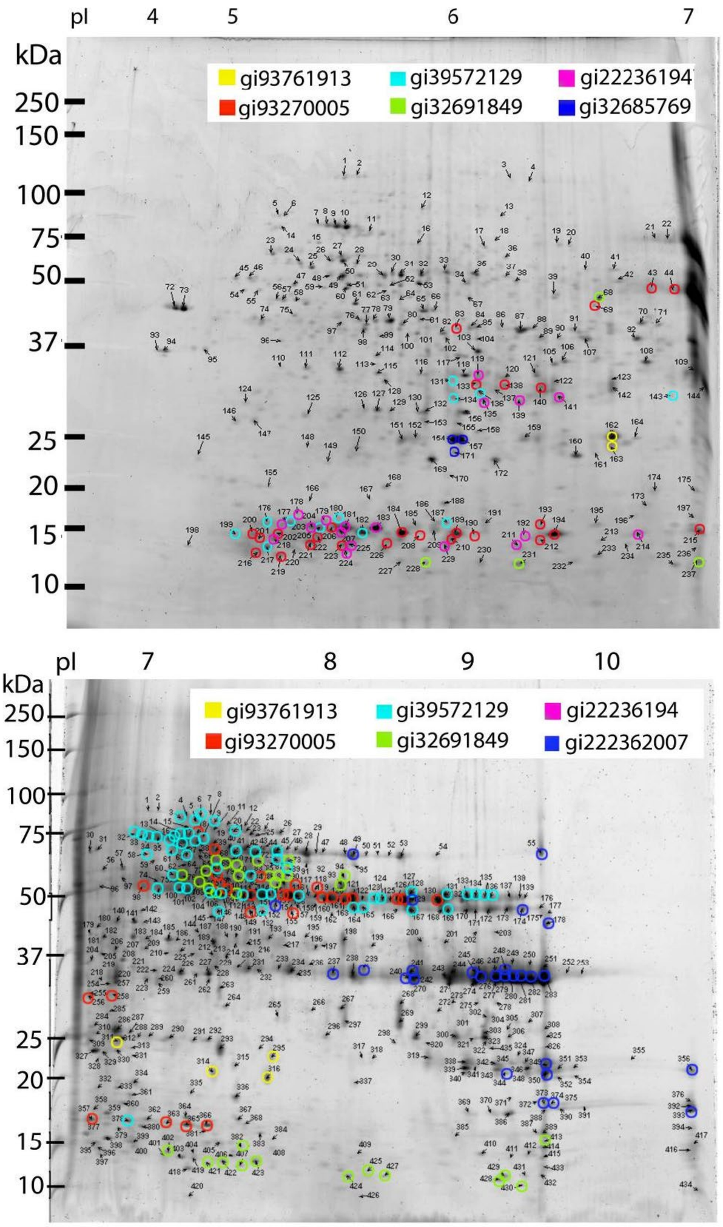


Figure 1.10. Example of a pair of 2-DE gels showing the pH ranges from 4-7 and 7-10 (image taken from Jerkovic, et al. (2010). The figure shows the different EST classes of the 7S globulin proteins identified with their corresponding EST GenBank gi number.

Many different methods have been applied to remove contaminants from protein extracts. They can be historically summarised as either additives to inactivate contaminants or solvents to inactivate and remove contaminants or a combination of both (Cremer and Van de Walle, 1985, Wang, et al., 2008). Proteolytic inhibitors such as phenylmethane sulfonyl fluoride (PMSF), polyvinylpyrrolidone (PVP) and polyvinylpolypyrrolidone (PVPP) were used to inactivate polyphenolic compounds, as well as sodium ascorbate to inactivate quinones. However, it was found that the inhibitors were not universally effective at inactivation and removal of contaminants (Cremer and Van de Walle, 1985). This led to the use of solvent based methods such as TCA, acetone, or phenol extraction in order to both inhibit contaminants and remove them from the protein sample (Cremer and Van de Walle, 1985).

Hari (1981) used acetone in a precipitation and wash step after protein extraction and was the first to report on successful 2-DE proteome analysis of green tobacco leaves. A number of other successful studies using either acetone, TCA precipitation or wash steps followed, culminating in the establishment of the TCA/acetone method (Damerval, et al., 1986). A few years earlier, Schuster and Davies (1983) had developed existing protocols into the phenol extraction method. Both protocols are now well established as they involve different physical processes. The salt and solvent in the TCA/acetone method forces water from the proteins and causes them to precipitate, while the phenol method forces hydrophilic regions of the protein inwards, while the hydrophobic regions position outwards interacting with the phenol, thereby dissolving proteins within the phenol phase.

Although the TCA/acetone and phenol protocols are commonly used, almost all studies that have made direct comparisons between the two methods observed different proteomic profiles when analysed by 2-DE. A greater protein spot number, resolution and streaking in 2-DE was observed for phenol-extracted proteins from tomato (Saravanan and Rose, 2004), grape berry

(Vincent, et al., 2006), sugarcane stalk tissue (Amalraj, et al., 2010), and maize leaf midrib Wang, et al. (2016). In contrast, the TCA/acetone method produced a higher overall quality 2-DE gel when examining tomato pollen and black locust phloem (Sheoran, et al., 2009, Zhang, et al., 2015).

An investigation of rice roots using 2-DE by Song, et al. (2012) reported that TCA/acetone gave more protein spots and less streaking than the phenol extraction method. Similarly, Zhang, et al. (2015) found TCA/acetone to be more suitable for 2-DE; however, they also observed that the phenol method yielded the best tandem mass spectrometry (MS/MS) results when identifying the protein spots. Generally, 2-DE comparison studies show that protein spots would be specific to either the TCA/acetone or phenol method, favouring either protein size and/or pI (Amalraj, et al., 2010, Saravanan and Rose, 2004, Song, et al., 2012, Vincent, et al., 2006, Zhen and Shi, 2011)

When considering the studies described above, both TCA/acetone and phenol methods were unable to adequately resolve a number of different plant tissues, and as such, led to the development of the TCA/acetone/phenol extraction method (Wang, et al., 2003, Wang, et al., 2008). This protocol combined elements of the TCA/acetone and phenol extraction methods. Wang, et al. (2006) demonstrated that the TCA/acetone/phenol protocol could deliver good results in 2-DE applications from tissues such as bamboo, grape, iris, olive, lemon, pine, redwood, sugar-cane, tobacco, posidonia grass, apple, banana, grape, kiwi, olive, orange, pear and tomato fruits. A comparison between the phenol and TCA/acetone/phenol by Maldonado, et al. (2008) using *Arabidopsis* leaf, showed that the TCA/acetone/phenol method resulted in slightly poorer 2-DE resolution than the phenol method; however, it yielded more protein spots, better protein molecular weight spread and pI range with similar results observed by

Domżalska, et al. (2016). However, other studies (Vincent, et al., 2006, Wang, et al., 2016) showed that for 2-DE, the phenol extraction method gave better results in maize leaf.

Further developments in protein extraction methods have been attempted to improve sample extraction in plant proteomics. Twenty per cent dimethyl sulfoxide (DMSO) added to TCA/acetone extraction for rice roots was found to give superior performance over the traditional TCA/acetone and phenol protocols (Song, et al., 2012). Possibly this is because DMSO disrupts non-covalent interactions between proteins and contaminants allowing more organic contaminants to be removed from the pellet. A study by Xiang, et al. (2010) on the use of MG/NP-40 extraction buffer gave better results than for TCA/acetone. The MG/NP-40 extraction buffer contains NP-40 (detergent), enzyme inhibitors, and chemicals to precipitate common contaminants such as phenolic compounds, and for samples containing photosynthetic cells can increase the number of resolvable spots on 2-DE gels by reducing the amount of (highly abundant) Rubisco extracted.

Finally, dissolving of the protein pellet is an important part of any extraction protocol with the resuspension buffer generally consisting of one or more combinations of urea and/or thiourea and/or a detergent such as SDS or NP-40 (Ashoub, et al., 2011). Physically dissolving the protein pellet is an important final step in any protein extraction protocol. For example, Gómez-Vidal, et al. (2008) found room-temperature shaking of the protein pellet in re-suspension buffer to be more effective than using either ultrasound or heat plus shaking. In summary, the studies show that extraction methods will vary depending on the sample being analysed. As such, the researcher may be guided by the previous studies or will be required to find the most suitable protein extraction method specific to the sample under study.

1.4. Proteomic methods

The word ‘proteome’ describes the entire complement of proteins expressed by a cell, tissue, or organism, and was introduced into the scientific lexicon by Mark Wilkins at a 1994 symposium (Dunn, 1995). The word ‘proteomics’ refers to the study of the proteome.

First developed in the late 1950s, SDS-PAGE has developed into the dominant technique for one-dimensional electrophoresis (1-DE), and remains a powerful and popular tool for looking at proteins. The technique of using polyacrylamide gel electrophoresis (PAGE) to separate charged particles was first described in the late 50s (Raymond and Weintraub, 1959) and 1-DE became a popular tool in discovering and classifying new sub-fractions of gliadin and glutenin proteins (Elton and Ewart, 1960, Graham, 1963, Woychik, et al., 1961). Summers, et al. (1965) added the denaturing effect of SDS (to PAGE) with these techniques further refined by Laemmli (1970) into the model most used today. SDS-PAGE is still applied today as a fast and reliable technique to analyse protein extractions. However, the separation is performed in one dimension, hence this technique is only useful for examining groups of proteins with significantly differing mass.

To this end, two-dimensional electrophoresis (2-DE) was developed (O'Farrell, 1975) to allow protein separation using two independent properties, pI and size (Brown and Flavell, 1981, Kasarda, et al., 1988). In the first dimension of separation, proteins are mobilised and separated along a pH gradient polyacrylamide strip via an electric field. Positively charged proteins will migrate towards the negative electrode and negatively charged proteins will migrate towards the positive electrode. Proteins become negatively charged when their pH environment is higher than the protein’s pI and alternatively, protein becomes positively charged when the pH environment is lower than the protein’s pI. As such, proteins migrate in the electric field until their mobility is interrupted and stop as a result of their overall charge becoming neutral. This happens when the pI of the protein matches the pH along the pH gradient strip.

This technique also allows for relative quantitation of proteins between samples via image analysis software. Furthermore, protein spots of interest can be excised from the second-dimension gel (size separation) for protein identification. Proteins are extracted from the polyacrylamide gel and then enzymatically digested into peptides using a proteolytic enzyme such as trypsin. The amino acid sequence of the peptides can be determined by Edman degradation (Rathmell, 2000), or by more modern techniques such as mass spectrometry.

1.4.1. Mass Spectrometry in plant research

The mass spectrometry technique of MALDI TOF-TOF is a powerful high throughput technique used to identify proteins (Vensel, et al., 2002). In recent years, continuing improvements in mass spectrometry (MS) and the wheat and barley proteomic databases including interconnectivity and data analysis has become a powerful tool in discovering and understanding biochemical systems in cereals. Analysis using mass spectrometry is done by one of two generally methods, unlabelled or labelled shotgun proteomics. The former involves peptides derived from enzymatic digestion by trypsin or another protease of the extracted proteins (Capriotti, et al., 2014). The proteins are fractionated in-line through liquid chromatography before being identified and quantified via mass spectrometry. Each protein sample extract is analysed separately and spectra and ion count are relatively compared in the analysis. The latter approach is similar except that a chemical label (called an isobaric tag) is covalently bonded to the peptides of each sample. Each differently labelled peptide sample is then pooled and analysed in a single pass through the mass spectrometer. The data is then used to identify and quantify both the peptide and the sample from which it came through the isobaric tag.

Two popular chemical labels for labelled mass spectrometry analysis of peptides are isobaric tags for relative and absolute quantification (iTRAQ; Figure 1.11), or Tandem Mass Tags

(TMTs; Figure 1.12). The TMT labels are suitable for up to ten samples, or for iTRAQ up to eight samples (Altenbach, et al., 2010).

Of course, more focussed proteomic studies require the target organ, tissue, cells, cellular components (or unique protein fractions) to be separated from other components. For example, all studies into wheat grain quality require separating gluten fractions prior to mass spectrometry analysis (Hurkman, et al., 2013, Pompa, et al., 2013, Skylas, et al., 2000). The four main protein fractions of gluten are HMW-GS, LMW-GS, gliadins and albumin/globulins. Gluten represents between 58-65 percent of all wheat grain proteins and consists of a 1:1 ratio of the glutenin and gliadin protein groups. In terms of food production (bread, pasta, etc.) it is the gluten proteins that are critical, with the ratios of HWM-GS to LMW-GS (gutenin proteins) and the overall ratio of glutenin to gliadins (as well as the proportions of different gliadin sub-fractions) that are largely responsible for the baking/cooking qualities of doe. In contrast the wheat albumin/globulin proteins have (in general) little effect on food properties, being mostly responsible for regulating growth and/or stress response of the germinating seed (MacRitchie, 2016, Žilić, et al., 2011). Indeed, studies investigating biotic and abiotic defence mechanisms require dissecting grain into tissue sections (Jerkovic, 2011). Thus isolating the tissues of interest (or protein fractions), will provide a more focussed picture of the proteome and how it differs or relates to neighbouring tissues.

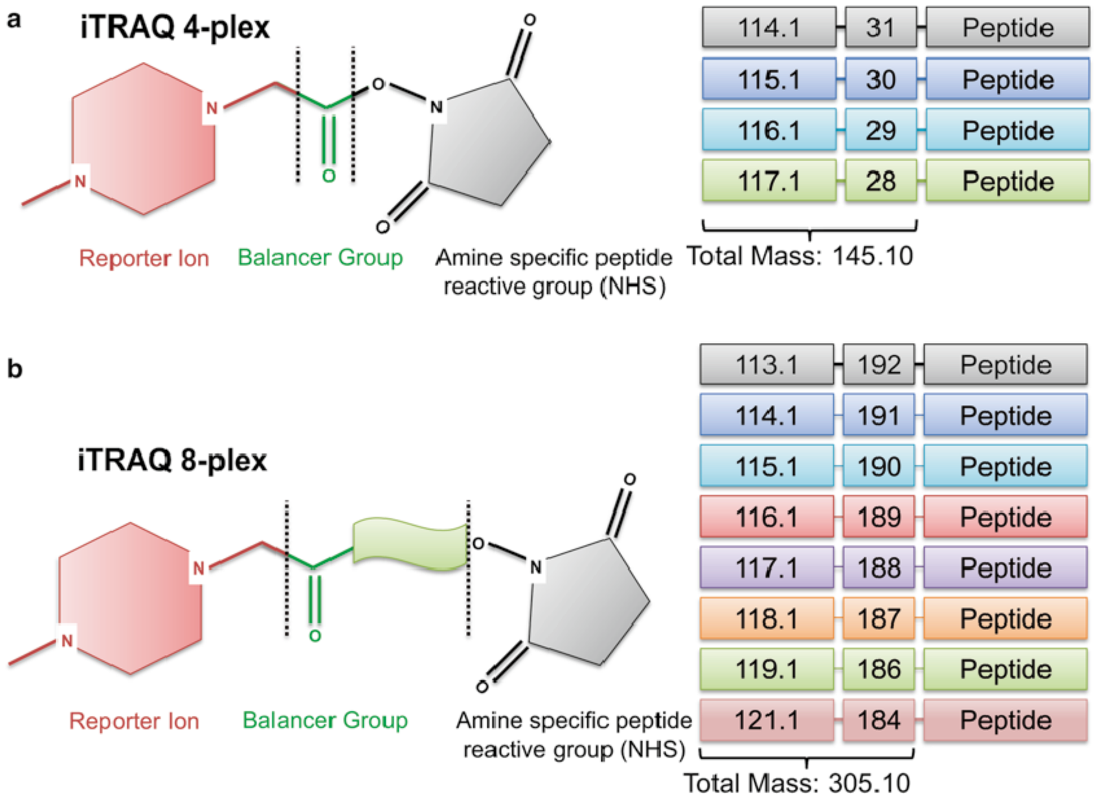


Figure 1.11. Isobaric tags for relative and absolute quantification. The two formats are a) 4-plex and b) 8-plex. Image taken from website on the 9/12/18 (https://www.creative-proteomics.com/blog/index.php/introduction-of-isobaric-tag-for-relative-and-absolute-quantitation-itraq/).

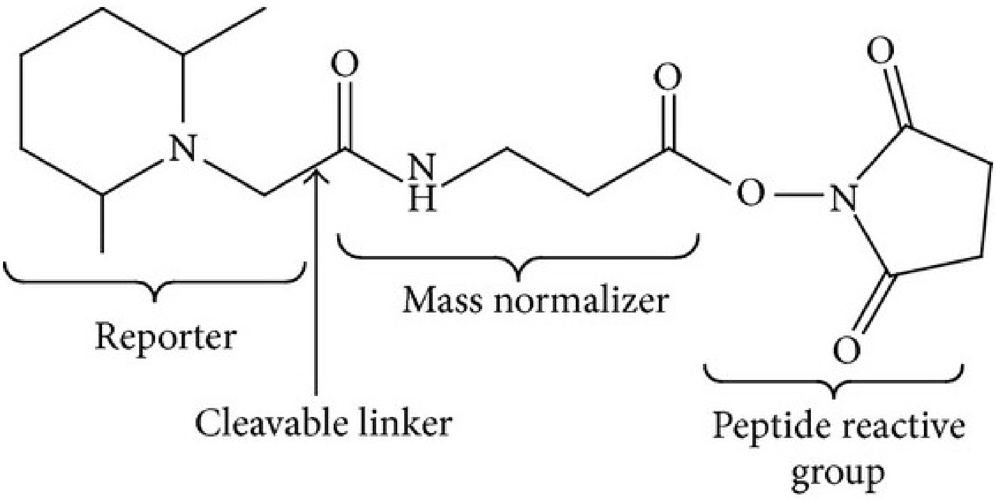


Figure 1.12. Diagrammatic representation of a tandem mass tag showing the reporter group, the cleavable linker, mass normalizer and peptide reactive group. Image taken from website on the 9/12/18 (https://www.thermofisher.com/order/catalog/product/90111)

Thus, decades of research into gluten proteins, as well as the constant evolution of laboratory technologies and protocols have led to a large dataset of results. Although some of these are contradictory, there is enough consistency to allow some general conclusions to be made on the quality of dough derived from wheat flour. Not only is dough quality seemingly dependent on a higher percentage of total protein within the wheat grain (Monaghan, et al., 2001), more specifically, dough quality appears to be dependent on a higher percentage of gluten within the grain (Guess, 1900, Halton, 1924). It has also been shown that within the gluten fraction, a lower proportion of gliadin to glutenin is an effective measure of quality (He, et al., 2013, Hoseney, et al., 1969), with Gupta and MacRitchie Gupta and MacRitchie (1994) suggesting gliadin as having more of a disrupting influence with the polymeric proteins. Yet, several other studies have demonstrated the importance of gliadin to dough quality (Branlard and Dardevet, 1994, Metakovsky, et al., 1997, Metakovsky, et al., 1997), although these conclusions may well be due to incorrect classification of some of the gliadin proteins. Similarly, for the glutenin fraction, a higher proportion of HMW-GS to LMW-GS is correlated with improved dough quality (MacRitchie, et al., 1991, Zhu and Khan, 2002); however, the proportion of HMW-GS to LMW-GS is not the only determinant of quality, as observed by Shewry, et al. (1992). They have shown that the cultivar Hereward with HMW glutenin subunits 3+12 and 7+9 resulted in a high-quality dough, whereas all other cultivars investigated with the same set of HMW-GS had a lower quality. Here either the gliadin portion, the LMW-GS, a combination of both, or possibly even an unknown protein or set of proteins resulted in the higher quality found in the Hereward cultivar.

In summary, the total protein fraction, absolute and relative amounts of various fractions and sub-fractions, as well as the quality of each protein isoform and its ability to form long polymers all have a quantifiable effect on gluten quality and are all under genetic control (He, et al., 2005). Thus, genetics in combination with environmental and biotic stress have influence on the quantity and quality of gluten (and other) proteins expressed, which in turn determines the overall dough and baking/cooking qualities of wheat flour (Fuertes-Mendizábal, et al., 2013, Payne, 1987, Zhu and Khan, 2002).

1.5. The biotic and environmental effects on wheat and barley proteomes

The following sections discuss the phylogenetic and proteomic changes of wheat and barley as a result of adaptations to biotic and environmental variability or stress.

1.5.1. Changes in the wheat proteome due to drought

According to the Food and Agriculture Organization of the United Nations (http://faostat3.fao.org), wheat and barley are the first and third most harvested grain in the world by weight respectively and have been subjected to a number of comprehensive studies into the effect of abiotic stress (Kosová, et al., 2011). Of these, drought represents an important sub-group, and is especially relevant due to a large percentage of wheat and barley crops being grown in areas prone to drought (Farooq, et al., 2014). Indeed, it has been reported that drought stress has the largest effect on the reproductive and grain-filling stages (Pradhan, et al., 2012), and thereby the yield.

Although the proteomics of wheat and barley were studied as early as 2001 and 2002 (Østergaard, et al., 2002, Skylas, et al., 2001), the first proteomic study into the effect of drought on wheat was by Hajheidari, et al. (2007). In this study, two drought susceptible genotypes (Arvand and Kelk Afghani) and one drought tolerant genotype (Khazar-1) were investigated to find proteins that could eventually be used to locate genetic markers for drought stress. Control and water deficit-treatments were irrigated at 75-and 150-mm evaporation, respectively. Mature seeds were harvested, and 121 differentially expressed proteins were detected using 2-DE, with 57 being identified by MALDI. The largest functional group of proteins that showed differential expression (27 out of 57 total), were involved in stress defence, and while most of these proteins were up-regulated in all cultivars, the expression levels were higher in the tolerant cultivar, especially for proteins involved in ROS scavenging. The authors also found several isoforms of α-amylase inhibitor were up-regulated in the tolerant genotype after drought treatment and down-regulated in the susceptible genotype after drought treatment. An increase in α-amylase inhibitor is known to protect grain starch from catabolism (Franco, et al., 2002). Protein synthesis and assembly (13 proteins), and metabolism (10 proteins) were the next largest functional groups detected. Interestingly, no proteins involved in other functions were detected, although there were 63 unidentified, yet differentially express proteins (Hajheidari, et al., 2007).

A similar study of durum wheat seedlings (*T. durum* cv. Ofanto) were subjected to drought stress at day eight and the first leaf harvested at day fourteen by Caruso, et al. (2009) who showed that thirty-six proteins were differentially expressed were evenly expressed across six functional groups; calvin cycle (9%), glycolysis and gluconeogenesis (18%), amino acid biosynthesis (12%), ROS scavenging (15%), defence mechanisms (6%), and post-transcriptional regulation (3%). Notably, the general expression pattern seen in this study was very similar to the proteomic studies examined below: proteins involved in photosynthetic mechanisms were differentially regulated, including up-regulation of photosystem II to counteract a loss of photosynthetic activity, while RuBisCO subunits and some associated proteins were generally down-regulated to reduce ROS production. In contrast, ROS scavenging proteins were all up-regulated, as were most proteins involved in amino acid and amine biosynthesis (S-adenosylmethionine synthetase, glutamine synthetase). Amino acid and amine biosynthesis are important in providing raw materials for protein production, antioxidants, as well as molecules involved in osmoregulation. The only major functional group not detected in this study were proteins involved in molecular repair or protection, such as heat shock proteins (Caruso, et al., 2009).

In the study by Rollins, et al. (2013), the response of the wheat root proteome to drought stress in a drought tolerant (Nesser) and drought sensitive cultivar (Opata) was examined by inducing a pseudo drought-response via the plant stress hormone abscisic acid (ABA). The wheat was grown from seed under normal conditions for 10 days, followed by exposure to ABA for 6 hours. The roots were harvested, proteins extracted and labelled with isobaric-tags (4-plex iTRAQ). The differentially expressed proteins were detected and estimated using LC-MS/MS. Notably, this is the first time that iTRAQ was applied to drought studies in wheat. Eight hundred and five differentially expressed proteins were detected after ABA treatment, and while 151 of these were common to both cultivars, 421 were cultivar-independent and ABA responsive. Of the latter, 131 proteins had a greater abundance in the tolerant cultivar, and represented the functional sets of proteins: defence, heat shock proteins, and signal transduction pathways (kinases, phosphatases, GTP-binding proteins, and 14-3-3 protein homologs). Notably, the proteins in both cultivars that showed differential expression (cultivar-specific and non-specific), had a much greater amount of significant expression in the tolerant cultivar (166) than for the susceptible (67). More specifically the tolerant cultivar was seen to have more types of heat shock proteins, proteins involved in secondary metabolism, and cell wall biogenesis, which suggested that Nesser had a greater number of pathways and a stronger response to drought stress.

1.5.2. Genetic diversity and changes in wheat proteome in response to drought

Peng, et al. (2009), compared the effects of drought on the root and leaf proteomes of a drought-tolerant hybrid wheat cultivar (Shanrong No 3) and its drought-susceptible wheat parent (Jinan

177). With the second parent of the hybrid being the wheatgrass *Thinopyrum ponticum,* the authors believed that the genetic diversity between the hybrid with its wheat parent would help them to understand the effect of drought on the root and leaf proteome of both cultivars. Shanrong No 3 and Jinan 177 were exposed to 24 hours of drought stress (18% polyethylene glycol) with differential expression of ninety-three root and sixty-five leaf proteins using 2-DE and MALDI-TOF-TOF. These results revealed that the majority of differentially expressed proteins were shared by the two cultivars, with the main groups being signal transduction, transport, detoxification, and carbon and nitrogen metabolism. However, the tolerant cultivar had a generally higher induction of differentially expressed proteins. Specifically, proteins involved in ROS scavenging (including antioxidant production) were higher in the tolerant cultivar, as were enzymes such as V-ATPases that helped maintain ion and water balance. In contrast, the susceptible cultivar displayed a more fragmented set of RuBisCO subunit isoforms, while the tolerant cultivar had an increased number of chlorophyll protector proteins. This indicated that the susceptible cultivar was less able to maintain normal photosynthetic mechanisms, while the tolerant was better able to protect these mechanisms. Lastly, a higher level of proteins involved in the gibberellin pathway and lower level for proteins involved in the ethylene pathway in the tolerant cultivar implied the promotion of growth. Meanwhile the susceptible cultivar exhibited the opposite effect in both pathways implying senescence was dominating.

The study by Budak, et al. (2013), was similar to that by Peng, et al. (2009), but looked at an even greater genetic range of plants, comparing the drought stressed proteomes of two wild emmer wheats (TR39477 and TTD22) with a modern durum wheat (Kiziltan cutivar). Seventy-five differentially expressed proteins were detected by 2-DE, and 66 protein spots were identified with nano LC-ESI-MS/MS. Of these, the most common functional groups were those involved in carbohydrate transport and metabolism, many of which were up-regulated during drought, including a number of proteins involved in photosynthesis. Proteins such as RuBisCO isoforms showed higher proteins levels in all cultivars, especially in the wild emmer wheats. The next largest group of proteins were those involved in energy production and conversion, with ion transporter proteins upregulated in wild emmer, which conversely downregulates parts of its photosynthetic machinery. Amino acid transport and metabolism was mostly affected in the more susceptible (modern Durum cultivar) variety with higher methionine synthase levels, which implies that growth was overriding stress tolerance. Lastly, it was found that polyamine oxidase was at a higher level in the tolerant cultivar despite being a source of H_2O_2, with the authors speculating that the wild tolerant variety may have an ancient alternative line of defence against abiotic stress.

1.5.3. Proteomic changes that favour more efficient mobilisation of nutrients in the wheat stem under drought

In an attempt to identify molecular mechanisms of wheat stem reserve mobilisation, Bazargani, et al. (2011) subjected two cultivars of wheat, with differing capacities to mobilise nutrients through the stem, to drought conditions. After anthesis, drought-susceptible (N14) and drought-tolerant wheat (N49) were exposed to drought by maintaining their soil at 50% field capacity. Then, the tillers were harvested at 10, 20, and 30 days after anthesis. One hundred-and thirty-six differentially expressed proteins were observed by 2-DE, eighty-two of which were identified by MALDI-TOF-TOF MS/MS. Overall, the tolerant cultivar (N49) showed a higher level of differentially expressed protein, with the peak for both cultivars being at twenty days after anthesis (DAA). The tolerant cultivar had 18, 74, and 23 variably expressed proteins at 10, 20, and 30 DAA (respectively), while for the same time series, the susceptible cultivar had 25, 38, and 21 proteins with differing expression. Although the tolerant cultivar had a greater number of differentially expressed proteins, most of these were down-regulated, with up-regulation in the tolerant cultivar only more prevalent in proteins involved in energy metabolism and ROS removal. RuBisCO large and small subunit, RuBisCO activase, and oxygen evolving proteins, were all lower in the tolerant cultivar, while SAM synthase, which is involved in ethylene production and senescence was up-regulated. Also, the tolerant cultivar showed an up-regulation of nine proteins involved in ROS removal compared to two for the susceptible variety. Both cultivars showed differential expression of signalling proteins, including 14-3-3, MFP-1, and MAF1. From these results, the authors suggest that the tolerant cultivar senesces faster and is more efficient at protecting cells within the stem, while the stem nutrients for the susceptible cultivar are more effectively transported from the stem to the developing grain.

1.5.4. Wheat proteome changes over time during drought

In a study by Ge, et al. (2012), the effect of drought on the wheat grain proteome over time was examined in developing wheat kernels. Ningchun 4 (drought-tolerant) and Chinese Spring (drought-susceptible) varieties were subjected to a watering regime of one-third the level of the control, starting 12 days prior to heading. Samples were collected at 10, 14, 18, and 26 days after flowering (DAF), and 152 significant proteins were detected by 2-DE, with 96 identified with MALDI-TOF. From these, the three largest functional groups observed were carbohydrate metabolism (39%), stress/defence (18%), and photosynthesis (13%). The common stress defence proteins were detected in both cultivars (SOD, CAT, APX), while the growth regulator TCTP was generally up-regulated over time only in the tolerant cultivar. In the early stages of grain development HSP70 increased, while LEA increased at later stages. Enzymes involved in glycolysis and the tricarboxylic acid cycle (TCA) were also up-regulated under drought, such as GAPDH, cytosolic 3-phosphoglycerate kinase, and malate dehydrogenase. β-amylase was up-regulated in both cultivars, while sucrose synthases were up-regulated more in the tolerant cultivar, the latter potentially giving the tolerant cultivar more energy for stress responses. Moreover, a higher expression of Triosephosphate isomerase (TPI) in the tolerant cultivar indicates it has a greater tolerance to water-stress. Homeostasis of photosynthesis seems more effective in the tolerant cultivar, with general up-regulation of the RuBisCO large subunit, and an early up-regulation of the oxygen evolving complex (OEC), which gradually decreases over time.

Similar to the paper by Ge, et al. (2012), a study by Jiang, et al. (2012) initiated a time–course study into the effects of drought on the developing wheat grain. The proteomic mechanisms of drought tolerance was examined in a drought tolerant wheat cultivar (Kauz) and a drought-susceptible wheat cultivar (Janz) from the middle spike collected at 10, 15, 20, and 25 days post-anthesis. The protein extracts analysed by 2-DE revealed 153 differentially expressed proteins. The drought-stressed, tolerant cultivar exhibited higher levels of detoxification and defence proteins (ROS scavengers, peptidase inhibitors, and salt-stress-responsive proteins), proteins involved in carbohydrate metabolism (AGPase, sucrose synthase, and ALR), and those involved in signal transduction proteins (WD40 and G-beta like protein). Moreover, compared to Ge et al. (2012), a similar but slightly different pattern was seen in photosynthetic proteins, including up-regulation of OEE1 in both cultivars at the first two sampling stages, and a lower expression of OEE1 in the susceptible cultivar at the last two stages. For the remaining functional groups, there was a general up-regulation under drought stress, but little difference between cultivars for all time periods.

1.6. Abiotic and biotic stress in barley

Several studies have reported that climate change and extreme weather events are likely to decrease the geographic growing range of barley, as well as negatively affecting its yield and quality (Anwar, et al., 2015, Dreccer, et al., 2018). It would be of great interest to determine how elements of the grain proteome respond to the abiotic and biotic stresses that could result from predicted changes, and whether biomarkers that would mark these changes. While the overall omic-complexity of barley (genome, transcriptome, and proteome) may not translate from biomarker discovery to the development of cultivars tolerant to climate-change stress, other benefits could be garnered. Aside from the potential of biomarker proteins to indicate climatic change, they could be developed as markers to identify cultivars and their provenance (see Chapter 3, Introduction).

Like abiotic stress, biotic stress also has a major impact on barley yield or quality, which may introduce toxins or other compounds that affect animal or human health. To better understand barley grain-microbe interactions, Sultan, et al. (2016) examined the proteomes of the barley grain surface proteins and the microbes found on the seed surface. Results indicated the grain surface shows a relatively specific response to microbial colonization, with glucanases, chitinases, and alginate lygases, all being expressed to degrade bacterial, fungal and algal cell walls. Lower levels of antifungal and apoptotic proteins, as well as low molecular weight microbial proteolytic-enzyme-inhibitors, which reduce mycelium growth and spore germination, were also found. Also found were a number of proteins involved in a more general stress response such as thioredoxin reductase, and a number of stress-related proteins.

Microbial proteins were detected at a much lower lever, but nonetheless, a number of proteins from the bacterial membrane or proteins with transmembrane domains were detected. Proteins involved in the control of small molecule diffusion (such as antibiotics) were also detected, as well as biosynthesis of secondary metabolites and toxins. Defence against the plant antimicrobial response was also seen in a number of microbial ROS scavenging enzymes and chaperones. In contrast, the fungal surface proteome showed a greater potential for plant tissue damage, with proteins involved in plant cell-wall and peptide degradation (including xylanases and peptidases), as were proteins necessary to assist fungal growth, primary metabolism, nutrient acquisition, virulence factors, energy metabolism, RNA/DNA synthesis, and proteins involved in steroid and secondary metabolite synthesis.

A similar study was performed by (Trümper, et al., 2016) involved infecting barley grain with the fungus *Fusarium graminearum* at anthesis and examined the grain proteome at different grain ripening stages. This fungus is economically important as it can reduce grain quality, and because fungal mycotoxins can accumulate in products such as beer, it can affect production and potentially human health. A number of changes in the grain proteome occurred throughout grain ripening. At the early to middle stages of infection, protease inhibitors and a putative chitinase were up-regulated, as were proteins involved in oxidative burst (flood of ROS leading to programmed cell death). This results in the neutralization of foreign hydrolytic enzymes and a physical barrier to fungal attack causing fungal cell death. At mid-infection stages, up-regulation of protein degradation and two thaumatin-like proteins involved in hyphal and spore lysis indicated that the plant cells were shifting towards survival and invader attack. Finally, at the latter stages of infection, all proteins just mentioned were down-regulated, including a notable decrease in those involved in oxidative burst and stress response. Thus, over a number of weeks of infection, the barley grain proteome moved from defence, to attack, and finally a down-regulated response either because the threat had been managed or cell resources were exhausted.

1.7. RNA and protein expression in plant cells

The genome stores information on how to react and direct resources under various environmental conditions and exposure to pathogens. Response to stress is measurable through changes in RNA and protein expression within the various plant cells. The role of RNA is primarily an intermediary that translates information from DNA into different types of proteins and expression levels. In contrast to most RNA, proteins not only act directly in keeping plants in a state of homeostasis, they are also responsible for actions within cells. For example, tissues and organs under biotic and/or abiotic stress mediate a direct return to homeostasis or at least prevent death until the crisis has passed. The variable speed and specificity of different enzymes and their isoforms, as well as their complex network of interactions, allows subtle or extreme variations in reactivity. They not only control the constituent processes of all parts of the organism but are the most immediate and effective agents of restoring cellular environments to homeostasis, or at least implement survival strategies that give plants a chance of survival under abiotic and/or biotic stress. Examining RNA and protein expression levels in various anatomical structures of the plant will give the greatest insight into how a plant survives, grows, reproduces and adapts within its environment.

1.8. Aims and hypotheses

1.8.1. Aims

The aims of this thesis are to 1) determine the optimal protein extraction method for wheat and barley grain that will both maximise the total protein yield and expand the number of differing protein types captured, 2) apply TMT-labelled shotgun proteomics to identify biomarker proteins in wheat and barley grain that can be used to identify the cultivar, farm origin, and potentially relate the results to grain quality, and 3) investigate whether the next generation DNA sequencing technique known as RNASeq can be used to either complement TMT labelled shotgun proteomics, and a potential alternative method of protein biomarker discovery.

1.8.2. Hypotheses

It has been shown by both ancient cultures and modern science that the composition of mature wheat and barley grain is remarkably stable under the correct storage conditions. Prior to maturity, grain composition is influenced during its growth and development by the interplay between the plant’s genes and the biotic and abiotic stresses on the plant. I hypothesise that, 1) at least a small subset of proteins within a grain proteome will have statistically significant differences in protein expression levels between different cultivars and/or farm locations, and between grains of differing quality and grades, 2) the protein extraction protocol can be optimised to improve both the yield and diversity of extracted proteins, thus improving detectability, and enabling the extracted protein samples to be as biochemically representative of the full grain proteome as possible, and 3) transcriptomics may be useful as an alternative or complementary technique to proteomics for the discovery of protein biomarkers to determine the provenance of wheat or barley grain.

Chapter 2. Optimisation and comparison of two common protein extraction methods for maximum yield and diversity of wheat grain proteins for high-throughput proteomic analysis

2.1. Introduction

Protein extraction methods for proteomic analysis, such as TCA, acetone and phenol, each have their limitations in extracting total proteins from plant material. This is mainly due to the variety of protein solubility and the tough cellulose cell walls found in plants. Applying any single protein extraction method will result in a loss in protein diversity when attempting to obtain a complete proteomic profile of a given sample. As such, protein extraction methods are chosen based on the target protein or proteins of interest within the proteome.

One of the first protein extraction methods developed enabled the targeted extraction of albumin, globulin and gluten protein sub-fractions from wheat grain (Osborne and Voorhees, 1893). This method has been further refined over many decades to extract gluten fractions described as low-molecular-weight glutenin subunits (LMW-GS), high-molecular-weight glutenin subunits (HMW-GS), α-like gliadin, γ-gliadin, and ω-gliadin (Barak, et al., 2015, Wieser, et al., 2006). This targeted approach has been highly effective in investigating gluten proteins in wheat grain, especially in relation to grain quality.

In the last two-to-three decades, there has been growing use of mass spectrometry and database centred high-throughput proteomic techniques in grain protein research. These techniques range from two-dimensional gel electrophoresis (2-DE) coupled with MALDI (Jiang, et al.,

2012) to current techniques such as high-throughput and increasingly sensitive, labelled and unlabelled shotgun proteomics (Zhang, et al., 2013). These high-throughput proteomic techniques ideally require a protein extraction method that both remove all impurities but also capture a high yield and diversity of proteins from the sample, so that the resulting dataset of unique proteins accurately reflects the proteome. Yet, as reported by many studies, the heterogeneity of protein chemistries within a proteome makes it difficult for any general extraction technique to completely capture all proteins within a proteome (Amalraj, et al., 2010, Saravanan and Rose, 2004, Song, et al., 2012, Vincent, et al., 2006). As such, for any given sample, it is important to invest time to develop an optimal extraction method. Ideally, a protein extract with the most representative protein yield and diversity can be delivered to the mass spectrometer, and in turn improve the quality of the subsequent downstream computational analysis.

Here we investigate both the optimal starting amount of sample, and whether the trichloroacetic-acid/acetone (TCA/acetone) or the TCA/acetone/phenol ('combined-phenol') protein extraction methods result in higher yields and greater diversity of proteins. These two methods perform differently in the way the denaturing agents (TCA in the TCA/acetone method, or phenol in the 'combined-phenol' method) work to purify the proteins. Phenol causes a much more complete denaturation, by forcing hydrophobic sections of the protein's amino acid chain outwards, occupying the outer surface of the protein, and hydrophilic sections inwards, occupying the centre region of the protein (Pusztai, A. 1965). On the other hand, TCA causes a gentler unfolding of proteins into a state that allows them to stick together by hydrophobic aggregation and then precipitate (Sivaraman et. al. 1997; Xu, Z. 2003). These chemical differences result in the extraction of unique protein subsets specific to each method.

As such, we classified the identified proteins into functional groups from each protein extraction method to determine the types of proteins you would expect from each method.

2.2. Methods

2.2.1. Environmental conditions, grain size and morphology

The three cultivars of wheat used in the following experiments (Gregory, Livingston, and Spitfire), were grown to maturity and grain harvested near the town of Wallendbeen in NSW, approximately 120 km north-west of Canberra. Care was taken to achieve uniform crop management and soil type, while the growing season included no extreme events. For each of the three cultivars there were three replicate plots. Unfortunately, information on the exact location of the farm, and the size and type of plots have been unattainable due to staff turnover at Grain Growers (commercial collaborator). After harvesting, the grains were dried and stored as per standard agricultural practice. The relatively uniform grain morphology and size for each cultivar reflected negligible biotic or abiotic stresses during grain development and maturity. From each plot a 1 kg grain sample was obtained resulting in a total of nine 1 kg wheat grain samples (3 cultivars x 3 replicates for each cultivar).

2.2.2. Weighing and grinding samples

The grain samples obtained from each farm were processed as follows. All samples for analysis were performed in biological triplicates. The 100-grain weight was calculated by weighing 500 grains and dividing by five. For each sample, 200-grain weight of grain was rough ground in a coffee grinder for 20-30 s. The grinder was cleaned with 70% ethanol after each use. From the roughly ground grain, 300 mg was added to a 2-mL plastic tube, together with ~30 mg of washed sand. This powder/sand mix was then thoroughly emptied into a mortar and pestle and ground to a fine powder. After each grind, the powder was scraped back into the 2-mL plastic tube. The finely ground powder was then used for protein extraction.

2.2.3. TCA/Acetone/Phenol protein extraction

Total proteins were extracted using a modification of the method used to extract bran from Jerkovic, et al. (2010). For each grain sample, the finely ground powder (10, 20, 50 and 100 mg) were each washed twice with 1.5 mL of cold acetone in 2-mL plastic tubes, vortexed, then centrifuged at 15,000 *g* for 3 min at 4°C. After the final centrifugation, the acetone was discarded. To each pellet was added 1 mL of cold 10% trichloroacetic acid (TCA) in acetone and vortexed for 30 s. The samples were then centrifuged at 15,000 *g* for 3 min at 4°C. This step was repeated, each time discarding the supernatant. Finally, 1 mL of cold 80% acetone was added, vortexed for 30 s and centrifuged at 15,000 *g* for 3 min at 4°C. The supernatant was discarded, and the pellets were left to dry at room temperature until almost all the acetone had evaporated. The pellets were then suspended in 0.8 mL phenol (tris-buffered at pH 8.0) and vortexed until fully re-suspended. If necessary, the pellets were mixed with a pipette tip to assist re-suspension. To the re-suspended pellets was added 0.8 mL of dense SDS buffer (30% sucrose, 2% SDS, 0.1 tris-HCl, pH 8.0, 5% 2-mercaptoethanol) and vortexed for 30 s. The mixture was then centrifuged for 3 min at 10,000 *g*. The upper phenol layer (~650 µL) was removed and placed into a 10-mL Falcon tube. Care was taken to not disturb the phenol/dense SDS buffer interface. The protein was precipitated from the phenol layer by adding 3.25 mL (~5 volumes) of cold methanol, 0.1 M ammonium acetate, and placed at -20°C overnight. The precipitated proteins were then centrifuged at 10,000 *g* for 5 min at 4°C. The protein pellets were washed with 1 mL of cold methanol, 0.1 M ammonium acetate by vortexing for 30 s and centrifugation at 10,000 *g* for 5 min at 4°C. The supernatant was discarded, and the wash step was repeated. Finally, 1 mL of cold 80% acetone was added, vortexed for 30 s and centrifuged at 10,000 *g* for 30 s. The supernatant was discarded, and the wash step was repeated. The pellets were air dried until the acetone had almost completely evaporated. To the semi-dried pellets was added 600 µL of 6 M urea. The tubes were then placed in an incubator shaker at 40°C to completely dissolve the pellet (~1 - 4 h).

2.2.4. TCA/acetone protein extraction

Protein was extracted using the TCA/acetone procedure commonly used for plant proteomics extraction (Wu, et al., 2016) with slight modifications. For each grain sample, the finely ground powder (10, 20, 50 and 100 mg) were each placed into separate 2-mL plastic tubes. To each tube was added 1.5 mL of cold TCA/acetone-DTT-PMSF solution (10% trichloroacetic acid, 10mM dithiothreitol, 1 mM phenylmethylsulfonyl fluoride, in acetone). The tubes were then vortexed for 30 s and then incubated at -20°C for 1 h to precipitate proteins. After incubation, the samples were centrifuged at 20,000 *g* for 30 min at 4°C. The pellets (comprising proteins and cell debris) were then washed in 80% cold acetone, vortexed for 30 s and centrifuged at 10,000 *g* for 10 min at 4°C. This was repeated twice before the pellets were dried in a speed-vac for 30 min at room temperature. One millilitre of 6 M Urea was then added to each pellet and the tubes were placed on an incubator shaker for 1 - 4 h at 40°C. If necessary, dissolution of the pellet was sped-up by breaking the pellet with a pipette tip, followed by aspiration. Once the pellets were re-suspended, they were then centrifuged at 15,000 *g* for 10 min to remove the debris from the solution. Following centrifugation, a large viscous ‘glue-like’ layer had formed above the pellet, leaving a thin supernatant layer. Approximately 100-200 µL of this layer was removed and placed into a 2-mL plastic tube. The pellets were re-suspended in 200 µL of 6 M urea and further centrifuged to yield approximately 100-200 µL of supernatant. The total amount of protein harvested from each sample was quantified using the Bradford assay.

2.2.5. Protein estimation using Bio-Rad protein assay reagent

Bio-Rad protein assay reagent was used to estimate the protein extracted as detailed below. The samples were measured in a PHERAstar FS plate reader (BGM Labtech), with software version 4.00 R4 and firmware version 1.13. Standard Bradford settings were used: 10 s of double orbital shaking motion (500 rpm) before reading, absorbance measured at 595 nm for 0.1 s, including 20 flashes per well. The plates used were clear 96-well Greiner, flat bottom plates (Sigma-Aldrich: M2936). The blank was 3 M urea. BSA Standards were 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, and 2.0 mg. mL^{-1} in 3 M urea, and samples were diluted so that their final urea concentration was 3 M. For each blank, standard, and sample, three 10 µL technical replicates were aliquoted into separate wells of a 96-well plate, followed by the addition of 200 µL of diluted Bio-Rad protein assay reagent into each well. Sample concentrations were calculated from a curvilinear regression graph with concentration (mg. mL^{-1}) vs absorbance at 595 nm. The final amount of protein is protein concentration multiplied by the sample volume.

2.2.6. Percent protein calculation

The protein concentration was calculated by determining the total amount of protein extracted, divided by the total amount of starting material (finely ground wheat grain).

2.2.7. SDS-PAGE

The extracted proteins were analysed by SDS-PAGE using Bolt® 4-12% Bis-Tris plus pre-cast gels (Catalogue number: NW04120BOX), and run in a Bolt ® mini gel tank (Catalogue number: A25977) using 1× MES buffer. All equipment and solutions (excluding MES buffer) were purchased from Thermo Fisher Scientific. Approximately 10-20 µL of 4× Bolt® (Cat.

Number: B0007) LDS sample buffer was added to fresh tubes containing sample aliquots (e.g. 2.5 µL sample buffer in 7.5 µL sample). Once the sample and sample-loading solutions were added together, they were vortexed and pulse centrifuged. The protein ladder used was "Broad-range" from Bio-Rad (Catalogue Number 161-0317) - prepared as per recommended protocol, with 10-20 µL of sample-loading solutions added to any remaining wells. Two technical replicates were loaded for each cultivar. Once loaded, electrophoresis was performed at 170 V until the leading dye front had just run off the bottom of the gel (~40 min). The gels were then placed into Coomassie stain on a gentle shaker overnight, followed by de-staining for at least 4 h in reverse osmosis (RO) water.

2.2.8. Trypsin in-gel digestion

In-gel trypsin digestion of proteins was performed using a modification of the procedure of Shevchenko, et al. (2007) as follows. Each sample lane of the de-stained gel was dissected into eight equal pieces. Each piece was then further sliced into 8–12 smaller pieces and placed into a well on a 96-well plate. The Coomassie stain was removed by washing twice with 200 µL of 50% acetonitrile (ACN), 100 mM NH_4HCO_3 for 20 min each time. The washed gel pieces were dehydrated by adding 200 µL of 100% acetonitrile, incubating for 5 min at room temperature, followed by air drying for 10 min. The proteins within the gel pieces were then reduced with 50 µL of 10 mM DTT in 50 mM ammonium bicarbonate for 1 h at 37°C. The proteins were then alkylated at room temperature in the dark for 45 min using 50 µL of 50 mM iodoacetamide in 50 mM ammonium bicarbonate. The gel pieces were then washed twice with 200 µL of 100 mM ammonium bicarbonate and 200 µL of 100% acetonitrile incubating for 5 min each time. The gel pieces were then air-dried. Trypsin digestion followed with the addition of 30 µL of trypsin (6.7 ng. μL^{-1} trypsin in 50 mM ammonium bicarbonate) to each well, keeping the plate at 4°C for 30 min before incubating overnight at 37°C. The trypsin digestion was stopped and peptides extracted by adding 50 µL of 2% formic acid in 50% acetonitrile incubating for 30 min at room temperature. The supernatant was removed and placed into 0.5-mL plastic tubes. The supernatant was then evaporated using a speed vac.

2.2.9. Peptide preparation for Mass Spectrometry

To each 0.5-mL plastic tube containing peptides was added 21 µL of 2% formic acid and then centrifuged at maximum speed (10,000 *g* or more) for 10 min to remove leftover debris from solution. From the supernatant, 10 µL from each sample was pipetted into corresponding wells of a 96-well plate that is compatible with the mass spectrometer auto sampler. Once all samples were aliquoted, the plate was sealed with a plastic seal and each of the sample wells was punctured with a sharp blade to allow the mass spectrometer to access the solution in each well.

2.2.10. Mass Spectrometry

The peptides from the trypsin digested proteins were analysed using nanoflow liquid chromatography tandem mass spectrometry (nano LC-MS/MS). The reverse phase columns were packed in-house to approximately 8 cm (100 µm i.d.) with Magic C18AQ resin (200 Å 5 µm, Michrom Bioresources, CA, USA) in a fused silica capillary with an integrated electrospray tip. Ten microliters of each sample was then injected into a C18 column using a "Surveyor" autosampler (Thermo Fisher Scientific). After injection, the C18 column was washed with buffer A (5% *v*/*v* ACN, 0.1 *v*/*v* formic acid) for 10 min at a flow rate of 1 µL. min^{-1}. The peptides were then eluted from the column with 0-50% buffer B (95% *v*/*v* ACN, 0.1 *v*/*v* formic acid) for 58 min at 500 nL. min^{-1} followed by a second elution of 50-95% buffer B over 5 min at 500 nL. min^{-1}. The eluate from the column was then directed towards the nanospray ionisation source of the LTQ-XL ion-trap mass spectrometer (Thermo Fisher Scientific). The spectra were scanned over the range 400-1500 amu. Automated peak recognition and the dynamic exclusion window was set to 90 s, and MS/MS of the top six most intense precursor ions at 35% normalisation collision energy were performed using Xcalibur software (version 2.06; Thermo Fisher Scientific).

2.2.11. Mass spectrometry analysis

The raw files output by the mass spectrometer were converted to the mzXML format using the open source proteoWizard windows application. Peptides and their associated proteins for the eight fractions of each sample were then determined through the global proteome machine (GPM) software with version 2.2.1 of the X!Tandem algorithm by searching against a trimmed Uniprot wheat proteome for *Triticum aestivum*, and including common human and trypsin peptide contaminants. Additional searching was performed against a reversed sequence database to evaluate the false discovery rate (FDR). The parameters of the search also included MS and tandem MS tolerances of ±2 Da and ± 0.2 Da, i.e. tolerance of up to three missed tryptic cleavages and K/R-P cleavages. Fixed modifications were set for carbamidomethylation of cysteine and variable modifications were set for oxidation of methionine. For each sample, non-redundant output files were generated that included identified proteins with log (e) values less than – 1: Including a file for each slice (8 total), and a merged data file for all 8 slices.

2.2.12. Gene Ontology Data

Gene ontology information for each protein, and gene ontology functional summaries (known as "Slims") were obtained respectively by the web application "GORetriever" and "GOSlimViewer" (http://agbase.msstate.edu/) maintained by the Mississippi State University (McCarthy, et al., 2006)

2.2.13. Collection of GO Slims

We first selected a subset of proteins that were identified in all three cultivars for all replicates and extraction methods. This allowed for a manageable list of Slims to which we could add expression data through a set of custom R-scripts containing three major stages. The first was the construction of a matrix consisting of rows made up of proteins and their associated data, while each column beyond the protein data is titled with a Slims identifier. A ‘zero’ or ‘one’ is placed in each cell under the identifier, depending on whether the protein to the left has any GO identifiers associated with it under the Slims identifier title of the column. The next stage was to total the number of ‘ones’ found in each column to get the number of proteins contained within each SLIM identifier, or to replace the ‘ones’ with expression data for each protein (row), sum those data, and then compare the sum to the total of all Slims expressed, and finally the expression data per SLIM with additional experimental information. The last stage consists of saving summary .csv files and creating various graphs which are then saved to disk.

2.3. Results

2.3.1. Protein extraction optimisation

In order to detect potential biomarkers in wheat grain that are expressed at very low-levels, we implemented a protein extraction protocol that maximised the protein yield and in turn, revealed a diverse set of proteins representative of the wheat grain proteome. For the TCA/acetone method, we observed no difference in protein yield from the starting material (10, 20, 50 or 100 mg), whereas the ‘combined-phenol’ method displayed a small increase in yield as the starting amount increased (Table 2.1). In contrast, when comparing the two methods for each of the four starting amounts, the ‘combined-phenol’ method had double the yield. In addition, there was a difference in protein extraction yield between cultivars using the ‘combined-phenol’ method (Figure 2.1).

Table 2.1. Comparison of percent by weight yields of protein between different starting amounts for both the TCA/Acetone and the combined phenol methods. Starting amounts were either 10 or 100 mg of finely crushed wheat grains, with three replicates for each starting weight.

Method	10 mg (% protein recovered)		100 mg (% protein recovered)		T-Test, p-value
	Mean	StDev	Mean	StDev	
TCA/acetone	1.432	0.150	1.386	0.028	0.628
'combined-phenol'	2.923	0.196	3.320	0.158	0.052

Table 2.2. Comparison of percent by weight yields of protein extracted between extraction methods. Each comparison looked at a starting amount of either 10, 20, 50 or 100 mg of finely crushed wheat grains.

Starting Weight (mg)	TCA/acetone (% protein recovered)		'combined-phenol' (% protein recovered)		T-Test, p-value
	Mean	St Dev	Mean	St Dev	
10	1.432	0.150	2.923	0.196	0.0004677
20	1.374	0.041	2.995	0.172	0.0002868
50	1.332	0.101	3.154	0.293	0.0005215
100	1.386	0.028	3.320	0.158	0.0000309

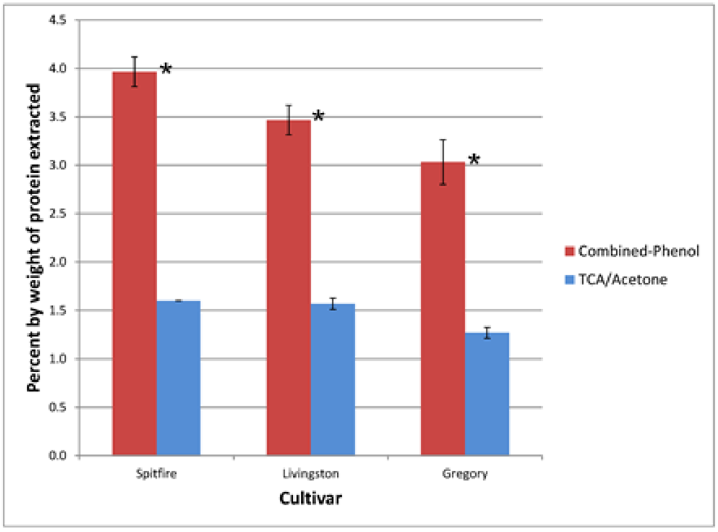


Figure 2.1. Comparison of the protein yield from three wheat cultivars using the 'combined-phenol' and TCA/acetone extraction methods. The yield was calculated in terms of the weight of total protein extracted from the initial starting weight expressed as a percentage. Statistically significant difference between the percentage of protein recovered using either the 'combined-phenol' or the TCA/acetone protein extraction methods.

The proteins extracted using the TCA/acetone and ‘combined-phenol’ methods produced contrasting banding patterns on the reduced SDS-PAGE gel (Figure 2.2). The ‘combined-phenol’ extraction generated more uniform banding in the upper half of the gel compared with TCA/acetone extractions, while TCA/acetone had more proteins below 20. There were also more intense bands between 36-55 kDa for the TCA/acetone method at the expected size and number of gluten sub-fractions, such as ω-gliadins, α–like-gliadins and LMW-GS. This suggests that the TCA/acetone method is more selective in extracting these sub-fractions of gluten proteins, although given the difference in total yield between extraction methods it is probably not as efficient.

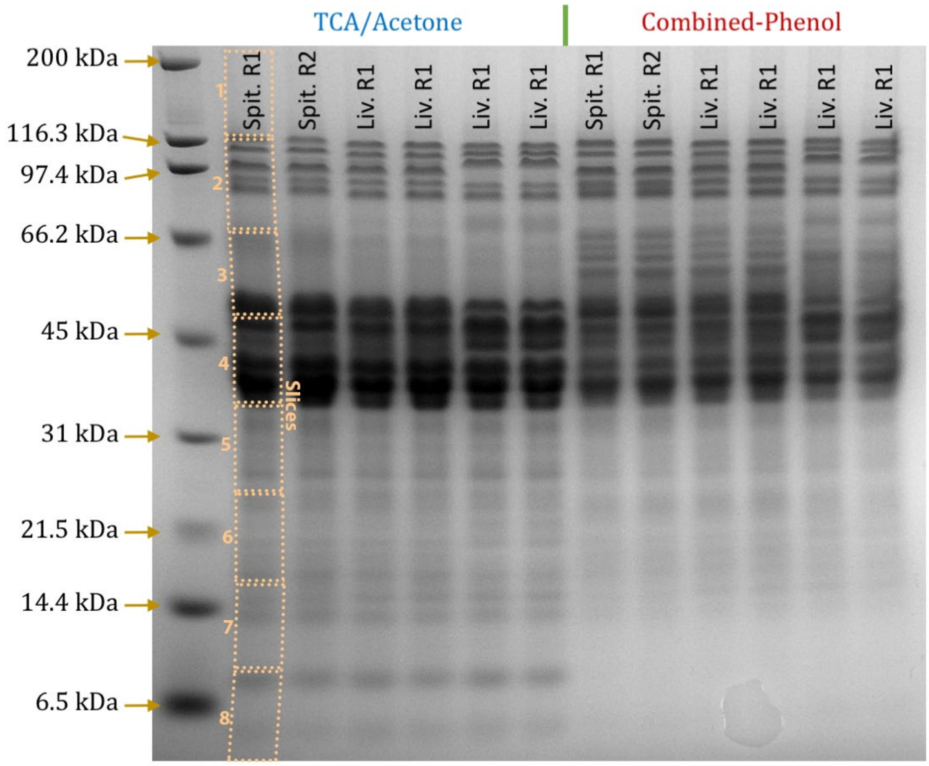


Figure 2.2. SDS-PAGE gel of wheat grain proteins extracted using either the TCA/acetone or ‘combined-phenol’ method.

The cultivar abbreviations are: Spit. = Spitfire; Liv. = Livingston; Greg. = Gregory; and “R” = Replicate. Standards used were the “Broadrange” by Bio-Rad. The illustrated slices indicate how each lane was sectioned to size fractionate the proteins prior to trypsin digestion and mass spectrometry of the resulting peptides.

The total number of proteins identified in all biological replicates in each variety using each extraction method is shown in Figure 2.3. The combined extraction method thus appears to yield a consistently higher diversity of proteins extracted than the TCA acetone method, which would appear to favour biomarker discovery.

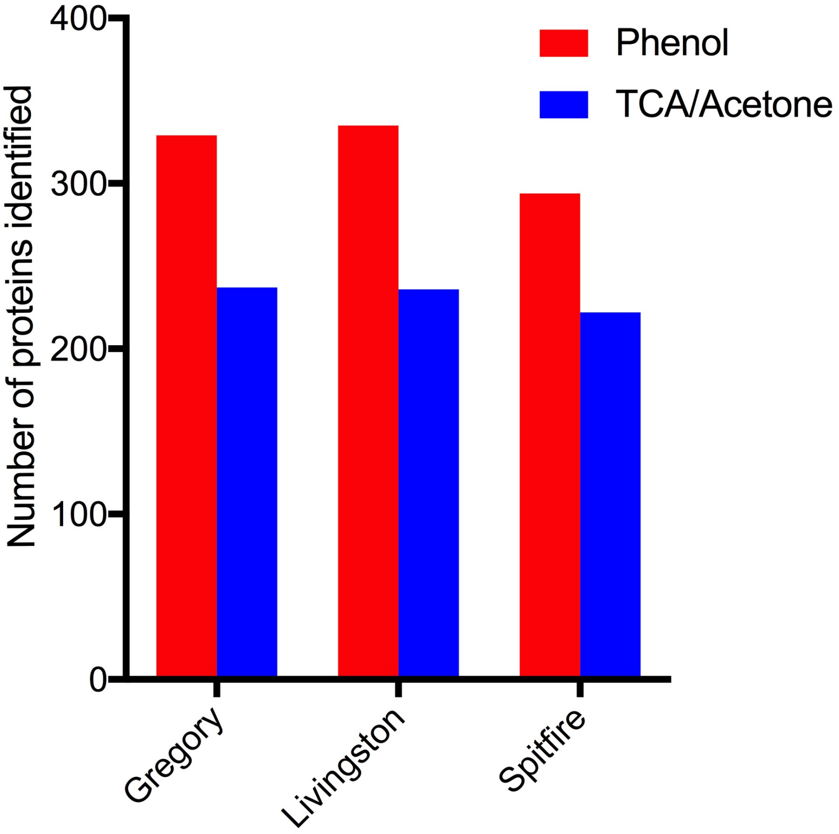


Figure 2.3. Comparison of TOTAL number of proteins identified in each variety by extraction method.
The total number of wheat grain proteins identified in all 3 biological replicates is shown. The difference in the number of proteins identified between the 'combined-phenol' and TCA/acetone methods were statistically significant.

There was a total of 537 proteins identified across all varieties and extraction methods. To be positively identified the protein had to be present in all biological replicates at statistically significant levels as described in the methods. There were 293 proteins that were common in ALL cultivars. As shown in Figure 2.4, 85 of these proteins were uniquely extracted by TCA/acetone, 154 were uniquely extracted by 'combined-phenol', but only 89 were extracted by both methods. The ‘combined-phenol’ extraction method was clearly superior in extracting more different types of proteins when compared with the TCA/acetone method as shown in Figure 2.3, and as such, this method has a greater potential for biomarker discovery, although there is obviously a selective extraction of proteins.

Quantitative assessment of the proteins expressed using normalised spectral abundance factors and by presence or absence of a protein between cultivars is shown in Table 2.3. There is a slight advantage in identification of differentially expressed proteins and presence/absent proteins between cultivars using phenol extraction compared to TCA/Acetone. This is despite the ‘combined-phenol’ method resulting in a higher yield of protein than the TCA/acetone, the latter method still did extract 85 proteins that were common in all cultivars and unique to this method (Figure 2.4).

Despite the greater diversity of proteins obtained by the ‘combined phenol’ method, the proportion of proteins discovered per cultivar (either common or unique) remained similar for both methods (Figure 2.5).

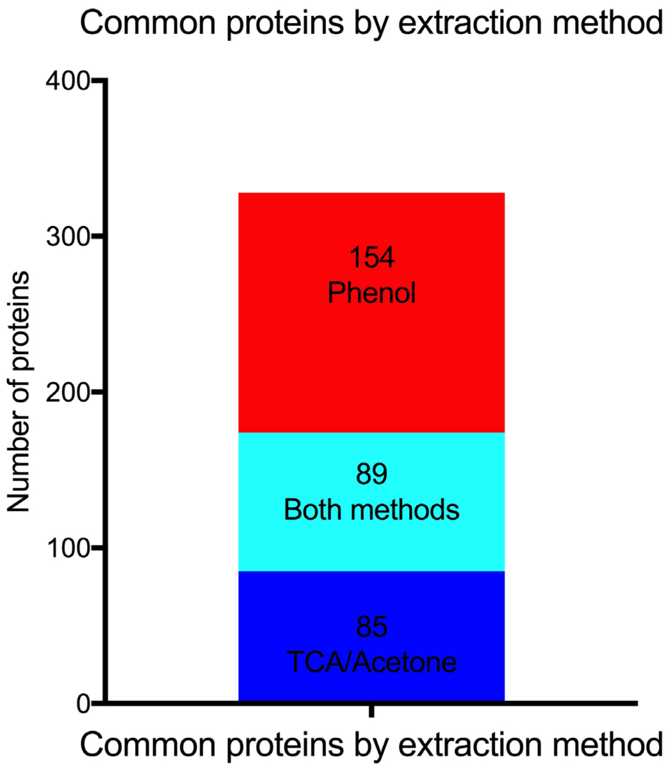


Figure 2.4. Classification of total number of common proteins present in ALL 3 replicates of ALL three wheat varieties by extraction method.

Table 2.3. Proteins displaying differential or present/absent expression between cultivars identified by each extraction method.

Protein extraction method	**Number of proteins**	
	Present or absent	Differentially expressed
TCA/Acetone	185	16
Combined Phenol	187	22
Combined between methods	241	38

Note: A comparison of the number of proteins between the two extraction methods in terms of biomarker discovery using unlabelled shotgun proteomics. The proteins of interest showed either statistically significant differential expression, or a presence/absence pattern between the three cultivars.

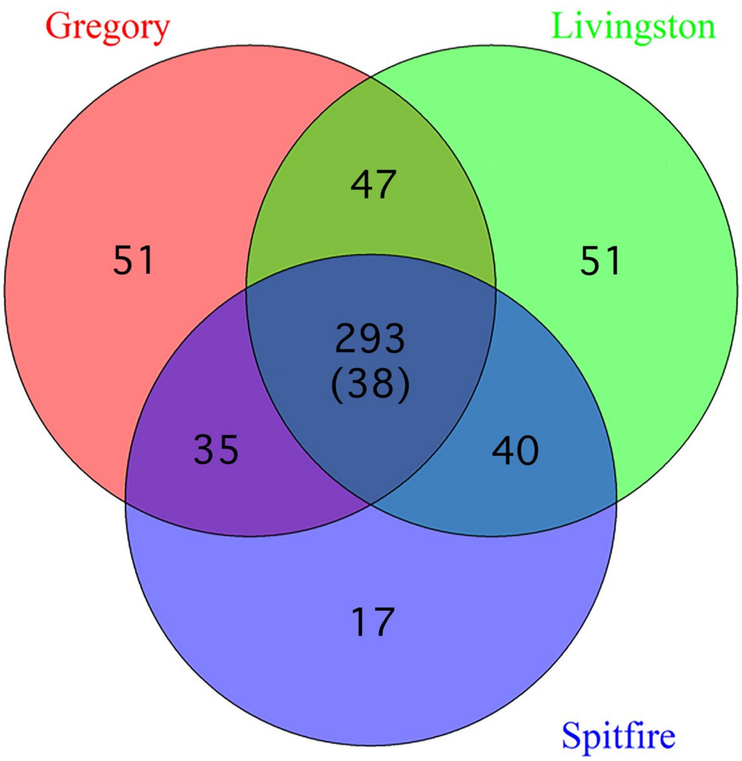


Figure 2.5. Venn diagram of identified proteins originating from Gregory, Livingston and Spitfire wheat cultivars. The 38 proteins in brackets in the central overlapping section are differentially expressed between all three cultivars.

2.3.1.1. Summary of protein functionality

Gene ontology (GO) identifiers assigned to each identified protein in the uniport database were used to further characterise the list of identified proteins. It should be noted that most proteins have multiple GO identifiers specified but also that not all proteins have functional (gene ontology) information associated with them. Consistent with the larger number of identified proteins in the 'combined-phenol' groups, these groups generated a more diverse list of protein functions. These proteins from the different extraction methods were analysed separately to determine if specific classes or functions of proteins were specifically extracted or excluded during the process.

In the list of GO identifiers any differing patterns in general functionality are potentially lost in the hundreds of specific functions attached to each GO identifier. To adjust for this, we used a subset of GO identifiers known as GO Slims, each of which describes a group of very similar functions. This allowed us to sort the large list of GO identifiers into a much smaller set of GO SLIM categories that resulted in a functional summary of the proteome (Lomax, 2005). Then, with the assistance of custom R scripting, we quantitated each category by either counting the number of unique proteins per Slims category, or also adding the protein expression to the total protein count in each Slims category. In general, both quantitative methods gave similar results, so we focused on the latter as it contained more information. Figure 6 shows the expression data for one of the three parent GO's, known as biological process (BP), as this was the most information rich. The findings for BP were comparable to the remaining parent ontologies of cellular component (CC) and molecular function (MF). Also, shown in Figure 6 is the expression patterns in several Slims categories that changes according to the extraction method, with all but two categories having statistically significant changes. For example, the GO terms of "response to biotic stimulus" and "response to external stimulus" derive from proteins that have a greater hydrophobicity, which may account for the higher amount of expression detected in TCA/acetone method.

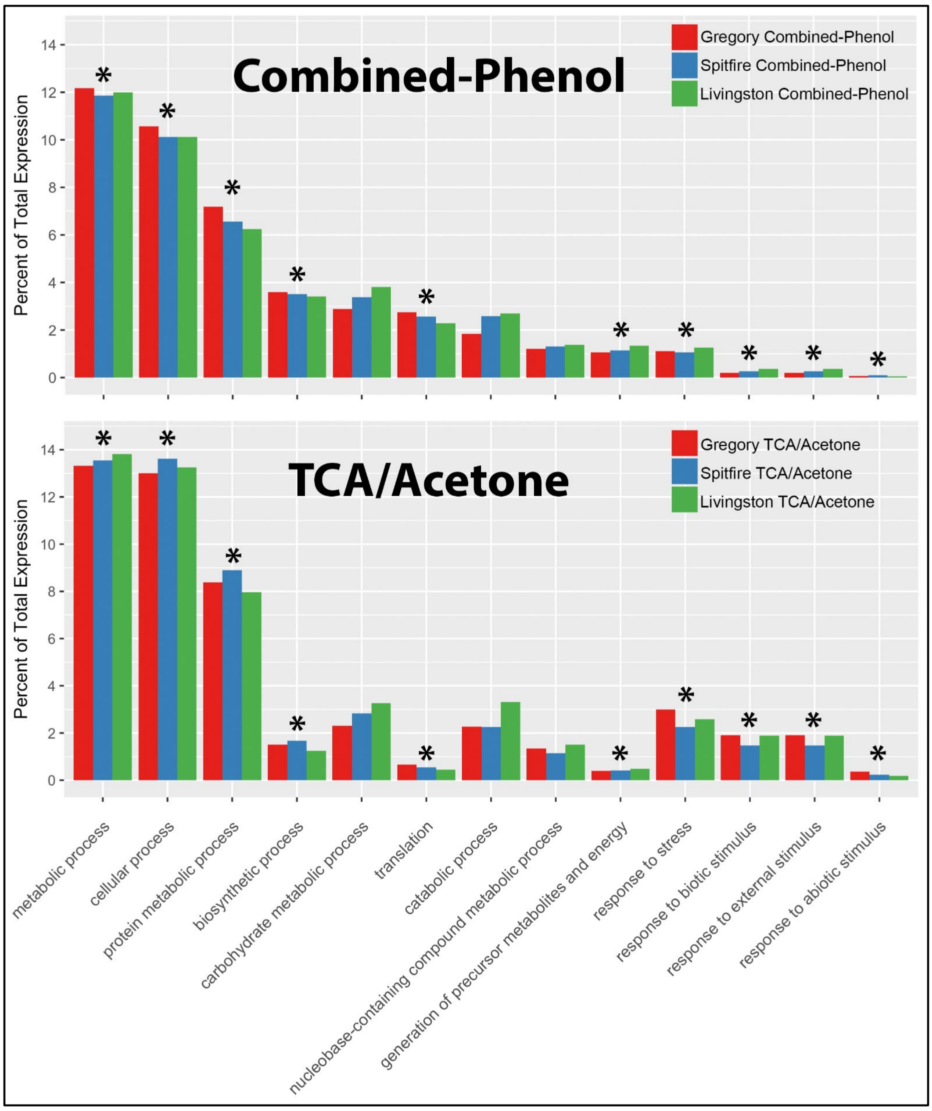


Figure 2.6. Total wheat grain protein expression from (A) 'combined-phenol' extracted samples, and (B) TCA/acetone extracted samples summarised into gene ontology Slims terms for the parent gene ontology category of Biological Process. *Statistically significant difference between functional summaries of proteins extracted by either the 'combined-phenol' or TCA/acetone protein extraction methods.

2.3.1.2. Protein hydrolysis from TCA

The SDS-PAGE gel lanes (Figure 2.2) of the TCA/acetone extracted proteins showed two small bands (7-9 kDa and 3-4 kDa) that were not present in the 'combined-phenol' extracted proteins. By applying unlabelled shotgun proteomics, together with the size fractionation of the SDS-PAGE gel, we examined the size profile of proteins found in the eight gel slices obtained from each of the sample lanes. Using R scripts, these eight slices per lane were then analysed and the corresponding dot-plots drawn of spectral count verses molecular weight (Figure 2.7). For slices 4-8 the majority of the observed protein sizes were within the expected size range, while for slices 1-3 most of the protein sizes were slightly smaller than the expected range. The difference in the expected protein sizes verses the majority of observed protein sizes were proportionally smaller from the top of the gel downwards (slice-1 to -3). This was not an uncommon observation and may have been due to a number of factors such as the amino acid sequence, phosphorylation, glycosylation, and methylation changing migration of larger fragments. Nonetheless, the majority of proteins for each slice were detected in the expected or near expected size range. However, for slices four to seven there were some proteins with a molecular size greater than the range expected from the slice. To a lesser degree this was also true for the proteins extracted by the 'combined-phenol' method, but the second peak arose later at around slice five. Thus, for both methods, it appears that there may have been some acid hydrolysis during the protein extraction process, cutting larger proteins and resulting in them in gel slices with smaller proteins. The hydrolysis observed appeared to be more pronounced in the TCA/acetone method, presumably because of the much longer TCA incubation periods.

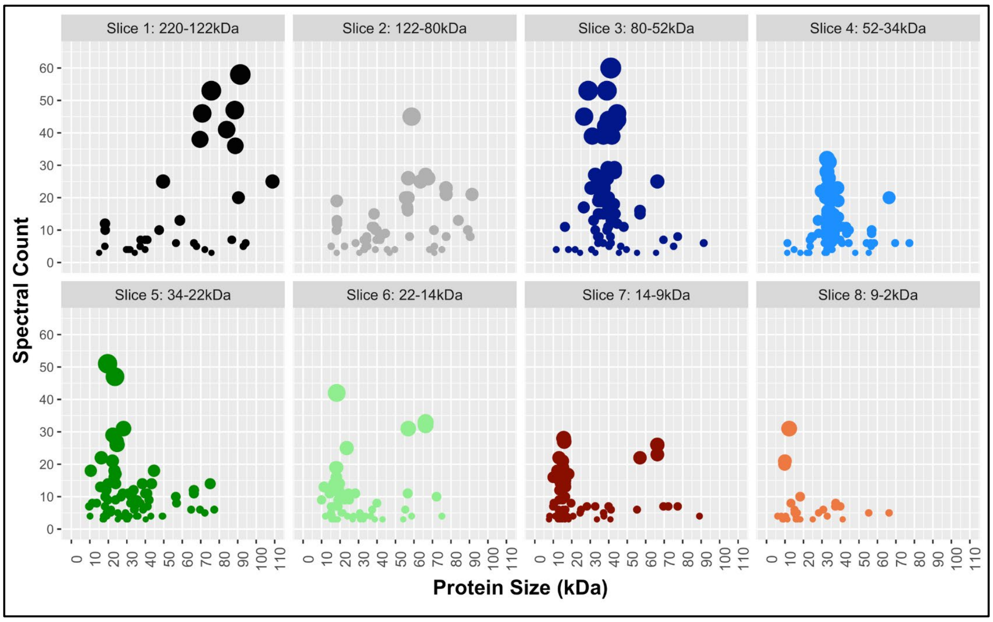


Figure 2.7. Normalised spectral counts (NSAF) of peptides extracted from the SDS-PAGE gel slices from the TCA/acetone extracted proteins from Gregory cultivar.

2.3.1.3. Gluten Fractions

Gluten is a well-studied set of proteins from wheat grain, with various protein fractions and sub-fractions observed to be critical for dough quality (bread, noodles, etc.). As such, we also investigated the effectiveness of the 'combined-phenol' and TCA/acetone methods in extracting gluten. Using an R script on the protein description column of the x-tandem files obtained after unlabelled shotgun proteomics, we found four gluten categories and sorted the remaining proteins into eight major non-gluten categories. The results from each of these fractions were quantitated in terms of spectral count and are summarised in Figure 2.8. We observed that all but one of the major gluten fractions was present. ω-gliadin remained undetected in both protein extraction methods (in all three cultivars), while α-like gliadin and γ-gliadin, as well as HMW and LMW glutenin were present. Also, it was observed that the TCA/acetone method was more effective and specific in extracting gluten fractions. I.e. being twice as efficient in extracting α-like gliadin and γ-gliadin proteins. In contrast, the 'combined-phenol' method is superior at extracting most of the non-gluten proteins, again showing a statistically significant increase in all the non-gluten proteins except serpins, which showed a decrease compared to the TCA/acetone method.

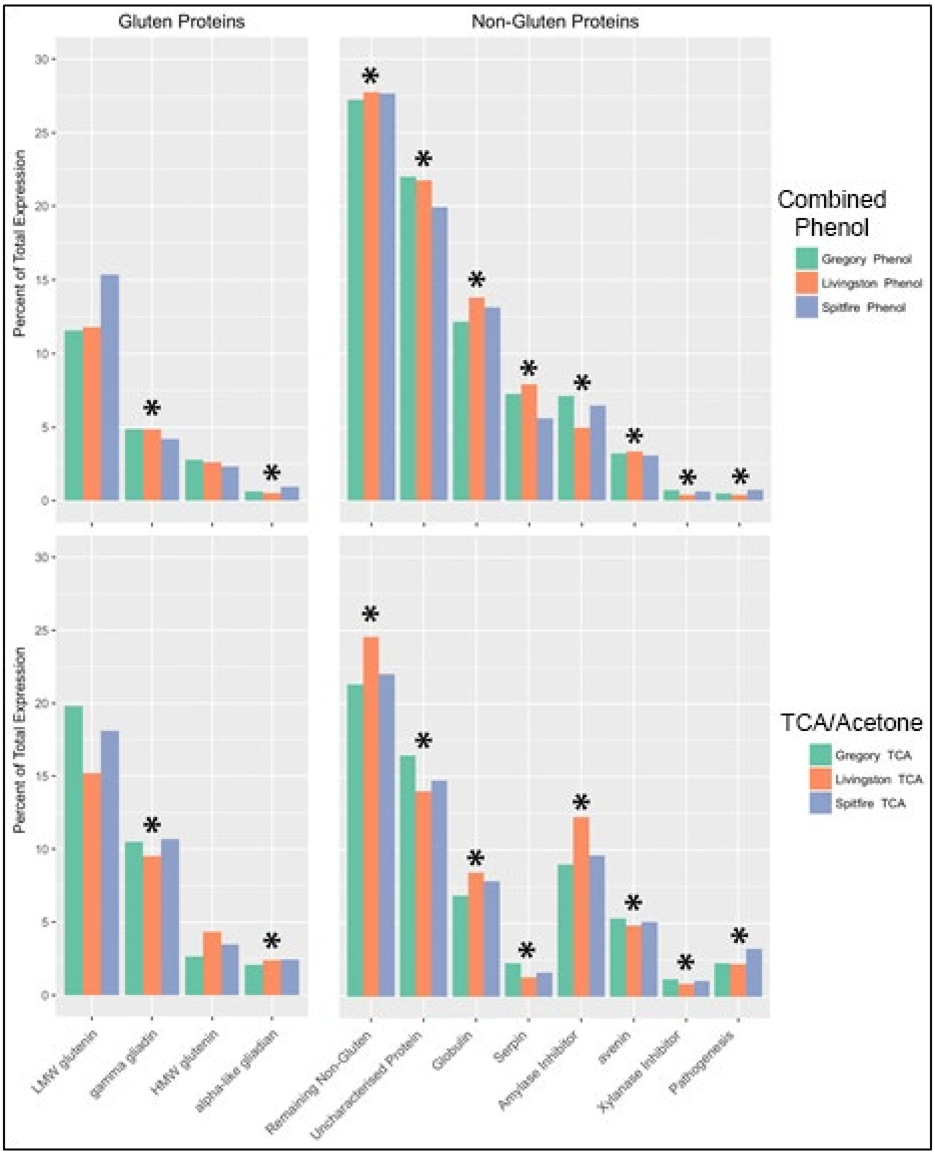


Figure 2.8. Total wheat grain proteins extracted by ‘combined-phenol’ (top two panels), or TCA/acetone (bottom two panels). Results are further sorted into gluten (left-most panels) and non-gluten fractions (right-most panels). *Statistically significant difference in expression between the two extraction methods for each category.

Table 2.4. Comparison between TCA/acetone and 'combined phenol' methods showing protein and gluten ratios and expected values used for grain quality.

Extraction Method	Protein Category	NSAF Sum	Expected
'combined-phenol'	*Glu/Total Protein*	*21%*	56-68%
'combined-phenol'	*Gli/Glu*	*0.345*	0.45-0.76
'combined-phenol'	*HMW/LMW*	*0.199*	0.37-1.44
TCA /acetone	*Glu/Total Protein*	*35%*	56-68%
TCA /acetone	*Gli/Glu*	*0.594*	0.45-0.76
TCA /acetone	*HMW/LMW*	*0.196*	0.37-1.44

2.4. Discussion

Our first investigation examined the starting amount of finely ground grain that would give the optimal yield of wheat grain protein (Table 2.2). Using TCA/acetone, there was no measurable increase in yield when sample size was increased from 10 to 100 mg. Although there was a slight increase in yield in the 'combined-phenol' method, it was not statistically significant (Table 2.1). When we compared the two methods between each of the 10, 20, 50, and 100 mg samples, the 'combined-phenol' method had almost double the protein yield (

Table **2.2**). Similarly, for all consecutive experiments performed with a starting amount of 50 mg for all cultivars, the 'combined-phenol' had the highest yields (Figure 2.1).

We observed that the TCA/Acetone method always formed an agarose-gel-like precipitate, or 'plug', which equated to ~60% of the final sample volume (precipitate and supernatant). This greatly reduced the volume of supernatant available for analysis. We speculate that the grey-white semi-translucent 'plug' consisted of polysaccharides, principally starch and/or glycoproteins, being major constituent of cereal grain extracts. This is consistent with its absence when extracted with phenol because of the ability of phenol to remove sugars in barley extracts (Hurkman and Tanaka, 1986). The similarity between wheat and barley in their starchy grain morphology, may also account for the presence of the 'plug' that we observed in the TCA/acetone extract from wheat (Bathgate and Palmer, 1972).

Since the ‘combined-phenol’ method showed the highest protein yield, we investigated whether it also delivered the highest diversity of proteins for protein biomarker detection. Summarised in Figure 2.3, we found that the ‘combined-phenol’ method extracted the greatest number of proteins for all cultivars, with the diversity between cultivars ranging from 10 - 28%. protein diversity between Gregory and Spitfire cultivars. Using either extraction method independently the difference between varieties was not statistically significant but it was different when comparing these cultivars to Livingstone. However, when a T-test was performed between all proteins identified with both methods, statistically significant differences were revealed between all cultivars, as summarised in Figure 2.3. Also, we found that the results derived from the ‘combined-phenol’ method identified more proteins of interest (Table 2.3) than that of TCA/acetone method. As such, the ‘combined-phenol’ method was more effective in identifying proteins that were either present in one or more cultivars and absent in the remaining cultivar(s), or identifying proteins that had statistically significant differential expression between cultivars. This provided the grounds for novel cereal grain biomarkers by virtue of the higher yield and greater diversity of proteins.

Due to the large chemical diversity of proteins found within a proteome, each protein extraction method extracts a unique subset of proteins (Saravanan and Rose, 2004, Zhang, et al., 2015, Zhen and Shi, 2011). As shown in the SDS-PAGE gel (Figure 2.2), the protein banding patterns are clearly different between the two methods. Specifically, looking at the size range of 55-36 kDa, the TCA/acetone protein extract bands were very prominent, possibly due to this method being better at extracting certain types of gluten proteins. The ω-gliadins, α-like-gliadins, and LMW-GS all fall within this size range (55-36 kDa) and together all three make up a substantial percentage of total grain protein (18-32%). When the list of identified proteins obtained from the TCA/acetone method for all three cultivars was compared with the larger list from the 'combined-phenol' method, a substantial number of proteins were found to be unique to the TCA/acetone method (Gregory: 131, Livingston: 141, and Spitfire: 111). Meaning that while the 'combined-phenol' method had higher yield and diversity of proteins in comparison to the TCA/acetone method, it still did not capture the full diversity of the wheat grain proteome.

Since each extraction method purifies a group of unique proteins, the two populations of proteins extracted by either method should each demonstrate different functional characteristics. We obtained gene ontology (GO) functional information from each set of proteins extracted from the two methods tested for the three wheat cultivars, and summarised these results into Figure 2.8. Again, the 'combined-phenol' method resulted in a statistically significant greater set of unique GO identifiers that each represented a specific function, which in total meant that a larger set of functionalities was found.

Although effective, the above GO identifier list is a crude measure of proteome functionality, primarily due to the hundreds of GO identifiers. Any attempt to describe or summarise functionality is lost in the detail. Hence, to reduce this complexity and determine whether a common functional theme existed between TCA/acetone or 'combined-phenol' protein extracts, we used a subset of GO identifiers known as GO Slims to summarise and compare the functionalities. With the assistance of custom R scripting, we determined a list of GO Slims that represented a functional summary of the proteome under study. Then through further R scripting, we added expression data to the Slims and graphed the results as a proportional comparisons of protein expression levels for each SLIM term. These results are displayed in Figure 6, which compares the functional summaries of the three cultivars, as well as the two extraction methods. Although only the parent GO category of biological process (BP) is represented here - as it is the most information rich - overall observations are similar for the graphs from the remaining parent categories of cellular component (CC) and molecular function (MF). In general terms, Figure 2.8 displays a clear difference in the Slims expression between the two methods, and while it visually may look as though the TCA/acetone method has the greater overall expression, when the numbers are summed in a table, the ‘combined-phenol’ reveals higher expression. It is also true that out of the thirteen Slims terms graphed, all but three (carbohydrate metabolic process, catabolic process, and nucleobase-containing compound metabolic process), do not show statistically significant differences between the two methods. We speculate that the two largest differences between the two methods, represented by the SLIM terms ‘response to biotic stimulus’ and ‘response to external stimulus’, correlates with the chemistries of the proteins that possess that functionality. Generally, the proteins that fall under those two functional groups are excreted proteins with long chain sugars covalently bound, making these proteins highly hydrophilic. These proteins are thus more likely to be extracted by the TCA/acetone method.

It is important to select a protein extraction method that does not excessively modify or damage proteins during the extraction process. As we found in this study, the 4 kDa and 8 kDa protein bands shown in Figure 2.2 indicate that the TCA/acetone method had potentially caused acid hydrolysis of proteins. Since TCA is a weak acid, it is possible that these small proteins are the by-products of acid hydrolysis of the larger proteins (Darragh, et al., 1996, Lugg, 1946, Sun, et al., 2014). Thus, it appears that several proteins have been hydrolysed, and their smaller hydrolysed products have moved down the gel, where after digestion, their peptides were detected and identified as larger proteins, outside of the expected MW range. Within this putative hydrolysed protein distribution, we also examined the identity of each protein and found that the majority were either unidentified proteins or non-gluten seed storage proteins such as globulins and triticins. Interestingly, although not visible on the SDS-PAGE gel, we found that protein hydrolysis had also occurred within the 'combined-phenol' protein extracts; however, it started at a larger MW range and had less intensity. This latter observation is similarly explained by the TCA used in the second wash step of the 'combined-phenol' method, where the low amount of hydrolysis is probably due to the minimal exposure of the proteins to TCA during use of this method.

The gluten protein fractions within wheat grain are known determinants of grain quality and food industry applications. We examined how effective the TCA/acetone and 'combined-phenol' protein extraction methods were at purifying these proteins. Although only a few unique gluten proteins are found in the wheat grain proteome, they are highly expressed and represent 60-70% of total protein within the grain. Gluten ratios are indicators for grain quality (gluten/total-grain-proteins, gliadin/glutenin, and HMW-GS/LMW-GS) and as such, each method was examined for the suitability of this purpose (Uthayakumaran, et al., 1999, Wrigley, 2006). This was achieved by writing a custom R-script that performed a text string search of the protein description, and then collated and grouped the data from the various gluten proteins, noting that protein description and GO information are bioinformatically different entities. For completeness, we also repeated the analysis for the most common types of non-gluten proteins, with the comparisons between the two methods for both gluten and non-gluten proteins summarised in Figure 2.8 and Table 2.4. We can see that for all the gluten protein categories, the comparative levels of expressed gluten proteins are higher for the TCA/acetone method, and specifically for α-like gliadins and γ-gliadins, this increase was statistically significant. However, there was a complete absence of ω-gliadin from both extraction methods, with the expected percentage within total protein should be roughly 4–8%. The comparatively greater efficiency that the TCA/acetone method extracts gluten may well account for the intense banding seen in the TCA/acetone extracts on the SDS-PAGE gel (Figure 2.2). This was observed at around 36-55 kDa, especially as the three gluten sub-fractions expected around this size range (α-like gliadins, γ-gliadins, and LMW-GS) make up 18-32% of total grain protein. In contrast, half of the non-gluten protein categories were more efficiently extracted by the 'combined-phenol' method, while the other half were more efficiently extracted by the TCA/acetone method – with all changes showing statistical significance. Although looking at the non-gluten proteins overall, the majority of non-gluten proteins were more efficiently extracted by the 'combined-phenol' method.

The gluten quality ratios in Figure 2.8 and Table 2.4 show important ratios of various grain protein fractions and sub-fractions. The resulting ratios are remarkably different between the two extraction methods. When comparing the ratios obtained from our results with those from a previous study on wheat grain, it appears that our results are far below those expected (Plessis, 2013). This is likely due to the different methodology used by Plessis (2013). In their study, they used Osborne-like extraction methods to extract gluten fractions and SE-HPLC for further sub-fractionation and quantification. It is assumed from the low protein percentages and gluten ratios observed from Table 2.4 in this study, that while the TCA/acetone is superior to the 'combined-phenol' method for gluten protein extraction, both methods result in a dramatic loss of gluten. There was a difference between the two protein extraction methods, where the TCA/acetone ratios were closer to the expected values for two of the ratios, and within the expected range for the gliadin:glutenin ratio. However, the gluten over the total protein ratio was less than 35%, well below the expected range. Lastly, except for the gliadin:glutenin ratio for TCA/acetone, the remaining ratios were all below the expected range for both extraction methods. This demonstrated that neither method is efficient in extracting gluten proteins and thus is not ideal for determining gluten ratios.

Chapter 3. Proteomic analysis of barley grain for the discovery of putative protein biomarkers to identify cultivar and farm origin

3.1. Introduction

In both domestic and international markets, the sale of barley grain in Australia is lucrative with exports alone worth $2.4 billion (AU) in the financial years 2016-2017 (http://www.agriculture.gov.au/about/commitment/portfolio-facts/grains). Yet, like any product in the modern world, barley grain must adapt to consumer trends. Whether that be organically grown barley grain, a grain with a low or high protein or carbohydrate content, or even a specific cultivar that uniquely delivers a certain texture and taste to a boutique beer. For these reasons barley grain sales are becoming less restricted to the quality and grading system in use within Australian states, or nationally. Container exports of barley grain and other cereals representing a relatively low volume but potentially higher profit, are increasing as farmers and/or distributors attempt to maximise their income by selling a premium product to markets with specific requirements.

Currently the only way of guaranteeing the provenance of the barley grain is by quality assurance methods that involve detailed record keeping of the harvest, every stage of transport, storage, sales, and processing. To date there is no laboratory-based test to determine the grains farm origin or cultivar.

It is well established that the biochemistry of a mature seed is determined by both the genetics of its parent plant and the biotic and abiotic influences on that plant as the seed develops

(Hurkman, et al., 2013, Hurkman, et al., 2009, Ma, et al., 2018). Proteins are at the core of seed biochemistry and determine the viability of seedlings. Thus, different cultivars from the same farm, or different farms using the same cultivar may potentially lead to differences in the grain proteome due to genetic variability or farm growing conditions (abiotic or biotic influences). The use of tandem mass tags (TMT)-labelled shotgun proteomics, and the subsequent bioinformatics analysis, is anticipated to enable the discovery of proteins with statistically significant differential expression. Thus, putative biomarkers may be identified for identifying the farm origin of the barley grain and cultivar.

3.2. Methods

3.2.1. Barley grain sample details and sample processing

The samples for this experiment were barley grain from the Commander, Gairdner and Hindmarsh cultivars. Commander is a malting barley that is high yielding with mid-season flowering and maturity. Gairdner is also a malting barley of a semi-dwarf variety that matures moderately late in the season and grows well in high rainfall cropping regions. In contrast, Hindmarsh is a semi-dwarf feed variety that is early maturing with a high yield, plump grains, and good grain weight. All three cultivars were at three different farms (Breeza, Trangie Agricultural Research Centre [TARC], and Terry Hie Hie [THH]), and at each farm were three replicate plots. The grain was harvested in 2013. The farm locations are separated by hundreds of kilometres (Breeza to Terry Hie Hie, 240 km; Terry Hie Hie to TARC, 406 km; and TARC to Breeza, 324 km) while the distance between the three plots on the farm was not provided by the farm research facilities.

Location coordinates: obtained from http://weather.mla.com.au (date accessed site: 13/12/18).

Above Mean Sea Level (AMSL)

Breeza location: 31.2442°S, 150.4579°E, 309m AMSL

TARC location: 32.0319°S, 147.9839°E, 234m AMSL

Terry Hie Hie location: 29.7956°S, 150.1511°E, 285m AMSL

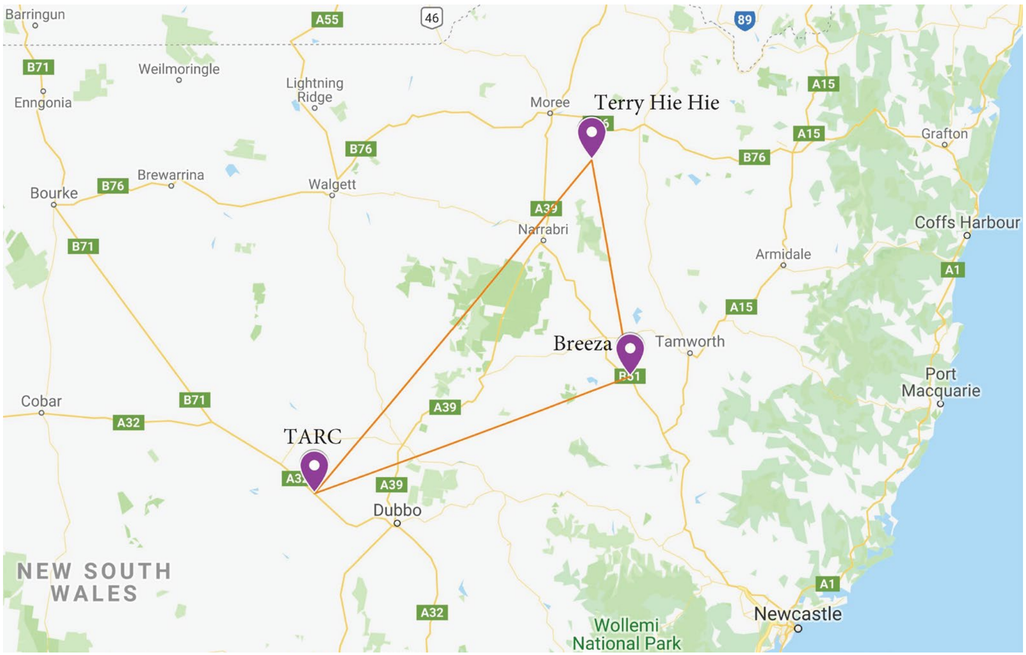


Figure 3.1.: Map showing the location of the three farms of Terry Hie Hie (THH), Trangie Agricultural Research Centre (TARC), and Breeza. Farm sources of the barley grain used in this study was generated from Google Maps.

According to data from the New South Wales government website known as "SEED" (Sharing and Enabling Environmental Data), an interactive mapping database (https://geo.seed.nsw.gov.au/Public_Viewer/index.html?viewer=Public_Viewer&locale=en-AU) there are differences in the three sites in terms of crop growing capacity. "SEED" describes land-use sustainability in terms of "soil capability" and scores the land from 1 (highest quality) to 8. Of the three sites, Breeza is in general the highest on this scale (level-2), while THH is second (level-3), and TARC being the lowest (level-4).

The barley was sown (germinated upon sowing) on the 29th of May 2013 and was harvested around late-October/early-November (exact date unknown). The crops were not watered (only from rainfall) and were treated with fungicides (details obtained from Rohan Brill from the Department of Primary Industries, NSW Government).

All weather observation data (average monthly rainfall and temperature during crop growth stages) was obtained from http://www.bom.gov.au/climate/data/index.shtml (date accessed: 13/12/18).

Weather station locations:

Rainfall:

TARC – weather station at location

Breeza – Gunnedah resource centre was the closest location (44 km) for weather observation

Terry Hie Hie – Bingara (40km) and Warialda PO (48km) closest weather stations

Temperature:

TARC – weather station at location

Breeza – Gunnedah resource centre was the closest location (44 km) for weather observation

Terry Hie Hie – Moree (52 km) closest weather stations

3.2.2. Experimental design

TMTs were selected for labelling sample peptides for proteomic analysis, as this approach is high-throughput with high sensitivity. The experiments in this study were organised into three major sample groups, each using a set of 10 different TMT labels per sample group. Within each sample group were three biological replicates. The tenth TMT label in each experiment represents the 'pool' of the nine samples within that group. The first sample group (TMT-set-1) is a comparison of three different cultivars (Commander, Gairdner, and Hindmarsh) all grown at TARC farm. The second sample group (TMT-set-2) is a comparison of the Commander barley cultivar grown at three different farms (TARC, THH, and Breeza). Finally, the third sample group (TMT-set-3) is a comparison of the Gairdner barley cultivar grown at three different farms (TARC, THH, and Breeza).

Table 3.1. Barley grain samples – experiment design for TMT-labelled proteomics.

TMT label No.	**TMT set-1**	**TMT set-2**	**TMT set-3**
1 (126)	Commander-TARC-BR1-2013	Commander-TARC-BR1-2013	Gairdner-TARC-BR1-2013
2 (127N)	Commander-TARC-BR2-2013	Commander-TARC-BR2-2013	Gairdner-TARC-BR2-2013
3 (127C)	Commander-TARC-BR3-2013	Commander-TARC-BR3-2013	Gairdner-TARC-BR3-2013
4 (128N)	Gairdner-TARC-BR1-2013	Commander-THH-BR1-2013	Gairdner-THH-BR1-2013
5 (128C)	Gairdner-TARC-BR2-2013	Commander-THH-BR2-2013	Gairdner-THH-BR2-2013
6 (129N)	Gairdner-TARC-BR3-2013	Commander-THH-BR3-2013	Gairdner-THH-BR3-2013
7 (129C)	Hindmarsh-TARC-BR1-2013	Commander-Breeza-BR1-2013	Gairdner-Breeza-BR1-2013
8 (130N)	Hindmarsh-TARC-BR1-2013	Commander-Breeza-BR2-2013	Gairdner-Breeza-BR2-2013
9 (130C)	Hindmarsh-TARC-BR1-2013	Commander-Breeza-BR3-2013	Gairdner-Breeza-BR3-2013
10 (131)	Pool of all TARC farm biological reps	Pool of all Commander Barley biological reps	Pool of all TARC farm biological reps

Legend: BR = Biological Replicate; Commander, Gairdner, and Hindmarsh = Barley cultivars; Breeza, TARC, and THH (Terry Hie Hie) = farms.

3.2.3. Barley grain sample processing

The processing of grain samples was performed as described in Chapter 2 (Methods section) using the ‘combined-phenol’ extraction method, with some modifications (described below) to prepare the peptides for TMT-labelling.

3.2.4. Reduction and alkylation of protein extract

The extracted proteins were reduced by adding 2.5 µL of 1 M dithiothreitol (DTT) to 500 µL of resuspended protein solution (5 mM DTT final concentration), followed by incubating for 15 minutes at room temperature (RT). Alkylation was then performed by addition of 5 µL of 1 M iodoacetamide (IDC) to the 500 µL resuspended protein solution (10 mM final IDC concentration), followed by incubating in the dark for 30 minutes at RT. Alkylation was quenched by addition of 2.5 µL of 1 M DTT and incubating for 15 minutes in the dark at RT. The final DTT concentration was 10 mM.

3.2.5. Methanol/chloroform precipitation

The samples were precipitated using methanol/chloroform method adapted from (Montealegre, et al., 2010). To the 0.5 mL sample was added 2 mL of methanol, then 0.5 mL of chloroform and finally, 1.5 mL of MilliQ water. Note: the samples were vortexed following each addition of the following solutions. The resulting ratio of this mixture was 1:4:1:3, (sample:methanol:choloroform:water). The samples were left to stand for 5 minutes at RT and then centrifuged at 14,000 rpm for 5 minutes. The top water/methanol layer on top of the interface was removed without disturbing the interface and then discarded. The protein pellet was washed with 2 mL (4 volumes) of cold methanol and vortexed. Finally, the sample was centrifuged at 14,000 rpm for 2 minutes, the supernatant removed and the pellet left to semi-dry. Care was taken to avoid allowing the pellet to completely dry in order to aid in resuspension. To further assist resuspension (presumably from the freeze thaw cycle), the pellet was placed in the freezer at -20°C for 30 min. After freezing, 500 µL of 8 M urea in 50 mM Tris-HCl (pH 8.8) was added to the pellet and placed on an orbital shaker at 35°C until it had completely dissolved, typically 60 min.

3.2.6. Protein digestion

To prepare for protein digestion, 150 µg of each protein sample extract was placed into a fresh 1.5-mL plastic tube. For the pooled samples, a total of 150 µg of protein extract was added to a fresh 1.5-mL plastic tube. Each sample was then diluted eight-fold by adding 7x the sample volume with 50 mM Tris-HCl, pH 8.8 to each tube. To each diluted sample was added 1.5 µg of Lys-C endoproteinase, resulting in a 1:100 ratio of Lys-C to sample (1.5 µg Lys-C to 150 µg protein). The samples were then left to incubate overnight at 37°C. Following incubation, 1.5µg of trypsin was added to each sample (again, at a 1:100 ratio) and incubated for 4 hours or overnight at 37°C. Trypsin was inactivated by addition of 100% TCA so that the final TCA concentration in each sample was 1%, then pH strips were used to make sure the samples pH was <= 2.0. This is important, as a low pH is needed for the next desalting step (stage tipping), hence more TFA was added when necessary. A small amount of each sample was analyzed by SDS-PAGE to ensure that the digestion was complete.

3.2.7. SDS-PAGE

The method for examining protein extraction quality using SDS-PAGE is described in chapter 2 (Methods section).

3.2.8. Constructing SDB-RPS tips for sample desalting (stage tipping)

A 1 mL pipette tip was cut at the 0.5 mL mark. Using a Hamilton 16-gauge needle (# 90516), a small disc was cut from a 3M SDB-RPS Empore Disc. One small disk was used for each 100 µg of sample peptide. Each small disc was cut so that the SDB-RPS disc/s is/are retained in the needle. The needle was placed into the opposite end of the cut pipette tip (the larger opening), and the small discs gently pushed out of the needle and into the pipette tip using the plunger. The discs were pressed into the pipette tip to the point of resistance. The plunger was used to gently pat the other end of the discs flat into the small end of the cut pipette tip. Finally, the plunger was placed back into the large end of the pipette tip and was used to gently push down further on the disc until further resistance was felt. Caution was taken to not compress the SDB-RPS too much as it could prevent the flow-through of the wash and elution solutions.

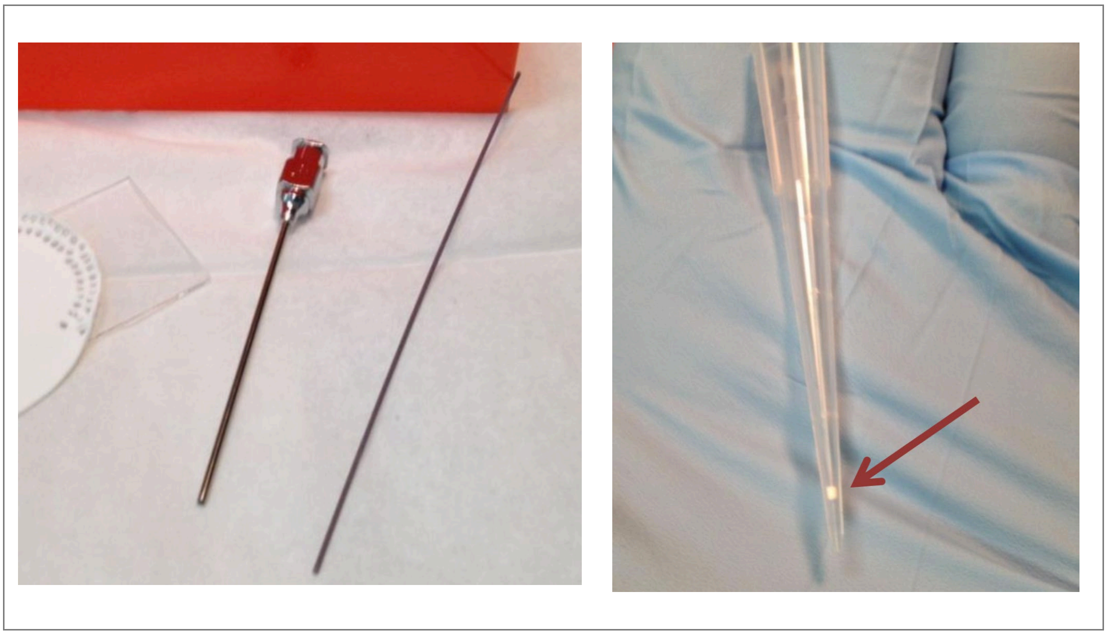


Figure 3.2. Elements of SDB-RPS tip construction.
16-guage needle and rod to fit the inside of the needle, and the pipette tip with gently compacted SDB-RPS disk that had been cut out by the needle and pushed out of the needle by the rod.
Image taken from:
https://cbs.umn.edu/sites/cbs.umn.edu/files/public/downloads/Stage_Tip_MCXProtocol_w_photos_CMSP_20150817.pdf

3.2.9. SDB-RPS de-salting (stage tipping)

The stage tips were constructed as described above and labelled. Each sample was loaded into its appropriately labelled stage tip. The samples were then centrifuged at 2,000 x *g* until each sample had passed through each tip (the peptides should be retained due to the low pH). The filter tips were washed twice with 200 µL of 0.2% trifluoroacetic acid (TFA). Each sample was eluted by adding 200 µL of 'Elution 3' solution (5% ammonium hydroxide, 80% acetonitrile [ACN]). The eluate was collected by centrifugation at 2,000 x *g*. Another 200 µL of 'Elution 3' solution was added and the centrifugation step was repeated. The 400 µL eluent samples were dried-down in a SpeedVac at RT overnight. Each sample was then resuspended in 150 µL of 100 mM HEPES buffer (pH 8).

3.2.10. Micro-BCA assay of desalted peptide samples for TMT labelling

Micro-BCA working solution was prepared as per manufacturer's instructions (Thermo Fisher Scientific Australia, Micro-BCA protein assay kit, #23235), with the total volume made to 25:24:1 of solutions A, B, and C, respectively. Micro-BCA standards (Bovine Serum Albumin [BSA]) were prepared in water at the following protein concentrations: 0.2, 0.04, 0.02, 0.01, 0.005, 0.0025, 0.001 and 0.0005 µg.µL^{-1}. The samples were diluted 1 in 4 with 100 mM HEPES-NaOH buffer (pH 8). The loadings on the 96-well Greiner flat-bottomed plate (Sigma-Aldrich: M2936) were 10 µL of standard and 5 µL of sample. To each well was added 150 µL of Micro-BCA working stock. Three technical replicates were loaded for each standard and sample. The plates were then measured using a PHERAstar FS plate reader (BMG LABTECH), with software version 4.00 R4 and firmware version 1.13. Standard BCA settings were applied: 10 seconds of orbital shaking at 500 rpm was performed prior to reading, and the absorbance measured at 562 nm for 0.1 s, at 20 flashes per well.

3.2.11. Sample preparation for TMTs

The peptide sample size (in µg) was chosen based on the lowest peptide sample amount. The maximum volume was calculated so that the appropriate volume of 100 mM HEPES-NaOH (pH 8) could be added to each tube to maintain a uniform volume and concentration of peptide across all samples. In the following procedure, extra dry, sealed, reagent quality ACN was used (100 mL Bottles, each packed in a septum sealed "Acroseal" bottle, Chem Supply Pty Ltd, #326811000). The 10 TMT labels were removed from the freezer (-20°C), making sure they were in the correct order. To each TMT label was added 84 µL of ACN. Each of the TMT tubes were vortexed for 10 s, followed by pulse centrifugation. To each tube was added 20 µL of the appropriate label. When performing this step, all of the contents of the first label was added to the appropriate samples until moving onto the second label (e.g. add TMT-126 to all samples requiring TMT-126, then TMT-127N to all those requiring TMT-127N, and so on). Once all of the TMT labels had been added, samples were briefly vortexed and pulse centrifuged. All samples were incubated at RT for 1 hour. The reaction was stopped following the addition of 8 µL of 5% hydroxylamine. Each TMT set was pooled into a 2-mL plastic tube, or split equally between two 2-mL plastic tubes when the sample volume was too large for one tube per set. The pooled samples were dried overnight in a SpeedVac. The dried sample pools were stored at -20°C. Finally, the dried peptides of each TMT set were resuspended with 1 mL of 1% formic acid (e.g. if there were two tubes for each TMT set, add 500 µL of 1% formic acid to each tube, then combine into one tube).

3.2.12. Preparing a test run of TMT labelled sample sets for mass spectrometry

From each TMT sample set was removed 10 µg (41.7 µL) of peptides to be prepared for a test run on the mass spectrometer to measure sample quality related to TMT labelling. The samples to be tested were desalted by stage tipping as described previously. For example, the SDB-RPS tips were constructed and then the peptides were desalted by stage tipping. Following this, the samples were dried in the SpeedVac and resuspended in 20 µL of 0.1% formic acid. From the resuspension was taken 10 µL and loaded into a vial ready for mass spectrometry.

3.2.13. SCX fractionation

Offline SCX fractionation was carried out to reduce the complexity of the mixture, using an Agilent 1260 quaternary HPLC pump with a PolyLC polysulfoethyl aspartamide column (200 mm×2.1 mm, 5µm, 200 Å; PolyLC, Columbia, MD). The column was equilibrated with buffer A (5 mM KH_2PO_4, 25% v/v ACN, pH 2.72), which is also used for sample resuspension, sample injection and peptide adsorption to the column. Peptide elution was achieved with a linear gradient of 10–45% buffer B (5 mM KH_2PO_4, pH 2.72, 350 mM KCl, 25 % ACN) for 70 minutes, which is then rapidly increased from 45 to 100% buffer B for 10 minutes at a flow rate of 300 $\mu L.min^{-1}$. The peptides were detected with an in-line UV detector at 210 nm. A total of 36 fractions of varying volumes were collected in a 96-well collection plate and dried down by vacuum centrifugation. To each of the 36 peptide-containing wells, 100 µL of 1% TFA was added and then vortexed thoroughly for 10 min at 4°C, before being combined into 12 fractions based on UV absorbance. These 12 fractions were desalted using SDB-RPS stage tips, dried down in SpeedVac and reconstituted in 0.1 % formic acid in preparation for LC-MS/MS.

3.2.14. Mass Spectrometry for TMT labelled samples

Peptide samples were separated on an EASY-nLC1000 liquid chromatography system (Thermo-Scientific) which was coupled to a Q Exactive Orbitrap mass spectrometer (Thermo-Scientific). Reversed-phase chromatographic separation was carried out on a 75 µm id.×100 mm, C18 HALO column, 2.7 µm bead size, 160 Å pore size. A linear gradient of 1-30% solvent B (99.9% ACN/0.1% FA) was run over 170 minutes. The mass spectrometer was operated in the data-dependent mode to automatically switch between Orbitrap MS and ion trap MS/MS acquisition. Survey full scan MS spectra (from *m/z* 350–1850) were acquired at a precursor isolation width of 0.7 *m/z*, resolution of 70,000 at an *m/z* of 400 and an AGC (Automatic Gain Control) target value of 1×10^6 ions. For the identification of the TMT labelled peptides, the ten most abundant ions were selected for higher energy collisional dissociation (HCD) fragmentation. HCD normalized collision energy was set to 35% and fragmentation ions were detected in the Orbitrap at a resolution of 70,000. Target ions that had been selected for MS/MS were dynamically excluded for 90 s. For accurate mass measurement, the lock mass option was enabled using the polydimethylcyclosiloxane ion (*m/z* of 445.12003) as an internal calibrant.

3.2.15. Making a multi-FASTA file for proteomics searches

A protein (FASTA) database was constructed for barley to analyse the peptide mass spectra following mass spectrometry. Firstly, an NCBI protein search was performed on barley (https://www.ncbi.nlm.nih.gov/protein/?term=txid112509), then the complete result downloaded as a FASTA file. Any poorly annotated and partial sequences from the FASTA file were removed by the program CD-Hit, which was downloaded from https://github.com/weizhongli/cdhit/. Command line instructions were as follows: “cd-hit –i sequence.fasta -o nr100 -c 1.00 -n 5 -d 120 -l 50”. Input filename is “sequence.fasta”, and the output is filename is “nr100.fa”. The sequence identity threshold was set to 1.0, word-length to 5, with a length of description set to 120, and all sequences less than 50 amino acids were discarded. The latter arbitrary value resulting in the loss of a few (if any) small functional proteins from the database, in favour of removing a greater number of unwanted expressed sequence tags (ESTs) that would slow down protein matching algorithms.

3.2.16. Protein identification and quantitation

The raw data files were generated by Xcalibur software (Thermo-Scientific) and processed using Proteome Discoverer V1.3 (Thermo-Scientific) through a local MASCOT server (version 2.3; Matrix Science, London, UK). The MS/MS spectra were searched against the appropriate custom barley FASTA database, which was constructed as described in section 3.2.15. The MS tolerance was set to ±10 ppm, MS/MS tolerance to 0.1 Da and trypsin digest settings enabling one missed cleavage. Carbamidomethylation of cysteine, 10-plex TMT tags on lysine residues and peptide N-termini were set as a static modification, while oxidation of methionine and deamidation of asparagine and glutamine residues were set as a variable modification. Search result filters were selected as follows: only peptides with a score >15 and below the Mascot significance threshold filter of p = 0.05 were included and single peptide identifications required a score equal to, or above the Mascot identity threshold. Protein grouping was enabled such that when a set of peptides from one protein was equal to, or completely contained within the set of peptides of another protein, the two proteins were contained together in a protein group. Proteins with at least two unique peptides were regarded as confident identifications. Relative quantitation of proteins was achieved by pairwise comparison of TMT reporter ion intensities. For example, the ratio of the labels for each of the treatment replicates (numerator) versus the labels of their corresponding control replicates (denominator). These results were then saved as a tab-delimited file for further analysis.

3.2.17. Statistical analysis of identified proteins using "TMTPrePro"

General statistical analysis and quality control of the data was performed using the TMTPrePro software package (developed by APAF), as described in the paper by Mirzaei, et al. (2017). There were two necessary file format inputs required to run this package. The first is a tab-delimited file of protein search results from Proteome Discoverer. The second, shown in Table

2, was a design file in Excel format that is made up of two tabs. The first of which contains the label, replicate, and grouping information in columns; and the second containing the protein file name and information on which the label is to be used as a reference (denominator). In TMTPrePRo, there were four adjustable parameters, which after some initial tests, were set to a 1.3-fold count cut-off (FCCutoff), a Z-score (ZScoreCutoff; 100×log(ratio)/Variability) cut-off of 2, a counts cut-off (CountsCutoff) of 1, and P-value cut off (PvalCutoff) of 0.05. After running TMTPrePro, a number of tables and graphs were produced that described the quality of the run and samples, the degree of similarity of sample sets, and whether any proteins had statistically significant differential expression. Two of the output files (ResultsOverall.xlsx, and the design file) were then used by various custom R-scripts for further analysis as will be described in chapters 4 and 5.

Table 3.2. Example design spreadsheet for TMTPrePro.

“xlsx” file - TAB 1 (Design)		
Label	**Replicate**	**Group**
126	Comm-BR1	1TARC-Comm
127_N	Comm-BR2	1TARC-Comm
127_C	Comm-BR3	1TARC-Comm
128_N	Gaird-BR1	2TARC-Gaird
128_C	Gaird-BR2	2TARC-Gaird
129_N	Gaird-BR3	2TARC-Gaird
129_C	Hind-BR1	3TARC-Hind
130_N	Hind-BR2	3TARC-Hind
130_C	Hind-BR3	3TARC-Hind
131	Pool	1Pool

“xlsx” file - TAB 2 (References)	
File	**UseReference**
160826_PW_06_proteinGroups_proteingroups.txt	131

Note: The spreadsheet contains two tabs as represented by the table above. Tab 1 is title “Design”, and tab 2 is titled “References”.

3.2.18. Custom R-script: PCA.R

This custom R-script gathers data from a number of sources: 1) the design file used by “TMTPrePro”, 2) the summary table of protein expression and other statistical results output by “TMTPrePro”, and 3) a FASTA database (file) of barley proteins. Other inputs (and script sections) are redundant and were only included for future expansion of the script. Minor manipulation of the script allowed the inclusion of either all detected proteins, or the selection of proteins that showed statistically significant differential expression. Thus, either the full data set or the reduced set (of putative biomarker proteins) were used for plots of principle component analysis (PCA), drawing heat-maps, as well as box-plots and violin-plots with or without the display of underlying data points (Appendix B, section B.6.1).

3.3. Results

3.3.1. TMT-barley grain experiment

The barley grain protein extract from the various samples (as described in the Materials and Methods section of this Chapter), was analysed by TMT-labelled shotgun proteomics, to identify and measure differentially expressed proteins. Three sample groups were compared to a pool of sample groups. Each sample group was made up of three biological replicates that originated from a specific farm and cultivar combination, and replicates from the same location and cultivar are defined as sample groups.

3.3.2. Barley grain sample weights

There was some variation in the measured starting weight of 200-grains for each sample, with a difference of 1.329 g between the highest and lowest sample result (Table 3.3; and Appendix B, table B.1.). The lowest weight-average was Commander grain harvested from THH farm resulting in 7.82 g per 200-grains, while the highest of 9.15 g came from Commander grain harvested at Breeza farm. The lowest farm average 200-grain weight, which includes all cultivars for a particular farm was 8.2 g from the THH farm and the highest was 9.03 g from the Breeza farm. The TARC farm had an average 200-grain weight of 8.36 g (Table 3.3). When looking at differences in 200-weight by either grouping samples according to location or cultivar, statistical significance was seen between the TARC-Breeza or THH-Breeza location comparisons (Appendix B, Table B.1.1). When 200-weight results were first grouped according to location followed by a comparison between cultivars at each location, statistical significance was seen between Commander and Gairdner comparisons at all locations (Appendix B, Table B.1.3). Hindmarsh was excluded as it was only grown at TARC.

Table 3.3. Averages of 200-grain weight measurements.

Sample group code	Farm location	Cultivar	Average 200-grain weight (g)	Average farm location 200-grain weight (g)
Br-Comm-2013	Breeza	Commander	9.15	9.03
Br-Gaird-2013	Breeza	Gairdner	8.91	
TARC-Comm-2013	TARC	Commander	8.80	8.36
TARC-Gaird-2013	TARC	Gairdner	7.86	
TARC-Hind-2013	TARC	Hindmarsh	8.44	
THH-Comm-2013	Terry Hie Hie	Commander	7.82	8.20
THH-Gaird-2013	Terry Hie Hie	Gairdner	8.59	

Location: Br = Breeza farm, TARC farm, THH farm
Cultivar: Comm = Commander barley, Gaird = Gairdner barley, Hind = Hindmarsh barley.

3.3.3. Protein quantitation

Following protein extraction, the proteins were quantified by the Bradford assay, which revealed a large variation in extraction efficiency. This was observed by the amount of protein recovered from each sample group and their replicates (Table 3.4). The lowest result (460 µg from TARC farm, Commander grain, replicate-3), could be explained by some loss during the protein extraction for that particular replicate. The remaining lesser variation possibly due to small errors during the multiple supernatant removal and wash steps, or the difficulty of pellet re-suspension. Variation was observed to be much reduced at the later SDB-RPS de-salting stage both in the measures of extracted protein weights (table 3.4), and highly similar protein banding patterns of replicate samples run on SDS-PAGE (eg. Figure 3.11).

The lowest average protein extract weight was from Commander grain harvested from TARC farm at 460 µg, while the highest of 1715 µg came from Hindmarsh grain harvested also at TARC farm. The lowest farm location average protein extract weight, which includes all cultivars from that farm was 828 µg from the Breeza farm and the highest was 1253 µg from the TARC farm. The THH farm had an average of 1024 µg protein extract weight (Table 3.4). For protein extract weight (Table 3.4) no statistical significance was seen, whether comparing either location, cultivar, or focussing on cultivar data at each of the three locations (Appendix B, Table B1b.1, Table B.1b.2, and Table B.1b.3).

Table 3.4. Bradford assay results showing the amount of protein recovered following the 'combined-phenol' extraction method.

Sample ID	Farm location	Cultivar	Biological replicate	Weight of sample protein extract (µg)	Average weight of sample protein extract per farm (µg)
TARC-Comm-BR1-2013	TARC	Commander	1	1535	1253
TARC-Comm-BR2-2013	TARC	Commander	2	1240	
TARC-Comm-BR3-2013	TARC	Commander	3	460	
TARC-Gaird-BR1-2013	TARC	Gairdner	1	725	
TARC-Gaird-BR2-2013	TARC	Gairdner	2	1660	
TARC-Gaird-BR3-2013	TARC	Gairdner	3	1400	
TARC-Hind-BR1-2013	TARC	Hindmarsh	1	1715	
TARC-Hind-BR2-2013	TARC	Hindmarsh	2	1265	
TARC-Hind-BR3-2013	TARC	Hindmarsh	3	1275	
THH-Comm-BR1-2013	Terry Hie Hie	Commander	1	1000	1024
THH-Comm-BR2-2013	Terry Hie Hie	Commander	2	1050	
THH-Comm-BR3-2013	Terry Hie Hie	Commander	3	740	
THH-Gaird-BR1-2013	Terry Hie Hie	Gairdner	1	875	
THH-Gaird-BR2-2013	Terry Hie Hie	Gairdner	2	1435	
THH-Gaird-BR3-2013	Terry Hie Hie	Gairdner	3	1045	
Br-Comm-BR1-2013	Breeza	Commander	1	625	828
Br-Comm-BR2-2013	Breeza	Commander	2	935	
Br-Comm-BR3-2013	Breeza	Commander	3	845	
Br-Gaird-BR1-2013	Breeza	Gairdner	1	690	
Br-Gaird-BR2-2013	Breeza	Gairdner	2	1065	
Br-Gaird-BR3-2013	Breeza	Gairdner	3	810	

*Total volume of resuspended protein was 500 µL.

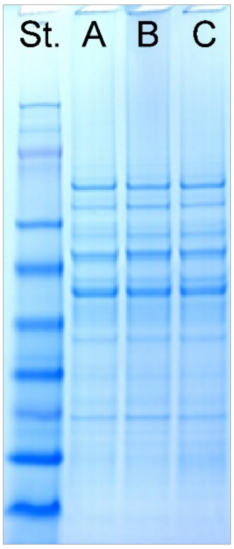


Figure 3.3. 'Combined-phenol' extracted proteins run by SDS-PAGE from barley grain of the Commander cultivar that was grown at TARC.

3.3.4. Preparation of samples for mass spectrometer

The trypsin and LysC protein digestion were successful, indicated by the missing protein bands in the qualitative SDS-PAGE gel (Figure 3.12). There were 71 µg to 150.7 µg of peptides per sample (Table 3.5). In terms of weight, this yield ranged between 5-33% of the total protein measured following the initial protein extraction (Table 3.4).

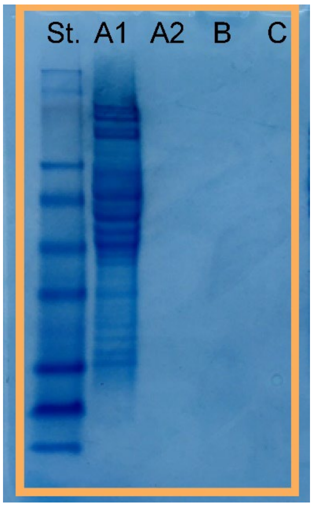


Figure 3.4. SDS-PAGE analysis of Lys-C and trypsin digested proteins from the barley grain samples that were extracted using the 'combined-phenol' method. St. = standards, A1 = undigested Commander barley proteins, A2 = digested Commander barley proteins, B = digested Gairdner barley proteins, and C = digested Hindmarsh barley proteins.

The lowest average peptide extract weight was from Hindmarsh grain harvested from TARC farm. This was 71 µg, while the highest of 150.7 µg came from Commander grain harvested also at TARC farm. The lowest farm location average peptide weight, which includes all cultivars from that farm, was 94.1 µg from the TARC farm and the highest was 103.1 µg from the Breeza farm. The THH farm had an average of 97.4 µg peptide weight (Table 3.4). Data on the weight of peptides recovered for each sample (Table 3.5) showed no statistical significance between Commander and Gairdner samples when each location was examined (Appendix B, Table B.1b.3; Hindmarsh data was not included as this cultivar was only grown in one location).

Table 3.5. Barley sample peptide concentration determined by micro BCA.

Sample ID	Farm location	Cultivar	Biological replicate	Total peptide (µg)	Total peptide per farm (µg)
TARC-Comm-BR1-2013	TARC	Commander	1	95.9	94.1
TARC-Comm-BR2-2013	TARC	Commander	2	99.9	
TARC-Comm-BR3-2013	TARC	Commander	3	150.7	
TARC-Gaird-BR1-2013	TARC	Gairdner	1	109.3	
TARC-Gaird-BR2-2013	TARC	Gairdner	2	83.0	
TARC-Gaird-BR3-2013	TARC	Gairdner	3	79.5	
TARC-Hind-BR1-2013	TARC	Hindmarsh	1	71.3	
TARC-Hind-BR2-2013	TARC	Hindmarsh	2	71.0	
TARC-Hind-BR3-2013	TARC	Hindmarsh	3	86.6	
THH-Comm-BR1-2013	THH	Commander	1	102.6	97.4
THH-Comm-BR2-2013	THH	Commander	2	75.2	
THH-Comm-BR3-2013	THH	Commander	3	103.6	
THH-Gaird-BR1-2013	THH	Gairdner	1	133.6	
THH-Gaird-BR2-2013	THH	Gairdner	2	83.2	
THH-Gaird-BR3-2013	THH	Gairdner	3	86.0	
Br-Comm-BR1-2013	Breeza	Commander	1	147.7	103.1
Br-Comm-BR2-2013	Breeza	Commander	2	90.6	
Br-Comm-BR3-2013	Breeza	Commander	3	109.1	
Br-Gaird-BR1-2013	Breeza	Gairdner	1	77.7	
Br-Gaird-BR2-2013	Breeza	Gairdner	2	87.6	
Br-Gaird-BR3-2013	Breeza	Gairdner	3	105.6	

3.3.5. Data quality – matched and filtered data

Following discovery of the matched and filtered proteins, this data were checked for quality. Density plots and box plots of protein expression data for each of the nine samples were examined with the tenth serving as a pool for comparison, and the general uniformity of the results showed the data to be of good quality (Appendix B, Figure B.1).

3.3.6. Sample uniformity/diversity for matched and filtered protein samples

Despite the general uniformity of results, the nine individual samples from each of the three TMT sets did show some differentiation in the matched and filtered proteins from each sample. Analysis of the heat-maps, PCAs and correlation plots of the protein expression data for each sample revealed that not all of the sample replicates clustered into their sample groups (Table 3.6). The PCAs for the three TMT sets showed a large variation within same sample groups while displaying very tight clustering of points for other sample groups (Appendix B, Figure B.2,Figure B.3 and Figure B.4).

Table 3.6. Summary of data quality for matched and filtered barley grain proteins.

TMT set	Sample group *(Each containing 3 biological replicates)*	Complete **heat-map** sample group clustering (☑ or ☒)	Complete **PCA** sample group clustering (☑ or ☒)	Complete correlation heat-map sample group clustering (☑ or ☒)
1	TARC-Comm-2013	☒	☑ Very Poor	☒
1	TARC-Gaird-2013	☒	☑ Poor	☒
1	TARC-Hind-2013	☑	☑	☑
2	TARC-Comm-2013	☒	☑ Very Poor	☒
2	THH-Comm-2013	☒	☒	☑
2	Br-Comm-2013	☒	☒	☒
3	TARC-Gairdner-2013	☒	☒	☒
3	THH-Gairdner-2013	☑	☒	☑
3	Br-Gairdner-2013	☑	☑	☑

Original data for the above table has been taken from Figure B.2,Figure B.3, and Figure B.4 in Appendix C.
Note: Each sample group is made up of three biological replicates.

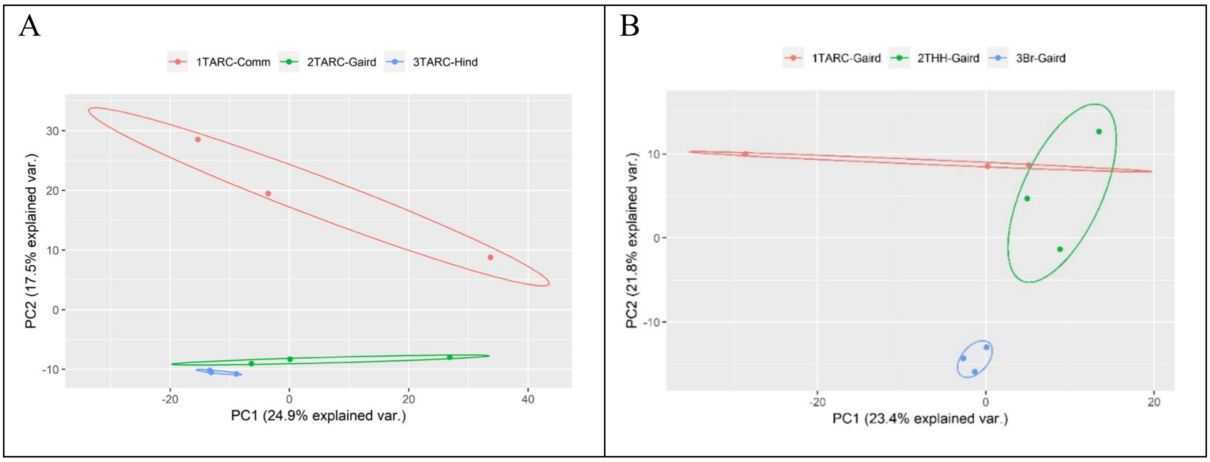


Figure 3.5. (A) PCA for TMT set-1 showing close plotting points for replicate samples from Hindmarsh barley grown at TARC farm, (B) PCA for TMT set-3 showing close plotting points for replicate samples from Gairdner barley grown at Breeza farm. The sample data is of matched and filtered proteins.

3.3.7. Sample clustering after putative biomarker discovery

The discovery of putative biomarker proteins was performed by 'TMTPRePro' (section 3.2.17). Looking at the expression data of putative biomarkers in each sample it was observed that there was great similarity between sample replicates, and little similarity between sample groups (compare Table 3.6 and Table 3.7). All heat-maps and correlation plots displayed clustering of samples into their groups (Table 3.6), and PCAs showed close plotting of replicate samples into their sample groups and good separation between these different groups (Appendix B, Figure B.5, Figure B**.6** and Figure B**.7**).

Table 3.7. Summary of data quality for putative biomarker discovery from samples.

TMT set	Sample group	Complete **heat-map** sample group clustering (☑ or ☒)	Complete **PCA** sample group clustering (☑ or ☒)	Complete correlation heat-map sample group clustering (☑ or ☒)
1	TARC-Comm-2013	☑	☑	☑
1	TARC-Gaird-2013	☑	☑	☑
1	TARC-Hind-2013	☑	☑	☑
2	TARC-Comm-2013	☑	☑	☑
2	THH-Comm-2013	☑	☑	☑
2	Br-Comm-2013	☑	☑	☑
3	TARC-Gairdner-2013	☑	☑	☑
3	THH-Gairdner-2013	☑	☑	☑
3	Br-Gairdner-2013	☑	☑	☑

Original data for the above table has been taken from Figure B.5, Figure B.6 and Figure B.7 in Appendix C.

Note: Each sample group is made up of three biological replicates.

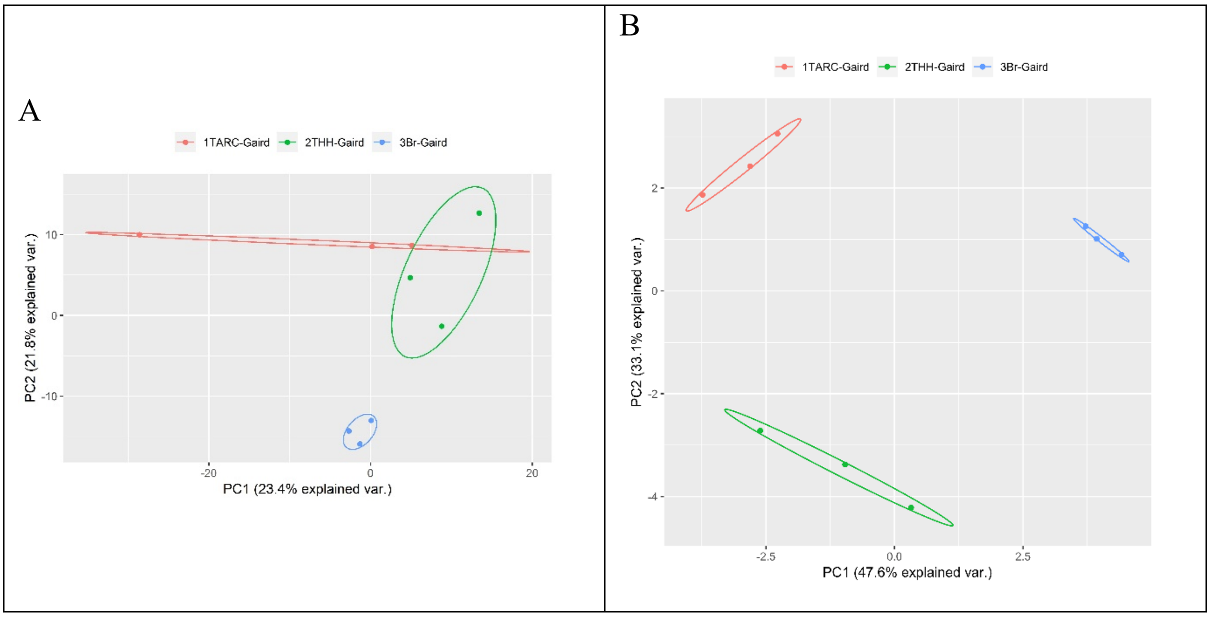


Figure 3.6. Example of the improvement in sample group clustering when sample expression data moves from (A) matched and filtered proteins to that of (B) putative biomarker proteins.

3.3.8. SCX fractionation

The TMT-labelled samples were run on a mass spectrometer prior to being fractionated through SCX. The number of proteins detected was 745 (TMT set-1), 789 (TMT set-2) and 745 (TMT set-3). In terms of matched and fractionated proteins, the addition of SCX fractionation almost doubled the number of proteins: 1,312 (TMT set-1), 1,200 (TMT set-2) and 1,314 (TMT set-3; Figure 8). Putative biomarker detection also increased with SCX fractionation: from 50 (TMT set-1), 26 (TMT set-2) and 18 (TMT set-3), to 80 (TMT set-1), 31 (TMT set-2) and 21 (TMT set-3; Table 3.9).

Table 3.8. Increase in numbers of matched and filtered proteins detected using SCX fractionation of samples.

Run type	Number of detected proteins		
	TMT set-1 Comparing proteomes across cultivars	TMT set-2 Comparing proteomes across farm locations	TMT set-3 Comparing proteomes across farm locations
TMT test runs (non-SCX fractionation of samples)	745	789	745
TMT full runs (SCX fractionation of samples)	1312	1200	1314

Table 3.9. Increased putative biomarker discovery using SCX fractionation.

Run type	Number of putative biomarkers		
	TMT set-1 Comparing proteomes across cultivars	TMT set-2 Comparing proteomes across farm locations	TMT set-3 Comparing proteomes across farm locations
TMT test runs (non-SCX fractionation of samples)	50	26	18
TMT full runs (SCX fractionation of samples)	80	31	21

3.3.9. Discovery of putative biomarkers

TMT set-1 revealed 80 putative biomarkers relating to proteomic differences between cultivars. The list of these putative biomarkers is displayed in Appendix, Table 3.11. TMT sets-2 and-3 revealed 31 and 21 putative biomarkers (respectively) relating to proteomic differences across farm locations (Appendix B, Table 3.12 and Table B.1). Only 4 putative biomarkers were found to be common in TMT sets-1 and-2 (two-homologous 22.0 kDa class IV heat shock proteins [M0UGW6 and M0Y7H8], two-homologues of Haegeman factor inhibitor proteins [M0ULY1 and M0YS73], with only 2 common putative biomarkers in TMT sets-2 and-3 (microtubule associated protein [F2CTI9]; and defensin D2 [F2EKF1]. There was only one putative biomarker common to TMT sets-1 and -3 (late embryogenesis abundant protein [M0VEJ0]). There were no putative biomarkers common to all of the TMT sets (TMT sets-1, -2 and -3).

Table 3.10. Number of putative biomarkers from the entire set of TMT-labelled barley samples.

Set	Study	Testing cultivar variation or location	Number of putative protein biomarkers
TMT set-1	Commander, Gairdner and Hindmarsh cultivars at TARC farm	Cultivar	80
TMT set-2	TARC, THH and Breeza farms growing Commander barley	Location	31
TMT set-3	TARC, THH and Breeza farms growing Gairdner barley	Location	21

3.3.10. Functionality of putative biomarker proteins

We investigated the putative biomarker proteins discovered from the three TMT sets. The Uniprot identifiers were input into 'GORetriever' and 'GOSlimViewer' at the 'AgBase' web site (http://agbase.arizona.edu/cgi-bin/tools/index.cgi) and a list of gene ontology summary functional categories (Slims) was obtained and counted. The three figures below (Figure 3.7; Appendix B, Figure B.8 and Figure B.**9**) graphically describe the Slims protein functionality results for the three lists of putative biomarkers.

The biomarker functionality of TMT set-1 is represented in Figure 3.7. Under the 'Biological process' gene ontology parent term, the functions of 'metabolic process' and 'cellular process' are particularly high, while for the parent term of 'molecular function' the number of proteins (biomarkers) counted to have the function 'binding' was up to 26. For this bar-graph the parent term of cellular component did not have any outstanding SLIM terms. TMT set-2 and TMT set-3 did not generate enough SLIM terms from their smaller list of proteins to show any functionality of interest (Appendix, Figure B.8 and Figure B.**9**)

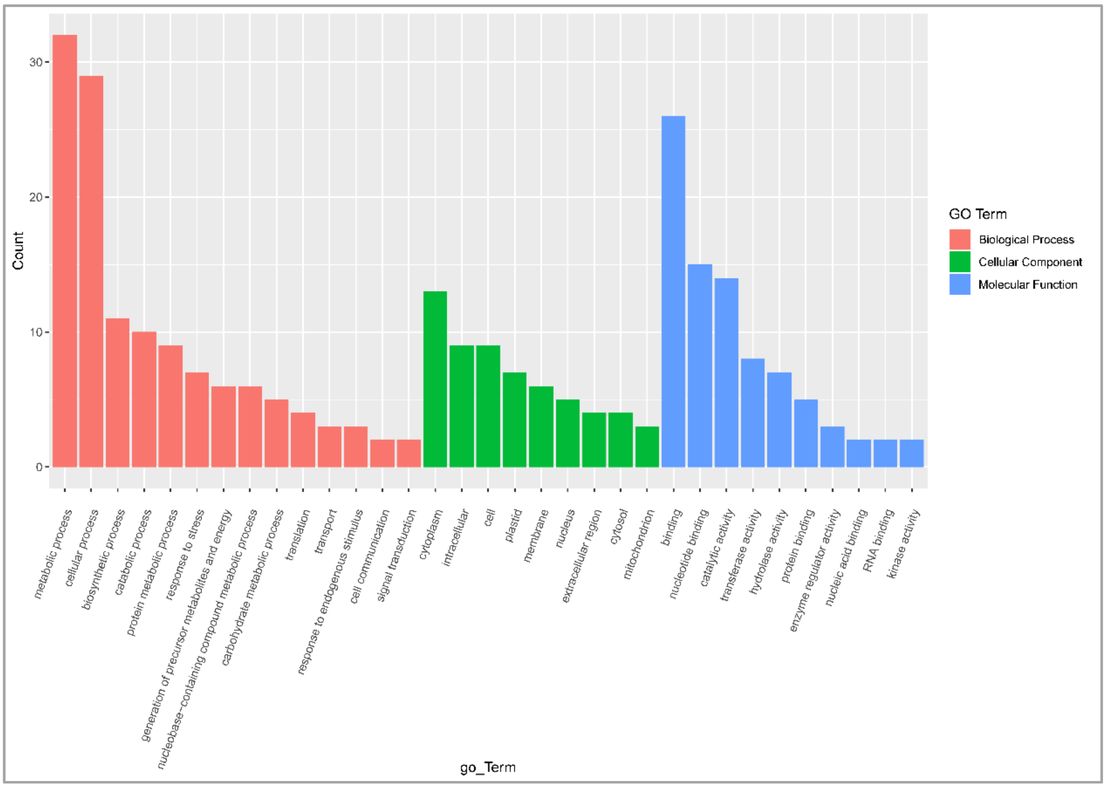


Figure 3.7. TMT set-1 GO-Slims functional summary of putative biomarkers in barley grain (proteome comparison across cultivars). Only GO Slims categories that are associated with 2 or more putative biomarker proteins are included.

3.4. Discussion

3.4.1. Overview

In this chapter, we applied TMT-labelled shotgun proteomics for the discovery of proteins that can be used as potential biomarkers for cultivar identification of farm origin and farm plot location. The proteomic analysis did reveal putative biomarkers. Many of the proteins identified as putative biomarkers were involved either directly or indirectly in stress response. Furthermore, GO Slims terms, or in other words functional summaries of the identified proteins also show that they have similar functional profiles whether we looked at proteomes comparing location (TMT sets-2 and -3) or cultivar (TMT set-1).

3.4.2. Barley grain sample variation

The variation in 200-grain weights between farms do not appear to match the differences in rainfall that each farm experienced during the growth of the crops. TARC experienced the highest total rainfall during this period, followed by THH and Breeza (Figure 3.8). When this data was compared with the 200-grain weight measurements of barley, there was an inverse trend in the pattern between the total water that each farm had received and the 200-grain weight (compare Figure 3.9 with Figure 3.8). A possible explanation is that drought stressed barley may have a higher growth rate than barley that is well-watered. The end result being complete grain fill, but each plant having less tillers, spikes and grains (H. Samarah, 2005). As such, the barley from the TARC and THH farms may have had larger total yields (this data was not supplied with the samples) with slightly less grain fill, resulting in the lower 200-grain weight for the Breeza crop. It appears that the amount of photosynthetically active radiation (PAR) didn't have an impact on either grain weight or the amount of protein extracted from seed samples. Data from the Australian Bureau of Meteorology (http://www.bom.gov.au/jsp/ncc/climate_averages/solar-exposure/index.jsp) showed that for the growing season of 2013, the highest and lowest average readings of solar radiation had a difference of only a 6% that was even lower (5%) for the expected period of seed development. Moreover, in terms of highest to lowest solar radiation the location order is: TARC, Breza, and THH; which is different to the grain protein yield for location (highest to lowest: TARC, THH, Breeza) and seed weight (highest to lowest: Breeza, TARC, THH).

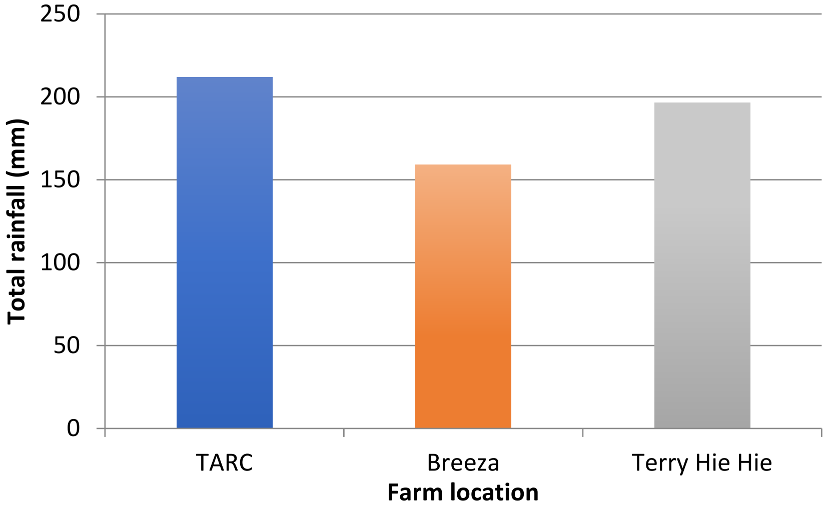


Figure 3.8. The average total rainfall during the crop growth period (May to October; 2013) obtained for the three different farm locations: TARC, Breeza and Terry Hie Hie.

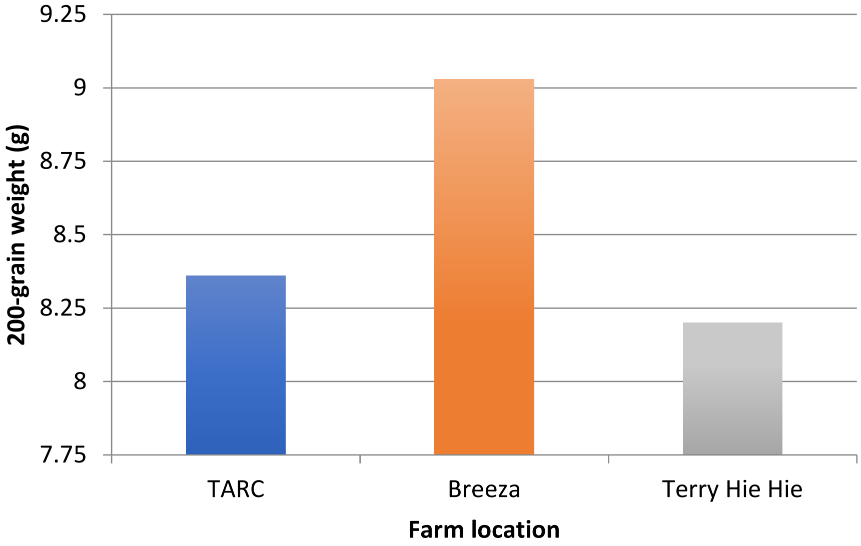


Figure 3.9. The average 200-grain weights of barley grain obtained from three different farm locations: TARC, Breeza and Terry Hie Hie.

The protein extract levels shown in Figure 3.10 displays an observable trending pattern with the total rainfall. Although the 200-grain weight was the highest for Breeza, the lower 200-grain weights for TARC and THH yielded more protein. This is likely due to the larger proportion of the grain (the starch) being overrepresented and the proteins from the aleurone and germ cells being underrepresented in the protein extract. The three locations were involved in nitrogen trials so soil nitrogen would have been equivalent at all locations for all cultivars and not a factor in differing protein levels.

There was also no observable correlation between grain fill and average monthly temperatures during this growth period (Figure 3.11). This is likely due to the average monthly temperature range difference being only 2.4ºC.

The proteins identified in barley from the TMT sets-2 and-3 (farm comparison) revealed proteins associated with drought stress, such as 20 kDa chaperonin, 10 kDa chaperonin, defensin D2, class IV heat shock protein, late embryogenesis abundant protein (Ge, et al., 2012, Stotz, et al., 2009, Wendelboe-Nelson and Morris, 2012). Since this study does not pinpoint which farm the differences in protein expression are from, it can only be speculated that the proteins are likely associated with samples obtained from the Breeza farm as it has experienced less rain (Figure 3.8).

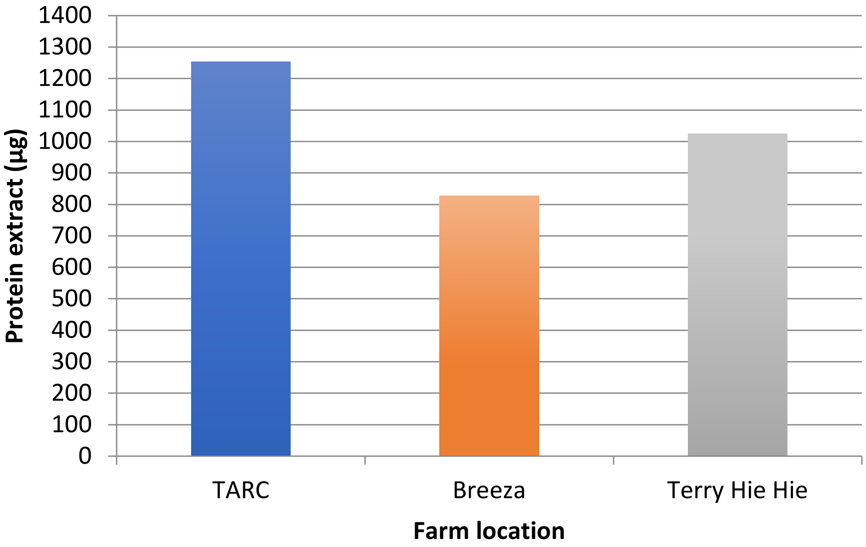


Figure 3.10. The average protein extract from barley grown at three different farms: TARC, Breeza and THH.

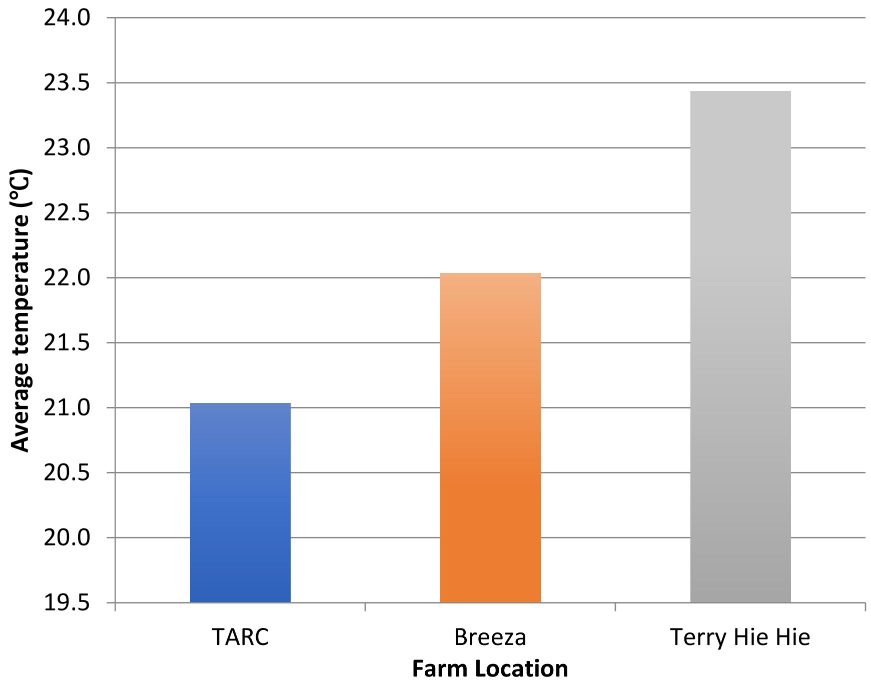


Figure 3.11. The average temperatures correlate with increasing latitude. TARC being the furthest south and THH being the furthest north.

3.4.3. Variability of Proteomes

Researchers have shown that seed proteomes can vary due to abiotic and biotic stress during seed development This is also observed in this study in the matched and filtered proteins detected in barley grain samples (Appendix B, Figure B.2, Figure B.3 and Figure B.4). Even though these proteins have been put through some basic data processing via Proteome Discoverer and the Mascot database, the sample does not always cluster into their sample groups. The PCAs, heat-maps and correlation plots often show one or two samples that have quite different expression data profiles than others in their sample group. Normally, this difference is eliminated with the discovery of putative biomarker proteins; however, in test runs following the same protocol as above, if the initial matched and filtered data is of poor quality characterized by high variability within sample groups, the number of putative biomarkers discovered is reduced. This was observed in a number of test runs with poor results (such as high intra-sample group variation), as well as for the barley TMTs being described in this chapter. This will be elaborated further in chapter 4.

3.4.4. Analysis of differentially expressed proteomes

After the matched and filtered proteins had been examined, this data was applied to the 'TMTPrePro' package, and putative biomarkers were discovered in all three TMT sets. The putative biomarkers found in each sample resulted in very tight clustering of samples into their sample groups. This observation indicates good sample quality control (QC) and gives weight to the assumption of biotic and abiotic stress changing the grain proteome.

For TMT set-2 and -3, because each set is looking at only one cultivar across multiple sites, theoretically biotic and abiotic stress make stress-response the only variable within these two sets. For example, if one cultivar is grown at Breeza, TARC and THH the greatest variation in grain proteins should be observed at the location where conditions (biotic and abiotic) are most challenging. Indeed while within-group sample data (biological replicates) should be similar the data between groups should be different for the proteins involved in stress response; as seen PCA, correlation plots and heat-map analyses (Appendix B, Figure B.5, Figure B**.6** and Figure B**.7**).

3.4.5. Putative biomarkers

3.4.5.1. Biomarker results

The analysis of each TMT-set resulted in the discovery of a number of putative biomarkers. The differences between cultivars (Commander, Gairdner, and Hindmarsh) in TMT set-1 were examined – which were all grown at the TARC farm. For this set, 80 putative biomarkers were discovered.

The differences between locations were examined in TMT sets-2 and -3. TMT set-2 compared Commander barley grown at three different farm locations (TARC, THH, and Breeza), and TMT set-3 compared Gairdner barley at these same locations. The results of this analysis identified 31 and 22 putative biomarkers respectively.

There was minimal overlap between the groups of proteins in the three experiments. The lack of significant overlap in the comparisons of cultivar differences with location differences is to be expected, as we would not necessarily expect the same proteins to differ between cultivars and locations. Considering there was only two overlapping proteins identified between cultivars in the location testing, comparing TMT sets-2 and -3, this suggests that any prospective provenance testing will likely be limited to individual cultivars as there is insufficient identification of proteins across all cultivars that could provide a protein profile for provenance that would work with all cultivars.

The proteomic differences between cultivars is unsurprising given that cultivars can be distinguished from each other using genetic testing. However, it is still somewhat surprising that 80 proteins were identified as being potentially able to distinguish between these cultivars. Further sampling in other growing seasons would help clarify if all 80 of these protein biomarkers can be consistently and robustly used for cultivar differentiation. Moreover, repeating these experiments over one or more seasons would also verify whether protein biomarkers exist that could consistently be used to determine the farm origin of the grain.

3.4.5.2. Stress response proteins

In TMT set-1, 80 putative protein biomarkers were discovered, of which 27 proteins are directly related to stress response (Table 3.11). For TMT set-2, there were 31 putative protein biomarkers discovered, of which seven are related to stress response (Table 3.12). Finally, there were 22 putative protein biomarkers discovered in TMT set-3, seven of these are known to have a stress-response role (Table 3.13).

Table 3.11. TMT set-1: List of putative biomarker proteins from barley grain.

Barley fasta ID	Uniprot	Protein Description
MLOC_5173.1	C3W8L1	Glucose-1-phosphate adenylyltransferase
AK375106	F2EG29	Late embryogenesis abundant protein
MLOC_6278.1	A0A0B4J2W8	Late embryogenesis abundant protein 1b
MLOC_24874.2	M0VEJ0	Late embryogenesis abundant protein
AK370848	F2E3X4	Proteasome subunit alpha type
AK374275	F2EDQ0	Gibberellin receptor GID1L2
AK363153	F2DGY6	Acyl CoA binding domain containing protein 6
MLOC_44325.2	F2CQQ1	2-dehydro-3-deoxyphosphooctonate aldolase putative expressed
AK367326	F2DTV8	Nuclear transcription factor Y subunit C 2
AK364296	F2DK78	Peroxidase 8
AK355507	F2CV55	Peroxidase 4
MLOC_5958.2	F2CWU8	Eukaryotic translation initiation factor 2 beta subunit putative expressed
AK364831	F2DLR3	Proteasome activator subunit 4 like
MLOC_21677.1	F2DXR4	Ripening related protein
MLOC_57898.2	F2EAZ9	Early nodulin like protein
MLOC_7780.1	M0Z6C2	Malate dehydrogenase
MLOC_2337.1	M0VDB7	17.4 kDa class I heat shock protein 3
MLOC_53175.2	M0WNE6	5-methyltetrahydropteroyltriglutamate homocysteine methyltransferase putative expressed
MLOC_61465.1	M0XMW5	Acyl [acyl carrier protein] desaturase
MLOC_78140.1	M0Z714	Serpin 2
MLOC_10834.2	M0UGW6	22.0 kDa class IV heat shock protein
AK248736.1	P20115	Citrate synthase
MLOC_79770.1	M0Z9Q3	Serpin 2
AK252829.1	Q9FXT9	26S protease regulatory subunit putative
MLOC_67715.1	M0YBZ9	Acidic endochitinase
MLOC_66477.1	M0Y7H8	22.0 kDa class IV heat shock protein
MLOC_67147.1	M0YA06	VIP1 protein
MLOC_81761.1	Q5URW6	Puroindoline B
MLOC_38476.1	M0VYA0	Non specific lipid transfer protein
AK373733	F2EC58	Cysteine proteinase inhibitor Cystatin
AK375664	F2EHN7	Late embryogenesis abundant protein
AK249268.1	B8B9K6	Ribosomal protein
MLOC_68101.1	F2D4L0	Glutathione transferase F5
AK374553	F2EEH7	Bifunctional inhibitor lipid transfer protein seed storage 2S albumin like protein
AK367973	F2DVQ5	Diacylglycerol kinase like protein
MLOC_4683.1	F2DML5	Peroxidase 12
AK362311	F2DEJ4	Eyes absent like protein
AK362492	F2DF25	Heat shock 70 kDa protein 5
MLOC_17045.1	M0V1Y1	F box WD 40 repeat containing protein
MLOC_46003.1	M0WA02	Low molecular weight glutenin subunit

Note: Rows highlighted in cream are proteins involved in stress response.

(Table continued next page)

(Continued from previous page [Table 3.11])

MLOC_12143.1	M0ULY1	Hageman factor inhibitor
MLOC_81977.1	M0ZD98	Elongation factor Tu
AK248705.1	P12782	Phosphoglycerate kinase
MLOC_72278.1	M0YS73	Hageman factor inhibitor
MLOC_73077.1	M0YUE3	Beta 1,3-glucanase 2
MLOC_21198.1	M0VAF1	Grain softness protein
MLOC_55594.1	M0WY96	Malonyl CoA acyl carrier protein transacylase containing protein expressed
AK252468.1	Q0JPT4	BES1 BZR1 homolog 4 LENGTH 325
AK250641.1	P36428	Alanine trna ligase
MLOC_10176.1	M0UEE6	Serpin 2
MLOC_34492.1	F2DW24	OTU domain containing protein 6B
AK355136	F2CU34	Heat shock protein 90
MLOC_73966.1	Q9FUM8	Homocysteine S methyltransferase 3
AK364178	F2D009	ATP dependent RNA helicase putative
AK372029	F2CVV8	Superoxide dismutase
AK368827	F2DY59	Chaperone protein htpG family protein LENGTH 780
AK376513	F2EK36	Chitinase
MLOC_12224.1	F2DR19	Histidinol dehydrogenase
AK358224	F2D2W7	Unknown protein
MLOC_5618.2	M0X173	Heat shock protein 90
MLOC_14295.2	M0UUA7	Late embryogenesis abundant protein
MLOC_5404.1	M0WRH5	Aldose 1 epimerase
AK250175.1	P08927	60 kDa chaperonin 2
MLOC_2934.1	M0VI12	NAD-dependent epimerase dehydratase
MLOC_10218.1	M0UEJ7	Aldehyde dehydrogenase putative
MLOC_8329.1	M0ZDY0	Defensin
AK252683.1	A5D8P8	UPF0510 protein INM02
MLOC_71318.1	M0YP02	DNA repair helicase rad5 16 putative
MLOC_70189.1	M0YKC6	Expansin protein
MLOC_17935.1	M0V3Y6	Citrate binding protein putative expressed
AK252423.1	Q9FMA3	Peroxisomal targeting signal 1 receptor
MLOC_61812.1	M0XP66	Pectin lyase like superfamily protein LENGTH 476
MLOC_52601.1	M0WL85	Zinc finger CCCH domain containing protein
AK365300	F2DN32	Transmembrane protein 87A
MLOC_63625.1	M0XW86	purple acid phosphatase 27 LENGTH 611
AK252610.1	Q4R347	Cell differentiation protein rcd1 putative expressed
AK252728.1	Q6Z351	Aldehyde oxidase
MLOC_25678.1	M0VF46	Cysteine proteinases superfamily protein LENGTH 361
MLOC_66422.2	M0Y7B9	WD repeat containing protein putative
MLOC_59653.1	M0XFS5	Gibberellin receptor GID1L2

Note: Rows highlighted in cream are proteins involved in stress response.

Table 3.12. TMT set-2: List of putative biomarkers proteins from barley grain and their functions.

Barley fasta ID	Uniprot	Protein Description
MLOC_56924.1	M0X4B4	Jasmonate induced protein
MLOC_5168.1	Q4VM11	Beta amylase
AK361994	F2DDM7	Tryptophan tRNA ligase
MLOC_77043.1	M0Z4S0	11S seed storage globulin
AK252834	A5BUU4	40S ribosomal protein SA
MLOC_38209.1	M0VXI9	RNA exonuclease
MLOC_43717.1	M0W4U0	S phase cyclin A associated protein in the endoplasmic reticulum
MLOC_15911.1	F2CPQ3	Unknown protein
MLOC_22184.1	F2CTI9	Microtubule associated protein family protein putative expressed
AK369479	F2E008	Aldose 1 epimerase
MLOC_6055.1	M0XJ70	Ubiquitin conjugating enzyme variant
AK366209	D5KWD4	starch synthase 2 LENGTH 792
MLOC_11338.3	M0UIV1	Nup98 protein
AK353926	F2CQM4	Cysteine proteinase
MLOC_4753.2	F2CRD9	Late embryogenesis abundant protein
MLOC_13881.1	F2CSZ1	Unknown protein
AK371185	F2E4W1	Magnesium and cobalt efflux protein corC putative
MLOC_71136.2	F2E7V8	Cinnamyl alcohol dehydrogenase
AK376628	F2EKF1	Defensin D2
MLOC_12143.1	M0ULY1	Hageman factor inhibitor
AK250295	Q43767	Non-specific lipid transfer protein
MLOC_72278.1	M0YS73	Hageman factor inhibitor
AK249385	Q07661	Nucleoside diphosphate kinase
AK252139	Q9M330	Proteasome inhibitor like protein
AK365973	F2CXV7	20 kDa chaperonin
AK364344	F2DKC6	Unknown protein
AK369476	F2E005	Unknown protein
MLOC_36316.1	F2DQP5	Annexin 2
MLOC_10834.2	M0UGW6	22.0 kDa class IV heat shock protein
AK248808	P93026	Vacuolar sorting receptor 1 putative
MLOC_66477.1	M0Y7H8	22.0 kDa class IV heat shock protein

Note: Rows highlighted in cream are proteins involved in stress response.

Table 3.13. TMT set-3: List of putative biomarkers proteins from barley grain, their functions, and a protein subset involved in stress response (cream colour).

Uniprot Ids	Protein description
M0VEJ0	Late embryogenesis abundant protein
F2DAA1	PEBP family protein
F2EG62	Succinate dehydrogenase iron sulfur subunit
F2D961	Lipoxygenase
M0VUI3	Lipoxygenase
F2CVY7	Phosphorylase
M0W7R3	Nucleolar protein 5
Q40004	Ribulose bisphosphate carboxylase small chain
F2CTD8	Defensin
P34893	10 kDa chaperonin
F2DMW8	UPF0061 protein
F2EKF1	Defensin D2 (Defence response)
M0YRS3	UDP glycosyltransferase
M0WDU5	Amine oxidase
F2CTI9	Microtubule associated protein family protein putative expressed
F2E6F3	Coatomer subunit gamma
F2CUZ5	Glycogen synthase
M0WCD7	rRNA N glycosidase
M0W6F2	Acidic endochitinase
M0X060	Adenine nucleotide alpha hydrolases like protein
M0UFT6	Inosine 5' monophosphate dehydrogenase
M0W433	Carbonic anhydrase

Note: Rows highlighted in cream are proteins involved in stress response.

3.4.6. Putative biomarkers and Slims

Functional summaries of all the biomarker proteins from the three TMT sets, in the form of GO Slims, are summarised in (Figure 3.7; Appendix B, Figure B.9 and Figure B.9). The bar graph shows that for the parent GO category of 'biological process' the most commonly associated functions were 'metabolic process' followed by 'cellular process'. For the parent category of 'molecular function' the most common function was 'binding'. For all three GO Slims, the counts were noticeably higher compared to the remaining GO Slims.

The results presented in this chapter reveal that labelled shotgun proteomics can potentially deliver a set of putative biomarkers that would be able to identify the farm of origin and cultivar of barley grain. In all three TMT sets tested, a number of putative biomarker proteins have been discovered, and many of these are involved either directly or indirectly with stress-response. Slims that are most represented by these putative biomarkers are those where stress response proteins are likely to reside such as 'metabolic process'. Moreover, these biomarkers could be candidates for an immune-based test such as an ELISA that use a panel of these proteins to deliver a fast and inexpensive test for the identification of the origin and cultivar of the grain.

Chapter 4. Proteomic analysis of wheat grain for the discovery of putative protein biomarkers to identify cultivar and farm origin

4.1. Introduction

The wheat industry is a substantial contributor to the Australian economy. In the years 2016-2017, the value of export wheat sales was a little over $6 billion AUD (http://www.agriculture.gov.au/abares/research-topics/agricultural-commodities/sept-2018/wheat). The wheat industry needs to adapt to the expectations of producers of wheat-based products to meet consumer demands in providing quality assurance procedures that track provenance and grain quality. The difficulties in providing quality assurance is that there are no advanced laboratory-based tests to determine the farm origin of wheat grain (or any other cereals). Such quality assurance parameters may include: the 'general health' of the grain, such as poor quality due to damage caused by rain at harvest, drought or pathogenic attack. Hence, the variable nature of the wheat proteome due to influences of biotic and abiotic stress in combination with natural genetic variation is anticipated to provide enough variability for the identification of proteins that can act as biomarkers to address these issues.

In this chapter, it is hypothesised that TMT-labelled proteomic analysis can be applied to wheat grain for the discovery of putative biomarkers. Similar to Chapter 3, I aim to discover potential biomarkers for identifying the farm location or cultivar of wheat grain using high throughput proteomic analysis of grain samples of three different wheat cultivars that were grown in three different farm locations.

4.2. Methods

4.2.1. Wheat grain sample details and sample processing

Tandem-Mass-Tags (TMTs) are utilized for the proteomic analysis in the identification and quantitation of putative biomarkers. The grain samples consisted of the three wheat cultivars: Gregory, Livingston, and Spitfire. Spitfire grain is classified as Australian Prime Hard (APH), making it a high protein milling wheat. It has an early to mid-maturity with early seedling vigour, is high yielding, and has a large grain size and high protein accumulation. Gregory is also ranked as APH with mid-season sowings and has a medium to slow maturity. Livingston is classified as an Australian Hard (AH) wheat with high to medium levels of protein. It has a high temperature tolerance and is an early maturing main season wheat in New South Wales. All three wheat cultivars were sampled using three biological replicates. Hence each replicate is from one of three different plots on the farm and three different farm locations (Breeza, TARC, and Terry Hie Hie [THH]).

The wheat was sown on the 8th of May 2013 and did not germinate until the 23rd of May. The crop was harvested around early to mid-November (exact date unknown). The crops were not watered (only from rainfall) and were treated with fungicides (details obtained from Rohan Brill from the Department of Primary Industries, NSW Government).

For location coordinates and weather observations, see 3.2.1

4.2.2. Experiment design

The proteomic analysis of the grain samples was divided into four separate TMT sets. Each TMT set consisted of ten TMT's. All of the following TMT sets include three biological replicates of each cultivar, and likewise each farm location. The experimental configuration for each TMT set and TMT-identification number allocation is outlined in Table 4.1. The sample comparisons are described as follows:

TMT set-1 – The Gregory cultivar was compared between three different farm locations (Breeza, TARC, and THH).

TMT set-2 – The Spitfire cultivar was compared between three different farm locations (Breeza, TARC, and THH).

TMT set-3 – The Livingston cultivar was compared between three different farm locations (Breeza, TARC, and THH).

TMT set-4 – The proteomes of the three different cultivars grown in the same location, were compared with three biological replicates of each cultivar (Gregory, Spitfire or Livingston) from sets-1 to -3 pooled. This pool represented the particular cultivar for that particular farm location. These pooled samples were then used to compare the different cultivars for each farm location. For example, the pooled samples Gregory, Spitfire and Livingston that were all grown in the Breeza farm location were each compared. This was repeated for the two other farm locations (Table 4.1).

Table 4.1. TMT allocations for proteomics analysis of samples.

TMT label No.	**TMT Set 1**	**TMT Set 2**	**TMT Set 3**	**TMT Set 4**
1 (126)	Gregory-Breeza-BR1-2013	Spit-Breeza-BR1-2013	Liv-Breeza-BR1-2013	Greg-Breeza-Pool
2 (127N)	Greg-Breeza-BR2-2013	Spit-Breeza-BR2-2013	Liv-Breeza-BR2-2013	Greg-TARC-Pool
3 (127C)	Greg-Breeza-BR3-2013	Spit-Breeza-BR3-2013	Liv-Breeza-BR3-2013	Greg-THH-Pool
4 (128N)	Greg-TARC-BR1-2013	Spit-TARC-BR1-2013	Liv-TARC-BR1-2013	Spit-Breeza-Pool
5 (128C)	Greg-TARC-BR2-2013	Spit-TARC-BR2-2013	Liv-TARC-BR2-2013	Spit-TARC-Pool
6 (129N)	Greg-TARC-BR3-2013	Spit-TARC-BR3-2013	Liv-TARC-BR3-2013	Spit-THH-Pool
7 (129C)	Greg-THH-BR1-2013	Spit-THH-BR1-2013	Liv-THH-BR1-2013	Liv-Breeza-Pool
8 (130N)	Greg-THH-BR2-2013	Spit-THH-BR2-2013	Liv-THH-BR2-2013	Liv-TARC-Pool
9 (130C)	Greg-THH-BR3-2013	Spit-THH-BR3-2013	Liv-THH-BR3-2013	Liv-THH-Pool
10 (131)	*Greg-THH-Pool-2013*	*Spit-THH-Pool-2013*	*Liv-THH-Pool-2013*	*Spit-All_Farms-Pool-2013*

Note: BR = Biological replicate, TR = Technical replicate; Greg = Gregory, Spit = Spitfire, Liv = Livingston; Breeza = Breeza farm, TARC = TARC farm, THH = Terry Hie Hie farm.

4.2.3. Wheat grain sample details and sample processing

The samples for this experiment were wheat grain from the Gregory, Livingston and Spitfire cultivars that were all grown in three different plots at the three different farms (Breeza, TARC, and Terry Hie Hie [THH]). The farm locations are separated by hundreds of kilometres (Breeza to Terry Hie Hie, 240 km; Terry Hie Hie to TARC, 406 km; and TARC to Breeza, 324 km) and the distance between the three plots on the farm was not provided by the farm research facilities.

4.2.4. Weighing and grinding samples

The weighing and grinding of samples was performed as described in Chapter 2 (Methods section).

4.2.5. TCA/acetone/phenol (‘combined-phenol’) protein extraction

The starting amount of fine ground grain was 100 mg per sample. The processing of grain samples was performed as described in Chapter 2 (Methods section) using the ‘combined-phenol’ extraction method, with some modifications (described below) to prepare the peptides for TMT-labelling.

4.2.6. Reduction and Alkylation

Samples were processed as described in Chapter 3 (Methods section).

4.2.7. Methanol/chloroform precipitation

Samples were processed as described in Chapter 3 (Methods section).

4.2.8. Bradford Assay

Sample protein concentration was determined as described in Chapter 2 (Methods section).

4.2.9. Percent protein calculation

The percent protein was determined as described in Chapter 2 (Methods section).

4.2.10. Protein digestion

Samples were processed as described in Chapter 3 (Methods section).

4.2.11. SDS-PAGE

The method for examining protein extraction quality using SDS-PAGE is described in Chapter 2 (Methods section).

4.2.12. Construct laboratory-made SDB-RPS tips for sample de-salting (stage tipping)

Samples were processed as described in Chapter 3 (Methods section).

4.2.13. SDB-RPS De-Salting (stage tipping)

Samples were processed as described in Chapter 3 (Methods section).

4.2.14. Micro-BCA assay of desalted peptide samples for TMT labelling

Samples were processed as described in Chapter 3 (Methods section).

4.2.15. Sample preparation for TMTs

Samples were processed as described in Chapter 3 (Methods section).

4.2.16. Preparing a test run of TMT labelled sample sets for Mass Spectrometry

Samples were processed as described in Chapter 3 (Methods section).

4.2.17. SCX fractionation

Samples were processed as described in Chapter 3 (Methods section).

4.2.18. Mass Spectrometry for TMT labelled samples

Samples were processed as described in Chapter 3 (Methods section).

4.2.19. Making a multi-FASTA file for use in proteomics searches

A protein (FASTA) database was constructed for barley to analyse the peptide mass spectra following mass spectrometry. Firstly, an NCBI protein search was performed on wheat (https://www.ncbi.nlm.nih.gov/protein/?term=txid4565), then the complete result downloaded as a FASTA file. Any poorly annotated and partial sequences were removed by the program CD-Hit, which was downloaded from https://github.com/weizhongli/cdhit/. The command line instructions were as follows: "cd-hit –i sequence.fasta -o nr100 -c 1.00 -n 5 -d 120 -l 50". The input filename is "sequence.fasta", and the output is filename is "nr100.fa". The sequence identity threshold was set to 1.0, word-length to 5, with a length of description set to 120, and all sequences less than 50 amino acids discarded.

4.2.20. Protein identification and quantitation

The MS/MS spectra were searched against the appropriate custom wheat FASTA database, which was constructed as described in section 4.2.19. Otherwise, the samples were identified and processed as described in 3.2.16.

4.2.21. Identification of significant proteins using TMTPrePro

Samples were processed as described in 3.2.17.

Table 4.2. Example design spreadsheet for TMTPrePro.

“xlsx” file - TAB 1 (Design)		
Label	**Replicate**	**Group**
126	Greg-Br-R1	1Greg-Br
127_N	Greg-Br-R2	1Greg-Br
127_C	Greg-Br-R3	1Greg-Br
128_N	Greg-TARC-R1	2Greg-TARC
128_C	Greg-TARC-R2	2Greg-TARC
129_N	Greg-TARC-R3	2Greg-TARC
129_C	Greg-THH-R1	3Greg-THH
130_N	Greg-THH-R2	3Greg-THH
130_C	Greg-THH-R3	3Greg-THH
131	Pool	4Pool-S7-9

“xlsx” file -TAB 2 (References)	
File	**UseReference**
160826_PW_06_proteinGroups_proteingroups.txt	131

Note: The spreadsheet contains two tabs as represented by the table above. Tab 1 is title “Design”, and tab 2 is titled “References”.

4.2.22. Custom R-script: PCA.R

A FASTA database (file) of wheat proteins was used; otherwise, the samples were processed as described in 3.2.18.

4.3. Results

4.3.1. Sample weights

The 200-grain weight of the samples ranged from a minimum weight of 6.50 g for Spitfire grain grown at Breeza farm to a maximum weight of 8.54 g for Spitfire grain grown at the TARC farm. The farm that had the lowest average 200-grain weight (includes all cultivars) was Breeza (6.65 g); the next highest was THH (7.46 g); and the highest was TARC (7.86 g; Table 4.3). Looking at differences in 200-weight by either location or cultivar a statistically significant difference was seen only between the TARC and Breeza farms, and none for any cultivar comparisons (Appendix C, Table C.1.3, Table C.1.4, and Table C.1.5). When this data was further divided into the three locations, and at each a comparison made between the three cultivars, only THH gown samples showed statistically significant difference in 200-weight between all three cultivars (Appendix C, Table C.1.3, Table C.1.4, and Table C.1.5).

Table 4.3. Averages of 200-grain weight measurements.

Sample code	Farm location	Cultivar	Average 200-grain weight (g)	Farm average 200-grain weight (g)
Br-Greg-2013	Breeza	Gregory	6.86	6.65
Br-Liv-2013	Breeza	Livingston	6.60	
Br-Spit-2013	Breeza	Spitfire	6.50	
TARC-Greg-2013	TARC	Gregory	7.29	7.86
TARC-Liv-2013	TARC	Livingston	7.75	
TARC-Spit-2013	TARC	Spitfire	8.54	
THH-Greg-2013	Terry Hie Hie	Gregory	6.55	7.46
THH-Liv-2013	Terry Hie Hie	Livingston	7.51	
THH-Spit-2013	Terry Hie Hie	Spitfire	8.33	

4.3.2. Protein extraction and SDS-PAGE

The average amount of protein extracted between all of the wheat samples was 1232 µg. The protein extract ranged from minimum of 945 µg (Gregory cultivar, THH farm) to a maximum of 1,679 µg (Spitfire cultivar, Breeza farm). The protein extract from the THH farm had the lowest average of 1,099 µg, TARC farm had an average of 1,211 µg, and the Breeza farm had the highest average of 1,386 μg (Table 4.4). Protein extraction quality was qualitatively examined by SDS-PAGE of the Gregory cultivar grown at TARC, which displayed good protein banding (Figure 4.1).

Examining the weight of sample proteins recovered from the three cultivars as well as the three cultivars at which they were grown (Table 4.4), statistical significance was only observed between the TARC and Breeza farms (Appendix C, Table C.1b.1, Table C.1b.2). Further dividing results the into the three locations and then comparing cultivar results at each location, the Livingston-Gregory and Spitfire-Gregory comparisons at Breeza farm showed statistical significance (Appendix C, Table C.1b.3, Table C.1b.4, and Table C.1b.5).

Table 4.4. **Amount of wheat protein extract recovered from the 'combined-phenol' method.**

Sample ID	**Farm location**	**Cultivar**	**Biological replicate**	**Total amount of sample proteins recovered (µg)**	**Average amount of sample proteins recovered (µg)**
Br-Greg-2013-BR1	Breeza	Gregory	1	1145	1,386
Br-Greg-2013-BR1	Breeza	Gregory	2	1231	
Br-Greg-2013-BR1	Breeza	Gregory	3	1092	
Br-Spit-2013-BR1	Breeza	Spitfire	1	1428	
Br-Spit-2013-BR1	Breeza	Spitfire	2	1665	
Br-Spit-2013-BR1	Breeza	Spitfire	3	1679	
Br-Liv-2013-BR1	Breeza	Livingston	1	1412	
Br-Liv-2013-BR1	Breeza	Livingston	2	1354	
Br-Liv-2013-BR1	Breeza	Livingston	3	1469	
TARC-Greg-2013-BR1	TARC	Gregory	1	1238	1,211
TARC-Greg-2013-BR1	TARC	Gregory	2	1145	
TARC-Greg-2013-BR1	TARC	Gregory	3	997	
TARC-Spit-2013-BR1	TARC	Spitfire	1	1433	
TARC-Spit-2013-BR1	TARC	Spitfire	2	1329	
TARC-Spit-2013-BR1	TARC	Spitfire	3	1290	
TARC-Liv-2013-BR1	TARC	Livingston	1	1128	
TARC-Liv-2013-BR1	TARC	Livingston	2	1101	
TARC-Liv-2013-BR1	TARC	Livingston	3	1234	
THH-Spit-2013-BR1	THH	Spitfire	1	1071	1,099
THH-Spit-2013-BR1	THH	Spitfire	2	1216	
THH-Spit-2013-BR1	THH	Spitfire	3	1149	
THH-Greg-2013-BR1	THH	Gregory	1	1053	
THH-Greg-2013-BR1	THH	Gregory	2	945	
THH-Greg-2013-BR1	THH	Gregory	3	1145	
THH-Liv-2013-BR1	THH	Livingston	1	1172	
THH-Liv-2013-BR1	THH	Livingston	2	1056	
THH-Liv-2013-BR1	THH	Livingston	3	1083	

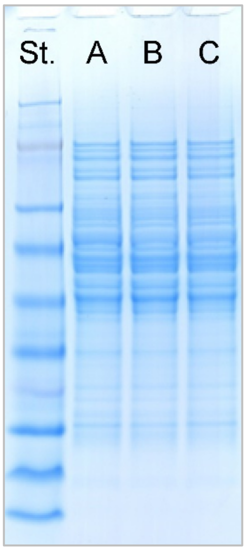


Figure 4.1. Example of 'combined-phenol' extracted proteins run on an SDS-PAGE gel, from three replicates (A, B and C) of Gregory wheat grain grown at TARC.

4.3.3. Protein digestion and SDS-PAGE

Enzymatic digestion of proteins was performed successfully as indicated by the qualitative SDS-PAGE gel in Figure 4.2. There was no protein banding in the peptide lanes of the gel for all samples of the digested proteins.

Quantitation of sample peptides was performed by Micro-BCA assay. The average amount of peptides from all of the wheat samples was 45.44 µg. The peptide concentrations between all samples ranged from minimum of 32.79 µg (from THH farm) to a maximum of 59.86 µg (from THH farm). The THH farm had the lowest average of 43.2 µg, TARC farm had an average of 44.1 µg, and Breeza farm had the highest average of 49.02 µg (Table 4.5).

For total peptide weight (Table 4.5), if the results were divided into three groups based on location, statistically significant differences between cultivars were seen at Breeza and THH (Appendix C, Table C.1b.6, Table C.1b.7, and Table C.1b.8). At Breeza this was only seen in the Livingston-Gregory and Spitfire-Gregory comparisons (Appendix C, Table C.1b.6), while at THH statistical significance was only seen in the Spitfire-Gregory and Spitfire-Livingston comparisons (Appendix C, Table C.1b.8).

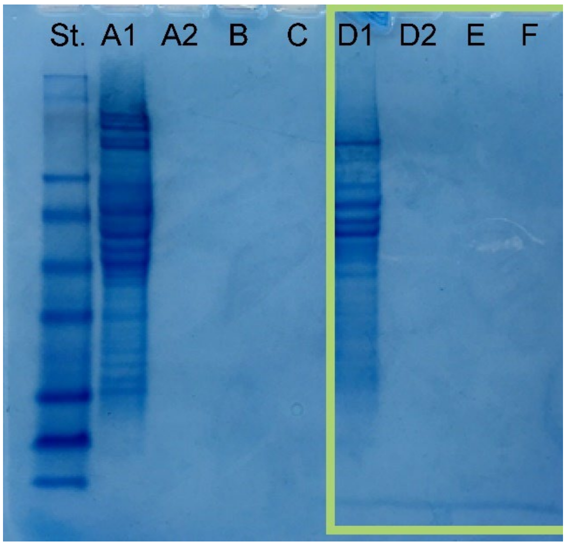


Figure 4.2. SDS-PAGE confirmation of Lys-C and trypsin digest of barley and wheat grain proteins (green rectangle). St. = Standards, A1 = undigested Commander barley proteins, A2 = digested Commander barley proteins, B = digested Gairdner barley proteins, and C = digested Hindmarsh barley proteins. D1 = Undigested Gregory proteins, D2 = Digested Gregory proteins, E = Digested Livingston proteins, F = Digested Spitfire proteins.

Table 4.5. Amounts of wheat sample peptides determined by Micro-BCA.

Sample ID	Farm location	Cultivar	Biological replicate	Total peptides (µg)	Average peptides (µg)
Br-Greg-2013-BR1	Breeza	Gregory	1	41.65	49.02
Br-Greg-2013-BR1	Breeza	Gregory	2	48.01	
Br-Greg-2013-BR1	Breeza	Gregory	3	42.04	
Br-Spit-2013-BR1	Breeza	Spitfire	1	52.42	
Br-Spit-2013-BR1	Breeza	Spitfire	2	48.93	
Br-Spit-2013-BR1	Breeza	Spitfire	3	51.35	
Br-Liv-2013-BR1	Breeza	Livingston	1	53.71	
Br-Liv-2013-BR1	Breeza	Livingston	2	53.96	
Br-Liv-2013-BR1	Breeza	Livingston	3	49.14	
TARC-Greg-2013-BR1	TARC	Gregory	1	54.08	44.10
TARC-Greg-2013-BR1	TARC	Gregory	2	42.62	
TARC-Greg-2013-BR1	TARC	Gregory	3	41.99	
TARC-Spit-2013-BR1	TARC	Spitfire	1	51.75	
TARC-Spit-2013-BR1	TARC	Spitfire	2	48.46	
TARC-Spit-2013-BR1	TARC	Spitfire	3	35.20	
TARC-Liv-2013-BR1	TARC	Livingston	1	46.74	
TARC-Liv-2013-BR1	TARC	Livingston	2	39.45	
TARC-Liv-2013-BR1	TARC	Livingston	3	36.60	
THH-Greg-2013-BR1	THH	Gregory	1	36.04	43.20
THH-Greg-2013-BR1	THH	Gregory	2	39.53	
THH-Greg-2013-BR1	THH	Gregory	3	44.51	
THH-Spit-2013-BR1	THH	Spitfire	1	49.53	
THH-Spit-2013-BR1	THH	Spitfire	2	58.01	
THH-Spit-2013-BR1	THH	Spitfire	3	59.86	
THH-Liv-2013-BR1	THH	Livingston	1	32.79	
THH-Liv-2013-BR1	THH	Livingston	2	34.20	
THH-Liv-2013-BR1	THH	Livingston	3	34.30	

4.3.4. Data quality – matched and filtered data

Following the identification of the matched and filtered proteins, the data was checked for quality. The general uniformity of the ‘density plots’ and ‘box plots’, representing protein expression data for each of the nine samples, were observed to be of good quality through close clustering of samples to their respective sample groups (Appendix C, Figure C.1).

4.3.5. Sample uniformity/diversity for matched and filtered protein samples

The wheat samples with their data of matched and filtered proteins showed some clustering into sample groups. For TMT sets-1 and -2, the clustering was limited (Appendix C, Figure C.2, Figure C.**3** and Figure C.**4**; parts A and B), However, for TMT sets-3 and -4, the samples all clustered into their sample groups for heat-maps, PCAs and correlation plots (Appendix C, Figure C.2, Figure C.**3** and Figure C.**4**; parts C and D). The difference between sample clustering of TMT sets-1 and -2 compared to TMT sets-3 and -4 is illustrated by the PCAs in Figure 4.3. TMT sets-1 to -3 are all based on the same design each having their variation based on farm location, whereas TMT set-4, compares the variation due to the three cultivars.

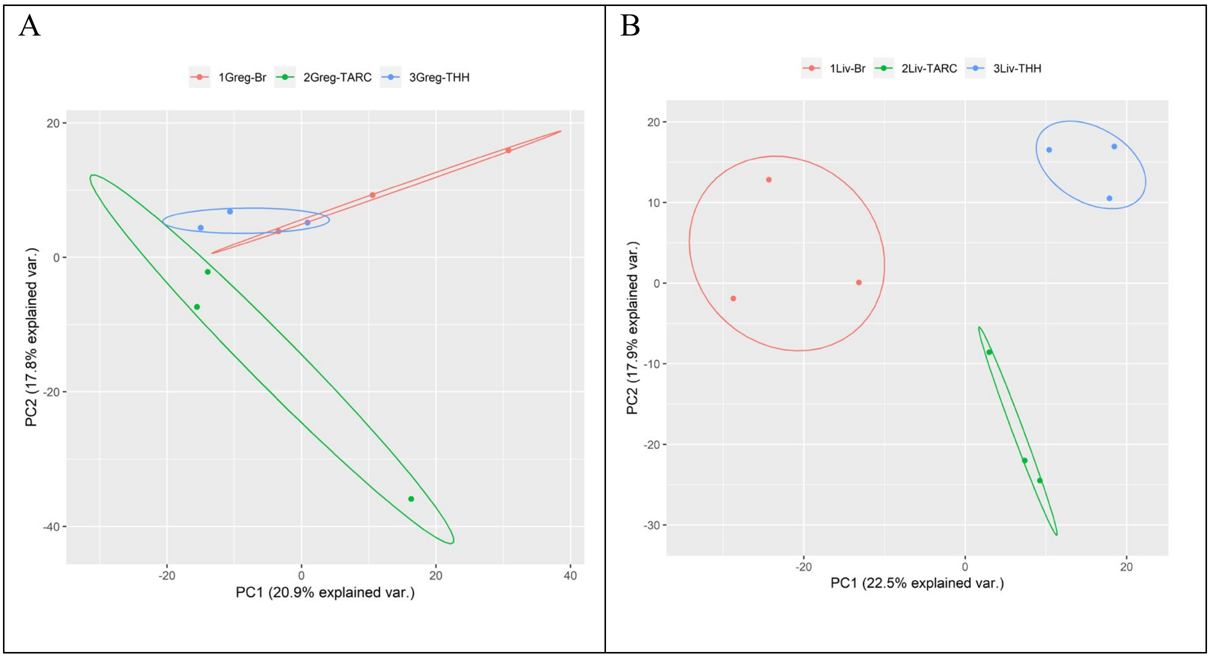


Figure 4.3. Comparison of PCAs from (A) TMT set-1 and (B) TMT set-3, from samples of matched and filtered proteins.

Table 4.6. Summary of data quality for matched and filtered wheat grain proteins.

TMT set	Sample group *(Each containing 3 biological replicates)*	Complete heat-map sample group clustering (☑ or ☒)	Complete PCA sample group clustering (☑ or ☒)	Complete Correlation heat-map sample group clustering (☑ or ☒)
1	Gregory-Breeza	☒	☒	☒
1	Gregory-TARC	☒	☒	☒
1	Gregory-THH	☑	☒	☒
2	Spitfire-Breeza	☑	☒	☒
2	Spitfire-TARC	☒	☒	☒
2	Spitfire-THH	☒	☒	☒
3	Livingston-Breeza	☑	☑	☑
3	Livingston-TARC	☑	☑	☑
	Livingston-THH	☑	☑	☑
	Gregory-All_Farms	☑	☑	☑
	Spitfire-All_Farms	☑	☑	☑
3	Livingston-All_Farms	☑	☑	☑

Original data for the above table is a summary of Figure C.2, Figure C.3, and Figure C.4 in Appendix C.
Note: One sample group is made up of three biological replicates.

4.3.6. Sample clustering after putative biomarker discovery

The 'TMTPRePro' software described in Chapter 3, section 3.2.17, performed the discovery of statistically significant differentially expressed proteins ('putative biomarkers'). From the expression data of the putative biomarkers, samples from all TMT sets were seen to cluster into their sample groups (Table 3.7), and PCAs showed increasing relatedness within sample groups and increasing un-relatedness between sample groups (Appendix C, Figure C.6). All heat-maps and correlation plots displayed clustering of samples into their groups (Table 4.7; Appendix C, Figure C.5 and Figure C.**7**). Only two sample groups failed to display complete clustering in heat-maps – Gregory wheat grown at TARC and THH. The sample groups of Gregory wheat grown at THH farm and Spitfire wheat as a pool of all three farms, failed to cluster in correlation plots (Table 3.7).

Table 4.7. Summary of data quality for putative biomarker discovery from samples.

TMT set	Sample group	Complete heat-map sample group clustering (☑ or ☒)	Complete PCA sample group clustering (☑ or ☒)	Complete Correlation plot sample group clustering (☑ or ☒)
1	Gregory-Breeza	☑	☑	☑
1	Gregory-TARC	☒	☑	☑
1	Gregory-THH	☒	☑	☒
2	Spitfire-Breeza	☑	☑	☑
2	Spitfire-TARC	☑	☑	☑
2	Spitfire-THH	☑	☑	☑
3	Livingston-Breeza	☑	☑	☑
3	Livingston-TARC	☑	☑	☑
3	Livingston-THH	☑	☑	☑
4	Gregory-All_Farms	☑	☑	☑
4	Spitfire-All_Farms	☑	☑	☒
4	Livingston-All_Farms	☑	☑	☑

Original data for the above table is a summary of Figure C.5, Figure C.6 and Figure C.7 in Appendix C.
Note: One sample group is made up of three biological replicates.

4.3.7. Putative Biomarker discovery

The TMT sets-3 and -4 displayed the better data, having more biomarkers discovered and better clustering of samples into the sample groups (Appendix C, Figure C.5, Figure C**.6** and Figure C**.7**). This difference in data quality corresponded to putative biomarker discovery, with TMT set-1 and TMT set-2 only detecting 9 and 13 putative biomarkers, respectively, while 56 putative biomarkers were detected in TMT set-3 and 173 from TMT set-4. Previously we had also performed a test TMT run with the same workflow and sample numbers minus the SCX fractionation. For the test run the numbers of putative biomarkers detected was higher in three out of four TMT sets and only lower in TMT set-3. Returning to the full run (using fractionated samples), only ten of the putative biomarkers detected were duplicated in two out of three TMT sets, and no proteins were detected in all four TMT sets.

Table 4.8. Numeric comparison of putative biomarkers from test TMT run (unfractionated samples) and full run (SCX sample fractionation) of TMT-labelled wheat grain sample sets.

Set	Study	Testing variation of cultivar or location	Number of putative protein biomarkers from 'full TMT run'	Number of putative biomarkers from 'test TMT run'
TMT Set 1	Breeza, TARC and THH farms growing Gregory wheat	Location	9	17
TMT Set 2	Breeza, TARC and THH farms growing Spitfire wheat	Location	13	15
TMT Set 3	Breeza, TARC and THH farms growing Livingston wheat	Location	56	49
TMT Set 4	Gregory, Spitfire and Livingston cultivars at all 3 farm locations	Cultivar	73	157

Table 4.9. Proteins common to selected TMT sets.

Common to TMT sets-1 and-2		Common to TMT sets-1, -3, and-4	
Identifier	**Description**	**Identifier**	**Description**
T1N5G8	Uncharacterized protein	P16851	Alpha-amylase/trypsin inhibitor CM2
B2CGM6	Triticin	Q41540	CM 17 protein
M7Z1Z4	Serpin-Z2B	C7C4X0	Alpha amylase inhibitor CM1 (Fragment)

4.3.8. Putative Biomarker functionality

The functionality of the putative biomarker proteins identified in the four TMT sets was investigated. Similar to the results in barley Chapter 3, a set of GO Slims were obtained to give a functional overview of the set of putative biomarker proteins discovered in each TMT set. For TMT sets-1 and -2, the number of protein biomarkers discovered was not sufficient to present any result of interest because the GO SLIM counts were too low (Appendix C, Table C.4 and Table C.**5**). For TMT sets-3 and -4, the functions summarised by the GO Slims terms 'molecular process' and 'cellular process' stood out (qualitatively higher) within the 'biological process' parent category (Figure 4.4). Meanwhile, the functions of 'catalytic activity' and 'binding' were noticeably higher within the 'Molecular Function' parent category for TMT set-4 (Figure 4.5).

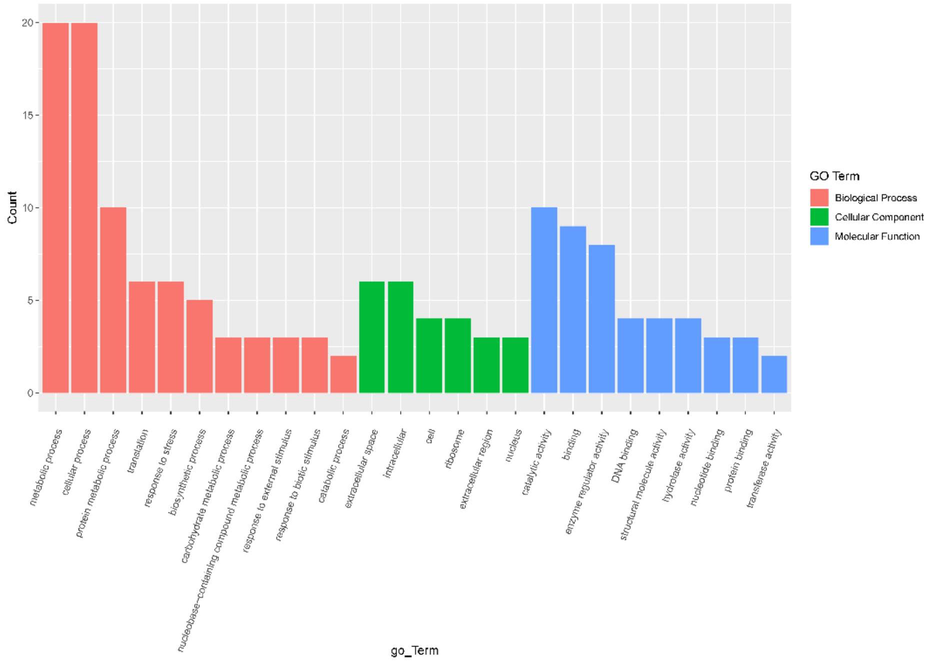


Figure 4.4. TMT set-3 Gene Ontology 'Slims' functional summary of putative biomarkers discovered. Only data with a count of 2 or more are included.

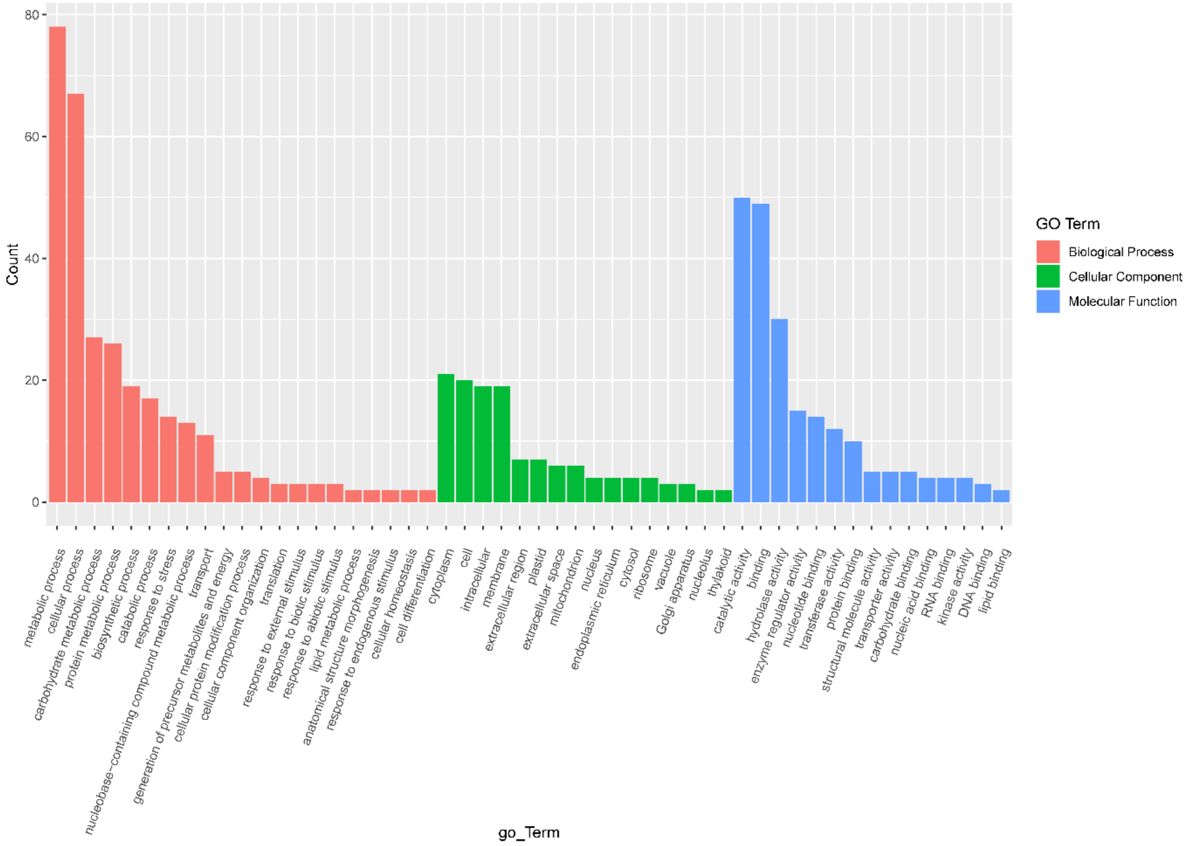


Figure 4.5. TMT set-4 Gene Ontology 'Slims' functional summary of putative biomarkers discovered. Only GO Slims categories that exist in 2 or more of the putative biomarker proteins are included.

4.4. Discussion

4.4.1. Overview

Tandem Mass Tag shotgun proteomics was shown to be a valid approach in putative biomarker discovery in Chapter 3 for barley grain. A modified experimental design (compared to the one in Chapter 3) to aid biomarker discovery was also successfully trialled (Table 4.1), many of which were similar stress response protein biomarkers to that of barley. The functionality of the discovered putative protein biomarkers was summarised for each TMT set (Figure 4.4 and Figure 4.5; Appendix C, Table C.4 and Table C.5).

4.4.2. Replicate variation

Wheat grain samples were visually examined for physical damage such as broken grain and degradation, weighed, and processed consistently between all experiments. The protein expression profile results for each sample showed some variation within sample groups, as well as the expected variation between sample groups. For example, replicate plots that were grown next to each other can be subjected to different conditions such as: differences in the exposure to rainfall, localised pathogen attack and soil composition. As a result, a greater sample variation tended to translate to the discovery of fewer putative biomarkers. This is mostly due to proteins that are close to the pass/fail threshold of biomarker discovery (statistically significant differential expression) being masked by higher levels of variation in sample replicates (i.e. too much data noise).

One of the causes of the variation between farms could potentially be due to the differences in rainfall that each farm experienced during the growth of the crops. TARC experienced the highest total rainfall during this period, followed by THH and Breeza (Figure 4.6). When this data was compared with the 200-grain weight measurements, there was an observable trend between the total water that each farm had received and the 200-grain weight (compare Figure

4.6 and Figure 4.7). It is well known that when wheat experience drought, the grain fill is compromised and result in low kernel weights (Farooq, et al., 2014, Madani, et al., 2010). There was however no observable correlation between grain fill and average monthly temperatures during this growth period (Figure 3.11). This may be likely due to the average monthly temperature range difference being only 2.4°C.

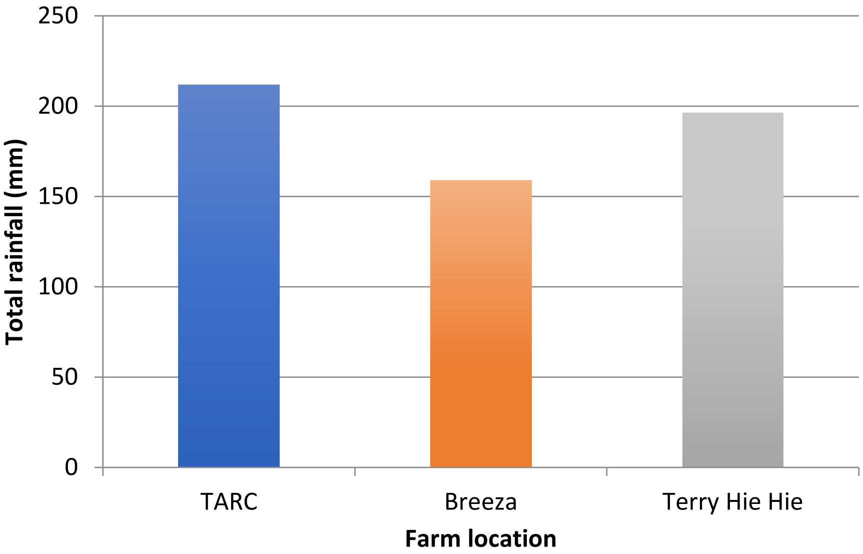


Figure 4.6. The average total rainfall during the crop growth period (May to October; 2013). Data obtained for the three different farm locations: TARC, Breeza and Terry Hie Hie.

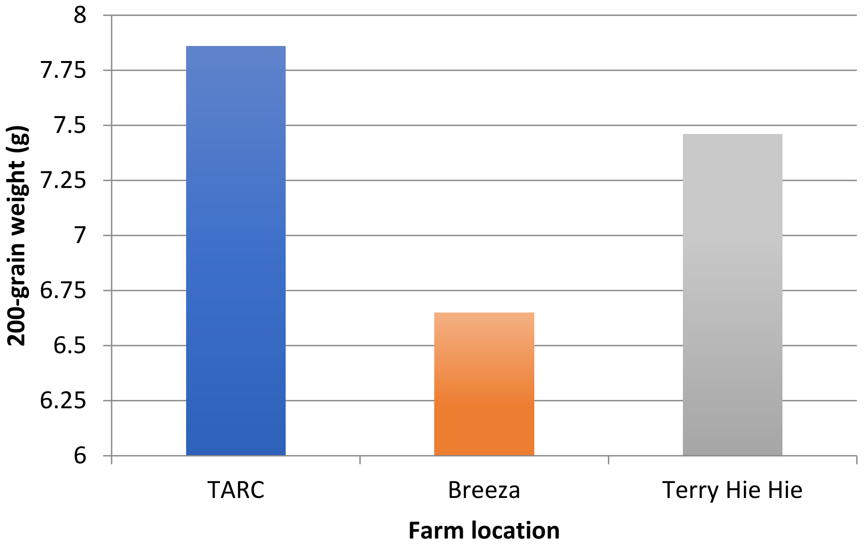


Figure 4.7. The average 200-grain weights of grain obtained from three different farm locations: TARC, Breeza and Terry Hie Hie.

The proteins identified in wheat from the farm comparison of TMT sets-1 to-3 revealed proteins associated with drought stress (Table 4.10, Table 4.11, Table 4.12). Since this study does not pinpoint which farm the differences in protein expression are from, it can only be speculated that the proteins are likely associated with samples obtained from the Breeza farm as it has experienced less rain.

However, the protein extract levels shown in Figure 4.8 display an observable inverse trend pattern with the total rainfall (Figure 4.6). Although the 200-grain weight was the lowest for Breeza, the higher 200-grain weights for TARC and THH yielded less protein. This is likely due to the larger proportion of the grain (the starch) in the TARC and THH samples being overrepresented and the proteins from the aleurone and germ cells being underrepresented in the protein extract.

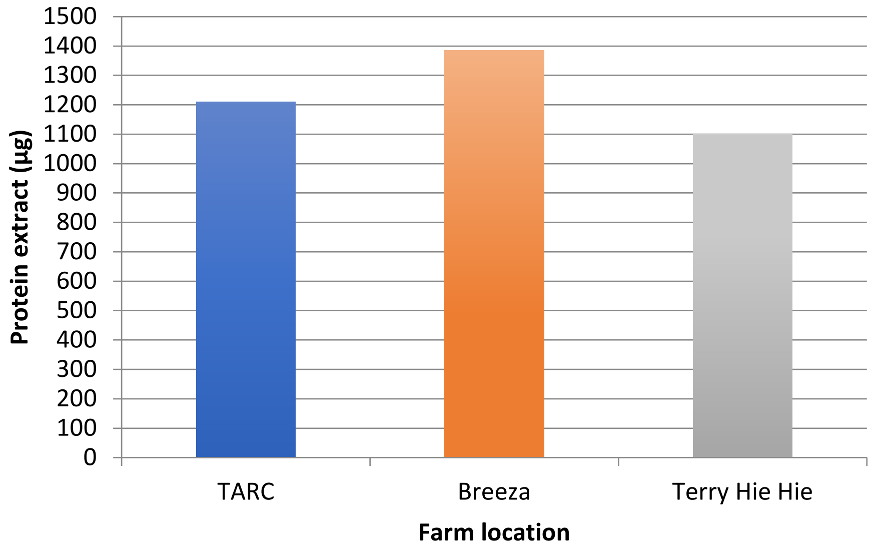


Figure 4.8. The average protein extract from wheat grown at three different farms: TARC, Breeza and THH.

4.4.3. Biomarker discovery

The experimental design was modified to increase statistical robustness, so that each individual sample was compared against one sample group. As described in Table 4.1, the rearranged design makes a comparison of nine individual samples against a pool of one sample group (three biological replicates), rather than a comparison against a pool of all nine samples as seen in the barley experiment of Chapter 3. This new design was successful in discovering putative biomarkers in all four TMT sets, although results were mixed. For TMT set-1 there were 9 putative biomarkers discovered and for TMT set-2 there were 13 putative biomarkers discovered. Both were comparing proteomes across three different locations. Compared with the barley results (TMT sets-1 and -2), which also compared proteomes across farm locations, these numbers represented a notable drop in numbers. It was possible that the different protein profile of barley compared to wheat was the likely source of this reported difference in putative biomarker discovery. However, TMT set-3 used the same design (variation of proteome across farms) as TMT sets-1 and -2 yet resulted in the discovery of 56 putative biomarkers while TMT set-4 resulted in the discovery of 173 putative biomarkers. Looking at all the data it appears highly likely that the main reason for the large differences in results for TMT sets-1 and -2 compared to TMT sets-3 and -4 was the quality of the data. Heat maps, PCAs, and correlation plots of unrefined data (matched and filtered proteins) for TMT sets-1 and-2, showed poor clustering of replicate samples into their respective groups, and larger sample variation within groups. In contrast, TMT set-3 (and TMT set-4) show good sample clustering into their respective groups, and the PCA displays smaller sample variation within sample groups. Thus, the lower data noise of TMT sets-3 and -4 seem to have resulted in a greater number of putative biomarkers discovered.

4.4.4. Putative biomarker functionality

The proposed idea that stress-response proteins are more variable within a proteome, is further reinforced by the fact that of the nine putative biomarkers detected in TMT set-1, six of these are known to be involved in stress response. Also, for TMT set-2, of the 13 putative biomarkers four are related to stress-response (Table 4.10 and Table 4.11). Due to the low number of putative biomarker proteins identified in TMT sets-1 and -2, it was difficult to identify proteins that were common in all four TMT sets. There was one serpin protein common to both TMT sets-1 and-2, which is a class of protein involved in stress response (Zhou, et al., 2016). Meanwhile TMT sets-3 and -4 had nine common proteins, six of which are involved in stress response (Table 4.12).

Table 4.10. TMT set-1: List of putative biomarkers from wheat grain.

Uniprot Identifier	Protein description
B8YLY9	Beta purothionin
B8YM21	Beta purothionin
E6Y2L2	Salt tolerant correlative protein
B8ZX17	High molecular weight glutenin subunit 1Bx13
B9A8E3	Protein disulfide isomerase
M7Z1Z4	Serpin-Z2B
M7ZK46	12S seed storage globulin 1
M8A7I9	Chitinase 2
H6S4F5	WAMP-3, antimicrobial peptide

Note: Rows highlighted in cream are proteins involved in stress response.

Table 4.11. TMT set-2: List of putative biomarkers from wheat grain.

Uniprot Identifier	Protein description
Q0Q5D8	High-molecular-weight glutenin By8
M8A6J5	Proteasome subunit beta type-6
Q9SQG8	Pathogenesis-related protein 4 (Fragment)
B2CGM6	Triticin
B2LXU4	Phosphorylase
M7YN66	Isocitrate dehydrogenase [NADP]
D2CPI7	HMW glutenin subunit
H9AXB3	Serpin-N3.2 (100% identity to Serpin-Z2B)
J3RHG6	Beta-glucosidase 4 (Fragment)
M7Z1Z4	Serpin-Z2B
Q07810	rRNA N-glycosidase
M7Y7F4	Glutamine synthetase
M8A4K5	Histone H2A

Note: Rows highlighted in cream are proteins involved in stress response.

Table 4.12. Proteins common to TMT sets-3 and -4.

Uniprot	Protein names
P17314	Alpha-amylase/trypsin inhibitor CM3 [Defense]
A4ZIX1	Monomeric alpha-amylase inhibitor (Fragment)
C7C4X0	Alpha amylase inhibitor CM1 (Fragment)
M7Z9L8	Uncharacterized protein
P83207	Chymotrypsin inhibitor WCI (Chloroform/methanol-soluble protein WCI) [Defense]
Q9ST57	Serpin-Z2A (TriaeZ2a) (WSZ2a)
A4ZIY9	Monomeric alpha-amylase inhibitor (Fragment)
A5HMG1	HMW glutenin subunit 1Bx13
Q8L6B4	Gamma gliadin

Note: Cream colour rows indicate proteins involved in stress response.

4.4.5. Functional summaries of biomarkers

To avoid describing long lists of proteins I have chosen to only comment on the functionality of proteins from TMT sets-3 and -4 in more general terms of functional summaries. As for Chapter 3, four TMT sets were examined in terms of gene ontology. Functional summaries

(GO Slims) from the 57 putative biomarkers of TMT set-3, showed that a few Slims functions were counted well above others (Figure 4.4). Namely, these ‘metabolic process’ and ‘cellular process’ for the parent category of ‘biological process’; and ‘catalytic activity’, ’binding’, and ‘enzyme regulator activity’ for the parent category of ‘molecular function’. TMT set-4 (Figure 4.5) has these same Slims but also includes ‘cytoplasm’, ‘cell’ and ‘intracellular, membrane’ from the ‘cellular component’ parent category.

4.4.6. Conclusion

The experiments presented above have shown that TMT-labelled shotgun proteomics is a powerful tool in putative protein biomarker discovery. Some tuning of sample input is needed to address the variations between biological replicate samples; however, the purpose of discovering protein biomarkers was to extract proteins and simulate what would be extracted and analysed in the field. On the other hand, for the purpose of better quality analysis for a more sensitive examination of biomarkers, an initial set of four to five replicate samples, per sample group, would be appropriate to perform a ‘test’ run through the mass spectrometer to determine the least variable sample for a full (fractionated sample) mass spectrometer run. This would reduce the intra-sample variation, or data noise, and allow more accurate discovery of statistically significant differential expression.

Finally, the most likely candidates for potential biomarkers for farm location are the ones related to biotic or abiotic stress. In an Australian context an important set of stress-response proteins would be those that respond to heat and/or drought (two abiotic stresses), with these two factors likely to play a large part in the mature grain during its growth period. Of course, these same proteomic studies would have to be repeated over several years to determine whether candidate biomarkers would indeed be predictive for farm location. In the current study, serpin and alpha-amylase proteins have been observed to be associated with drought (Zhou, et al., 2016), and identification of specific proteins from these two groups appear to correlate with the 200-grain weight and weather conditions experienced on the farm during the growth period.

The proteins serpin and alpha-amylase inhibitor are associated with drought or heat and appear to correlate with the 200-grain weight and weather conditions experienced at the Breeza plots during the growth period (Zhou, et al., 2016). For the cultivar identification, proteins associated with phenotype and endosperm composition (non-stress related proteins listed in Table 4.10, Table 4.11and Table 4.12), such as gluten and metabolic process proteins will be the most likely candidates.

Chapter 5. Detection of differentially transcribed mRNA transcripts in wheat grain for putative protein biomarker discovery and comparison with proteomic analysis

5.1. Introduction

The quantitation of types of messenger RNA (mRNA) is a commonly used technique to examine gene expression in organelles, cells, tissues, organs and whole organisms (Rangan, et al., 2017). With the advent and development of next generation sequencing for RNA (RNASeq), measuring the levels of individual RNA transcripts between samples has become increasingly accurate and cost effective (Lowe, et al., 2017). While broadly similar to proteomics in the sense of sorting and counting individual molecules (mRNA for RNASeq and peptide for proteomics), the technique of RNASeq has the advantage of amplifying the mRNA signal. Although the mRNA itself is not amplified, it is transcribed into cDNA which is subsequently amplified. While amplification may increase the chance of detecting low-copy transcripts that presumably code for low-copy proteins that are currently-undetectable via proteomics, it also introduces the potential for errors during the transcription and amplification stages. Moreover, mRNA transcripts do not necessarily get translated directly to protein, as they may be subject to post-transcriptional modification or degraded. Researchers of maize have shown that the levels of mRNA transcription are not always equivalent protein expression levels (Barros, et al., 2010). It may be that some of the proteins in mature grain are required for an immediate response – both during dormancy and germination – while the mRNA represents a longer term and more nuanced response to germination and seedling survival. Despite these potential differences between mRNA and protein levels, differentially transcribed mRNA is a potential candidate for biomarker discovery.

The aim was to measure and identify differentially expressed mRNA transcripts in wheat grain as a method to identify biomarkers that would have the potential to determine the provenance and farm location. Sample transcriptomes of the spitfire wheat grain examined and the proteins they encode compared to the differentially expressed proteins from the wheat grain results of Chapter 4.

5.2. Methods

5.2.1. Wheat grain samples

Refer to Chapters 3 and 4 methods for the details of the grain samples.

5.2.2. Testing RNA extraction methods

In determining the most suitable RNA extraction method in terms of yield and integrity/quality, three different published RNA extraction methods were compared. In all tests, fresh, finely ground wheat grain powder was prepared from the Spitfire sample. Extraction 'Method 1' was adapted from Li and Trick (2005), 'Method 2' was adapted from a combination of Holding, et al. (2007), and Reyes, et al. (2011), and 'Method 3' was the standard method for RNA extraction using TRI reagent as described by the manufacturer (Sigma-Aldrich).

5.2.2.1. RNA extraction 'Method 1'

Adapted from the RNA extraction method described in Li and Trick (2005), 400µl of 'Extraction Buffer I' (100 mM Tris-HCl, pH8.0; 150 mM LiCl; 500 mM EDTA, 1.5% SDS, and 1.5% 2-Mercaptoethanol) was added to each 2 mL sample tube containing finely ground grain (~0.1 g, which includes ~ 10% washed sand). The mixture was vortexed until the powder was fully suspended into the buffer. Phenol-chloroform (400 μL; 1:1, pH 4.7) was added to the tube and mixed well by inversion and centrifuged immediately at 13,000 x *g* for 15 minutes at 4°C. The upper aqueous phase (~400 μL) was transferred to new 2 mL plastic tube. TRI Reagent (1.2 mL; 3x sample volume) was added and the sample was mixed by gentle inversion followed by incubation at room temperature for 10 minutes. Following incubation, 208 μL (13% of the total volume) of chloroform-isoamyl alcohol (24:1) was added making 1.6 mL final volume. The sample was centrifuged at 13,000 x *g* for 15 minutes at 4°C. After centrifugation, the upper aqueous layer was placed into a fresh plastic tube. Isopropanol was added to a final concentration (sample volume + isopropanol + NaCl) of 30%, and NaCl to ~0.28 M. For example, 1.24 mL of sample was removed after centrifugation. To this volume was added 420 μL of 100% isopropanol and 109 μL of 5 M NaCl. The sample was mixed by inversion and placed on ice for 15 minutes, followed by centrifugation at 13,000x g for 15 minutes at 4°C. The supernatant was discarded and the pellet (containing RNA) was washed with 1 mL of 70% cold ethanol. The pellet was air-dried for 20 minutes at 4°C, then resuspended in 100 μL of RNase free water. An aliquot (10 μL) of the resuspended RNA was put aside for quality control. All samples were stored at -80°C.

5.2.2.2. RNA extraction 'Method 2'

'Method 2' was adapted from Holding, et al. (2007), and Reyes, et al. (2011). The Tris-buffered phenol was shaken to equilibrate at pH 8.0 and then stored at 4°C to settle. To approximately 0.1 g of finely ground wheat grain was added 500 μL of NTES buffer (20 mM Tris-HCl, pH 8.0; 100 mM NaCl; 10 mM EDTA, pH 8.0; and 1% SDS), followed by 250 μL of Tris- buffered phenol (pH 8.0), and finally 250 μL of 100% chloroform. The sample was aspirated using a pipette and vortexed followed by centrifugation at 10,000 x *g* for 10 minutes at 4°C.

Chloroform (500 µL) was added to each tube, shaken vigorously for 15 seconds and then placed on ice for 2.5 minutes prior mixing by inversion. The tubes were placed back on ice for another 2.5 minutes and then centrifuged at 10,000 x *g* for 10 minutes at 4°C. The upper aqueous phase was pipetted into a fresh 2 mL plastic tube (~ 300 µl), followed by addition of 900 mL of TRI Reagent (3x sample volume), and vigorously shaken for 15 seconds. The mixture was incubated for 5 minutes at room temperature before addition of 60 µL of chloroform (1/5 sample volume), incubation for a further 3 minutes at room temperature, followed by centrifugation at 10,000 x *g* for 10 minutes at 4°C. The upper aqueous phase (700 µL) was pipetted into a fresh plastic tube and to this was added 700 µL of 2-propanol (1x sample volume), mixed well and incubated for 10 minutes on ice. The sample was centrifuged at 10,000 x *g* for 10 minutes at 4°C, the supernatant removed and the pellet washed with 1 mL of cold ethanol (70%). The pellet containing RNA was air-dried for 20 minutes at 4°C, then resuspended in 100 µL of RNase free water. An aliquot (10 µL) of the resuspended RNA was put aside for quality control. All samples were stored at -80°C.

5.2.2.3. RNA extraction 'Method 3'

The third method was adapted from the standard Sigma-Aldrich protocol for using TRI Reagent. To approximately 0.08 g of finely ground wheat grain was added 1 mL of TRI Regent, then centrifuged at 12,000 x *g* for 10 minutes at 4°C. The supernatant was transferred to a fresh plastic tube and incubated at room temperature for 5 minutes. Chloroform (0.2 mL) was added to the sample and shaken vigorously by hand for 15 seconds. The mixture was left to incubate at room temperature for 3 minutes, followed by centrifugation at 12,000 x *g* for 15 minutes at 4°C. The upper aqueous phase of each sample was placed in a fresh plastic tube and to each tube was added 0.5 mL of 100% isopropanol (IPA), incubated at room temperature for 10 minutes, followed by centrifugation at 12,000 x *g* for 10 minutes at 4°C (the resulting pellet appears gel-like). The supernatant was discarded, and the pellet was washed with 75% cold ethanol briefly vortexed and centrifuged at 7,500 x *g* for 5 minutes at 4°C. The pellet was air dried at 4°C for 15 minutes then resuspended in 100 µL of RNase-free water. An aliquot (10 µL) of the resuspended RNA was put aside for quality control. All samples were stored at -80°C.

5.2.3. Phenol/chloroform RNA extraction, post DNase treatment

Phenol/chloroform (phenol at pH 4.5, ratio 1:1) was added to the DNase extraction at a 1:1 ratio, mixed by inversion, and then centrifuged at 14,000 rpm for 2 min. The upper aqueous layer was removed and placed in a fresh plastic tube. To this was added two volumes of 100% cold ethanol. Precipitation of RNA occurred after incubating at -20°C for approximately 1 hour. This was then centrifuged at 14,000 rpm for 5 minutes. The supernatant was removed and the pellet was rinsed with 70% ethanol. Finally, the washed pellet was resuspended in molecular grade purity RNase-free water.

5.2.4. RNA preparation for Illumina sequencing

RNA was extracted from grain samples of the Spitfire wheat cultivar using ‘Method 2’ (section 5.2.2.2). Three replicates for each of the three sites (Breeza, TARC, and Terry Hie Hie [THH]) were prepared. The quality of the RNA extraction was checked on an agarose gel. Samples were prepared for gel loading by adding 400 ng or more of RNA extract to a fresh plastic tube and mixing with RNase free loading dye (50% high-grade glycerol, 1 mM EDTA, and 0.4% bromophenol blue). The agarose gel was prepared by making a hot 1% agarose solution to which 2% (v/v) gel red (Biotium) was added before being poured into a mould. Once solidified, the gel was placed into an electrophoresis tank containing 1x TAE buffer. The molecular weight ladder (Quick-Load® 1 kb DNA Ladder - N0468S - by New England Biolabs) and samples were then loaded and run at 100-volts until a visible dye front from the loading buffer had almost run off the gel closest to the anode. The gel was then visualised by UV and digitally captured by a Gel Logic 2200 pro (manufactured by Carestream). Once the RNA samples passed visual inspection, the concentration of RNA was determined by using a NanoDrop™ 2000/2000c Spectrophotometer (Thermo Scientific™). DNase digests were then set up in a total volume of 250 µL. The volume of RNA sample and buffer were kept constant, while the volumes of DNase and water were adjusted so that the ratio of DNase to DNA/RNA was 1 Unit Enzyme to 1 µg Nucleic Acid. The DNase treated RNase samples were purified using the phenol/chloroform extraction method described in section 5.2.3. Samples were checked again by agarose gel and the RNA concentration and quality was determined using a NanoDrop as described above. These samples and their details were then sent to AGRF on dry ice to be sequenced on the Illumina HiSeq2500 sequencing platform, using a strand sensitive protocol and sequenced to a 100 bp length for both forward and reverse sequences.

5.2.5. Custom R-Script: "tximportAfterKallisto.R"

The initial quantification of mRNA transcript abundance was performed by the program Kallisto (Bray, et al., 2016). The R-script "tximportAfterKallisto.R" (Appendix E, section 0; which incorporated the R-package "tximport") converted the estimated expression counts, generated by Kallisto, into raw transcript counts. This conversion was necessary to fit the negative binomial model assumed by many downstream RNASeq analysis programs such as the R-package "DESeq2". Other R-packages - "dplyr" (Wickham, et al., 2017), "DESeq2" (Love, et al., 2014), "readr" (Wickham, et al., 2017), "AnnotationDbi" (Pagès, et al., 2017), "ReportingTools" (Huntley, et al., 2013) - are also included within this script to assist access to databases, re-formatting, as well as data preparation and manipulation.

5.2.6. DESeq2 R-package - summary of methods

DESeq2 is a software package that allows differential expression analysis of RNASeq data. To summarise, a table of mRNA raw counts (transcripts) is input into Deseq2 from the workflow discussed above, as is a design file that serves as a source for metadata. First, library size and mRNA composition bias are corrected by internal normalisation, and for each gene/transcript the geometric mean is calculated across all samples (rows of data). In each sample, counts for every gene are divided by this mean, and the median of these results (each sample), giving the size factor for that sample. Shrinkage estimation is used to calculate dispersion and fold-changes, with a model fit procedure calculating a dispersion value for each gene. A negative binomial model is fitted for each gene, and the Wald or LRT test is then used to calculate significance for each gene. DESeq2 also uses several data filters, such as outlier counting, the elimination of genes with normalised count means below a DESeq2 determined threshold, and removal through Cooks' distance (a method to determine important datapoints [usually outliers] that may warrant investigation, validation, or possibly deletion). For more detail on DESeq2 and its underlying algorithms see Love, et al. (2014).

5.2.7. Custom R-Script: "DESeq2_Functions.R"

The R-script "DESeq2_Functions.R" uses data either generated directly or imported from a saved file from the previous R-script "tximportAfterKallisto.R" to quality check the mRNA expression data and find the genes or RNA transcripts that show statistically significant differential expression. At the core of this script are a number of DESeq2 functions including quality control graphs, estimation of size factors, estimation of dispersion, and tests for significance through the likelihood ratio test (LRT) or Wald test. The R-packages "ggplot2" (Wickham, 2016), "RColorBrewer" (Neuwirth, 2014), "DESeq2" (Love, et al., 2014), "pheatmap" (Kolde, 2015), and "ReportingTools" (Huntley, et al., 2013), are additionally used within this script for data manipulation and production of graphical and tabular summaries. Quality control plots, and graphical and text summaries are written to new folders. The full “DESeq2_Functions.R” script can be found in the Appendix E, section 0

5.2.8. GO-Slims and the AgBase web service

The “AgBase” website, hosted by the Mississippi State University, has a number of web-based tools available to investigate protein functionality through gene ontology. Output from the AgBase website is used by several of my custom R-scripts. A list of UniProt identifiers found to have statistically significant differential expression is generated and is manually fed into the web tool “GO-Retriever”, where the gene ontology terms are retrieved for each protein. The resulting tabular file is then manually fed into the “GO SLIM Viewer” tool to summarise the functionality of the initial protein list in the form of GO Slims. A final zipped file is produced, which is manually saved and extracted to a target directory for analysis.

5.2.9. Custom BLAST database construction

To perform custom blast searches, the ‘blast’ suite of programs was installed from NCBI (‘ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/’). A custom BLAST database for wheat was then constructed using the ‘customblastdb’ using the default parameters and the wheat amino acid FASTA file as described in section 4.2.19 (Chapter 4).

5.2.10. Custom R-Script: “gtfToFastaThenBlastx.R”

This R-script “gtfToFastaThenBlastx.R” gathers information from the wheat Gene Transfer Format (GTF) file, the wheat genome sequence file (FASTA file), and the list of significant differentially expressed genes determined by the “DESeq2_Functions.R” script to ultimately enable a BLASTX search. Prior to running this script, the GTF and wheat-genome-FASTA file were obtained from the Ensembl Plants FTP site for wheat: ftp://ftp.ensemblgenomes.org/pub/plants/release-36/gtf/triticum_aestivum/Triticum_aestivum.TGACv1.36.gtf.gz. ftp://ftp.ensemblgenomes.org/pub/plants/release-36/fasta/triticum_aestivum/dna/Triticum_aestivum.TGACv1.dna.toplevel.fa.gz The script loads these ensemble plant files into memory, as well as the list of significant transcripts determined by the DESeq2_Functions.R script. This R-script then invokes and directs two command line programs. The first, "gnugrep", creates a wheat GTF subset file that only maps transcripts that match the previously determined (by "DESeq2_Functions.R") transcripts of interest. The second, "gffread", from the cufflinks package (Trapnell, et al., 2010), uses this GTF subset to produce a wheat FASTA file that only contains genes which match significant transcripts (i.e. genes and their isoforms). The "seqinr" R-package (Charif and Lobry, 2007) and custom scripting then subsets the FASTA file further so that it only contains genes of interest (no isoforms). This final FASTA file is then submitted to a BLASTX search that transcribes the DNA sequence of each gene into an amino acid sequence that is then matched against the same protein FASTA file that has been used for previous protein searches for wheat (Chapter 4, 4.2.19). The resulting tabular data is then filtered and summarised using the "Dplyr" R-package (Wickham, et al., 2017) and the data saved as both a ".csv" file and a text file list of UniProt protein identifiers, the latter of which is used to gather Gene Ontology information through the "GO-Retriever" web service (5.2.8). The script "gtfToFastaThenBlastx.R" can be found in Appendix D, section D.8.3.

5.2.11. Custom R-Script: "extraGOScript.R"

The "extraGOScript.R" uses a number of custom R-functions, as well as the "GO.db" (Carlson, 2016) and "Annotation.Dbi" (Pagès, et al., 2017), to gather gene ontology and GO-Slims data, and add this data to the tabular result of the BLASTX search (Custom R-Script:

"gtfToFastaThenBlastx.R", see Chapter 5, section 5.2.10). The GO-Slims data is first merged into a single table and then manipulated into a form that allows the merging with the BLASTX summary data file. This extended table of BLAST and gene ontology results for each identifier is then saved to file.

5.2.12. Custom R-Script: "makeFastaFromTable.R"

This script loads the "…blastxPlusFullInfo.csv" file that was created from the custom R-script 'extraGOScript.R' described above section 5.2.11. The UniProt identifier and its sequence is read and used by the script to create a FASTA amino acid file, which is then saved to disk.

5.2.13. Custom BLASTP

In order to extract the same set of 'UniProt' identifiers described in Chapters 3 and 4, a custom 'blastp' was then run using the custom protein database created by 'makeblastdb' from NCBI (download from: "ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/."), in conjunction with the FASTA amino acid file created by the R-script 'makeFastaFromTable.R' (5.2.12). The tabular output was then saved to disk for further analysis. This output included the type of protein identifiers described in Chapters 3 and 4, as well as their brief functional descriptions.

5.3. Results

5.3.1. RNA extraction method selection

In order to extract the purest and least degraded mRNA, we tested three different common RNA extraction methods as described in section 5.2.2. Based on the qualitative analysis of the agarose gel of the RNA extracts, all three methods were found to provide good quality extractions. Of the three methods, 'method 2' was selected as it appeared to have the least amount of RNA degradation (Figure 5.1).

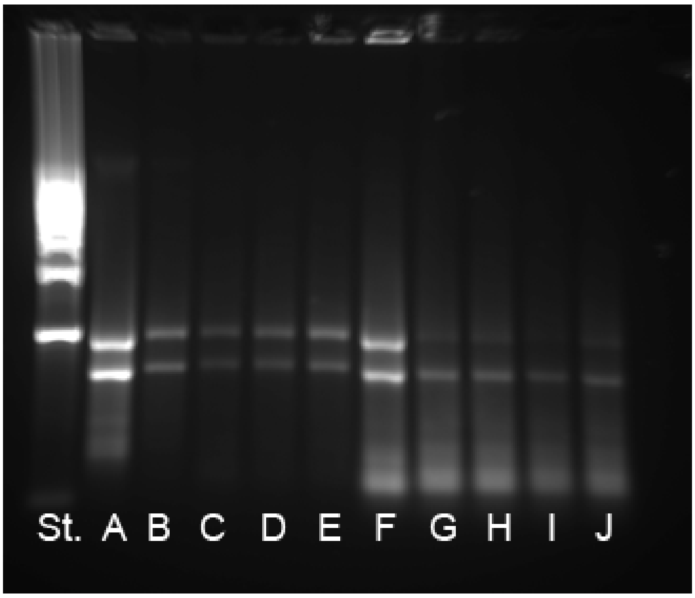


Figure 5.1. RNA sample runs on a 1% agarose gel, visualised with gel red showing three different RNA extraction methods (see section 5.2.2 for details). Samples A and B were extracted using 'method 1'; C and D from 'method 2'; and F, G, H, I and J from 'method 3'. St = molecular weight marker.

5.3.2. RNASeq – data quality

RNASeq was applied to examine the mRNA found in wheat grain from the Spitfire cultivar that had been grown at three different geographical/farm locations (Breeza, TARC, and Terry Hie Hie [THH]). After receiving the sequencing analysis results from the Australian Genome Research Facility (AGRF), they were put through our bioinformatics pipeline until we had counts of raw transcripts. Examination of the raw read count revealed some variation between samples and sample groups, ranging from 7 to 9.2 million, which based on previous experience seems acceptable. Normalisation of the raw read count was applied, however, it made little difference (Figure 5.2). Density plots of the data showed a good result with Spitfire wheat samples from Breeza, TARC and THH, each neatly overlaying one another, thus indicating that the sample preparation and RNA sequencing run was relatively consistent for all samples (Figure 5.3). Correlation plots of each sample's raw counts showed that the samples clustered into their sample groups, except for one of the TARC replicates (Figure 5.4). The principal component analysis (PCA) plot of the raw count data was acceptable, although the samples from the TARC and THH farms plotted further apart than those from Breeza (Figure 5.5).

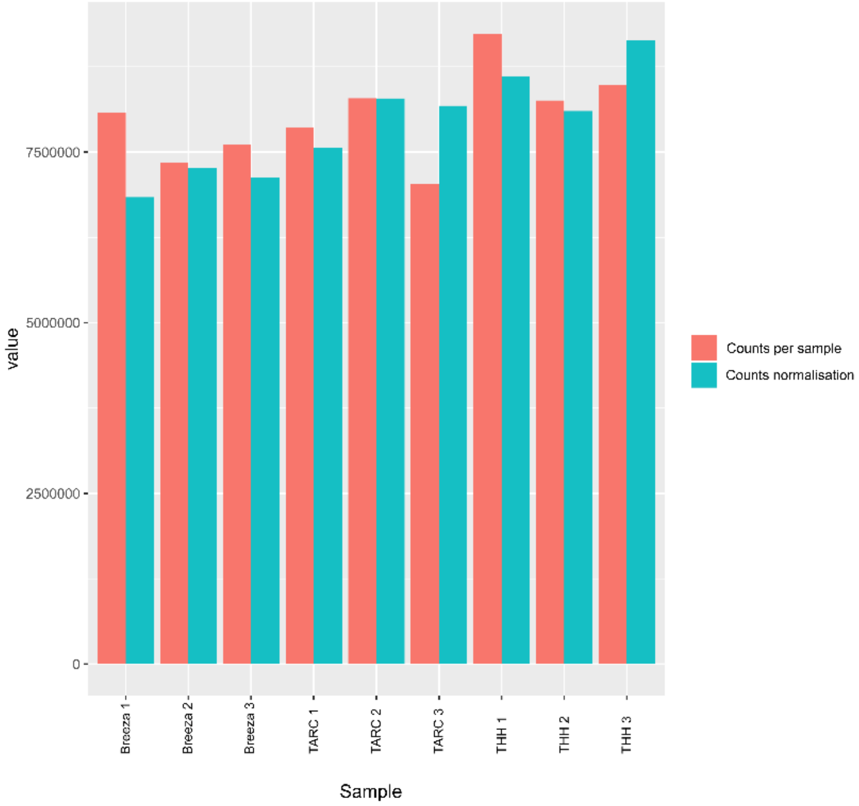


Figure 5.2. Raw and normalised transcript counts for all Spitfire wheat grain samples grown at three different farm locations.

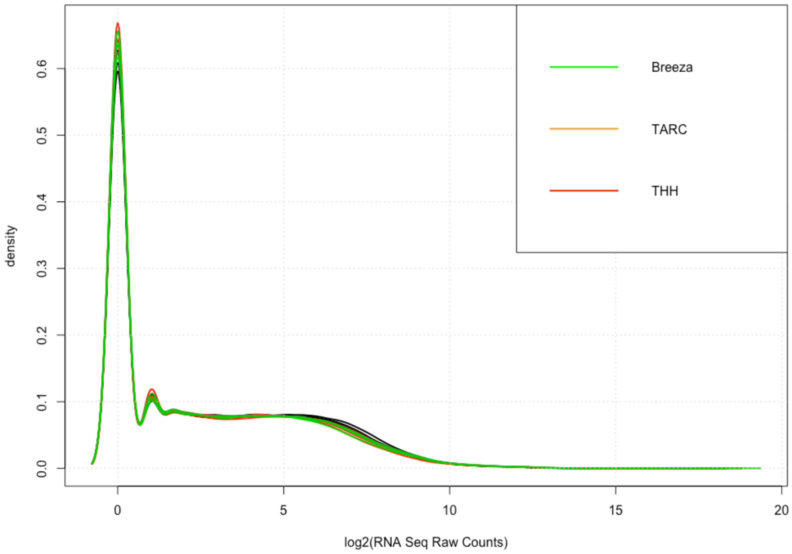


Figure 5.3. Density plots of transcript counts for all nine Spitfire wheat samples grown at three different farm locations.

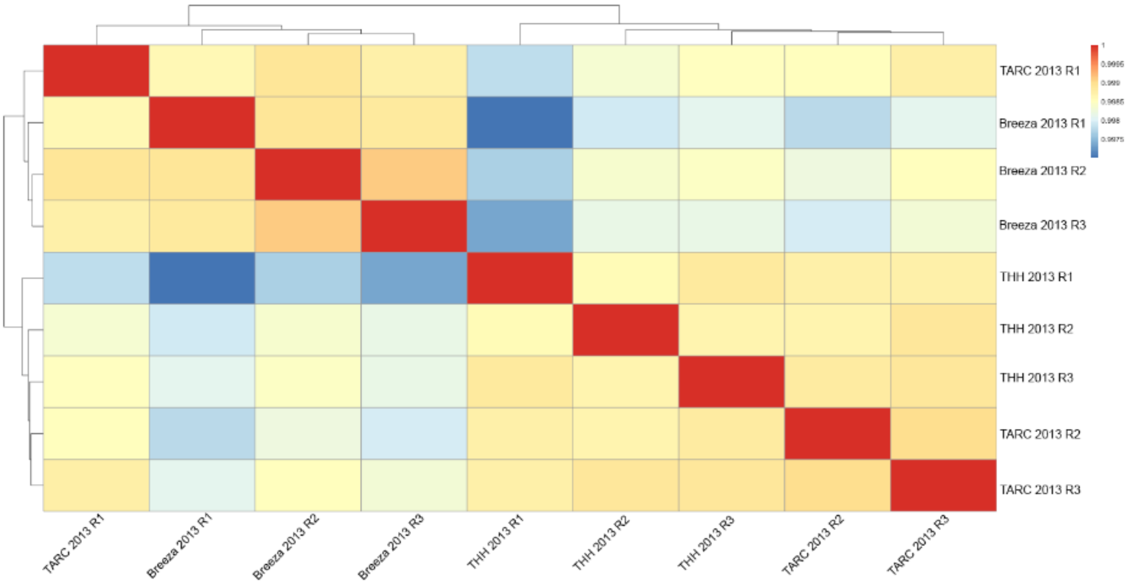


Figure 5.4. Correlation heat-map for the nine Spitfire wheat samples grown at three different geographical/farm locations. Note: only replicate-1 grown at the TARC farm clusters outside of its sample group.

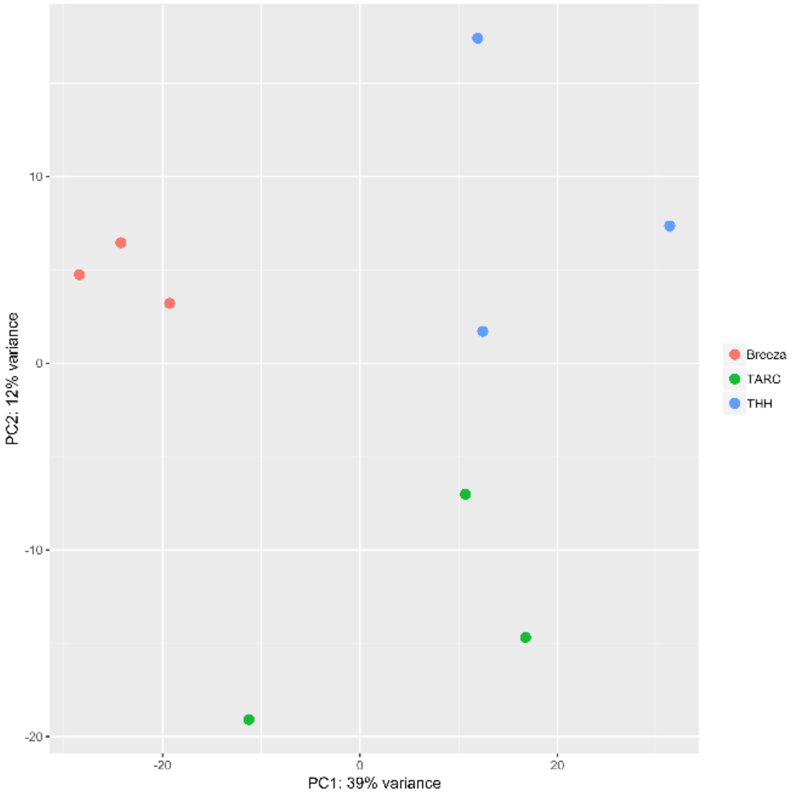


Figure 5.5. PCA of raw RNA transcript counts from the nine Spitfire wheat samples grown at Breeza, TARC, and THH farms.

The scatterplot of counts displaying mean versus variance (Appendix D, Figure D.1) shows that the data followed a negative binomial distribution as required by the DESeq2 R-package to examine the differential expression of RNA transcripts. Similarly, the plot of dispersion estimates revealed that the data is appropriate for assumptions made by DESeq2 (Figure 5.6).

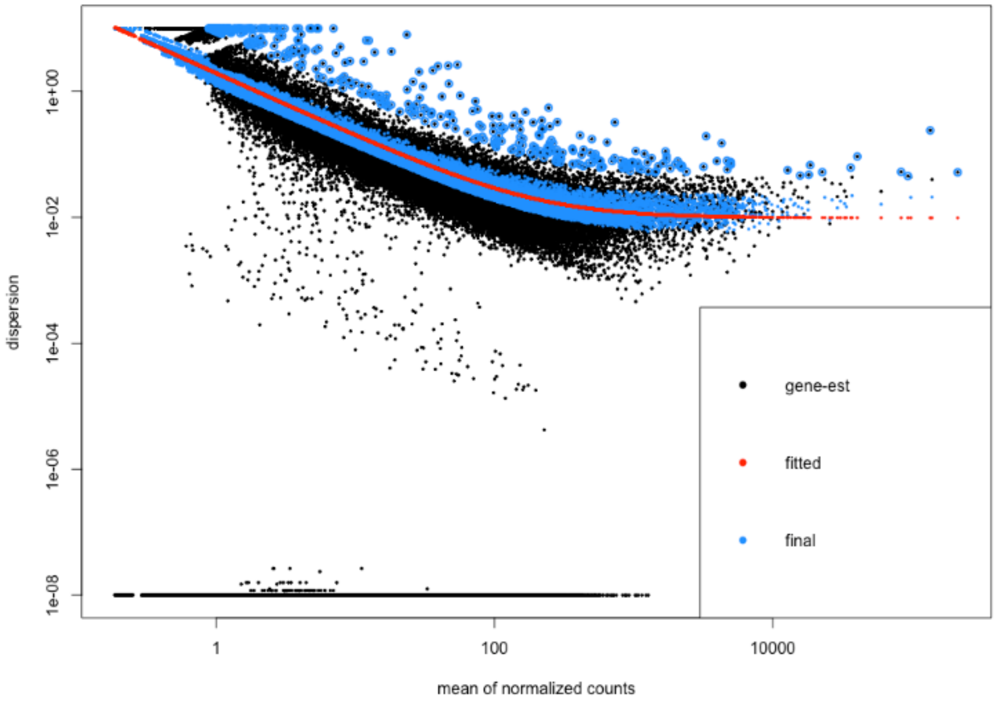


Figure 5.6. Plot of dispersion estimates for all transcripts

Functions within DESeq2 were used to calculate differential expression of RNA transcripts. Two sample groups (with three biological replicates each group) were used as 'input' – to make a comparison. Following data manipulation (as described in the methods), the differential expression of transcripts in all nine samples is 'output' and the likelihood-ratio-test (LRT), or Wald test, was applied to determine statistical significance. Initially all possible combinations of the three sample groups (farm locations) were compared, ignoring the reversal of combinations (Table 5.1). The samples grown at Breeza were compared against those grown at TARC using the Wald test. The comparison revealed 338 transcripts with a p-adjusted value below the 0.05 threshold, while the same sample comparison using LRT resulted in 613 transcripts of statistical significance (p = <0.05). When Breeza was compared against THH, it resulted in even higher numbers of transcripts – 1,264 (Wald test) and 2,350 (LRT) – that were statistically significant. The TARC comparison with THH resulted in 17 (Wald test) and 41 (LRT) statistically significant transcripts.

5.3.2.1. Breeza against TARC, and Breeza against THH comparisons

The Breeza and TARC comparisons, as well as the Breeza and THH comparisons listed in the first four rows of Table 5.1, shows a large range in significant transcripts detected (from 238 up to 2,350). However, the underlying data of statistically significant transcripts seemed poor. There was little difference between the PCA from raw transcript counts and the PCA of counts showing statistical significance (Figure 5.5; Appendix D, Figure D.6). Heat-maps had incomplete clustering of samples into groups, and volcano plots showed an abrupt transition from non-significant transcripts to significant transcripts (Figure D.2, Figure D**.3**, and Figure D**.5**; Parts A to D).

5.3.2.2. TARC against THH comparison

In the remaining comparisons shown in Table 1 (TARC and THH), there were relatively small numbers of significant transcripts detected, with 17 for the LRT and 41 for the Wald test. The sample groups clustered as expected in PCAs and heat-maps (Appendix D, Figure D.6 andFigure D.4. Same as for Figure D.2, except the comparison was between TARC and THH transcript expression data. LRT (A) or the Wald test (B) were also used to discover transcripts with statistically significant differential expression.

5.3.3. D.3 Volcano plots of sample-transcript expression

In the remaining comparisons shown in Table 1 (TARC and THH), there were relatively small numbers of significant transcripts detected, with 17 for the LRT and 41 for the Wald test. The sample groups clustered as expected in PCAs and heat-maps (Appendix D, Figure D.7,Figure D.**8**,Figure D.**9** and Figure D.**11**- parts E and F), and the volcano plots had a smoother transition from points of non-significant transcripts to significant transcripts (Appendix D, Figure D.10 – parts E and F).

Table 5.1 Results of tests for significance of transcripts using the LRT or Wald test.

Comparison	Test for significance	Significant transcripts count
Breeza compared to TARC	LRT	238
Breeza compared to TARC	Wald	613
Breeza compared to THH	LRT	1,264
Breeza compared to THH	Wald	2,350
TARC compared to THH	LRT	17
TARC compared to THH	Wald	41

There were large differences in the numbers of significant transcripts detected between the six comparisons (Table 5.1). The same comparisons were then examined in reverse orientation. Similarly, the differences between largest and smallest counts were substantial. The obtained results of the significant transcript counts are shown in Table 5.3. TARC compared to Breeza yielded 581 significant transcript counts using LRT (to determine significance) while the Wald test of the same comparison resulted in 974. The comparison between THH and Breeza yielded the identification of 954 (LRT) and 1769 (Wald test). Finally, the comparison between THH and TARC yielded the identification of 17 (LRT) and 41 (Wald test) significant transcripts.

Table 5.2. Summary of sample group clustering for heat-maps, PCAs, and spread of significant data on volcano plot.

Comparison	Complete **heatmap** sample clustering (☑ or ☒)	Complete **PCA** sample clustering (☑ or ☒)	Spread of significant data on **volcano plot** (☑ or ☒)
Breeza compared to TARC	☒	☒	☒
Breeza compared to TARC	☒	☒	☒
Breeza compared to THH	☒	☒	☒
Breeza compared to THH	☒	☒	☒
TARC compared to THH	☑	☑*	☑
TARC compared to THH	☑	☑*	☑

* = One sample outlier for both TARC and THH replicates
For original data refer to Appendix D, Figure D.2 toFigure D.6

Table 5.3. Results of tests for significance using the reverse orientation of sample comparisons as shown in Table 1.

Comparison	Test for significance	Significant transcripts count
TARC compared to Breeza	LRT	581
TARC compared to Breeza	Wald	974
THH compared to Breeza	LRT	954
THH compared to Breeza	Wald	1769
THH compared to TARC	LRT	17
THH compared to TARC	Wald	41

The resulting data from the comparisons of Table 5.3 was highly analogous to that of Table 5.1. The first four comparisons listed in Table 5.3 showed a large range of statistically significant transcripts varying from 581 to 1,769. The average to poor quality of this data was also reflected in the PCA, heat-map and volcano plots as summarised in Table 5.4 (original data in Appendix, Figure D.7 to Figure D.11). In the comparison from the last two rows of Table 5.3 (THH compared to TARC), 17 transcripts with a p-value 0.05 or less were detected by LRT, while 41 were detected with the Wald test. Note: the result of the previous comparison (in the reverse order) of TARC compared to THH (Table 5.1, rows 5 and 6), also resulted in 17 and 41 statistically significant transcripts.

Table 5.4. Summary of sample group clustering for heat-maps, PCAs, and spread of significant data on volcano plot.

Comparison	Complete **heat-map** sample clustering (☑ or ☒)	Complete **PCA** sample clustering (☑ or ☒)	Spread of significant data on **volcano plot** (☑ or ☒)
TARC compared to Breeza	☒2	☒	☒
TARC compared to Breeza	☒2	☒	☒
THH compared to Breeza	☒1	☒	☒
THH compared to Breeza	☒1	☒	☒
THH compared to TARC	☑	☑	☑
THH compared to TARC	☑	☑	☑

* = One sample outlier for both TARC and THH replicates
For original graphs refer to Appendix, Figure D.7 to Figure D.11.

The DESeq2 was used to examine all three-sample groups in an ANOVA-like approach to determine differential expression. Only LRT was used to calculate significance using this method, revealing 1767 transcripts that had p-values less than 0.05. The analysis was repeated with removing transcripts that had less than 10 counts across all nine samples. This resulted in a slight increase in significant transcripts to 1880. In reducing the complexity of the data, we picked an arbitrary p-adjusted value of 0.0001, which resulted in 253 significant transcripts across the nine samples. Despite the reduced number, the clustering of samples did not improve. A heat-map of the results is shown in Figure 5.7 and the volcano plot and PCA in Appendix D, Figure D.13 and Figure D.14.

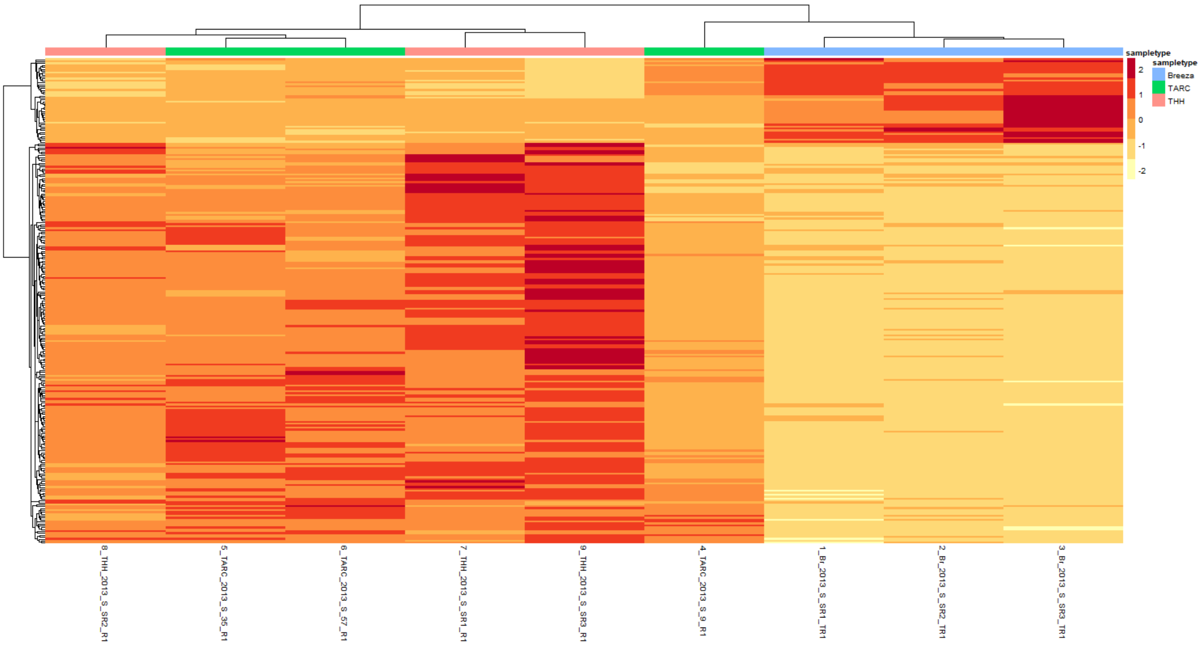


Figure 5.7. Heat-map of differentially expressed transcripts after AVOVA-like comparison across three sample groups and selecting data with a p-adjusted value of 0.0001 or lower after testing for significance using LRT.

Twelve-sample group comparisons were made, including one comparison across all three-sample groups. These comparisons and the resulting counts of significant transcripts are summarised in (Figure 5.8). For all of the sample comparisons involving wheat grain that was grown and harvested at Breeza farm, the counts ranged from high (338) to extremely high (2350). None of these results were symmetrical. When applying the Wald test, Breeza compared with TARC gave a different result to TARC compared against Breeza. In contrast, samples that did not include Breeza in the comparison, and using the same test for significance, gave the same answer no matter the order of comparison. Samples grown on TARC farm compared against those grown at THH farm resulted in 17 transcripts with differential expression and for the same comparison using the Wald test 41 transcripts with differential expression were found. The reverse order (THH compared to TARC) gave the same results.

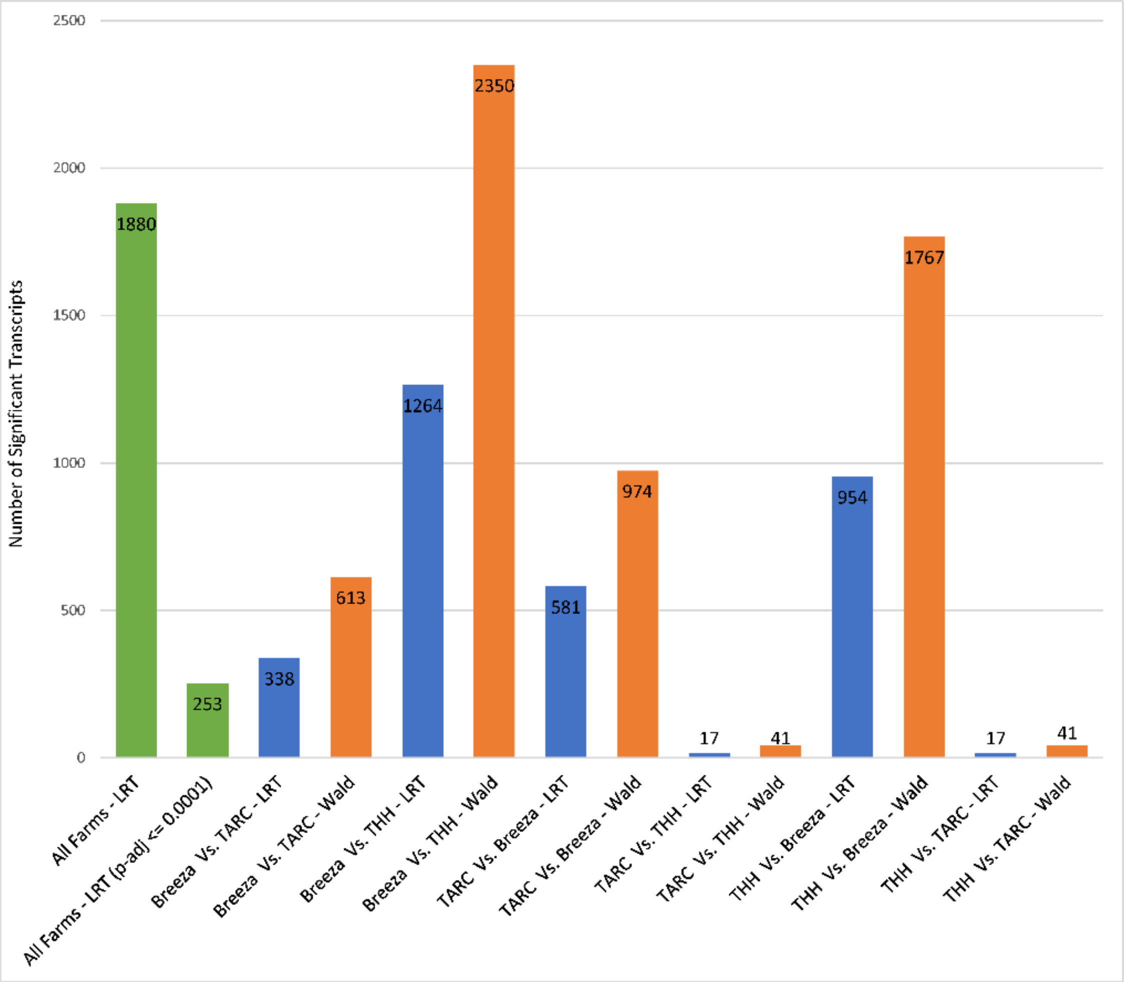


Figure 5.8. Summary bar graph of significant transcripts found after various comparisons that used either the LRT or Wald test.

The higher quality of data for the THH and TARC farm comparison shown in Figure 5.8, led us to examine further the 41 transcripts detected with differential expression. This was done due to the fact that they give the same number of significant transcripts no matter the comparison order. After a BLASTX search on each transcript sequence, we detected 30 unique proteins from the initial list of 41 transcripts. The results are listed in Table 5.5 (rows with no protein match for a transcript were removed), including a few examples of multiple transcripts that code for one protein (each highlighted in a colour). The UniProt Identifiers were then input into 'GORetriever' and 'GOSlimViewer' at the 'AgBase' web site (http://agbase.arizona.edu/cgi-bin/tools/index.cgi) and a list of gene ontology 'Slims' was obtained and counted. Figure 5.9 is a summary of the Slims result, showing protein functionality for the 30 significant proteins derived from the 41 transcripts. We observe that functionality associated with the categories of metabolic process, cellular process, and carbohydrate metabolic process are in the top three for the parent gene ontology term of Biological Process. The parent term of cellular component, membrane, intracellular, and cell categories are in the top three categories, while catalytic activity, binding and transferase activity are in the top three for the gene ontology parent term of Molecular Function.

Table 5.5. List of unique transcripts and their uniport identifier equivalents. Duplicate uniprot identifiers are highlighted in the same colour.

Transcript Id	Uniprot Id	Protein names
TRIAE_CS42_7AL_TGACv1_557592_AA1783720.1	M7YYP5	Copper transport protein ATOX1
TRIAE_CS42_6AL_TGACv1_473156_AA1529060.1	Q9ZP25	Small heat shock protein Hsp23.5
TRIAE_CS42_5DL_TGACv1_435700_AA1453740.1	M8A900	Putative aldehyde oxidase-like protein
TRIAE_CS42_2AL_TGACv1_094696_AA0301730.1	Q9ZPJ1	S-adenosylmethionine decarboxylase proenzyme (EC 4.1.1.50)
TRIAE_CS42_5BL_TGACv1_404534_AA1303340.1	M8A900	Putative aldehyde oxidase-like protein
TRIAE_CS42_1DL_TGACv1_063686_AA0229870.1	M7YXH3	Cytochrome P450 71D10
TRIAE_CS42_6AL_TGACv1_471999_AA1516750.1	Q94KM0	HSP17 (Small heat shock protein HSP17.8)
TRIAE_CS42_3DS_TGACv1_273985_AA0934140.1	Q94KM0	HSP17 (Small heat shock protein HSP17.8)
TRIAE_CS42_5BL_TGACv1_404407_AA1298710.1	M7Z1U5	Uncharacterized protein
TRIAE_CS42_2DL_TGACv1_158033_AA0506710.1	A0A1D5UH06	Uncharacterized protein
TRIAE_CS42_3DS_TGACv1_274198_AA0935000.1	M8A6E3	Bowman-Birk type trypsin inhibitor
TRIAE_CS42_5AL_TGACv1_374738_AA1207960.1	O82072	Phospoenolpyruvate carboxylase
TRIAE_CS42_1DL_TGACv1_061207_AA0188800.1	M8A7A1	UDP-glucose 4-epimerase GEPI48
TRIAE_CS42_2AS_TGACv1_113492_AA0356900.1	M8A8W1	Premnaspirodiene oxygenase
TRIAE_CS42_2BL_TGACv1_130219_AA0406550.1	M7Z617	26S proteasome non-ATPase regulatory subunit 1 homolog
TRIAE_CS42_6DS_TGACv1_544490_AA1748530.1	A0A1D6BGV3	Uncharacterized protein
TRIAE_CS42_1AL_TGACv1_002363_AA0041170.1	M7YXH3	Cytochrome P450 71D10
TRIAE_CS42_5AL_TGACv1_375837_AA1227780.1	Q84XZ4	Mitogen-activated protein kinase
TRIAE_CS42_6BL_TGACv1_500556_AA1606450.1	Q43210	Phenylalanine ammonia-lyase (EC 4.3.1.24)
TRIAE_CS42_6BS_TGACv1_513261_AA1636450.1	A0A1D6AVB2	Uncharacterized protein
TRIAE_CS42_U_TGACv1_640869_AA2077820.1	M7Z5S2	Uncharacterized protein
TRIAE_CS42_5DL_TGACv1_434877_AA1443000.1	M7ZZ82	Uncharacterized protein
TRIAE_CS42_6BS_TGACv1_514543_AA1661230.1	A0A1D6AZL6	Uncharacterized protein
TRIAE_CS42_6BL_TGACv1_500879_AA1611210.1	A0A1D6AR04	Uncharacterized protein
TRIAE_CS42_2AL_TGACv1_093415_AA0279740.1	A0A1D6RIE6	Uncharacterized protein
TRIAE_CS42_4DS_TGACv1_361964_AA1174830.1	A5A8T7	17.6kDa heat-shock protein
TRIAE_CS42_7BL_TGACv1_577339_AA1872660.1	Q6T484	Class I chitinase (EC 3.2.1.14)
TRIAE_CS42_6DL_TGACv1_527510_AA1705040.1	M8A8A8	Protein WAX2
TRIAE_CS42_5AL_TGACv1_376126_AA1232370.1	M7ZFG4	Cellulose synthase-like protein E6
TRIAE_CS42_1AL_TGACv1_001031_AA0023780.1	M8AGD8	Monosaccharide-sensing protein 2
TRIAE_CS42_3DS_TGACv1_271732_AA0906990.1	M7YJ37	60S ribosomal protein L9
TRIAE_CS42_5BL_TGACv1_418722_AA1368920.1	M7ZZ82	Sucrose-phosphate synthase 1
TRIAE_CS42_3DS_TGACv1_273985_AA0934130.1	A5A8T7	17.6kDa heat-shock protein
TRIAE_CS42_3DS_TGACv1_272328_AA0919130.1	M8ARL2	Sucrose-phosphate synthase 1
TRIAE_CS42_5AS_TGACv1_392807_AA1264980.1	M7ZT04	Germin-like protein 8-4
TRIAE_CS42_4AL_TGACv1_290422_AA0985250.1	M8A472	Elongation factor 2

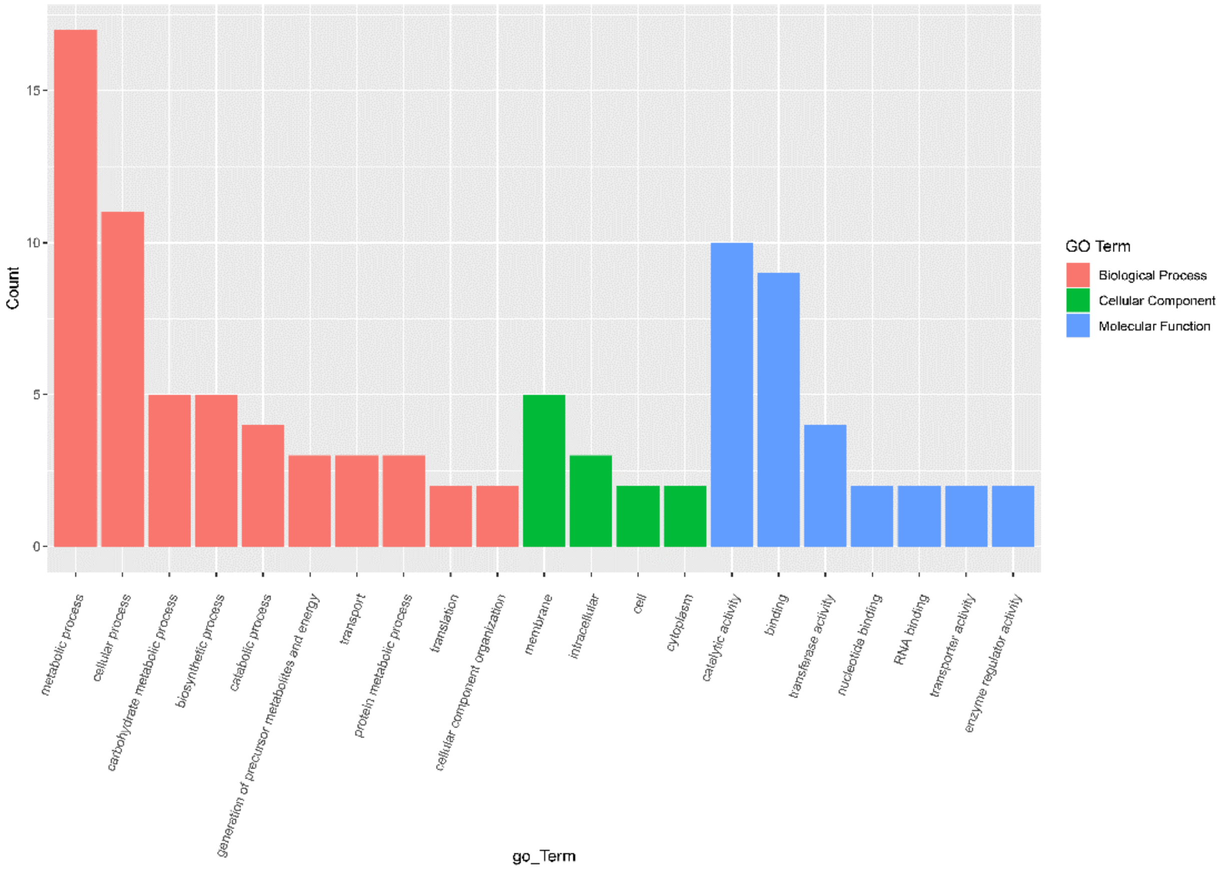


Figure 5.9. Bar graph of gene ontology 'Slims' counts (y-axis) per go summary term (x-axis), derived originally from the 41 RNA transcripts detected from the THH and TARC comparison using the Wald test.

5.4. Discussion

5.4.1. Sample choice

In this analysis, the mRNA expression from Spitfire wheat that was grown at three different farms locations (Breeza, TARC, and THH) was examined. There were three biological replicates per sample group from each farm location. This was done in the hope to correlate proteomic data from Chapters 2 and 4. While appreciating a previous study on grain samples having limited success in comparing proteomic results with transcriptomic results (Barros, et al., 2010), another study was more successful (Garcia-Seco, et al., 2017). The potential knowledge gain and the reality of ever improving technology and software, as well as expanding databases, made this attempt to find a correlation between significantly expressed RNA transcripts and proteins (putative biomarkers) an investigation worth attempting.

5.4.2. Quality of extracted RNA

As for any experiment, a critical factor in a reliable result is the quality of the initial input. This is especially true of next generation RNA sequencing (also known as RNASeq) due to the vast amounts of amplified product that results from the initial input of RNA. Degraded RNA, or an incomplete extraction of RNA will result in an unrepresentative pool. This error will then be magnified by the subsequent amplification of RNA fragments. To avoid this problem, three methods of RNA extraction were investigated. The first, by Li and Trick (2005), reported improvements in avoiding sample solidification (due to starch) and the problem of starch co-precipitating with RNA. The second used the same principle of keeping RNA in separate phases but used a different solubilisation buffer and other chemicals. The third was simply the Tri-Reagent manufacturers (ThermoFisher) recommended protocol. As seen in Figure 5.1, there was little difference in the quality of RNA extracted by the three methods. After reviewing the possible choices described above, the second method was chosen because it had the least RNA degradation.

5.4.3. Initial RNA-Seq data quality

The extracted RNA was sent to Australian Genomic Research Facility (AGRF) for next generation sequencing of the sample mRNA and the results were delivered in the form of FASTQ files which were then further analysed as described in the methods section. All the initial analyses led to the conclusion that the data was of good quality. The FASTQ files were of a suitable and consistent size (approximately 5 Mb), which is a simple but effective marker of quality, and the density plots showed little variation (Figure 5.3). All indications pointed to there not being an issue with the RNA input or the sequencing run. However, the correlation plot and PCA (Figure 5.4 and Figure **5.5**) of each sample's transcript counts did show some variation between samples. The spacing of sample point in the PCA showed greater variation within each of the THH and TARC sample groups and greater similarity between sample groups, especially when compared to Breeza. Although this variation was not excessive, it was something to remain aware of due to the comparative method employed by the R-package (DESeq2) that was used for determining differential expression and assigning significance to the thousands of transcripts detected in each sample. Potential issues with sample variation will be discussed below.

5.4.4. The form of sample data

Another critical point in RNA-Seq analysis is determining whether the data is suitable for the software that will be used to calculate statistically significant differential expression. Numerous prior experimental results had shown that RNA data was generally observed to fit the negative bionomical distribution, and many software packages used to search for differentially expressed RNA transcripts have this as underlying core assumptions. Together with the R-package, we have chosen to use DESeq2. For this experiment, scatter plots of the three sample groups (Figure D.1) and a plot of dispersion estimates (Figure 5.6) all verified that the data presented here does indeed display a negative binomial distribution.

5.4.5. Statistically significant differential expression

Although DESeq2 uses multiple algorithms and is capable of looking at multiple factors within the data, for this experiment DESeq2 was used to determine statistically significant differential expression of transcripts in only two ways as described in the methods section of this chapter. The first was via a comparison of two sample groups (similar to a T-test) to determine differential expression, followed by the use of the likelihood ratio test (LRT) or Wald test to assess significance. The second is a comparison of all three sample groups (ANOVA-like comparison) and can only use the LRT to test significance. Initially, only six combinations of two sample groups each were tested (Table 5.1), with a resulting large variation in differentially expressed transcripts being detected for all six combinations, ranging from a high of 2,350 to a low of 17 (Table 5.1). The higher results (in the thousands) of differentially expressed transcripts, especially 2,350 for Breeza compared to THH, seemed too high to be possible. Current proteomic technology has only enabled the discovery of up to 3,000 different proteins from seeds and only a fraction of these will be differentially expressed between samples due to differing genetics or environmental effects (biotic and abiotic). Moreover, as described in the results section, some of the underlying data is not robust, with inconsistent heat-maps, correlation plots and PCAs of results from comparisons. The exception being the comparison between samples from TARC and THH farms. For these two locations, the samples within their respective sample groups cluster well together in heat-maps, correlation plots and PCAs, although the number of significantly expressed transcripts is low, with only 17 detected using LRT and 41 using the Wald test.

5.4.6. Sample group comparisons

5.4.6.1. Reverse comparisons

The unusual results described in section 5.4.5 were further investigated by looking at the same comparisons in reverse order. A similarly large variation in the number of significantly expressed proteins was found for all (reversed) comparisons except that of THH compared TARC. Between the two sample groups of THH and TARC, the direction of comparison made no difference to the number of differentially expressed transcripts. Whether the comparison was TARC compared to THH, or THH compared to TARC, the result was always 17 when the LRT was used to determine significance and 41 when the Wald test was invoked. Moreover, when comparing these two sample groups, the resulting heat-maps, correlation plots and PCAs (Appendix D, Figure D.9; Figure D.10 and Figure D.11, Parts E and F), all displayed clustering of samples into their sample groups.

5.4.6.2. ANOVA-like comparison

An ANOVA-like comparison of all three sample groups was also performed but resulted in 1,880 significantly expressed transcripts. Even when the adjusted p-value for significance was reduced to 0.0001 the quality of the data quality of these new results did not improve as this did not improve the PCAs, heat-maps and correlation plots (Figure D.12, Figure D.13 and Figure D.14). Indeed, the sample clustering on the graphs looked equal whether the adjusted p-value was set to 0.05 (1,880 significant transcripts detected) or down to 0.0001 (253 transcripts detected).

5.4.6.3. Breeza sample group

The Breeza sample group comparisons results appeared unreliable. Figure 5.8 shows that when the Breeza sample group is involved in comparisons to determine differentiation and significance, the result has either too many false positives or too many false negatives. This can be a result of the variation in the raw transcript data observed in the PCA of Figure 5.5. It is likely that the tight plotting of Breeza samples to each other in the PCA, and the much looser plotting of TARC and THH samples to their own sample groups is unbalancing the underlying data. A similar example of this in much greater detail is given by the DESeq2 authors (http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#if-i-have-multiple-groups-should-i-run-all-together-or-split-into-pairs-of-groups). To this end, because of the good underlying data, only the differentially expressed transcripts resulting from the TARC and THH comparison with the Wald test were further investigated for function and relatedness to putative biomarkers reported in Chapter 2 and 4.

5.4.7. Functional summaries to assist biomarker discovery

Having found 41 transcripts with differential expression, we then investigated whether the proteins that they coded for were related to the proteins that originated from the same sample group tested in Chapter 4 (Spitfire wheat grown at Breeza, THH, and TARC farms). No common proteins between the two studies could be found, however, there was only one with good homology found (chitinase; approximately 95%). The 30 proteins derived from the 41 significant transcripts did diminish the probability of successful matches, considering the small number of putative biomarkers discovered in the proteomics analysis (Table 4.8).

5.4.8. Conclusion

Although differentially expressed transcripts of statistical significance were observed in this experiment, sample expression data variation within sample groups negatively influenced the results. The number of transcripts of interest was either unreliable (in terms of data quality), or, had their numbers reduced due to data noise. Nonetheless, with the correct comparisons, reliable data of statistically significant differential expression of RNA transcripts was obtained and the protein products determined. From functional descriptions derived from UniProt, there were many similarities between the putative biomarkers described in Chapter 4 and the derived proteins from the transcripts of interest found in this chapter. However, when the amino acid sequences were compared (blastp), only one chitenase was found to be homologous (95%) between proteins identified in Chapter 4 and 5.

Chapter 6. Conclusion and Future Directions

6.1. Conclusion

This thesis firstly introduced the history of grasses and their path to the successor crops wheat and barley that we consume today. The challenges in extracting wheat proteins for analysis was been addressed in Chapter 2, with a comparative study validating the ‘combined-phenol’ as the preferred choice for extracting proteins for proteomic analysis. This method gave the best yield and protein diversity. The samples were analysed by unlabelled shotgun proteomics which identified differential expression and the presence/absence of proteins. The optimised protein extraction method was applied in the labelled proteomics analysis of barley and wheat (Chapter 3 and 4, respectively). Despite not being able to detect proteins with presence/absence across proteins groups, the labelled proteomic analysis was worth the gain in the number of potential biomarkers and the higher throughput in processing the grain samples. In both chapters, there was success in identifying many candidate proteins for biomarkers, which will allow for future work to be conducted in how these biomarkers can be applied in the test-kits for barley and wheat through ELISA. Transcriptomics (RNAseq) identified statistically significant differentially expressed mRNA transcripts encoding proteins in the wheat grain transcriptomes of Spitfire cultivar, which showed variation between the three different farm locations. However, no correlation was found between the putative biomarkers identified by transcriptomics compared with TMT-labelled proteomics. In the following section (Future Directions), some potential approaches and considerations for the development of test-kits utilising the discovered protein biomarkers and potential applications for the mRNA analysis of samples will be discussed.

6.2. Future directions

6.2.1. Protein biomarker-based testing

The potential biomarkers identified in this study indicate variably expressed proteins within each cultivar and the same cultivar at different farm locations caused by the specific conditions on that farm at that point in time. The growing conditions will likely change for every consecutive crop grown at that farm and thus the levels of biomarker will most likely change. To address this, a reference sample of the same cultivar for that particular harvest date will be required in order to create a 'fingerprint' using multiple protein biomarkers to match to the reference sample. Cultivars from farms that have had identical or very similar growing conditions such as temperature, soil nutrients, rain and plant pathogens will likely show similar proteomic profiles, which may require additional analysis. This may require an additional 'fingerprint' assay that utilises a different set of biomarkers. The annual change in conditions will require a yearly survey of biomarker expression levels, from samples representing standard locations and cultivars. However, once the appropriate biomarkers have been determined they themselves should not change, and hence the wheel will not need to be reinvented but only maintained.

The selection of potential biomarkers for identifying cultivars is anticipated to be different from the biomarkers used to identify farm location, since the latter are most likely to be related to biotic and/or abiotic stress response. Also, biomarker proteins that have overlap (similar proteins) between the farm location and cultivar identification biomarker proteins may not be very useful for either purpose. Considering the above, the potential biomarkers for cultivar identification are likely to be proteins that are associated with grain characteristics such as endosperm proteins and general house-keeping proteins. Biotic and abiotic stress proteins may also be applied in the cultivar identification as long as they do not overlap with those proteins specific to farm location; especially cultivars that are known to be more resistant to biotic and/or abiotic stress.

Following the results gathered in this thesis, it is suggested the future direction of this project is to develop a test-kit assay that can detect levels of a small number of the most likely candidate putative biomarkers. The test-kits could involve an immuno-chemistry assay-based test, such as an ELISA, or a colour strip test (antibody-based) similar to a pregnancy test. In the case of the ELISA, multiple wells of a 96-well plate can be utilised with each well containing a different antibody generating a specific biomarker fingerprint. A positive result may be where each well in the plate needs to indicate antigen binding. Similarly, the colour strip test method is envisaged to be lined with a number of antibodies and the fingerprint compared to a previously determined reference chart of a particular cultivar or farm. Moreover, as mentioned above, once chosen the biomarker proteins should not need to be changed from year to year, but baseline readings will need to be established annually for all farms and the cultivars they grow. These tests will need to be simple and robust so that someone in the field (such as a farmer) could perform it with the minimal amount of test equipment or laboratory.

6.2.1. RNA biomarker based testing

Investigations into the wheat grain transcriptome have shown that while discovery of statistically significant differential expression of mRNA is possible, the encoded proteins were not matched with the protein biomarkers that were discovered in the proteomics analysis. Hence more work will need to be done since the transcriptomic results were rather limited as only the Spitfire cultivar was analysed. Further work in this approach may provide mRNA sequences for RNA-based assays test-kits that could be developed that target the most promising putative mRNA biomarkers. At the moment, the disadvantage of this approach is that it requires laboratory-based equipment and specially trained staff capable of performing cDNA synthesis, PCR amplification and qPCR. However, with the advancement of sequencing technology, future analysis may be done using a hand-held sequencing device.

Appendix A.

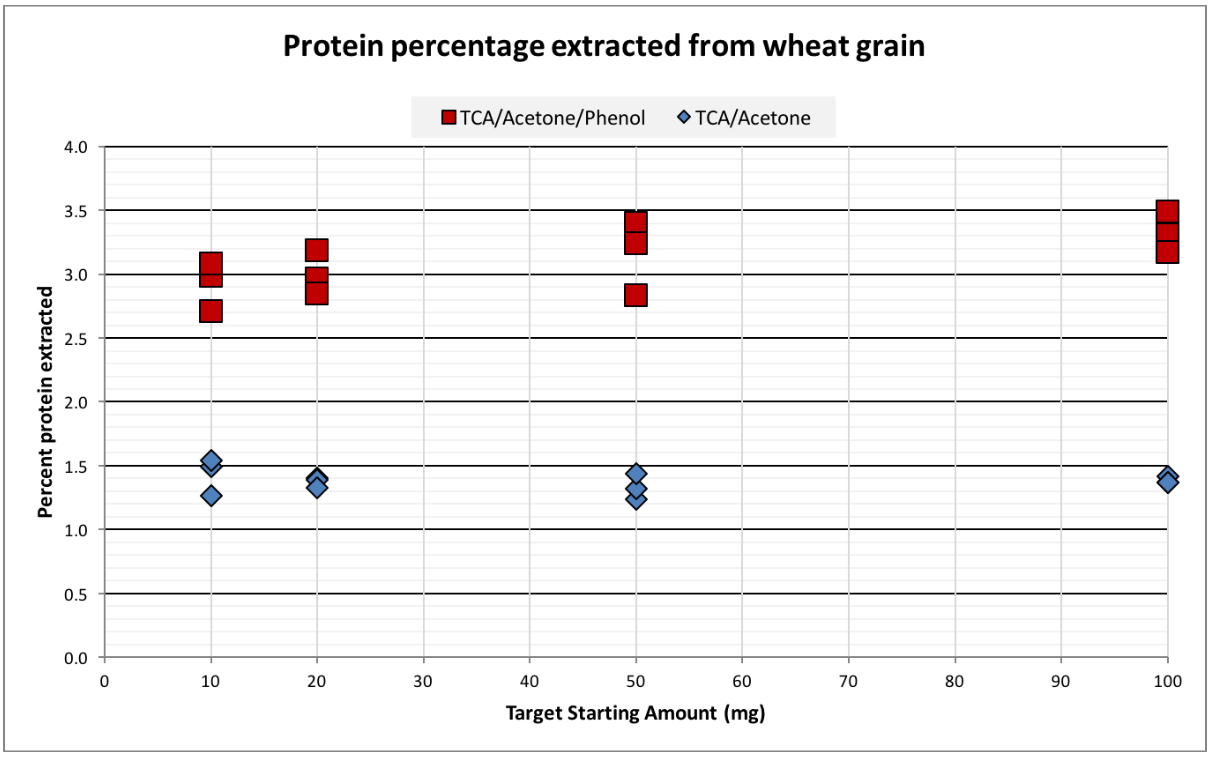


Figure A.1. Percentage of protein extracted from starting amount of grain powder

Appendix B.

B.1 Barley grain weights

Table B.1. 200 grain sample weights for all barley cultivars samples grown at Breeza, TARC, and THH.

Sample Code	Farm Location	Cultivar	Biological Replicate	Harvest Year	Measured 200-weight (g)
Br-Comm-BR1-2013	Breeza	Commander	1	2013	9.151
Br-Comm-BR2-2013	Breeza	Commander	2	2013	9.166
Br-Comm-BR3-2013	Breeza	Commander	3	2013	9.198
Br-Gaird-BR1-2013	Breeza	Gairdner	1	2013	8.907
Br-Gaird-BR2-2013	Breeza	Gairdner	2	2013	8.996
Br-Gaird-BR3-2013	Breeza	Gairdner	3	2013	8.779
TARC-Comm-BR1-2013	TARC	Commander	1	2013	8.797
TARC-Comm-BR2-2013	TARC	Commander	2	2013	8.896
TARC-Comm-BR3-2013	TARC	Commander	3	2013	8.841
TARC-Gaird-BR1-2013	TARC	Gairdner	1	2013	7.857
TARC-Gaird-BR2-2013	TARC	Gairdner	2	2013	7.857
TARC-Gaird-BR3-2013	TARC	Gairdner	3	2013	7.834
TARC-Hind-BR1-2013	TARC	Hindmarsh	1	2013	8.44
TARC-Hind-BR2-2013	TARC	Hindmarsh	2	2013	8.317
TARC-Hind-BR3-2013	TARC	Hindmarsh	3	2013	8.444
THH-Comm-BR1-2013	Terry Hie Hie	Commander	1	2013	7.822
THH-Comm-BR2-2013	Terry Hie Hie	Commander	2	2013	7.812
THH-Comm-BR3-2013	Terry Hie Hie	Commander	3	2013	7.818
THH-Gaird-BR1-2013	Terry Hie Hie	Gairdner	1	2013	8.587
THH-Gaird-BR2-2013	Terry Hie Hie	Gairdner	2	2013	8.573
THH-Gaird-BR3-2013	Terry Hie Hie	Gairdner	3	2013	8.64

Table B.1.1. P-value results, examining 200-weight differences of barley samples between farm locations, based on table B.1.

Comparison (farms)	P-adj
TARC-Breeza	0.009391905
Terry Hie Hie-Breeza	0.003742214
Terry Hie Hie-TARC	0.717354511
ANOVA (overall):	0.002765008

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table B.1.2. P-value results, examining 200-weight differences of barley samples between cultivars, based on table B.1.

Comparison (cultivars)	P-adj
Gairdner-Commander	0.782016199
Hindmarsh-Commander	0.814506869
Hindmarsh-Gairdner	0.989548869
ANOVA (overall):	0.7393875

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table B.1.3. P-value results, examining 200-weight differences between Commander and Gairdner barley cultivars grown at a single location (either Breeza, THH, or TARC).

Comparison (cultivars)	P-adj
Commander-Gairdner	0.04226
Commander-Gairdner	0.0005444
Commander-Gairdner	0.0003966

Note: P-value calculated by t-test. Hindmarsh barley was not included as it was grown only at one location

B.1b Barley extracted protein weights

Table B.1b. Barley - weight of extracted proteins in µg

Sample Code	Farm location	Cultivar	Biological replicate	Harvest Year	Sample protein weight_µg
TARC-Comm-BR1-2013	TARC	Commander	1	2013	1535
TARC-Comm-BR2-2013	TARC	Commander	2	2013	1240
TARC-Comm-BR3-2013	TARC	Commander	3	2013	460
TARC-Gaird-BR1-2013	TARC	Gairdner	1	2013	725
TARC-Gaird-BR2-2013	TARC	Gairdner	2	2013	1660
TARC-Gaird-BR3-2013	TARC	Gairdner	3	2013	1400
TARC-Hind-BR1-2013	TARC	Hindmarsh	1	2013	1715
TARC-Hind-BR2-2013	TARC	Hindmarsh	2	2013	1265
TARC-Hind-BR3-2013	TARC	Hindmarsh	3	2013	1275
THH-Comm-BR1-2013	Terry Hie Hie	Commander	1	2013	1000
THH-Comm-BR2-2013	Terry Hie Hie	Commander	2	2013	1050
THH-Comm-BR3-2013	Terry Hie Hie	Commander	3	2013	740
THH-Gaird-BR1-2013	Terry Hie Hie	Gairdner	1	2013	875
THH-Gaird-BR2-2013	Terry Hie Hie	Gairdner	2	2013	1435
THH-Gaird-BR3-2013	Terry Hie Hie	Gairdner	3	2013	1045
Br-Comm-BR1-2013	Breeza	Commander	1	2013	625
Br-Comm-BR2-2013	Breeza	Commander	2	2013	935
Br-Comm-BR3-2013	Breeza	Commander	3	2013	845
Br-Gaird-BR1-2013	Breeza	Gairdner	1	2013	690
Br-Gaird-BR2-2013	Breeza	Gairdner	2	2013	1065
Br-Gaird-BR3-2013	Breeza	Gairdner	3	2013	810

Table B.1b.1. P-value results, examining protein sample weight (µg) differences of barley samples between farm locations, based on table B.2.

Comparison (farms)	P-adj
TARC-Breeza	0.049883874
Terry Hie Hie-Breeza	0.540867339
Terry Hie Hie-TARC	0.374182754
ANOVA (overall):	0.05863339

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table B.1b.2. P-value results, examining protein sample weight (µg) differences of barley samples between cultivars, based on table B.2.

Comparison (cultivars)	P-adj
Gairdner-Commander	0.636321
Hindmarsh-Commander	0.097082
Hindmarsh-Gairdner	0.288946
ANOVA (overall):	0.1149416

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table B.1b.3. P-value results, examining protein sample weight (µg) differences between Commander and Gairdner cultivars grown at a single location (either Breeza, THH, or TARC).

Comparison (cultivars)	P-adj
Commander-Gairdner (Breeza)	0.7303
Commander-Gairdner (THH)	0.3938
Commander-Gairdner (TARC	0.6887

Note: P-value calculated by t-test.

Peptide sample weights: P-values

Table B.1b.4. P-value results of peptide sample weight (µg) differences between Commander and Gairdner cultivars grown at a single location (either Breeza, THH, or TARC).

Comparison (cultivars)	P-adj
Commander-Gairdner (Breeza)	0.2999
Commander-Gairdner (THH)	0.7286
Commander-Gairdner (TARC	0.269

Note: P-value calculated by t-test. Hindmarsh barley was not included as it was grown only at one site.

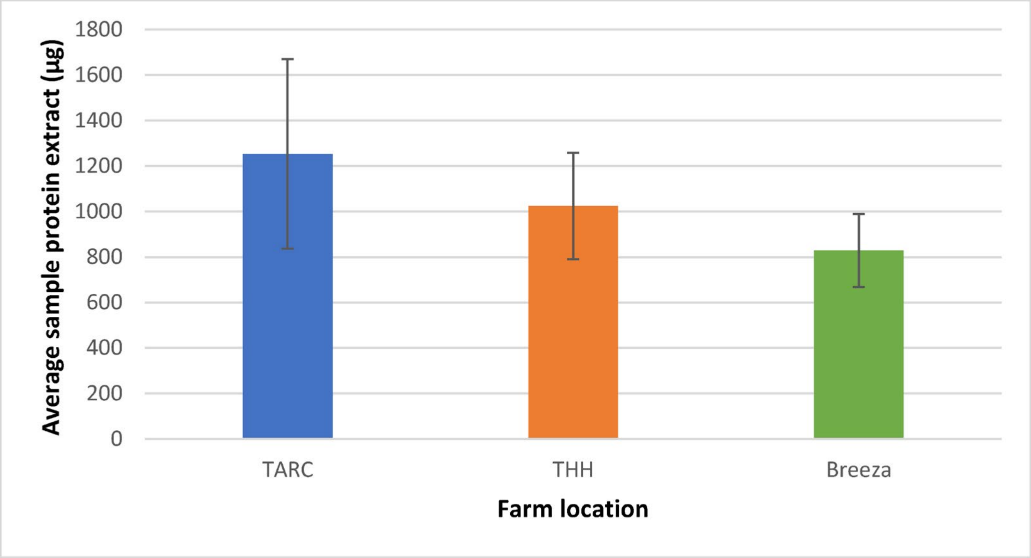


Figure B.a. The average weight of sample protein extract for barley grain samples grown at TARC, THH, and Breeza farms (error bars are ± standard deviation).

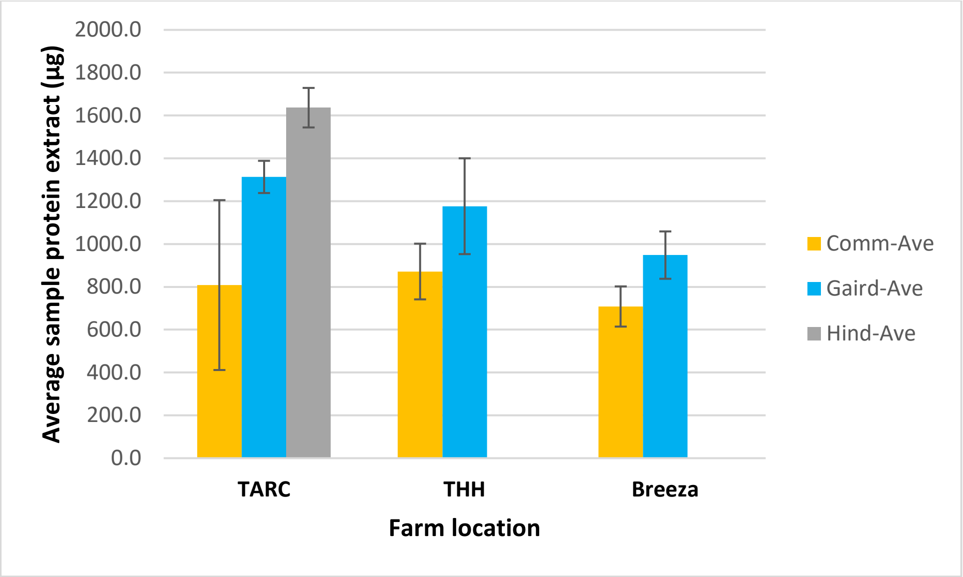
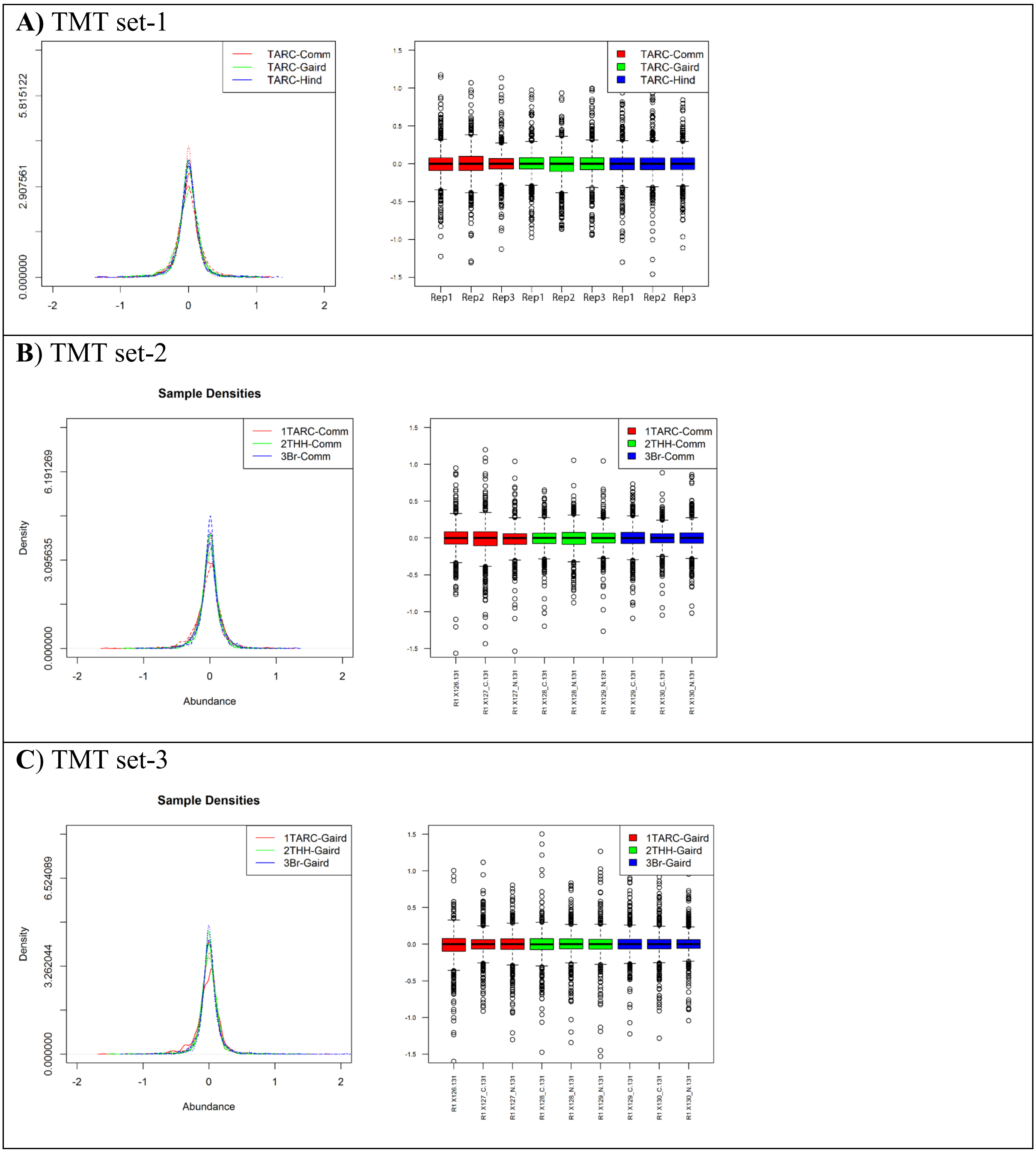


Figure B.b. The average weight of sample protein extracts for Commander, Gairdner, and Hindmarsh barley cultivars grown at TARC, THH, and Breeza farms (error bars are ± standard deviation).

B.2 Density plots and box plots of initial protein expression data


A) TMT set-1

B) TMT set-2

C) TMT set-3

Figure B.1. Density plots of matched and filtered sample proteins from barley: A = TMT set-1, B = TMT set-2, C = TMT set-3.

B.3 Data summary for TMT sets 1 to 3: matched and filtered proteins.

B.3.1 TMT set-1 matched and filtered protein data

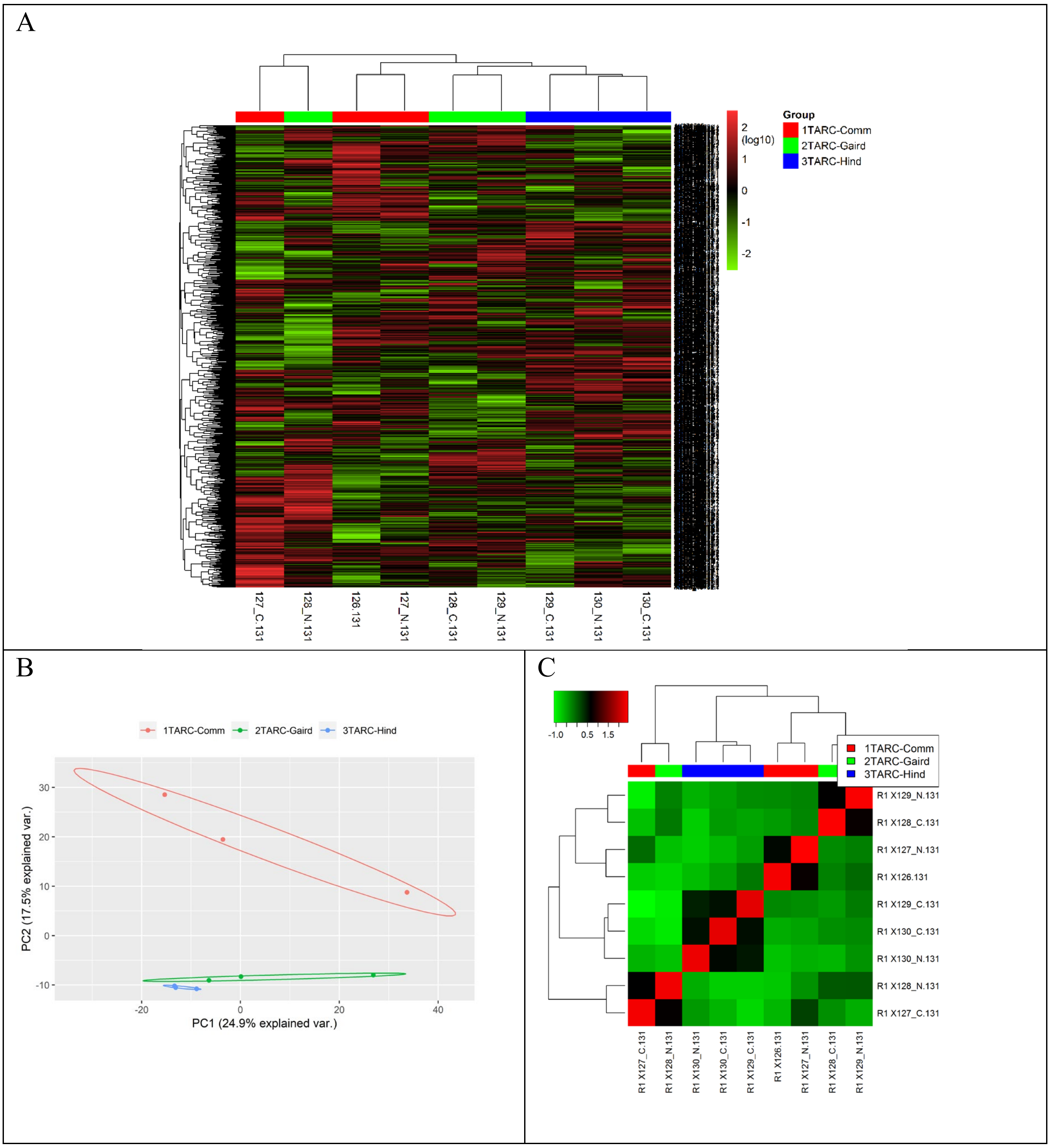


Figure B.2. (A) Heatmap, (B) PCA, and (C) Correlation Plot summaries for TMT set-1 consisting of matched and filtered protein expression (before putative biomarker discovery) from barley grain samples.

B.3.2 TMT set-2 matched and filtered protein data

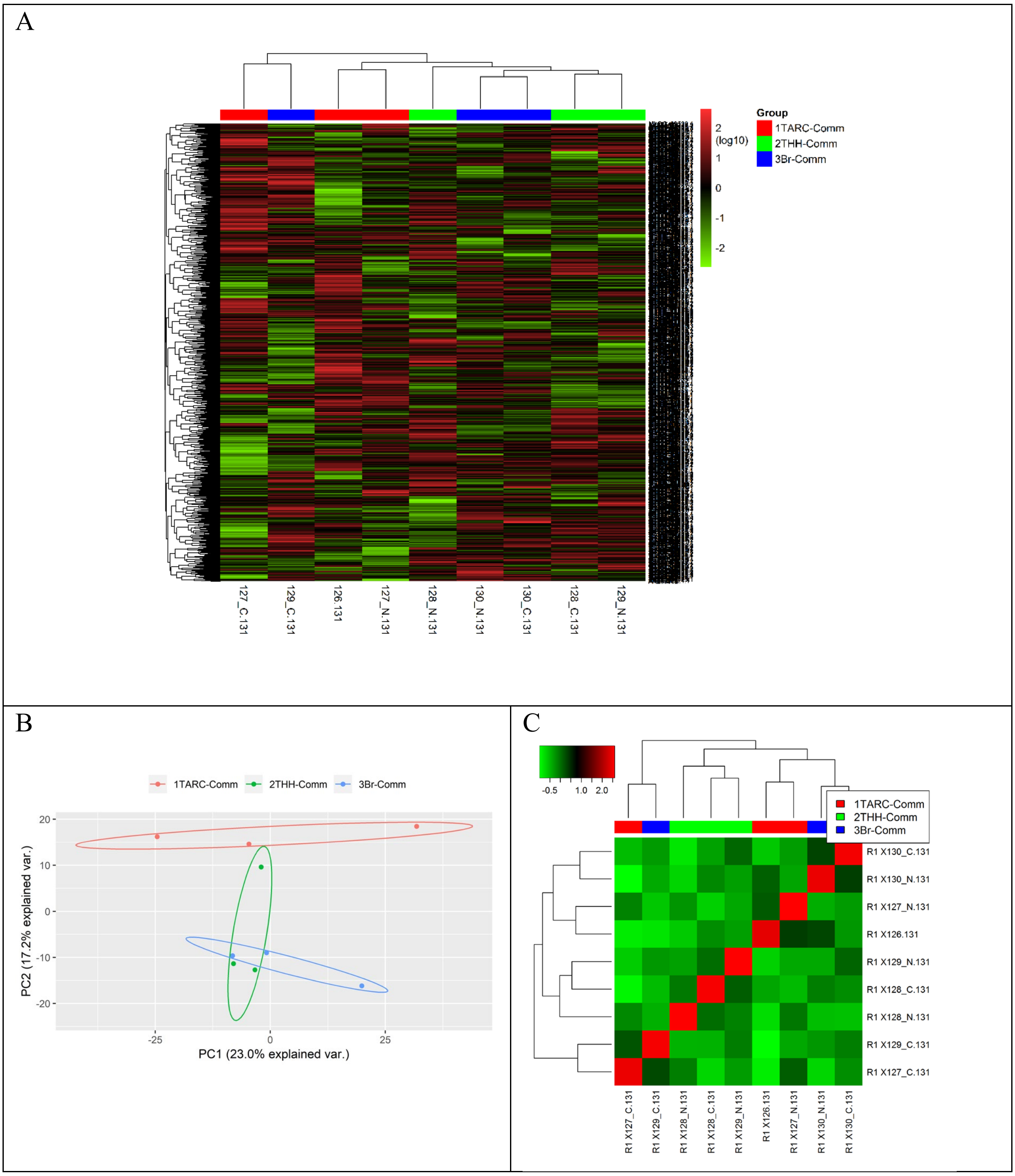


Figure B.3. (A) Heatmap, (B) PCA, and (C) Correlation Plot summaries for TMT set-2 sample data consisting of matched and filtered protein expression (before putative biomarker discovery) from barley grain samples.

B.3.3 TMT set-3 matched and filtered protein data

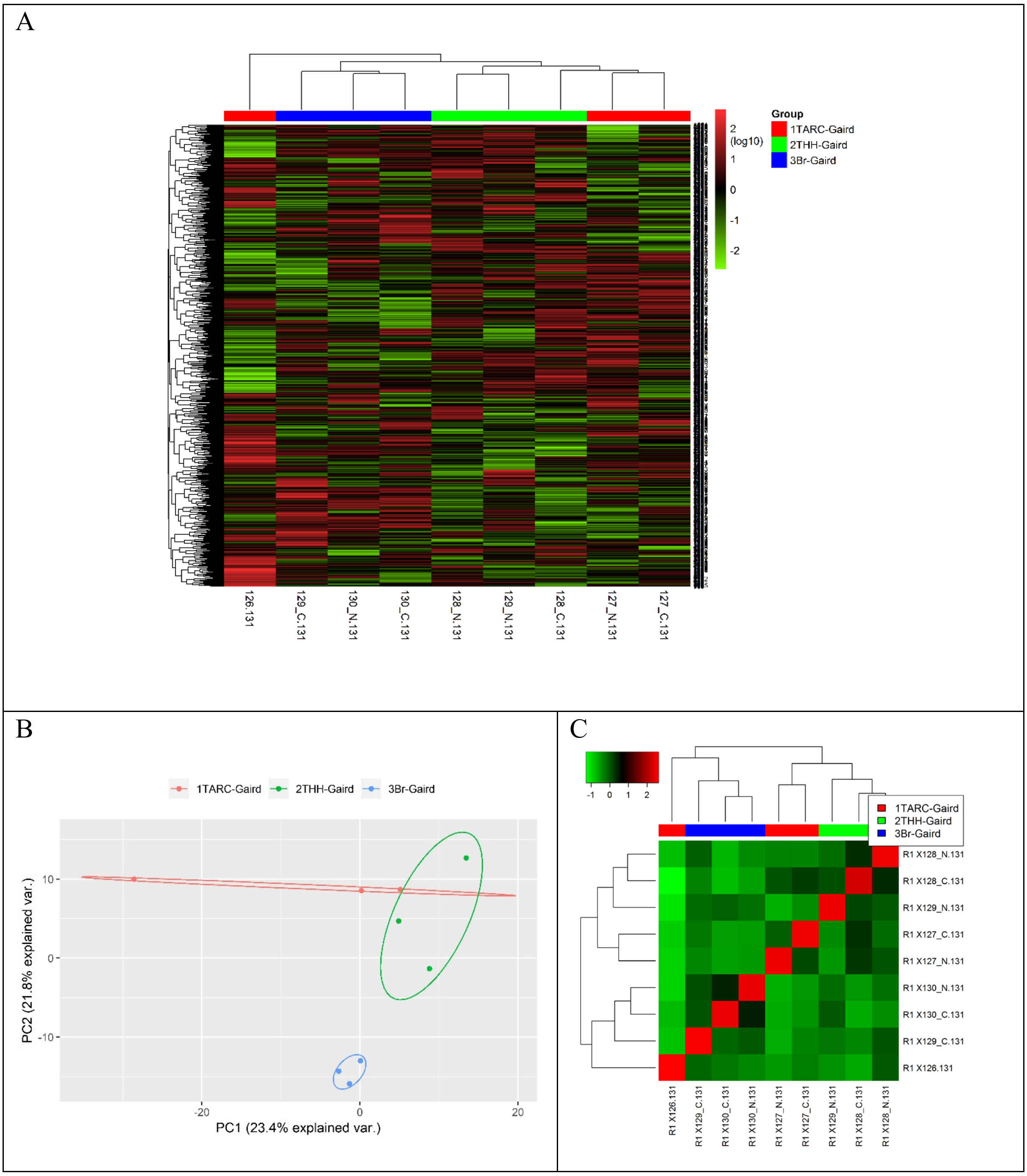


Figure B.4. (A) Heatmap, (B) PCA, and (C) Correlation Plot summaries for TMT set-3 sample data consisting of matched and filtered protein expression (before putative biomarker discovery) from barley grain samples.

B.4 Data summary for TMT sets 1 to 3: putative biomarkers.

B.4.1 TMT set-1: Putative biomarker proteins

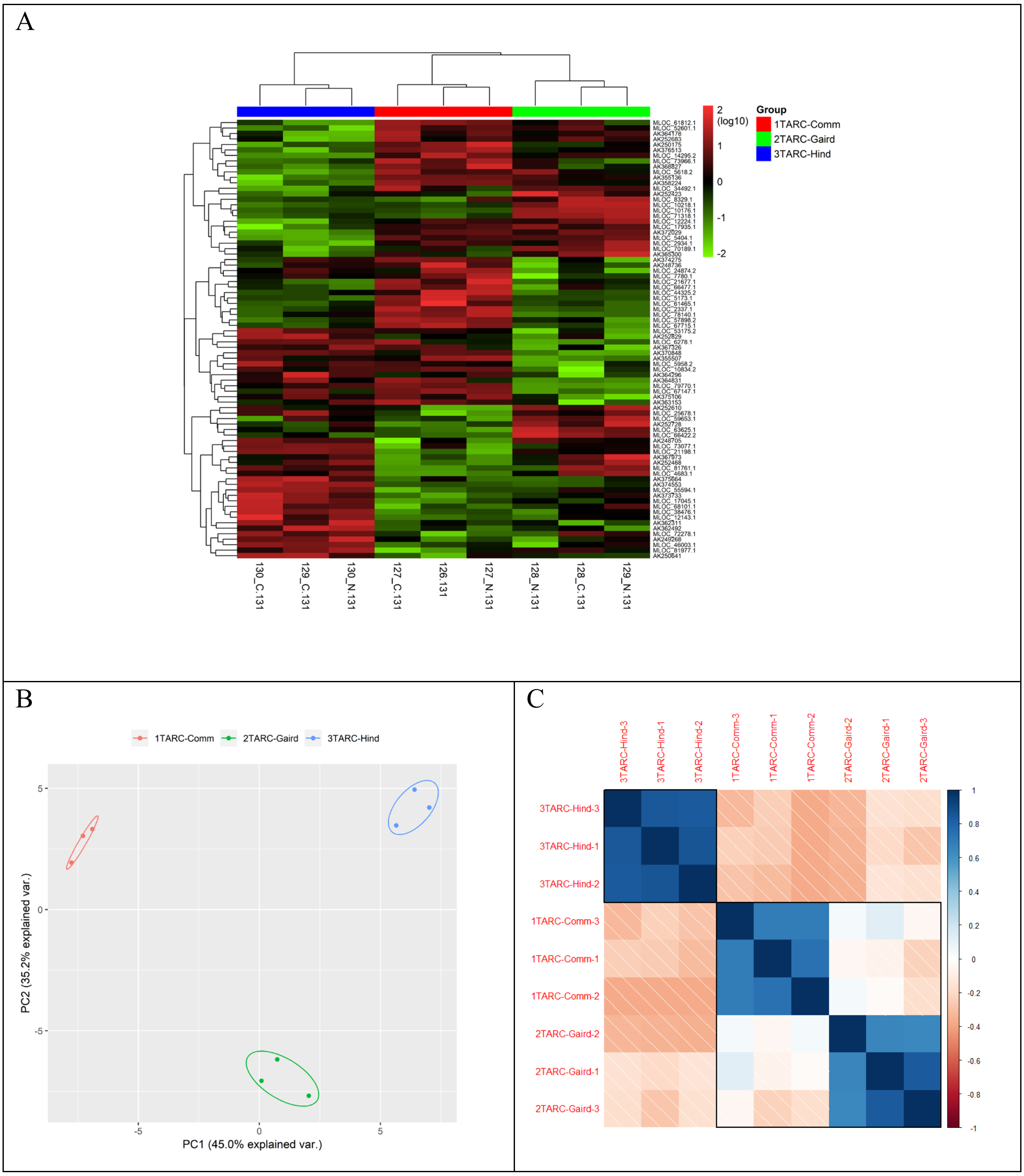


Figure B.5. Heatmap and PCA summaries of TMT set-1 sample data consisting of putative biomarker protein expression.

B.4.2 TMT set-2: Putative biomarker proteins

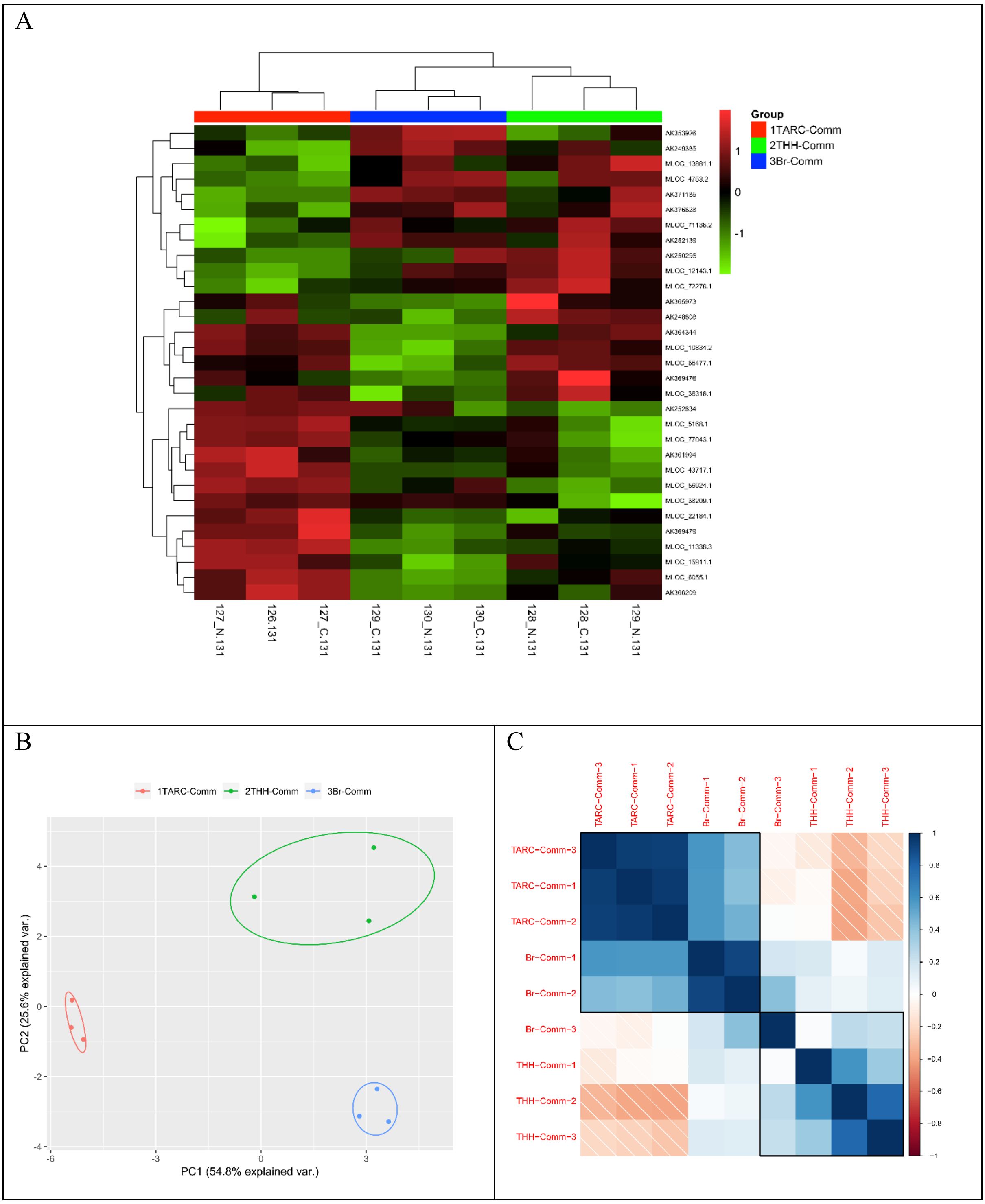


Figure B.6. Heatmap and PCA summaries of TMT set-2 sample data consisting of putative biomarker protein expression extracted from barley grain samples.

B.4.3 TMT set-3: Putative biomarker proteins

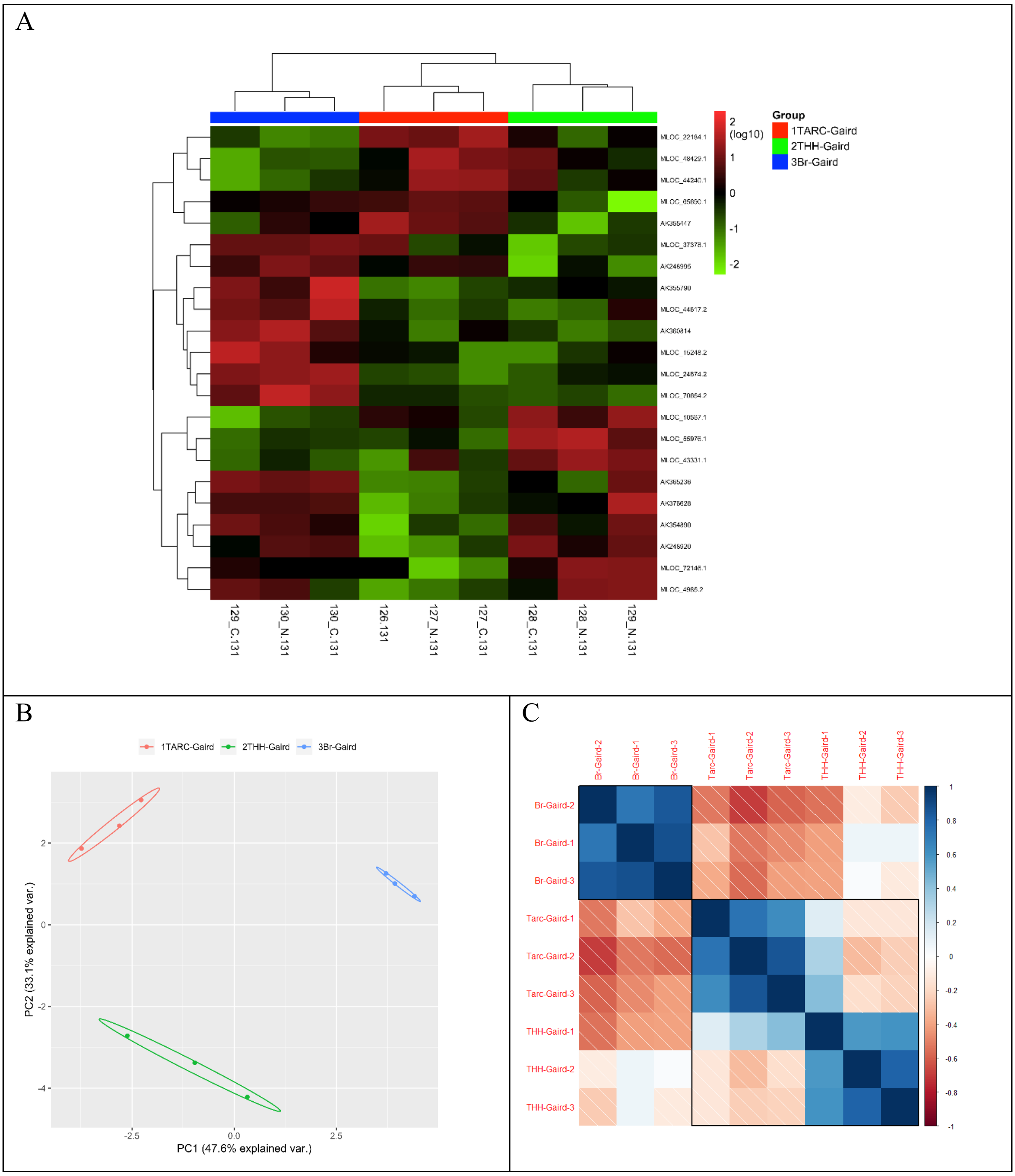


Figure B.7. Heatmap and PCA summaries of TMT set-3 sample data consisting of putative biomarker protein expression extracted from barley grain samples.

Table B.1. TMT set-3: List of putative biomarker proteins from barley grain

Barley fasta ID	Uniprot	Protein Description
MLOC_24874.2	M0VEJ0	Late embryogenesis abundant protein
AK360814	F2DAA1	PEBP family protein
MLOC_15248.2	F2EG62	Succinate dehydrogenase iron sulfur subunit
MLOC_70664.2	F2D961	Lipoxygenase
MLOC_37378.1	M0VUI3	Lipoxygenase
AK355790	F2CVY7	Phosphorylase
MLOC_44617.2	M0W7R3	Nucleolar protein 5
AK248995	Q40004	Ribulose bisphosphate carboxylase small chain
AK354890	F2CTD8	Defensin
AK248920	P34893	10 kDa chaperonin
AK365236	F2DMW8	UPF0061 protein
AK376628	F2EKF1	Defensin D2
MLOC_72146.1	M0YRS3	UDP glycosyltransferase
MLOC_4986.2	M0WDU5	Amine oxidase
MLOC_22184.1	F2CTI9	Microtubule associated protein family protein putative expressed
MLOC_65690.1	F2E6F3	Coatomer subunit gamma
AK355447	F2CUZ5	Glycogen synthase
MLOC_48429.1	M0WCD7	rRNA N glycosidase
MLOC_44240.1	M0W6F2	Acidic endochitinase
MLOC_55976.1	M0X060	Adenine nucleotide alpha hydrolases like protein
MLOC_10567.1	M0UFT6	Inosine 5' monophosphate dehydrogenase
MLOC_43331.1	M0W433	Carbonic anhydrase

B.5 GO Slims – Summary of functionality for putative biomarker proteins

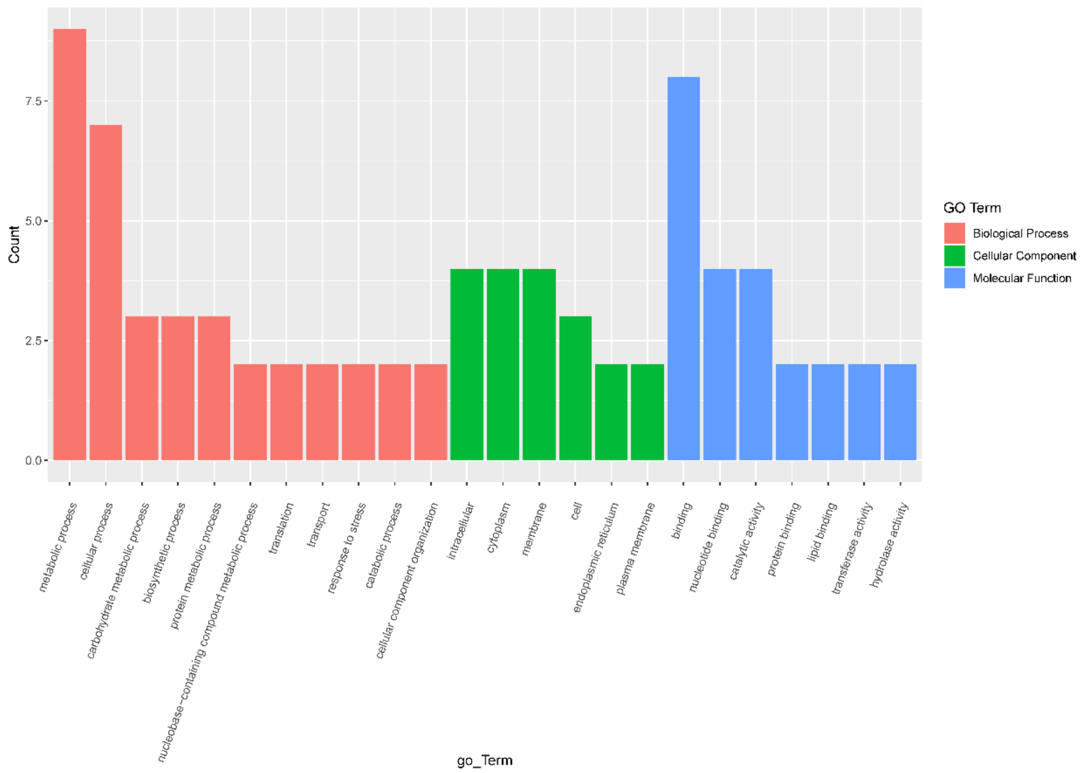


Figure B.8. TMT set-2 GO-Slims functional summary of putative biomarkers discovered (proteome comparison across cultivars). Only data with a count of 2 or more are included.

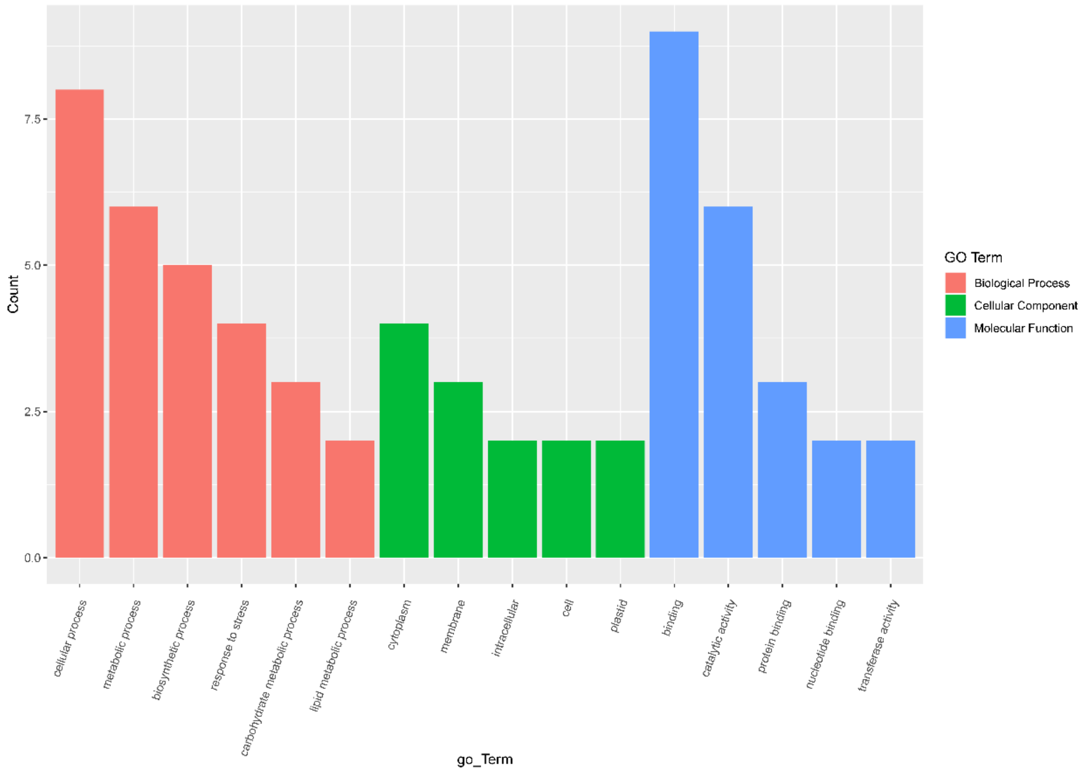


Figure B.9. TMT set-3 GO-Slims functional summary of putative biomarkers discovered (proteome comparison across cultivars). Only data with a count of 2 or more are included.

B.6 R-scripts mentioned in Chapter 3

B.6.1 ‘PCA.R

```
# BiocManager::install("UniProt.ws", version = "3.8")
# # Install "ggbiplot" if not already installed
# install_github("vqv/ggbiplot")
options(scipen=999)
if(.Platform$OS.type == "windows"){
      Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jdk1.8.0_191')
}
library("UniProt.ws")
library("pheatmap")
library("tidyr")
library("rJava")
detach("package:rJava", unload=TRUE)
library("ggbiplot")
library("dplyr")
library("RColorBrewer")
library("gridExtra")
library("grid")
library("tcltk")
library("rChoiceDialogs")
library("ggplot2")
library("readxl")
```

```
#################### SAVE HEATMAP Function ####################
#Save pheatmap function
save_pheatmap <- function(x, filename, width=1500, height=800) {
      stopifnot(!missing(x))
      stopifnot(!missing(filename))
      png(filename = filename, width = width, height=height)
      grid::grid.newpage()
      grid::grid.draw(x$gtable)
      dev.off()
}

unfactorize <- function(df){
      for(i  in  which(sapply(df,  class)  ==  "factor"))  df[[i]]  =
as.character(df[[i]])
      return(df)
}

####!Function to split a dataframe of "AK" and "MLOC" barley identifiers
into two vectors
splitBarleyIds = function(idInput){
      #Find only the idInput beginning with "AK"
      akIdentifiers  =  as.character(idInput[grep("AK[0-9]*|AK[0-9]*\\..*",
idInput[[1]]), ])
      #If you need to you can remove the decimal point from the Identifier
      akIdentifiers = gsub("(AK.*)(\\.[0-9]*$)", "\\1", akIdentifiers)
      #Find only the identifiers beginning with "MLOC"
      mlocIdentifiers   =   as.character(idInput[grep("MLOC_[0-9]*|MLOC_[0-
9]*\\..*", idInput[[1]]), ])
      #Remove the decimal point from the Identifier
      #mlocIdentifiers = gsub("(AK.*)(\\.[0-9])", "\\1", mlocIdentifiers)
      allList  =  list(akIdentifiers  =  akIdentifiers,  mlocIdentifiers  =
mlocIdentifiers)
      return(allList)
}


getUniprotFromMlocAk = function(idsList){
      #Load the "UniProt.ws" package into R
      library("UniProt.ws")
      #Get "ak" identifers from the list named "idsList"
      #From "splitBarleyIds" function
      akIdentifiers = unlist(idsList["akIdentifiers"])
      #Get "MLOC" identifers from the list named "idsList"
      #From "splitBarleyIds" function
      mlocIdentifiers = unlist(idsList["mlocIdentifiers"])
      #Set the Taxon number for Barley (Wheat = 4565)
      speciesId <- UniProt.ws(taxId=112509)
      #Key Type or Database the program looks into for "AK" identifiers
      ak_kt = "EMBL/GENBANK/DDBJ"
      #Key Type or Database the program looks into
      mloc_kt = "ENSEMBL_GENOMES PROTEIN"
      #Data columns that will be output
      columns <- "UNIPROTKB"
      #The command to retrive UniProt Identifiers from "AK" Identifiers (if
they exist)
      akRetrieve  <-  UniProt.ws::select(speciesId,  akIdentifiers,  columns,
ak_kt)
      #Change the name of the first column in the akRetrieve data.frame
```

```
     names(akRetrieve)[1] = "Identifier_Input"
     #The command to retrive UniProt Identifiers from "MLOC" Identifiers
(if they exist)
     mlocRetrieve = UniProt.ws::select(speciesId, mlocIdentifiers, columns,
mloc_kt)
     #Change the name of the first column in the mlocRetrieve data.frame
     names(mlocRetrieve)[1] = "Identifier_Input"
     #Join the tables together
     retrieveAll = rbind(akRetrieve, mlocRetrieve)
     return(retrieveAll)
}

convertBarleyIds = function (AllIdsColumn) {
     idInput = AllIdsColumn
     #List of MLOC and AK Identifiers
     idsList = splitBarleyIds(idInput)
     #Get Uniprot Ids from MLOC and AK Identifier List
     #Result is a data.frame
     idsListSplit = getUniprotFromMlocAk(idsList)
     #Find the rows of the data.frame that have NULL (missing) values in
UniProt column
     getIdsWithMissingUniprot                                        =
idsListSplit[is.na(idsListSplit$UNIPROTKB),]
     #Remove the UniProt column and turn the AK and MLOC Ids into a character
vector
     getVectorOfMissingIds = getIdsWithMissingUniprot$Identifier
     if (length(getVectorOfMissingIds) > 0) {
          #Make a small FASTA file from the vector of AK and MLOC Ids
          mkSmallFASTAList  =  makeSmallFASTAList(getVectorOfMissingIds)
#Input a vector of characters ("Identifiers")
          #Use the "mkSmallFASTA" list in memory to perform a BLASTP
          #The result is a dataframe of MLOC and AK Identifiers in one
column
          #UniProt identifiers in the other column
          blastpTable = blastpResult(mkSmallFASTAList)
          # Remove the descriptions and keep identifiers
          blastpTable$Identifier_Input   =   gsub("(MLOC_[0-9]*|MLOC_[0-
9]*\\.[0-9]*|AK[0-9]*|AK[0-9]*\\.[0-9]*)(_)(.*)",

                                        "\\1",
blastpTable$Identifier_Input)
          # Remove any decimal numbers from identifiers so the vector will
match
          blastpTable$Identifier_Input   =   gsub("(AK[0-9]*)\\.[0-9]*",
"\\1", blastpTable$Identifier_Input)
          # The blastpTable and idsListSplit are combined, giving a full
list of Uniprot and AK, MLOC Identifiers
          tableOfIdsAndBlast    =    idsListAndBlastptable(blastpTable,
idsListSplit)
          # Remove any rows with missing data
          # tableOfIdsAndBlast = na.omit(tableOfIdsAndBlast)
          return(tableOfIdsAndBlast)
     }
     tableOfIdsAndBlast = idsListSplit
     return(tableOfIdsAndBlast)
}


getBarleyDescription = function(barleyIds, proteinFastaFile) {
     library("seqinr")
     #Load the "fasta" file into a list format via the "seqinr" package
```

```
      fastaFile <- read.fasta(file = proteinFastaFile, seqtype = "AA",
as.string = TRUE)
      # Seqinr function "getName" to get sequence names
      seqNames = getName(fastaFile)
      #Tidy up the names so that "getAnnot" will work well
      listNamesFasta = gsub("(MLOC_[0-9]*|MLOC_[0-9]*\\.[0-9]*|AK[0-
9]*|AK[0-9]*\\.[0-9]*)(_)(.*)", "\\1", names(fastaFile))
      names(fastaFile) = listNamesFasta
      trimmedFasta = fastaFile[charmatch(barleyIds, names(fastaFile))]
      #Get annotations from each listed item (protein)
      trimmedFastaAnnot = getAnnot(trimmedFasta)
      #Remove any NULL entries from the "trimmedFastaAnnot" list
      trimmedFastaAnnot = trimmedFastaAnnot[!sapply(trimmedFastaAnnot,
is.null)]
      trimmedFastaAnnot = unlist(trimmedFastaAnnot)
      trimmedFastaAnnot = gsub("(MLOC_[0-9]*|MLOC_[0-9]*\\.[0-9]*|AK[0-
9]*|AK[0-9]*\\.[0-9]*)(_)(.*)", "\\1\\|\\3", trimmedFastaAnnot)
      trimmedFastaAnnot = gsub(">", "", trimmedFastaAnnot)
      trimmedFastaAnnot = strsplit(trimmedFastaAnnot, split = "\\|", fixed =
FALSE, perl = FALSE, useBytes = FALSE)
      ##### Function Call #####
      finalIdAndDescriptionDF = getDescriptionDF(trimmedFastaAnnot)
      #####
      finalIdAndDescriptionDF["description"] =
gsub(".*(unknown.protein).*|.*(unknown.function).*", "Unknown protein",
ignore.case = TRUE, finalIdAndDescriptionDF$description)
      return(finalIdAndDescriptionDF)
}


getWheatDescription = function(wheatIds, proteinFastaFile){
      library("seqinr")
      #Load the "fasta" file into a list format via the "seqinr" package
      fastaFile <- read.fasta(file = proteinFastaFile, seqtype = "AA",
as.string = TRUE)
      # Seqinr function "getName" to get sequence names
      seqNames = getName(fastaFile)
      #Tidy up the names so that "getAnnot" will work well
      listNamesFasta = gsub("^.*\\|(.*)\\|.*", "\\1", names(fastaFile))
      names(fastaFile) = listNamesFasta
      trimmedFasta = fastaFile[charmatch(wheatIds, names(fastaFile))]
      #Get annotations from each listed item (protein)
      trimmedFastaAnnot = getAnnot(trimmedFasta)
      #Remove any NULL entries from the "trimmedFastaAnnot" list
      trimmedFastaAnnot = trimmedFastaAnnot[!sapply(trimmedFastaAnnot,
is.null)]
      trimmedFastaAnnot = unlist(trimmedFastaAnnot)
      idList = gsub("^.*\\|(.*)\\|.*", "\\1", trimmedFastaAnnot)
      trimmedFastaDesc = gsub("(^.*\\|.*\\|)(.*) OS=.*$", "\\2",
trimmedFastaAnnot)
      trimmedFastaDesc = gsub("^[A-Z|0-9]*_[A-Z|0-9]* ", "",
trimmedFastaDesc)
      #trimmedFastaDesc = gsub(" ", "_", trimmedFastaDesc)
      finalIdAndDescriptionDF = data.frame(idList, trimmedFastaDesc)
      names(finalIdAndDescriptionDF) = c("Identifier_Input", "description")
      return(finalIdAndDescriptionDF)
}

#--------------------- END FUNCTIONS ---------------------#
#-------------------------------------------------------#
```

```
if(.Platform$OS.type == "windows"){
      designPath = tk_choose.files(default =
"C:/Users/paul_/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall",
caption = "Select the heatmap design",

            multi = FALSE, filters = NULL, index = 1)
} else {
      designPath = tk_choose.files(default =
"~/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall", caption =
"Select the heatmap design",

            multi = FALSE, filters = NULL, index = 1)
}

if(.Platform$OS.type == "windows"){
      dataTablePath = tk_choose.files(default =
"C:/Users/paul_/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall",
caption = "Select the table of values for heatmap",

                       multi = FALSE, filters = NULL, index = 1)
} else {
      dataTablePath = tk_choose.files(default =
"~/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall", caption =
"Select the table of values for heatmap",

                       multi = FALSE, filters = NULL, index = 1)
}

if(.Platform$OS.type == "windows"){
      fastaDatabaseFile = tk_choose.files(default =
"C:/Users/paul_/Google_Drive/PhD/DataBases/Protein/Databases_used_in_Mascot
_Search", caption = "FASTA database path",

                       multi = FALSE, filters = NULL, index = 1)
} else {
      fastaDatabaseFile = tk_choose.files(default =
"~/Google_Drive/PhD/DataBases/Protein/Databases_used_in_Mascot_Search",
caption = "FASTA database path",

                       multi = FALSE, filters = NULL, index = 1)
}

if(.Platform$OS.type == "windows"){
      SlimsSummaryAllFile = tk_choose.files(default =
"C:/Users/paul_/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall/Sl
ims_Summary_Outputfiles", caption = "Full Slims table (fullSlimsTable.csv)",

                                  multi = FALSE, filters = NULL, index = 1)
} else {
      SlimsSummaryAllFile = tk_choose.files(default =
"~/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall/Slims_Summary_O
utputfiles", caption = "Full Slims table (fullSlimsTable.csv)",

                                  multi = FALSE, filters = NULL, index = 1)
}

studyDirPath = dirname(dataTablePath)
studyName = gsub(".*/(.*)$", "\\1", studyDirPath)
outDir = paste0(studyDirPath, "/Pheatmap_Outputfiles")
```

```
if (!file.exists(outDir)) {
      dir.create(outDir)
} else {
      print("File Exists!")
}


designFile = gsub("^.*\\/(.*)\\.xlsx$", "\\1", designPath)
dataTableFile = gsub("^.*\\/(.*)\\.xlsx$", "\\1", dataTablePath)


TMTDesign = read_excel(path = designPath, sheet = "design", range =
cell_cols("A:C"))
TMTDesign = TMTDesign[colSums(!is.na(TMTDesign)) > 0]



dataTable = read_excel(path = dataTablePath, sheet = "AllData", range =
cell_cols("A:O"))
dataTable = dataTable[order(dataTable$Clusters),]
# dataTable = na.omit(dataTable)
names(dataTable)[1] = "Identifier_Input"






#### Determine whether the identifiers are Wheat or Barley and start the
initial tidy of data
prepPro_AllIds = dataTable
if (any(grepl("^MLOC_|^AK[0-9]*", prepPro_AllIds[[1]]))) {
      species = "Barley"
      proteinFastaFile = fastaDatabaseFile
} else {
      species = "Wheat"
      proteinFastaFile = fastaDatabaseFile
}
print(species)
#prepPro_AllIds = prepPro_AllIds[order(prepPro_AllIds$Clusters),]
########## Make a dataframe of IDs only ##########
AllIdsColumn = prepPro_AllIds[1]
names(AllIdsColumn) = "Identifier_Input"
# Original Identifiers and then add uniprot Ids plus descriptions
start.time <- Sys.time()
if (species == "Barley") {
      SlimsSummaryAllsub = read.csv(SlimsSummaryAllFile, stringsAsFactors =
FALSE)
      SlimsSummaryAllsub = SlimsSummaryAllsub[ , c("Identifier_Input",
"uniprot", "description")]
      Identifier_Input = AllIdsColumn
      Identifier_Input$Identifier_Input = gsub("(^AK.*)\\..$", "\\1",
Identifier_Input$Identifier_Input)
      idsTable = left_join(Identifier_Input , SlimsSummaryAllsub, by =
"Identifier_Input")
} else if (species == "Wheat") {
      uniprot = AllIdsColumn
      names(uniprot) = "uniprot"
      idsTable = cbind(AllIdsColumn, uniprot)
      Identifier_Input = idsTable$Identifier_Input
      descriptionTable = getWheatDescription(Identifier_Input,
proteinFastaFile)
      idsTable$description                                         =
descriptionTable$description[match(idsTable$Identifier_Input,
descriptionTable$Identifier_Input)]
}
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken



sampleNames = names(dataTable)[grep("^R1.X(.*$)", names(dataTable))]
comparisonLabel = unique(gsub("^.*\\.(.*$)", "\\1", sampleNames))
# Save the row to be deleted if needed
TMTDesignLabelRow = TMTDesign[grep(comparisonLabel, TMTDesign$Label), ]
TMTDesign = TMTDesign[grep(comparisonLabel, TMTDesign$Label, invert = TRUE),
]


sampleDF = data.frame(Sample = sampleNames, stringsAsFactors = FALSE)
sampleDF$Label = gsub("^R1.X(.*$)", "\\1", sampleDF$Sample)
sampleDF$Label = gsub("(^.*)\\..*$", "\\1", sampleDF$Label)

fullDesign = TMTDesign %>% full_join(sampleDF, by = "Label")
fullDesign = fullDesign[ ,c(ncol(fullDesign), 1:3)]
fullDesign[ , "Sample"] = gsub("^R1\\.X(.*)$", "\\1", fullDesign$Sample)


                                              # #This is the full design to
aid in making other plots
                                              #      #such as violin plots
                                              # fullDesign = read.csv(file =
designPath, header = TRUE, stringsAsFactors = FALSE)
                                              #    names(fullDesign)[1]    =
"Sample"
                                              #  fullDesign[  ,  "Sample"]  =
gsub("^R1\\.X(.*)$", "\\1", fullDesign$Sample)



#The design to be used for making a matrix for the heatmap
Design = fullDesign
rowNamesDesign = Design$Sample
                  rowNamesDesign = gsub("^R1.X(.*$)", "\\1", rowNamesDesign)
                  Design[,"Sample"] = NULL
                  rownames(Design) = rowNamesDesign


#Merge datatable and dataTable to get identifiers
dataTable$Identifier_Input       =       gsub("(^AK.*)\\..$",        "\\1",
dataTable$Identifier_Input)
dataTable = dataTable %>% full_join(idsTable, by = "Identifier_Input")
idsAndDescriptions = dataTable[ ,c(1,ncol(dataTable))]
                  names(dataTable)     =     gsub("^R1.X(.*$)",      "\\1",
names(dataTable))
namesWanted = names(dataTable[c(1,ncol(dataTable))])
dataTable = dataTable[ ,c(namesWanted, rowNamesDesign)]
dataTable = unfactorize(dataTable)
dataTable = na.omit(dataTable)
dataTableNums = dataTable[,rowNamesDesign]
#dataTable = cbind(idsAndDescriptions, dataTableNums)
```

```
write.csv(dataTable, file = paste0(outDir, "/All_", dataTableFile,
"_dataTable.csv"), row.names = FALSE)
dataTableWriteLog = cbind(dataTable[ , c(1,2)], log10(dataTable[ ,
c(3:11)]))
write.csv(dataTableWriteLog, file = paste0(outDir, "/All_", dataTableFile,
"_dataTable_log10_All.csv"), row.names = FALSE)

# dataTable = read.csv(file = dataTablePath, row.names = 1, header = TRUE,
stringsAsFactors = FALSE)
dataTable.log.ids = log10(dataTable[ ,rowNamesDesign])
row.names(dataTable.log.ids) = dataTable$Identifier_Input

### Annotate our heatmap (optional)
annotation <- data.frame(Group = Design[,"Group"],
row.names=row.names(Design))

# Reorder Density levels
annotation$Group = factor(annotation$Group, levels =
unique(annotation$Group))
choiceColours = c("red", "green", "blue", "yellow", "orange", "purple")

AnnCol = choiceColours[1:(length(levels(annotation$Group)))]
names(AnnCol) <- levels(annotation$Group)


AnnColour <- list(
      Group = AnnCol)



      # Clustering Distance options = "euclidean", "maximum", "manhattan",
"canberra",
      # "binary" or "minkowski"
      # CLustering Method options = "ward.D", "ward.D2", "single",
"complete", "average",
      # "mcquitty", "median", "centroid"
sigProteins_heatmap = pheatmap(dataTable.log.ids, color =
colorRampPalette(c("lawngreen", "black", "firebrick1"))(100),
                        cluster_rows = TRUE, show_rownames=TRUE, annotation=
annotation, annotation_colors = AnnColour, border_color = NA,
                        width = 10, height = 8, fontsize = 10, fontsize_row
= 6, scale = "row", annotation_names_col = FALSE,
                        clustering_distance_cols = "manhattan",
clustering_method = "complete",
                        legend_breaks = c(-2,-1,0,1,2,2.1), main = "",
legend_labels = c("-2", "-1", "0", "1", "2", "       \n\n(log10)"))


# Draw grobs to improve the look of the graph
fillerRectangle = grid.rect(width = 0.5, height = 0.5, gp = gpar(fill =
"white", col = "white", alpha = 0.8))
grid.arrange(fillerRectangle, sigProteins_heatmap[[4]], fillerRectangle,
nrow=1, widths = c(1,20,1))

# Save the grobs (including the pheatmap heatmap) to disk
png(paste0(outDir, "/All_", dataTableFile, "_ids.png"), width = 10, height
= 8, res = 600, units = "in") # Open a new pdf file
grid.arrange(fillerRectangle, sigProteins_heatmap[[4]], fillerRectangle,
nrow=1, widths = c(1,20,1))
```

```
dev.off()

# With protein descriptions in the Y-axis instead of identifiers
dataTable.log.desc = log10(dataTable[ ,rowNamesDesign])
rowNamesHeatmapDesc  =  make.names(gsub("  ",  "_",  dataTable$description),
unique = TRUE)
row.names(dataTable.log.desc) = rowNamesHeatmapDesc

# The below graph is too messy with the descriptions
            #  sigProteins_heatmap_with_Desc  =  pheatmap(dataTable.log.desc,
color = colorRampPalette(c("lawngreen", "black", "firebrick1"))(100),
            #
                             cluster_rows    =    TRUE,    show_rownames=TRUE,
annotation= annotation, annotation_colors = AnnColour, border_color = NA,
            #
                             width   =   10,   height   =   8,   fontsize   =   10,
fontsize_row = 6, scale = "row", annotation_names_col = FALSE,
            #
                             clustering_distance_cols      =      "manhattan",
clustering_method = "complete",
            #
                             legend_breaks = c(-2,-1,0,1,2,2.1), main = "",
legend_labels = c("-2", "-1", "0", "1", "2", "       \n\n(log10)"))
            #
            #
            # # Draw grobs to improve the look of the graph
            #  fillerRectangle = grid.rect(width = 0.5, height = 0.5, gp =
gpar(fill = "white", col = "white", alpha = 0.8))
            #    grid.arrange(fillerRectangle,    sigProteins_heatmap[[4]],
fillerRectangle, nrow=1, widths = c(1,20,1))
            #
            # # Save the grobs (including the pheatmap heatmap) to disk
            # png(paste0(outDir, "/All_", dataTableFile, "_desc.png"), width
= 10, height = 8, res = 600, units = "in") # Open a new pdf file
            #                                  grid.arrange(fillerRectangle,
sigProteins_heatmap_with_Desc[[4]],   fillerRectangle,   nrow=1,   widths   =
c(1,20,1))
            # dev.off()


#-------------------- PCA --------------------#
dataTable2 = dataTable[c(1,3:11)]
dataTableTranspose = t(dataTable2)
dfNames = dataTableTranspose[1,]
dataTableTranspose   =   data.frame(dataTableTranspose,   stringsAsFactors   =
FALSE)
dataTableTranspose = dataTableTranspose[c(2:nrow(dataTableTranspose)), ]
names(dataTableTranspose) = dfNames
dataTableTranspose$TMT = row.names(dataTableTranspose)
dataTableTranspose     =     dataTableTranspose[c(ncol(dataTableTranspose),
1:(ncol(dataTableTranspose)-1))]
row.names(dataTableTranspose) = NULL
dataTableTranspose[        ,        c(2:ncol(dataTableTranspose))]         =
as.data.frame(lapply(dataTableTranspose[,c(2:ncol(dataTableTranspose))],
as.numeric))


dataTableTranspose["Group"] = NA

for (i in 1:nrow(dataTableTranspose)) {
      dataTableTranspose$Group[i]                                            =
fullDesign$Group[match(dataTableTranspose$TMT[i], fullDesign$Sample)]
}
dataTableTranspose  =  dataTableTranspose[c(1,  c(ncol(dataTableTranspose),
2:(ncol(dataTableTranspose)-1)))]
# Generate PCA
dataTablePCA_Result         <-         prcomp(dataTableTranspose[          ,
c(3:ncol(dataTableTranspose))],
                                                          center = TRUE,
                                                          scale. = TRUE)

# Prepare graph
ir.species = dataTableTranspose[ ,"Group"]

library(ggbiplot)
PCA_Plot  <-  ggbiplot(dataTablePCA_Result,  obs.scale  =  1,  var.scale  =  1,
var.axes = FALSE,
                                        groups = ir.species, ellipse = TRUE,
                                        circle = TRUE)
PCA_Plot <- PCA_Plot + scale_color_discrete(name = '')
PCA_Plot <- PCA_Plot + theme(legend.direction = 'horizontal',
                                         legend.position = 'top')

print(PCA_Plot)
PCA_Plot_Figure  =  ggsave(file  =  paste0(outDir,  "/All_",  dataTableFile,
"_PCA_plot.png"), plot = PCA_Plot, h=6, w=6, units="in",dpi=600)

#-------------------- END PCA --------------------#
#------------------------------------------------#




################### VIOLIN AND BOX PLOTS ###################
# prepare data for Violin Plot
#New dataframe name
dataTable.log.df = dataTable.log.ids
#Add column
dataTable.log.df$Protein = row.names(dataTable.log.ids)
# Re-arrange columns to improve table
dataTable.log.df     =     dataTable.log.df[     ,c(ncol(dataTable.log.df),
1:(ncol(dataTable.log.df)-1))]
# Remove row names
row.names(dataTable.log.df) = NULL
# Make a wide table into a long table
dataTable.log.df_long = dataTable.log.df %>%
      gather(TMT, Log_10_value, 2:10)
# Add a new column called "Group" and fill it with empty values
dataTable.log.df_long["Group"] = NA


# Refer to the design table to replace "TMT" with the appropriate "Group"
name
for (i in 1:nrow(dataTable.log.df_long)) {
      dataTable.log.df_long$Group[i]                                      =
fullDesign$Group[match(dataTable.log.df_long$TMT[i], fullDesign$Sample)]
}

# Get order of x-axis from pheatmap
col.order = sigProteins_heatmap$tree_col$order

# Turn the "TMT" character variable into a factor and define levels
dataTable.log.df_long$TMT  =  factor(dataTable.log.df_long$TMT,  levels  =
unique(dataTable.log.df_long$TMT))
                               # # Change the "TMT" levels to be the same as the x-
axis from the "pheatmap"
                               #            dataTable.log.df_long$TMT            =
factor(dataTable.log.df_long$TMT,
levels(dataTable.log.df_long$TMT)[col.order])

# Draw a violin plot
violin_plot = ggplot(dataTable.log.df_long, aes(x = TMT, y = Log_10_value,
fill = Group)) +
      geom_violin()  +  ylab("Expression  (Log10)")  +  theme(axis.text.x  =
element_text(angle = 90))
violin_plot
ggsave(paste0(outDir, "/All_", dataTableFile, "_violin_plot.png"))

#Draw a boxplot with points of data underlaying it
box_plot = ggplot(dataTable.log.df_long, aes(x = TMT, y = Log_10_value, fill
= Group)) +
geom_boxplot(outlier.shape   =   NA)   +   ylab("Expression   (Log10)")   +
theme(axis.text.x = element_text(angle = 270))
box_plot = box_plot + geom_point(colour = "black", size = 4, alpha = 0.2)
box_plot
ggsave(paste0(outDir, "/All_", dataTableFile, "_box_plot.png"), dpi = 600,
units = "cm", width = 29.7, height = 21)

#Draw a violin plot with points of data underlaying it
dot_violin_plot   =   ggplot(dataTable.log.df_long,   aes(x   =   TMT,   y   =
Log_10_value, fill = Group)) +
geom_violin() +
ylab("Expression (Log10)") + theme(axis.text.x = element_text(angle = 90))
dot_violin_plot = dot_violin_plot + geom_point(colour = "black", size = 4,
alpha = 0.2)
dot_violin_plot
ggsave(paste0(outDir, "/All_", dataTableFile, "_dot_violoin_plot.png"), dpi
= 600, units = "cm", width = 29.7, height = 21)
```

Appendix C.

C.1 Additional Figures and Tables for Chapter 4

Table C.1. 200-grain weight measurements of samples

Sample Code	Location	Variety	Biological Replicate	200-Weight (g)
Br-Greg-BR1-2013	Breeza	Gregory	1	6.64
Br-Greg-BR2-2013	Breeza	Gregory	2	6.74
Br-Greg-BR3-2013	Breeza	Gregory	3	7.20
Br-Liv-BR1-2013	Breeza	Livingston	1	6.83
Br-Liv-BR2-2013	Breeza	Livingston	2	6.65
Br-Liv-BR3-2013	Breeza	Livingston	3	6.34
Br-Spit-BR1-2013	Breeza	Spitfire	1	6.98
Br-Spit-BR2-2013	Breeza	Spitfire	2	6.28
Br-Spit-BR3-2013	Breeza	Spitfire	3	6.25
TARC-Greg-BR1-2013	TARC	Gregory	1	8.01
TARC-Greg-BR2-2013	TARC	Gregory	2	5.78
TARC-Greg-BR3-2013	TARC	Gregory	3	8.07
TARC-Liv-BR1-2013	TARC	Livingston	1	8.23
TARC-Liv-BR2-2013	TARC	Livingston	2	7.21
TARC-Liv-BR3-2013	TARC	Livingston	3	7.81
TARC-Spit-BR1-2013	TARC	Spitfire	1	7.72
TARC-Spit-BR2-2013	TARC	Spitfire	2	8.75
TARC-Spit-BR3-2013	TARC	Spitfire	3	9.15
THH-Greg-BR1-2013	Terry Hie Hie	Gregory	1	6.41
THH-Greg-BR2-2013	Terry Hie Hie	Gregory	2	6.64
THH-Greg-BR3-2013	Terry Hie Hie	Gregory	3	6.60
THH-Liv-BR1-2013	Terry Hie Hie	Livingston	1	7.56
THH-Liv-BR2-2013	Terry Hie Hie	Livingston	2	7.32
THH-Liv-BR3-2013	Terry Hie Hie	Livingston	3	7.64
THH-Spit-BR1-2013	Terry Hie Hie	Spitfire	1	8.30
THH-Spit-BR2-2013	Terry Hie Hie	Spitfire	2	8.24
THH-Spit-BR3-2013	Terry Hie Hie	Spitfire	3	8.45

Table C.1.1. P-value results, examining 200-weight differences of wheat samples between farm locations, based on table C.1.

Comparison (farms)	P-adj
TARC-Breeza	0.005753
Terry Hie Hie-Breeza	0.073735
Terry Hie Hie-TARC	0.501337
ANOVA (overall):	0.00690326

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table C.1.2. P-value results, examining 200-weight differences of wheat samples between cultivars, based on table C.1.

Comparison (cultivars)	P-adj
Livingston-Gregory	0.583322
Spitfire-Gregory	0.075571
Spitfire-Livingston	0.41124
ANOVA (overall):	0.09108327

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table C.1.3. P-value results, comparing 200-weight results between replicate samples of cultivars at Breeza only

Comparison	P-adj
Livingston-Gregory	0.6326875
Spitfire-Gregory	0.4286640
Spitfire-Livingston	0.9219048

Table C.1.4. P-value results, comparing 200-weight results between replicate samples of cultivars at TARC only

Comparison	P-adj
Livingston-Gregory	0.8148216
Spitfire-Gregory	0.2876709
Spitfire-Livingston	0.5714488

Table C.1.5. P-value results, comparing 200-weight results between replicate samples of cultivars at THH only

Comparison	P-adj
Livingston-Gregory	0.000314869
Spitfire-Gregory	0.000008341
Spitfire-Livingston	0.000721514

Table C.1b. Weight of extracted wheat proteins in µg

Sample Code	Farm location	Cultivar	Biological replicate	Harvest Year	Sample protein weight_µg
Br-Greg-2013-BR1	Breeza	Gregory	1	2013	1145
Br-Greg-2013-BR1	Breeza	Gregory	2	2013	1231
Br-Greg-2013-BR1	Breeza	Gregory	3	2013	1092
Br-Spit-2013-BR1	Breeza	Spitfire	1	2013	1428
Br-Spit-2013-BR1	Breeza	Spitfire	2	2013	1665
Br-Spit-2013-BR1	Breeza	Spitfire	3	2013	1679
Br-Liv-2013-BR1	Breeza	Livingston	1	2013	1412
Br-Liv-2013-BR1	Breeza	Livingston	2	2013	1354
Br-Liv-2013-BR1	Breeza	Livingston	3	2013	1469
TARC-Greg-2013-BR1	TARC	Gregory	1	2013	1238
TARC-Greg-2013-BR1	TARC	Gregory	2	2013	1145
TARC-Greg-2013-BR1	TARC	Gregory	3	2013	997
TARC-Spit-2013-BR1	TARC	Spitfire	1	2013	1433
TARC-Spit-2013-BR1	TARC	Spitfire	2	2013	1329
TARC-Spit-2013-BR1	TARC	Spitfire	3	2013	1290
TARC-Liv-2013-BR1	TARC	Livingston	1	2013	1128
TARC-Liv-2013-BR1	TARC	Livingston	2	2013	1101
TARC-Liv-2013-BR1	TARC	Livingston	3	2013	1234
THH-Spit-2013-BR1	THH	Spitfire	1	2013	1071
THH-Spit-2013-BR1	THH	Spitfire	2	2013	1216
THH-Spit-2013-BR1	THH	Spitfire	3	2013	1149
THH-Greg-2013-BR1	THH	Gregory	1	2013	1053
THH-Greg-2013-BR1	THH	Gregory	2	2013	945
THH-Greg-2013-BR1	THH	Gregory	3	2013	1145
THH-Liv-2013-BR1	THH	Livingston	1	2013	1172
THH-Liv-2013-BR1	THH	Livingston	2	2013	1056
THH-Liv-2013-BR1	THH	Livingston	3	2013	1083

200-weights: P-values

Table C.1b.1. P-value results, examining the differences of the weights of extracted proteins for wheat samples, between farm locations (based on table C.1).

Comparison (farms)	P-adj
TARC-Breeza	0.005753
Terry Hie Hie-Breeza	0.073735
Terry Hie Hie-TARC	0.501337
ANOVA (overall):	0.00690326

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table C.1b.2. P-value results examining the differences of the weights of extracted proteins for wheat samples, between cultivars (based on table C.1).

Comparison (cultivars)	P-adj
Livingston-Gregory	0.583322
Spitfire-Gregory	0.075571
Spitfire-Livingston	0.41124
ANOVA (overall):	0.09108327

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Protein extraction weights: P-values

Table C.1b.3. P-value results, comparing extracted protein weight (µg) of replicate samples between cultivars at Breeza only

Comparison	P-adj
Livingston-Gregory	0.040825534
Spitfire-Gregory	0.003657993
Spitfire-Livingston	0.137896215

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table C.1b.4. P-value results, comparing extracted protein weight (µg) of replicate samples between cultivars at TARC only

Comparison	P-adj
Livingston-Gregory	0.92824275
Spitfire-Gregory	0.05476470
Spitfire-Livingston	0.08711153

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table C.1b.5. P-value results, comparing extracted protein weight (µg) of replicate samples between cultivars at THH only

Comparison	P-adj
Livingston-Gregory	0.6813164
Spitfire-Gregory	0.3537254
Spitfire-Livingston	0.8036222

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Peptide extraction weights: P-values

Table C.1b.6. P-value results, comparing extracted peptide weight (µg) of replicate samples between cultivars at Breeza only

Comparison (cultivars)	P-adj
Livingston-Gregory	0.02404184
Spitfire-Gregory	0.04937812
Spitfire-Livingston	0.82401919

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table C.1b.7. P-value results, comparing extracted peptide weight (µg) of replicate samples between cultivars at TARC only

Comparison	P-adj
Livingston-Gregory	0.6504982
Spitfire-Gregory	0.9805424
Spitfire-Livingston	0.7571680

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

Table C.1b.8. P-value results, comparing extracted peptide weight (µg) of replicate samples between cultivars at THH only

Comparison	P-adj
Livingston-Gregory	0.220239889
Spitfire-Gregory	0.007378158
Spitfire-Livingston	0.001342270

Note: P-value calculated by one-way ANOVA, followed by the Tukey test.

C.2 Density plots and box plots of initial protein expression data

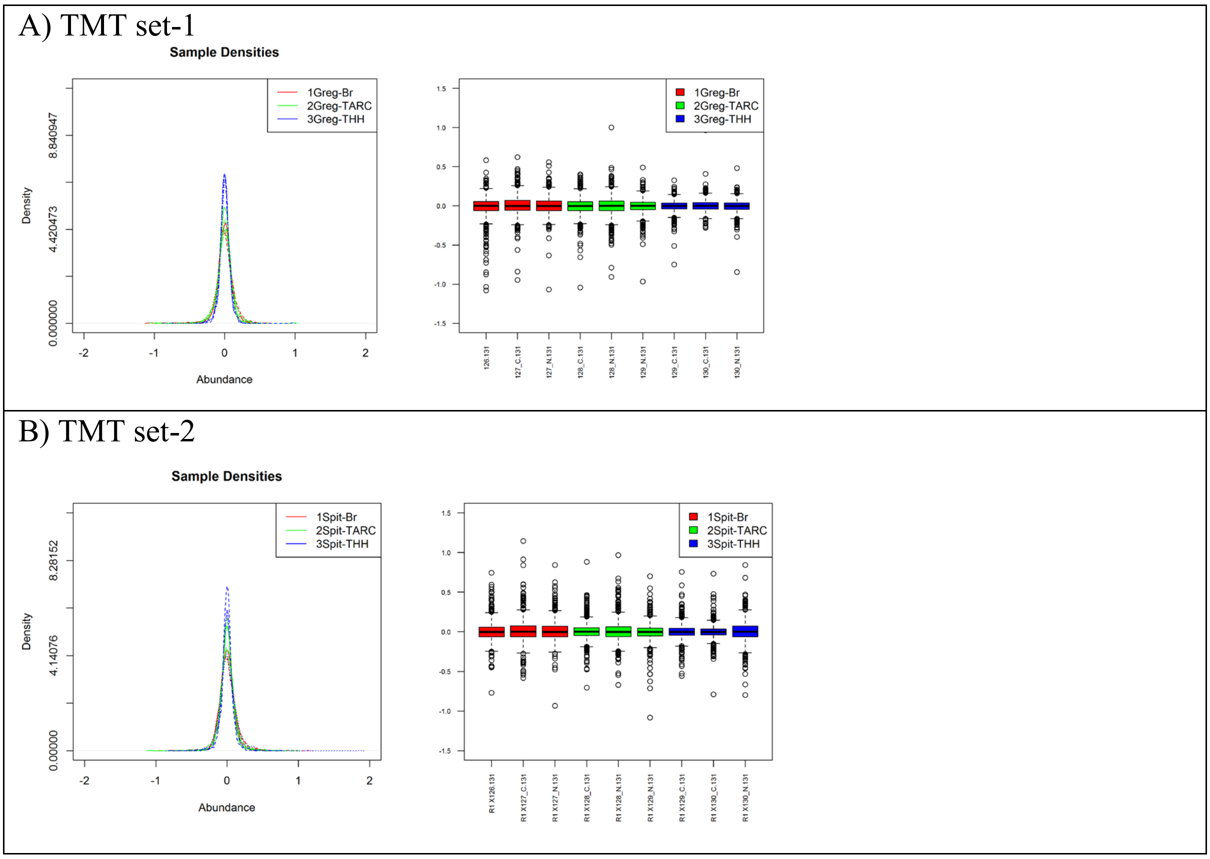
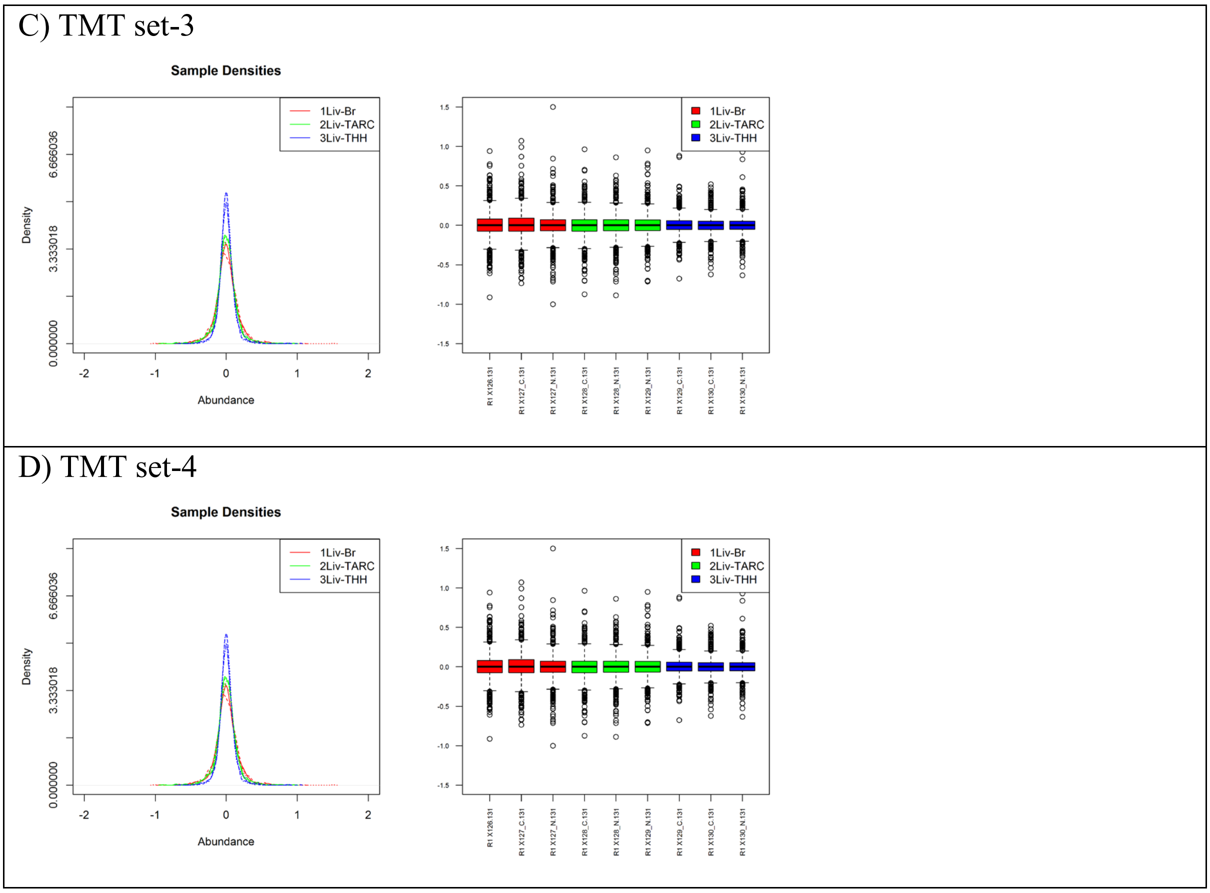


Figure C.1. Density and box plots to check quality of data for TMT experiment 1, 2013 harvest for matched and filtered proteins from wheat grain.



C) TMT set-3

D) TMT set-4

Figure continued from above (Figure C.1)

C.3 Data summary for TMT sets 1 to 4: matched and filtered proteins.

C.3.1 Heatmaps for TMT sets 1 to 4 for matched and filtered proteins

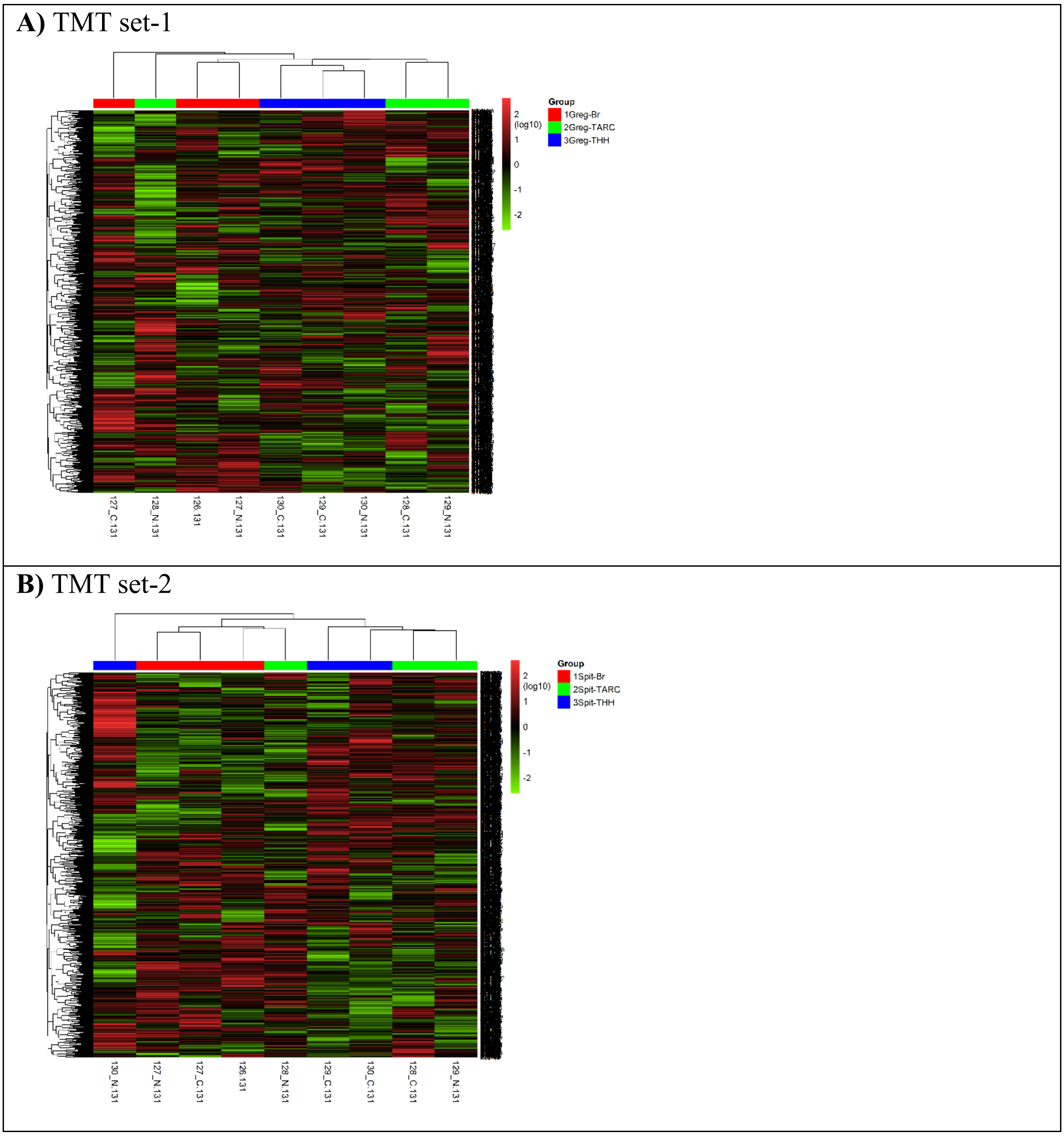


Figure C.2. Heat-maps to check quality of data for TMT experiment 1, 2013 harvest for matched and filtered proteins from wheat grain.

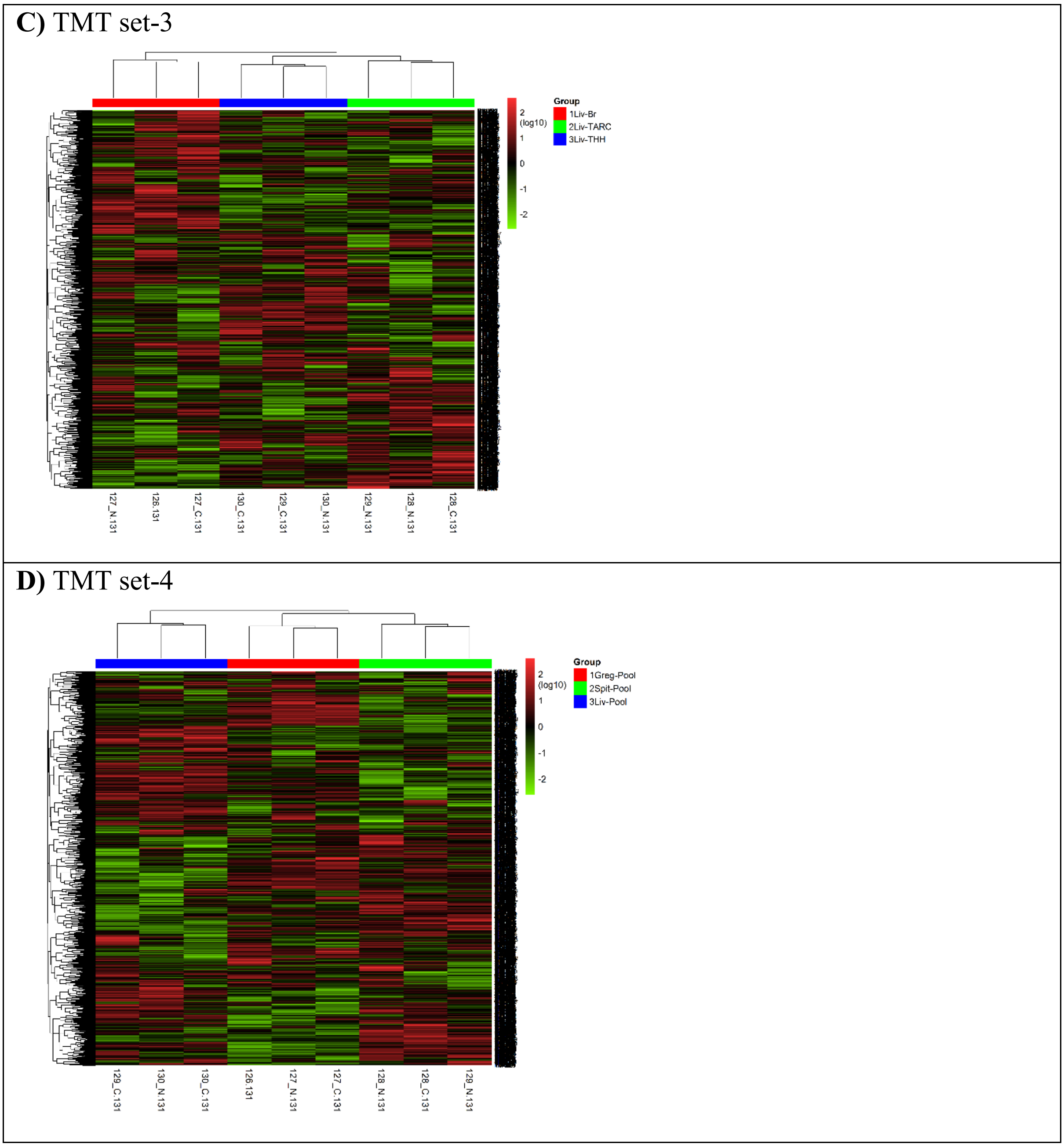


Figure continued from above (Figure C.2)

C.3.2 PCAs for TMT sets 1 to 4 for matched and filtered proteins

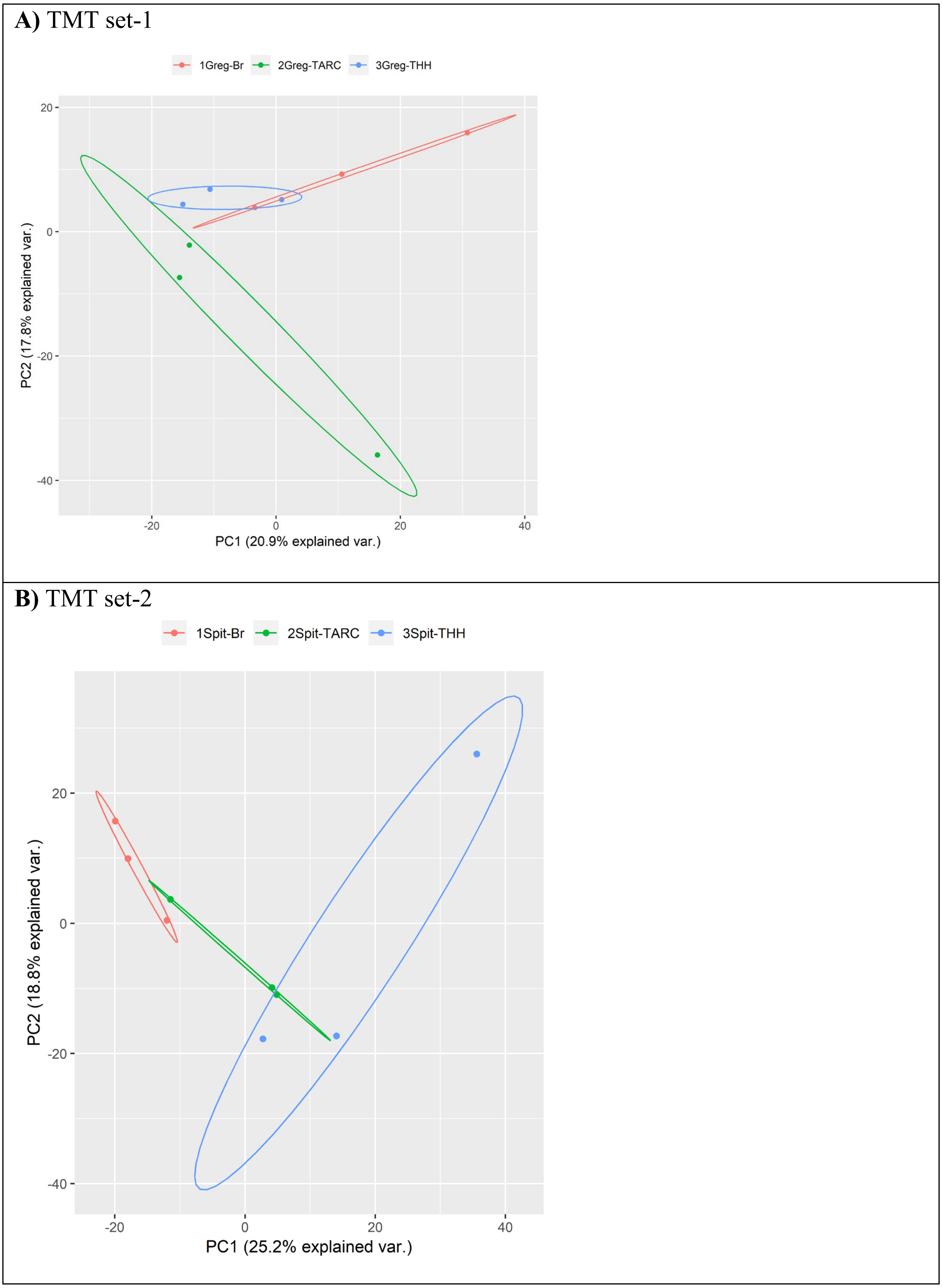
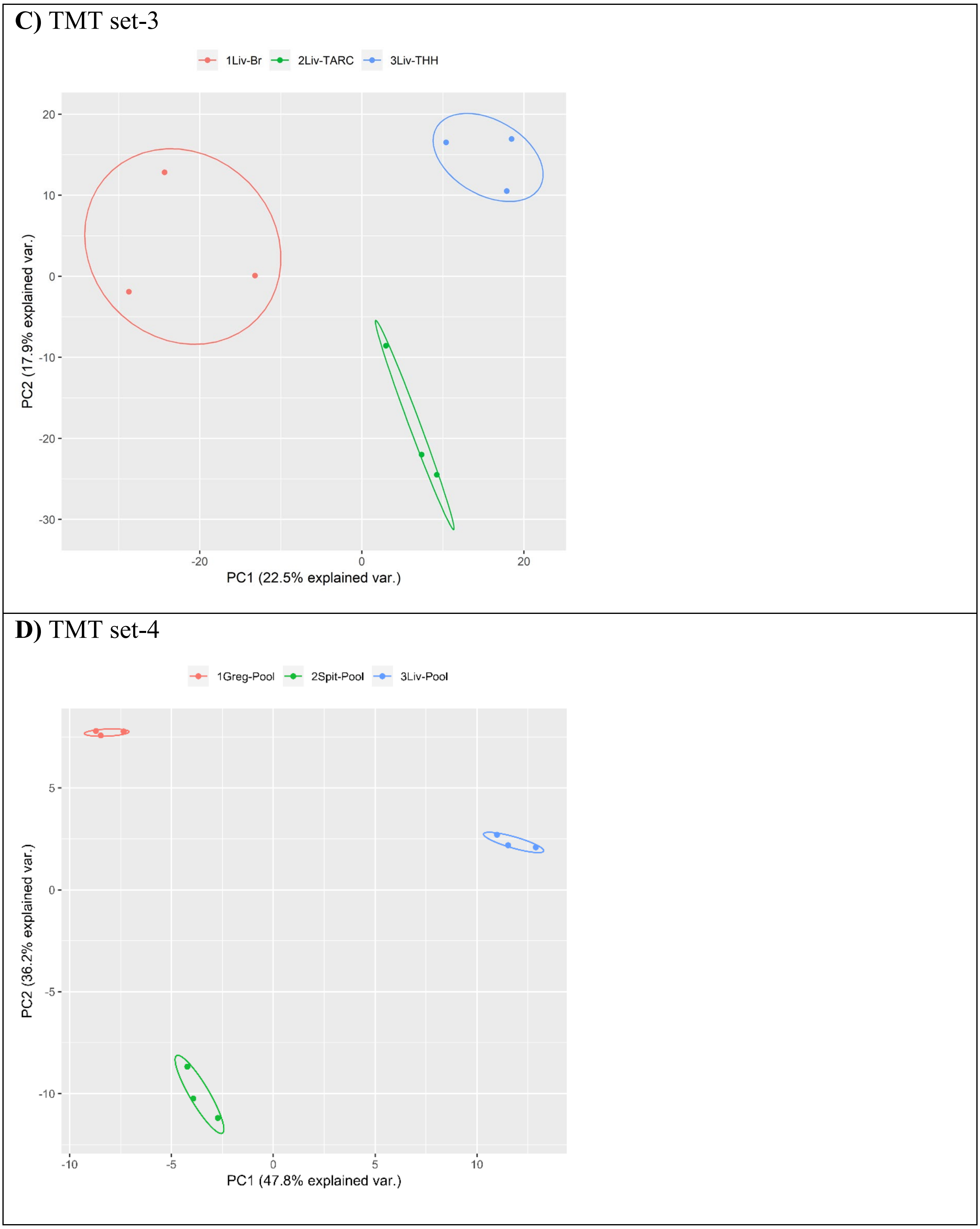


Figure C.3. PCAs to check quality of data for TMT experiment 1, 2013 harvest for matched and filtered proteins from wheat grain.

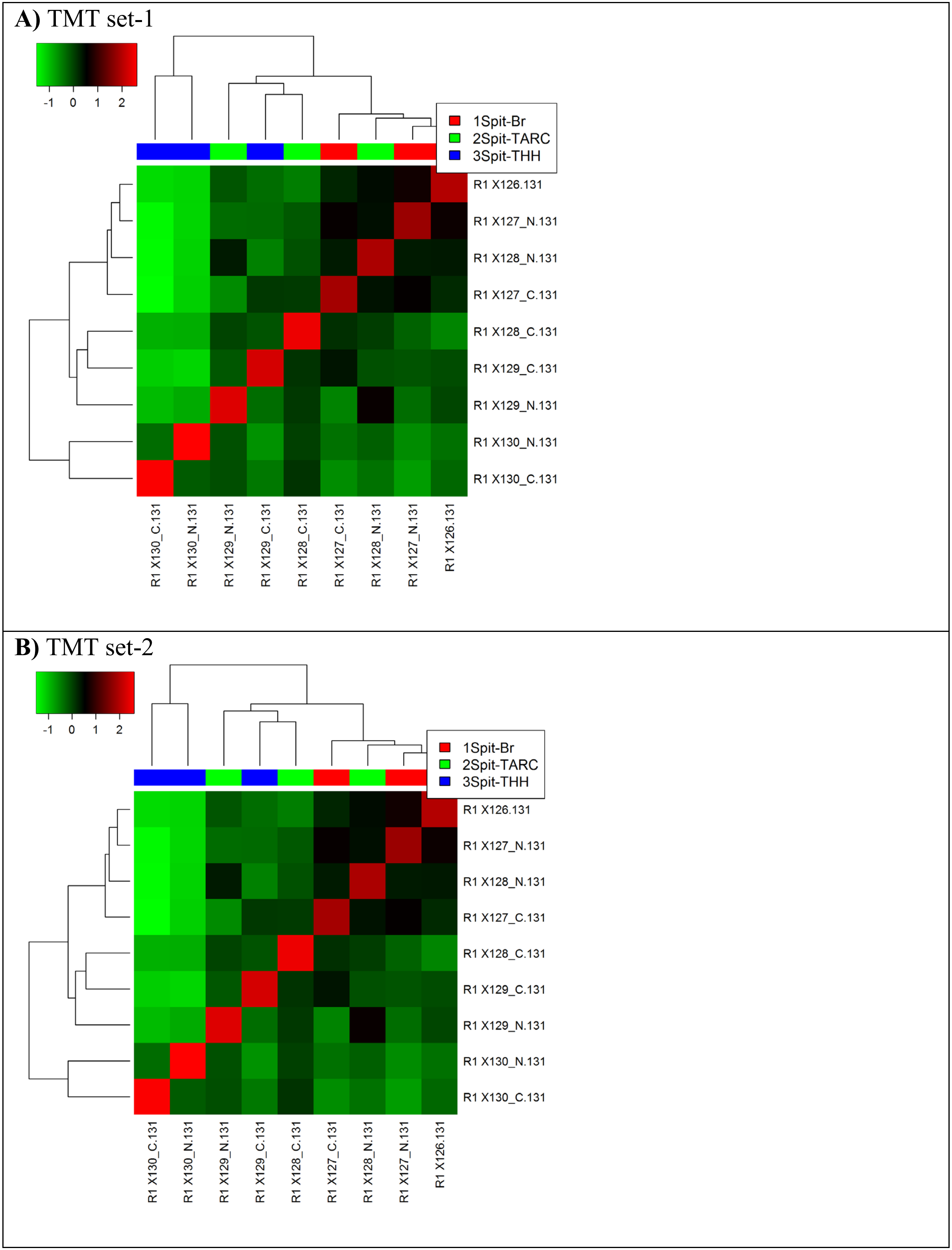


C) TMT set-3

D) TMT set-4

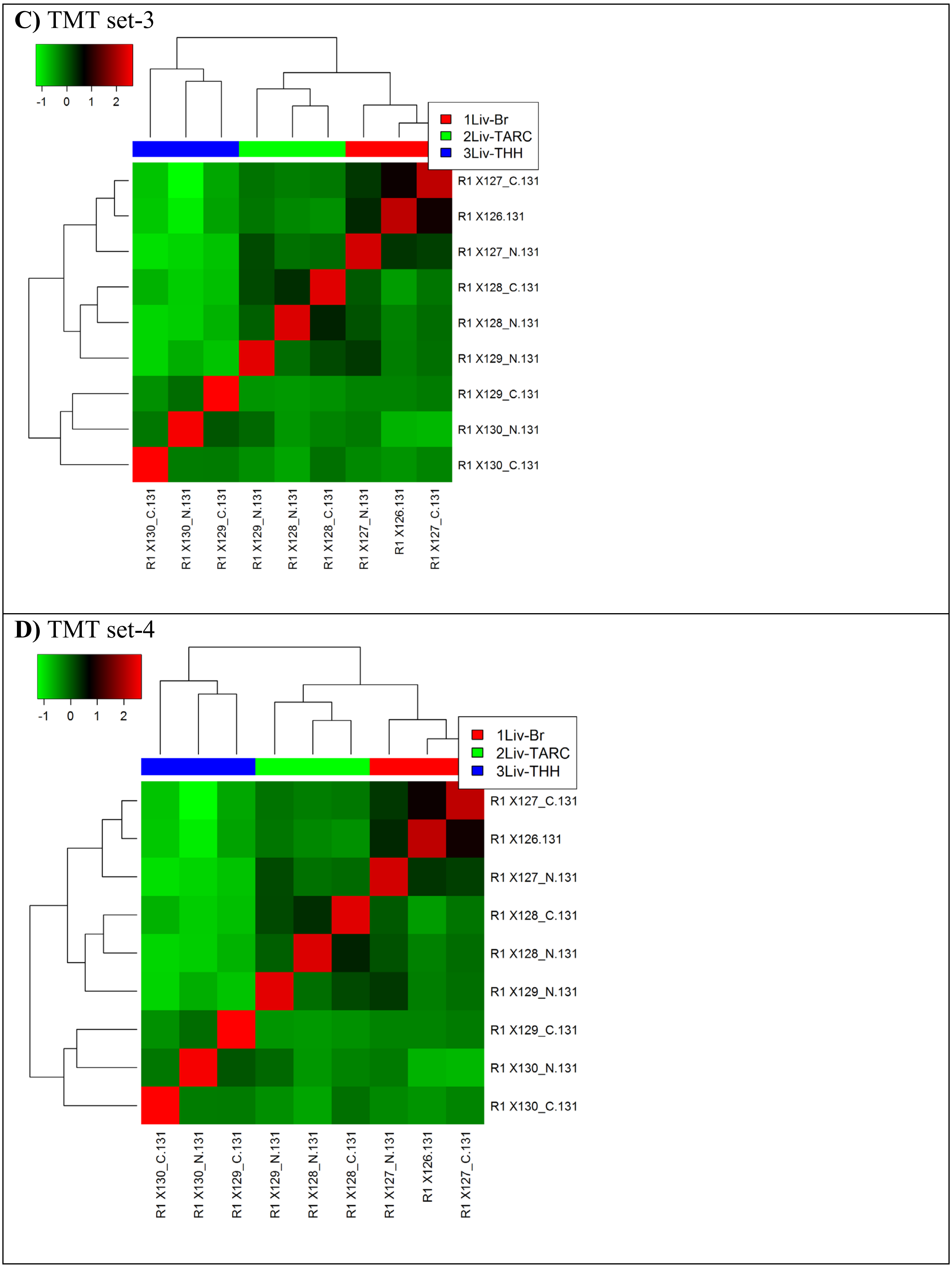
Figure continued from above (Figure C.3)

C.3.3 Correlation plots for TMT sets 1 to 4 for matched and filtered proteins


A) TMT set-1

B) TMT set-2

Figure C.4. Correlation plots to check quality of data for TMT experiment 1, 2013 harvest for matched and filtered proteins from wheat grain.



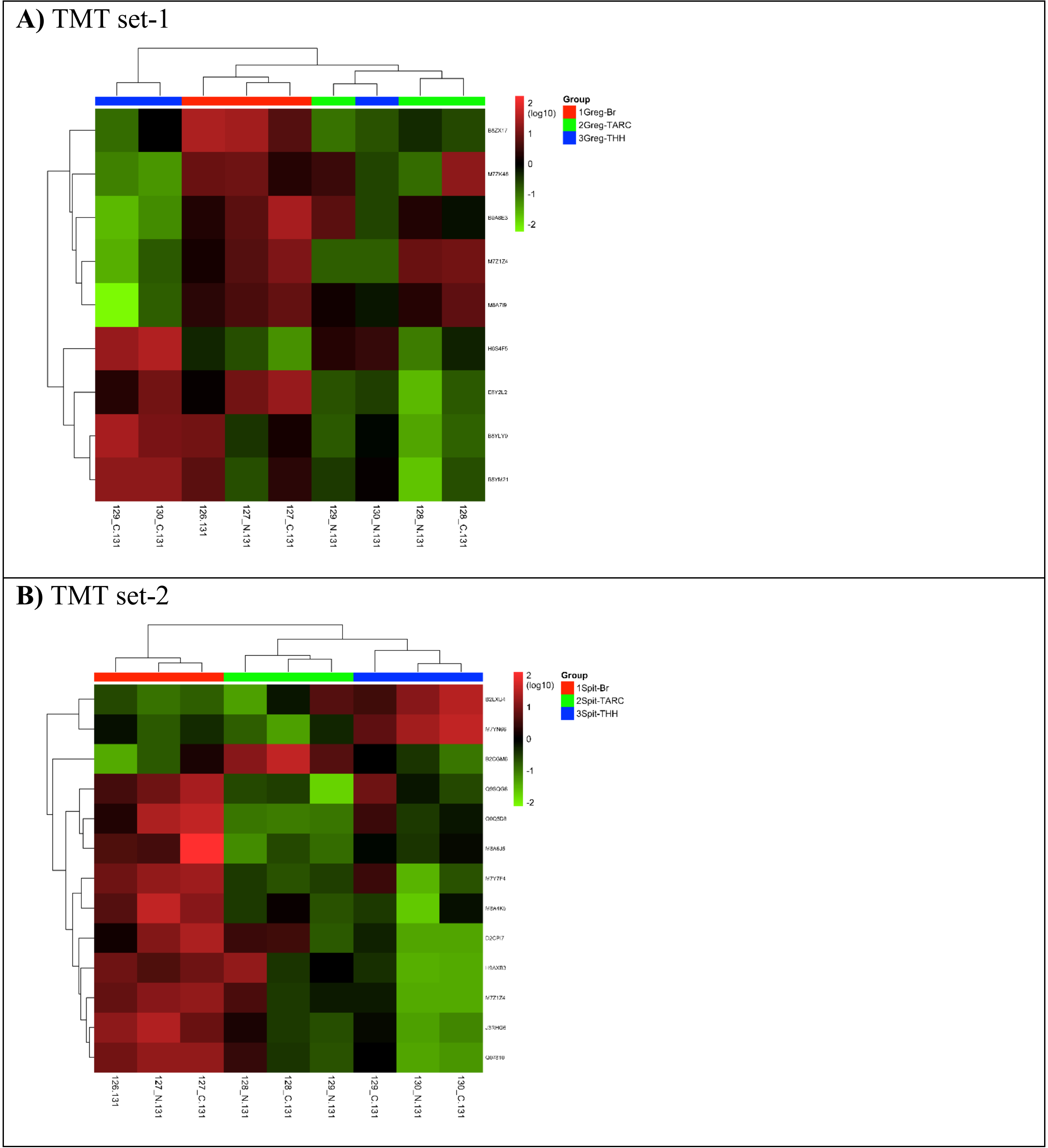
C) TMT set-3

D) TMT set-4

Figure continued from Figure C.4.

C.4 Data summary for TMT sets 1 to 4: Putative biomarker proteins.

C.4.1 Heatmaps for TMT sets 1 to 4 for putative biomarkers

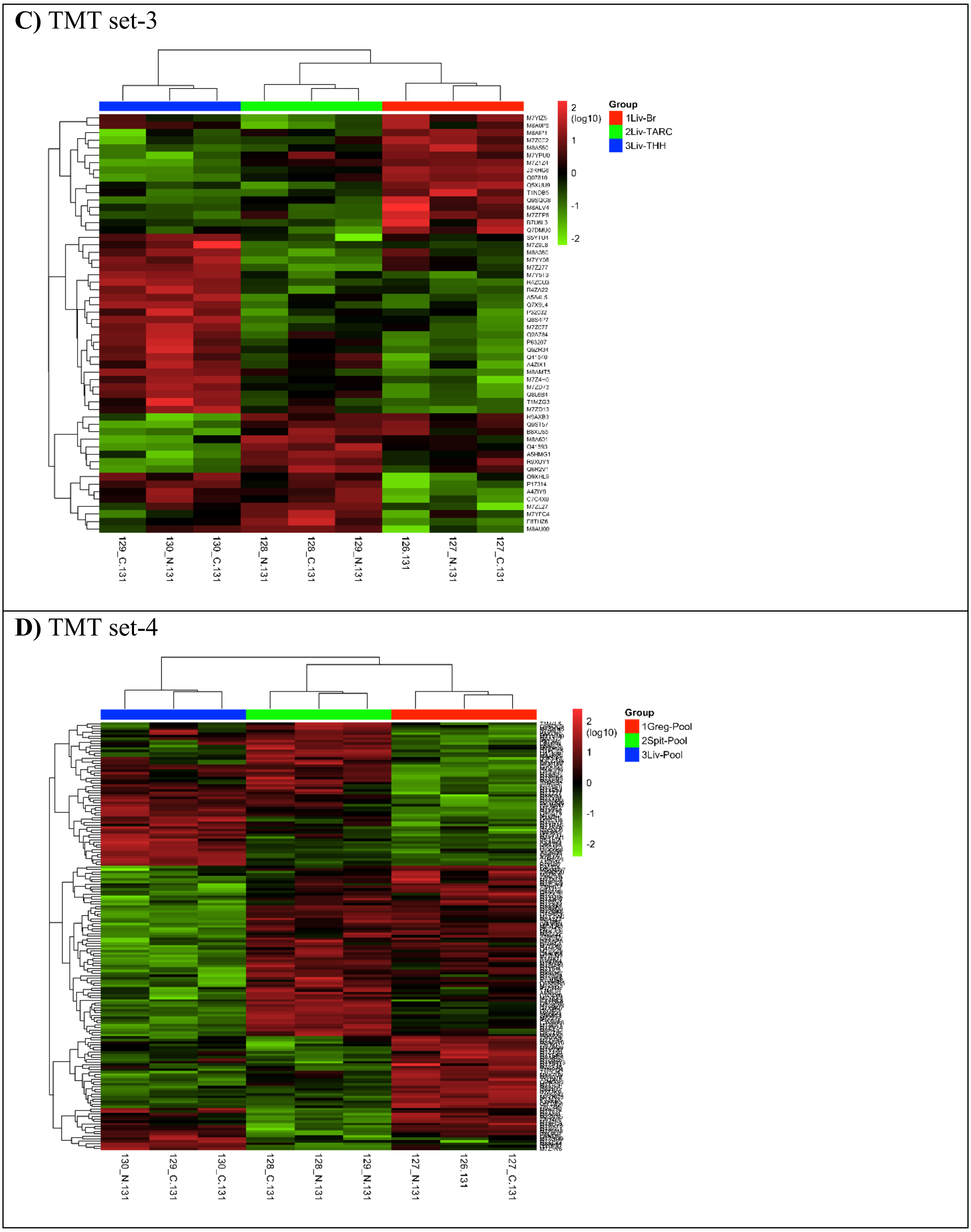


A) TMT set-1

B) TMT set-2

Figure C.5. Heat-maps to check quality of data for TMT experiment 1, 2013 harvest for putative biomarker proteins.

Greg = Gregory wheat, Spit = Spitfire wheat; THH = Terry Hie Hie farm, Br = Breeza farm, TARC = TARC farm



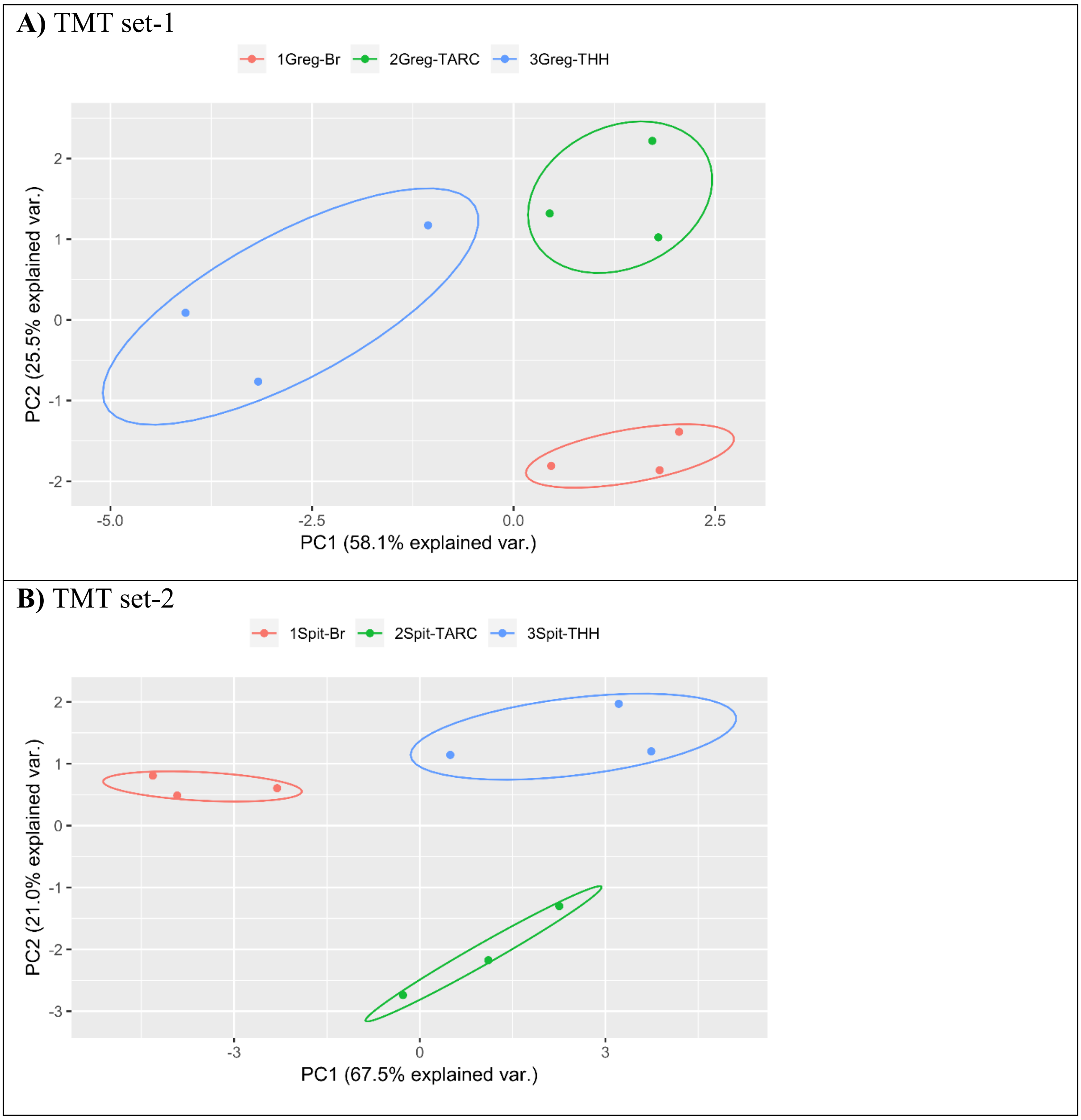
C) TMT set-3

D) TMT set-4

Figure continued from above (Figure C.5)

Greg = Gregory wheat, Spit = Spitfire wheat, Liv = Livingston wheat; THH = Terry Hie Hie farm, Br = Breeza farm, TARC = TARC farm

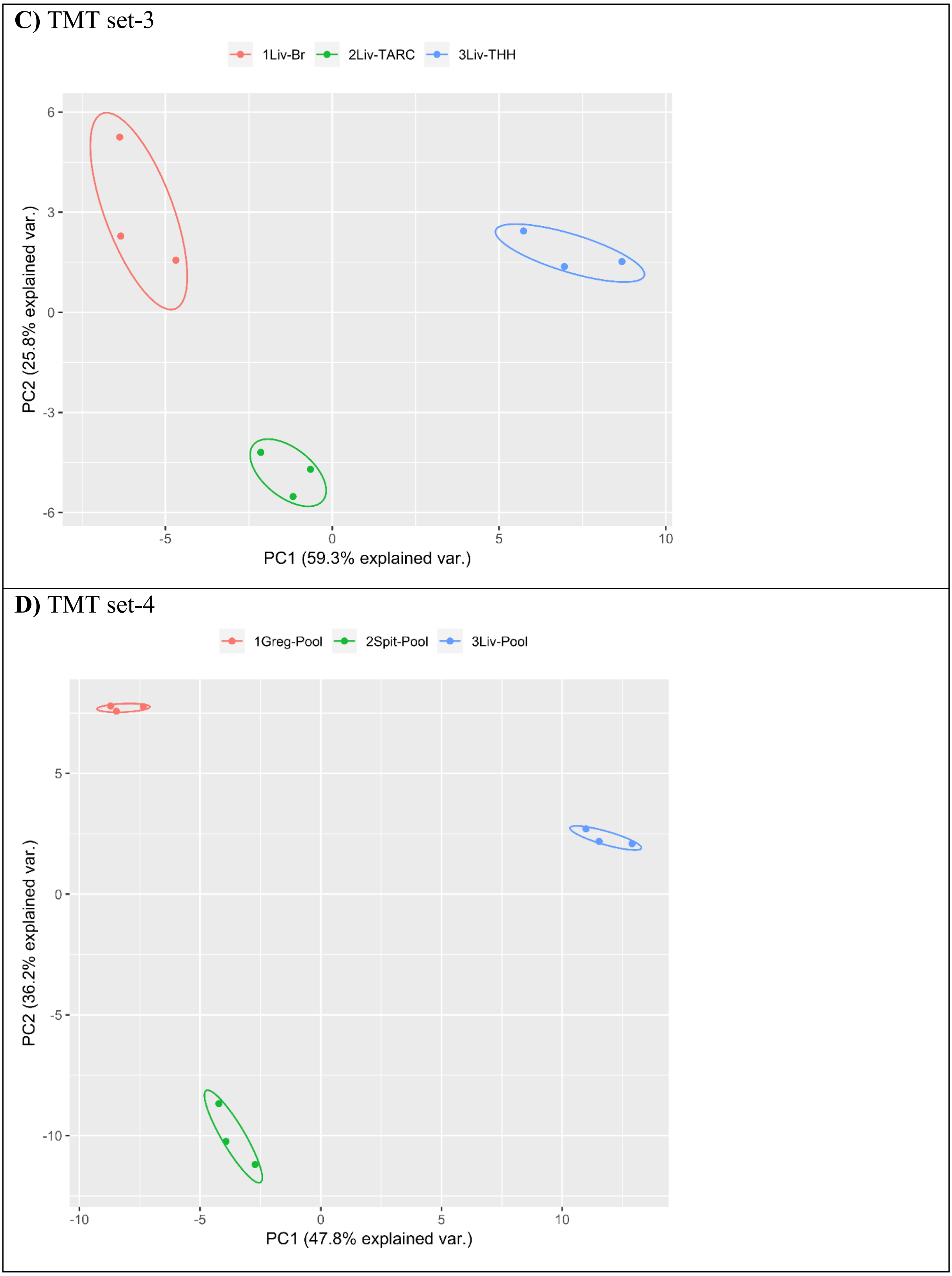
C.4.2 PCAs for TMT sets 1 to 4 for putative biomarkers



A) TMT set-1

B) TMT set-2

Figure C.6. PCAs to check quality of TMT data for putative biomarker proteins.
Greg = Gregory wheat, Spit = Spitfire wheat; THH = Terry Hie Hie farm, Br = Breeza farm, TARC = TARC farm



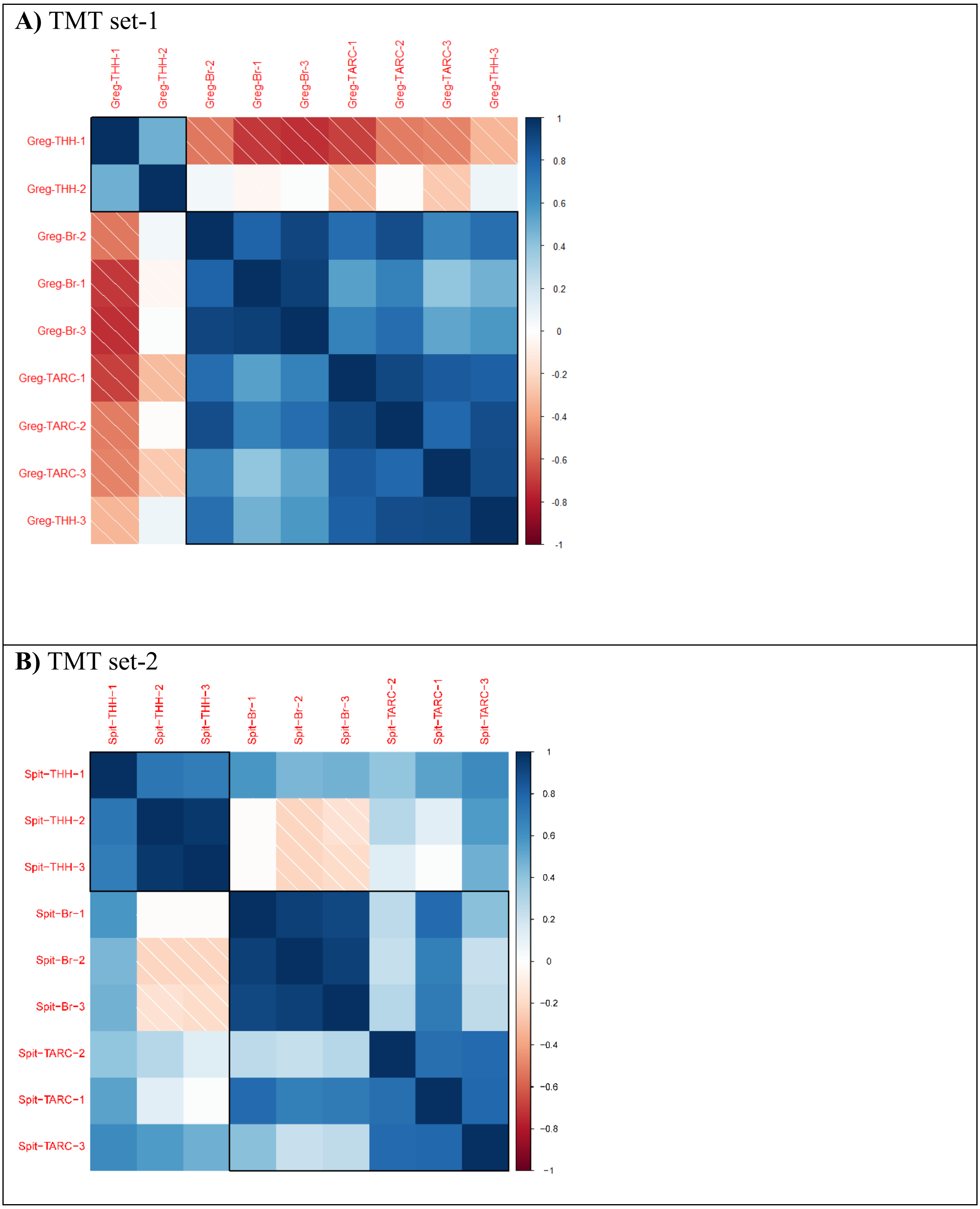
C) TMT set-3

D) TMT set-4

Figure continued from above (Figure C.6).

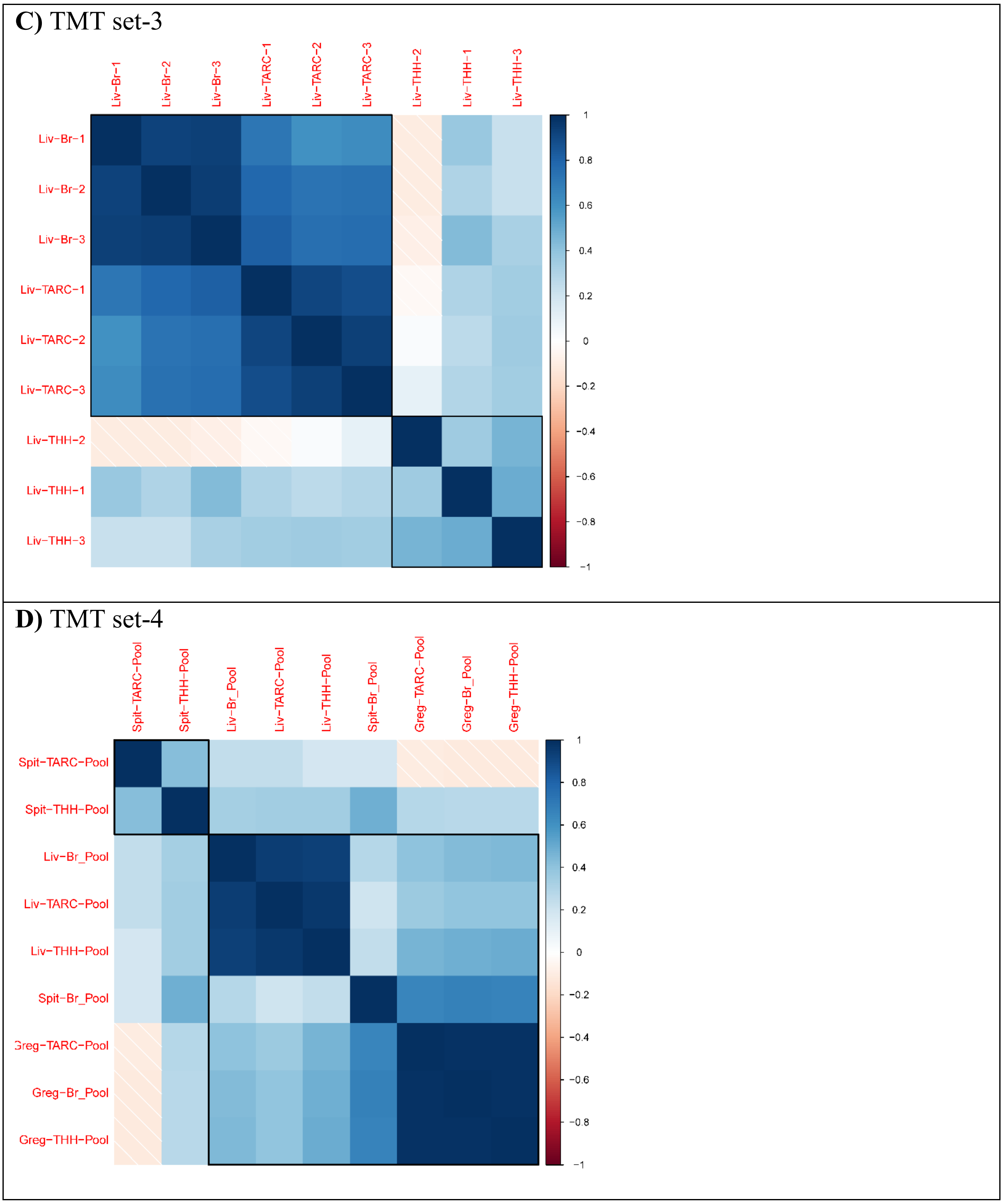
Greg = Gregory wheat, Spit = Spitfire wheat, Liv = Livingston wheat; THH = Terry Hie Hie farm, Br = Breeza farm, TARC = TARC farm

C.4.3 Correlation plots for TMT sets 1 to 4 for putative biomarkers


A) TMT set-1

B) TMT set-2

Figure C.7. Correlation plots to check quality of TMT data for putative biomarker proteins.
Greg = Gregory wheat, Spit = Spitfire wheat; THH = Terry Hie Hie farm, Br = Breeza farm, TARC = TARC farm



C) TMT set-3

D) TMT set-4

Figure continued from above (Figure C.7).

Greg = Gregory wheat, Spit = Spitfire wheat, Liv = Livingston wheat; THH = Terry Hie Hie farm, Br = Breeza farm, TARC = TARC farm

C.5 Putative biomarker proteins discovered in TMT sets 3 to 4.

Table C.2. TMT set-3: List of putative biomarkers from wheat grain.

Uniprot Identifier	Protein description
P17314	Alpha-amylase/trypsin inhibitor CM3
A4ZIY9	Monomeric alpha-amylase inhibitor (Fragment)
C7C4X0	Alpha amylase inhibitor CM1 (Fragment)
F8THZ6	Protein disulfide isomerase
Q9XHL9	Histone H1 WH1B.1
M7ZL27	Ribonuclease 3-like protein 3
M7YFC4	Acyl-CoA dehydrogenase family member 10
M8AU00	Putative alpha,alpha-trehalose-phosphate synthase [UDP-forming] 7
P83207	Chymotrypsin inhibitor WCI
P32032	Alpha-2-purothionin
Q2A784	Avenin-like a1
A5A4L5	Avenin-like b4
Q8S4P7	Thaumatin-like protein
Q41540	CM 17 protein
Q7X9L4	Proteinase inhibitor Rgpi9 (Fragment)
A4ZIX1	Monomeric alpha-amylase inhibitor (Fragment)
M7Y5T3	NAD(P)H-dependent 6'-deoxychalcone synthase
M8A380	Globulin-1 S allele
M7YY08	Aspartyl-tRNA synthetase, cytoplasmic
S5YTU4	NADH-dependent glutamate synthase
M7Z4H0	11S globulin seed storage protein 2
T1MZG3	Uncharacterized protein (Fragment)
Q9ZR34	Amylogenin
R4ZCU3	M-2
M7ZD73	Uncharacterized protein
R4ZA22	L-1
Q8L6B4	Gamma gliadin
M8AMT5	Primary amine oxidase
M7Z277	Putative O-methyltransferase 2
M7Z077	Uncharacterized protein
M7ZD13	Uncharacterized protein
M7Z9L8	Uncharacterized protein
Q41593	Serpin-Z1A
Q9ST57	Serpin-Z2A
A5HMG1	HMW glutenin subunit 1Bx13
B8XU65	High molecular weight glutenin x-type (Fragment)
H9AXB3	Serpin-N3.2
M8A601	Protein strawberry notch-like protein 1
R9XUY1	Omega-gliadin
Q6R2V1	High-molecular-weight glutenin subunit 1Dx2.1
M7Z0E2	Histone H2A
B7U6L3	Globulin 3C (Fragment)
J3RHG6	Beta-glucosidase 4 (Fragment)
M7YIZ5	Histone H4
M8A0P8	60S ribosomal protein L35-2
M7Z1Z4	Serpin-Z2B
Q7DMU0	Storage protein
Q07810	rRNA N-glycosidase
M8A580	60S ribosomal protein L4-1
Q5XUU9	Cytoplasmatic ribosomal protein S13
M8AIP1	DnaJ homolog subfamily C member 2
Q9SQG8	Pathogenesis-related protein 4 (Fragment)
M8ALV4	Chitinase 5
M7YPU0	Uncharacterized protein
T1NDB5	Uncharacterized protein
M7ZFP8	Uncharacterized protein

Table C.3. TMT set-4: List of putative biomarkers from wheat grain.

Uniprot Identifier	Protein description
Q41629	ADP, ATP carrier protein 1, mitochondrial
Q43312	Protein H2A.7
P16851	Alpha-amylase/trypsin inhibitor CM2
P17314	Alpha-amylase/trypsin inhibitor CM3
P81713	Bowman-Birk type trypsin inhibitor
P46274	Mitochondrial outer membrane porin
Q8H0K8	Xylanase inhibitor (Precursor)
Q8GZB0	Non-specific lipid-transfer protein (Precursor)
Q8RWR5	Beta-D-glucan exohydrolase
Q7X9K5	ATP synthase (Fragment)
Q5XUV7	Proteasome subunit beta type
A4ZIX1	Monomeric alpha-amylase inhibitor (Fragment)
A6N862	Puroindoline b
A7UME6	Xylanase inhibitor 801NEW
B1Q3K4	Basic region leucine zipper protein
B7U6L4	Globulin 3
C3VWL8	Dimeric alpha-amylase inhibitor
C3VWP8	Dimeric alpha-amylase inhibitor
C7C4X0	Alpha amylase inhibitor CM1 (Fragment)
D0PRB4	Peroxiredoxin
I3NM23	Lipoxygenase
M8AIC9	Prohibitin-1, mitochondrial
M7ZD89	Nucleolin
M7ZLM2	Uncharacterized protein
M8A3F7	Anthocyanidin reductase
M7ZXD7	2,3-bisphosphoglycerate-independent phosphoglycerate mutase
M7ZZV1	Uncharacterized protein
M7YXZ5	Uncharacterized protein
M7Z0X1	Aldose reductase
Q9ZSR6	Heat shock protein HSP26
M7YLR0	Protein IN2-1-like protein B
M7ZS22	Uncharacterized protein
Q8GV48	LEA2 protein
M7ZIQ2	Putative glutathione S-transferase GSTF1
T1MF37	Uncharacterized protein (Fragment)
M7YRL6	Cytochrome c
M7ZXI3	Uncharacterized protein
M8A2Z1	Aldose 1-epimerase
M7Z222	Uncharacterized protein
Q75RZ3	Putative beta-xylosidase (Fragment)
M8ATC6	Lysosomal alpha-mannosidase
M7YZW0	Uncharacterized protein
T1NKC7	Uncharacterized protein
M7YU16	ATP synthase subunit d, mitochondrial
M7YIB4	Reticulon-like protein
M7ZPE5	Ubiquitin thioesterase otubain-like protein
M7ZQK6	Uncharacterized protein
M7Z2P5	Ubiquitin carboxyl-terminal hydrolase
Q7XAP6	Uncharacterized protein
M8AAX5	Multiprotein-bridging factor 1a
M7ZAS8	Peroxidase 52
T1MQQ6	Uncharacterized protein
M7ZDV3	Uncharacterized protein

Part 2 of 4 for Table C.3, continued from above.

M7YKK8	Prohibitin-2
M7ZBW9	Uncharacterized protein
S5A8C3	S-formylglutathione hydrolase-like protein
M7ZVK3	60S ribosomal protein L10a-3
M7ZMS4	Endoglucanase 11
M7ZM54	Putative O-methyltransferase 2
M8AMW7	Bifunctional dihydroflavonol 4-reductase/flavanone 4-reductase
M7Z5S9	Basic 7S globulin
T1M621	Uncharacterized protein (Fragment)
M8AU47	U1 small nuclear ribonucleoprotein A
T1NCT0	Uncharacterized protein (Fragment)
M8AD20	Peroxisomal membrane protein 11-5
M8A410	Putative 6-phosphogluconolactonase 4, chloroplastic
M7Z9L8	Uncharacterized protein
M8AIQ3	Putative NADP-dependent oxidoreductase P1
T1LH91	Uncharacterized protein (Fragment)
P22701	Em protein CS41
P83207	Chymotrypsin inhibitor WCI
Q9ST57	Serpin-Z2A
A5JPR2	Peroxisomal ascorbate peroxidase
Q9FS79	Triosephosphate isomerase
Q41539	Endochitinase (Precursor)
Q8RVZ1	Putative xylanase inhibitor protein (Precursor)
Q7XYB5	Pyruvate orthophosphate dikinase (Fragment)
Q0Q5E3	Globulin 1
Q0Q5D9	Globulin 1
A4ZIY9	Monomeric alpha-amylase inhibitor (Fragment)
A5HMG1	HMW glutenin subunit 1Bx13
A7BJ77	Xylanase inhibitor
B2CGM6	Triticin
B5B0D5	Major allergen CM16
B5B1F8	Aspartate aminotransferase (Fragment)
B7U6L5	Globulin 3B
B8XU40	Gamma gliadin
D3KVP5	27k protein (Fragment)
D8L9S2	Glutamate decarboxylase, putative, expressed
E3W165	Waxy B1 (Fragment)
F4Y5A7	Heat shock protein 90
I0JTW3	Cystatin, expressed
J9Q8Q6	High molecular weight glutenin subunit 1Ay protein
M8ABB2	Tubulin beta-4 chain
M8A8M0	Glyceraldehyde-3-phosphate dehydrogenase, cytosolic 3
M7YUQ6	Aspartic proteinase oryzasin-1
M8APZ6	Nucleoside diphosphate kinase
M7ZMA4	Glutamate decarboxylase
M7ZK46	12S seed storage globulin 1
T1M4L5	Uncharacterized protein (Fragment)
Q5BLQ9	Grain softness protein-1B1 (Fragment)
T1M2W9	Uncharacterized protein
M7YVM9	Chitinase 1
T1MN05	Uncharacterized protein
M7ZRL7	Alanine aminotransferase 2
M7ZY01	Pullulanase 1, chloroplastic
M7ZB42	Calreticulin

Part 3 of 4 for Table C.3, continued from above.

M7ZY11	Uncharacterized protein
Q7M1M7	High-molecular-weight glutenin
Q8L6B4	Gamma gliadin
M7ZIE1	Calreticulin
M7ZCZ3	Uncharacterized protein
M7ZFH6	Alpha-galactosidase
Q2PCC5	Type 2 non specific lipid transfer protein (Precursor)
M7YLW1	Uncharacterized protein
M7Y803	Ubiquitin carboxyl-terminal hydrolase
M7ZZZ7	Coatomer subunit beta'-2
Q9SQG3	PR-4 (Fragment)
M7YIW7	Uncharacterized protein
M7ZEB5	Sulfurtransferase
M7Y780	Aldose 1-epimerase
M8AYL7	Alpha-L-arabinofuranosidase 1
M8A555	Uncharacterized protein
M7ZMH5	Isoamylase 1, chloroplastic
M7ZPU9	DEAD-box ATP-dependent RNA helicase 27
M7YRA6	Uncharacterized protein
M7YHL1	Uncharacterized protein
P33432	Puroindoline-A
Q6J160	S-type low molecular weight glutenin L4-55 (Fragment)
Q0Q5D8	High-molecular-weight glutenin By8
Q0WX49	Xylanase inhibitor TL-XI (Precursor)
A3FKE5	Superoxide dismutase (Fragment)
D2T2K0	Non-specific lipid-transfer protein (Fragment)
T1MDN2	Uncharacterized protein
M7ZX56	60S ribosomal protein L10-1
M7YQ69	Uncharacterized protein
M8A7U9	Enolase
M8A0V0	Dihydroxy-acid dehydratase
M7ZIZ5	Peroxidase 66
M8ARD9	Spermatogenesis-associated protein 20
T1LYJ4	Uncharacterized protein
M7ZB26	Isocitrate dehydrogenase [NADP]
M7Z1S0	Glycyl-tRNA synthetase 1, mitochondrial
T1L8G1	Uncharacterized protein
M7ZAQ9	Oxalate oxidase 2
T1LTB6	Uncharacterized protein (Fragment)
T1NA49	Uncharacterized protein
M7ZVA5	Uncharacterized protein
T1MPQ4	Uncharacterized protein
M8A0Y6	Uncharacterized protein
M7YYS2	Sister chromatid cohesion protein PDS5-like protein B
T1LDN5	Uncharacterized protein (Fragment)
M8A6W6	50S ribosomal protein L14
M7YGW3	Dihydroorotate dehydrogenase (Quinone), mitochondrial
M7YRC9	Uncharacterized protein
M7ZP14	Protein MAM3
M7ZM38	Actin-97
Q6RUI9	Glutamine synthetase
D3YE92	70 kDa heat shock protein
D8L9G7	Phosphorylase
M7ZVX5	Histone H2B.3

Part 4 of 4 for Table C.3, continued from above.

M7ZSG9	Catalase
M7ZIW6	RuBisCO large subunit-binding protein subunit beta, chloroplastic
M8A1S2	Trypsin/alpha-amylase inhibitor CMX1/CMX3
M7ZMF2	Uncharacterized protein
M8A8I0	Nucleoside diphosphate kinase
M7ZKL1	Uncharacterized protein
T1LIW9	Uncharacterized protein
M7ZU03	Mitochondrial import inner membrane translocase subunit TIM44
M8A584	Uncharacterized protein
M7YIH7	Prolyl 4-hydroxylase subunit alpha-2
M7ZH92	E3 ubiquitin-protein ligase UPL3
M7Z2F8	Uncharacterized protein

C.6 Putative biomarker proteins discovered in TMT set-1 and -2

Table C.4. GO Slims identifiers and their counts for putative biomarker proteins discovered from TMT set-1.

Slims GO ID	GO Term	Count	Process
GO:0006950	response to stress	2	Biological Process
GO:0005975	carbohydrate metabolic process	1	Biological Process
GO:0008152	metabolic process	1	Biological Process
GO:0009987	cellular process	1	Biological Process
GO:0019725	cellular homeostasis	1	Biological Process
GO:0005615	extracellular space	1	Cellular Component
GO:0005623	cell	1	Cellular Component
GO:0003824	catalytic activity	1	Molecular Function
GO:0005488	binding	1	Molecular Function
GO:0016787	hydrolase activity	1	Molecular Function

Table C.5. GO Slims identifiers and their counts for putative biomarker proteins discovered from TMT set-2.

Slims GO ID	GO Term	Count	Process
GO:0008152	metabolic process	4	Biological Process
GO:0009987	cellular process	4	Biological Process
GO:0005975	carbohydrate metabolic process	2	Biological Process
GO:0006950	response to stress	2	Biological Process
GO:0009058	biosynthetic process	2	Biological Process
GO:0019538	protein metabolic process	2	Biological Process
GO:0006412	translation	1	Biological Process
GO:0009056	catabolic process	1	Biological Process
GO:0009605	response to external stimulus	1	Biological Process
GO:0009607	response to biotic stimulus	1	Biological Process
GO:0005622	intracellular	3	Cellular Component
GO:0005623	cell	3	Cellular Component
GO:0005615	extracellular space	2	Cellular Component
GO:0005634	nucleus	1	Cellular Component
GO:0005737	cytoplasm	1	Cellular Component
GO:0005739	mitochondrion	1	Cellular Component
GO:0009536	plastid	1	Cellular Component
GO:0009579	thylakoid	1	Cellular Component
GO:0003824	catalytic activity	4	Molecular Function
GO:0005488	binding	4	Molecular Function
GO:0016787	hydrolase activity	3	Molecular Function
GO:0000166	nucleotide binding	2	Molecular Function
GO:0003677	DNA binding	1	Molecular Function
GO:0005515	protein binding	1	Molecular Function
GO:0016740	transferase activity	1	Molecular Function

C.7 R-script: 'PCA.R'

```
# BiocManager::install("UniProt.ws", version = "3.8")
# # Install "ggbiplot" if not already installed
# install_github("vqv/ggbiplot")
options(scipen=999)
if(.Platform$OS.type == "windows"){
      Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jdk1.8.0_191')
}
library("UniProt.ws")
library("pheatmap")
library("tidyr")
library("rJava")
detach("package:rJava", unload=TRUE)
library("ggbiplot")
library("dplyr")
library("RColorBrewer")
library("gridExtra")
library("grid")
library("tcltk")
library("rChoiceDialogs")
library("ggplot2")
library("readxl")
```

```
#################### SAVE HEATMAP Function ####################
#Save pheatmap function
save_pheatmap <- function(x, filename, width=1500, height=800) {
      stopifnot(!missing(x))
      stopifnot(!missing(filename))
      png(filename = filename, width = width, height=height)
      grid::grid.newpage()
      grid::grid.draw(x$gtable)
      dev.off()
}

unfactorize <- function(df){
      for(i in which(sapply(df, class) == "factor")) df[[i]] =
as.character(df[[i]])
      return(df)
}

####!Function to split a dataframe of "AK" and "MLOC" barley identifiers
into two vectors
splitBarleyIds = function(idInput){
      #Find only the idInput beginning with "AK"
      akIdentifiers = as.character(idInput[grep("AK[0-9]*|AK[0-9]*\\..*",
idInput[[1]]), ])
      #If you need to you can remove the decimal point from the Identifier
      akIdentifiers = gsub("(AK.*)(\\.[0-9]*$)", "\\1", akIdentifiers)
      #Find only the identifiers beginning with "MLOC"
      mlocIdentifiers = as.character(idInput[grep("MLOC_[0-9]*|MLOC_[0-
9]*\\..*", idInput[[1]]), ])
      #Remove the decimal point from the Identifier
      #mlocIdentifiers = gsub("(AK.*)(\\.[0-9])", "\\1", mlocIdentifiers)
      allList = list(akIdentifiers = akIdentifiers, mlocIdentifiers =
mlocIdentifiers)
      return(allList)
}


getUniprotFromMlocAk = function(idsList){
      #Load the "UniProt.ws" package into R
      library("UniProt.ws")
      #Get "ak" identifers from the list named "idsList"
      #From "splitBarleyIds" function
      akIdentifiers = unlist(idsList["akIdentifiers"])
      #Get "MLOC" identifers from the list named "idsList"
      #From "splitBarleyIds" function
      mlocIdentifiers = unlist(idsList["mlocIdentifiers"])
      #Set the Taxon number for Barley (Wheat = 4565)
      speciesId <- UniProt.ws(taxId=112509)
      #Key Type or Database the program looks into for "AK" identifiers
      ak_kt = "EMBL/GENBANK/DDBJ"
      #Key Type or Database the program looks into
      mloc_kt = "ENSEMBL_GENOMES PROTEIN"
      #Data columns that will be output
      columns <- "UNIPROTKB"
      #The command to retrive UniProt Identifiers from "AK" Identifiers (if
they exist)
      akRetrieve <- UniProt.ws::select(speciesId, akIdentifiers, columns,
ak_kt)
      #Change the name of the first column in the akRetrieve data.frame
      names(akRetrieve)[1] = "Identifier_Input"
      #The command to retrive UniProt Identifiers from "MLOC" Identifiers
```

```
(if they exist)
      mlocRetrieve = UniProt.ws::select(speciesId, mlocIdentifiers, columns,
mloc_kt)
      #Change the name of the first column in the mlocRetrieve data.frame
      names(mlocRetrieve)[1] = "Identifier_Input"
      #Join the tables together
      retrieveAll = rbind(akRetrieve, mlocRetrieve)
      return(retrieveAll)
}

convertBarleyIds = function (AllIdsColumn) {
      idInput = AllIdsColumn
      #List of MLOC and AK Identifiers
      idsList = splitBarleyIds(idInput)
      #Get Uniprot Ids from MLOC and AK Identifier List
      #Result is a data.frame
      idsListSplit = getUniprotFromMlocAk(idsList)
      #Find the rows of the data.frame that have NULL (missing) values in
UniProt column
      getIdsWithMissingUniprot                                          =
idsListSplit[is.na(idsListSplit$UNIPROTKB),]
      #Remove the UniProt column and turn the AK and MLOC Ids into a character
vector
      getVectorOfMissingIds = getIdsWithMissingUniprot$Identifier
      if (length(getVectorOfMissingIds) > 0) {
            #Make a small FASTA file from the vector of AK and MLOC Ids
            mkSmallFASTAList   =   makeSmallFASTAList(getVectorOfMissingIds)
#Input a vector of characters ("Identifiers")
            #Use the "mkSmallFASTA" list in memory to perform a BLASTP
            #The result is a dataframe of MLOC and AK Identifiers in one
column
            #UniProt identifiers in the other column
            blastpTable = blastpResult(mkSmallFASTAList)
            # Remove the descriptions and keep identifiers
            blastpTable$Identifier_Input    =    gsub("(MLOC_[0-9]*|MLOC_[0-
9]*\\.[0-9]*|AK[0-9]*|AK[0-9]*\\.[0-9]*)(_)(.*)",

                                          "\\1",
blastpTable$Identifier_Input)
            # Remove any decimal numbers from identifiers so the vector will
match
            blastpTable$Identifier_Input    =    gsub("(AK[0-9]*)\\.[0-9]*",
"\\1", blastpTable$Identifier_Input)
            # The blastpTable and idsListSplit are combined, giving a full
list of Uniprot and AK, MLOC Identifiers
            tableOfIdsAndBlast     =     idsListAndBlastptable(blastpTable,
idsListSplit)
            # Remove any rows with missing data
            # tableOfIdsAndBlast = na.omit(tableOfIdsAndBlast)
            return(tableOfIdsAndBlast)
      }
      tableOfIdsAndBlast = idsListSplit
      return(tableOfIdsAndBlast)
}


getBarleyDescription = function(barleyIds, proteinFastaFile) {
      library("seqinr")
      #Load the "fasta" file into a list format via the "seqinr" package
      fastaFile  <-  read.fasta(file  =  proteinFastaFile,  seqtype  =  "AA",
as.string = TRUE)
```

```
    # Seqinr function "getName" to get sequence names
    seqNames = getName(fastaFile)
    #Tidy up the names so that "getAnnot" will work well
    listNamesFasta = gsub("(MLOC_[0-9]*|MLOC_[0-9]*\\.[0-9]*|AK[0-
9]*|AK[0-9]*\\.[0-9]*)(_)(.*)", "\\1", names(fastaFile))
    names(fastaFile) = listNamesFasta
    trimmedFasta = fastaFile[charmatch(barleyIds, names(fastaFile))]
    #Get annotations from each listed item (protein)
    trimmedFastaAnnot = getAnnot(trimmedFasta)
    #Remove any NULL entries from the "trimmedFastaAnnot" list
    trimmedFastaAnnot = trimmedFastaAnnot[!sapply(trimmedFastaAnnot,
is.null)]
    trimmedFastaAnnot = unlist(trimmedFastaAnnot)
    trimmedFastaAnnot = gsub("(MLOC_[0-9]*|MLOC_[0-9]*\\.[0-9]*|AK[0-
9]*|AK[0-9]*\\.[0-9]*)(_)(.*)", "\\1\\|\\3", trimmedFastaAnnot)
    trimmedFastaAnnot = gsub(">", "", trimmedFastaAnnot)
    trimmedFastaAnnot = strsplit(trimmedFastaAnnot, split = "\\|", fixed =
FALSE, perl = FALSE, useBytes = FALSE)
    ##### Function Call #####
    finalIdAndDescriptionDF = getDescriptionDF(trimmedFastaAnnot)
    #####
    finalIdAndDescriptionDF["description"] =
gsub(".*(unknown.protein).*|.*(unknown.function).*", "Unknown protein",
ignore.case = TRUE, finalIdAndDescriptionDF$description)
    return(finalIdAndDescriptionDF)
}


getWheatDescription = function(wheatIds, proteinFastaFile){
    library("seqinr")
    #Load the "fasta" file into a list format via the "seqinr" package
    fastaFile <- read.fasta(file = proteinFastaFile, seqtype = "AA",
as.string = TRUE)
    # Seqinr function "getName" to get sequence names
    seqNames = getName(fastaFile)
    #Tidy up the names so that "getAnnot" will work well
    listNamesFasta = gsub("^.*\\|(.*)\\|.*", "\\1", names(fastaFile))
    names(fastaFile) = listNamesFasta
    trimmedFasta = fastaFile[charmatch(wheatIds, names(fastaFile))]
    #Get annotations from each listed item (protein)
    trimmedFastaAnnot = getAnnot(trimmedFasta)
    #Remove any NULL entries from the "trimmedFastaAnnot" list
    trimmedFastaAnnot = trimmedFastaAnnot[!sapply(trimmedFastaAnnot,
is.null)]
    trimmedFastaAnnot = unlist(trimmedFastaAnnot)
    idList = gsub("^.*\\|(.*)\\|.*", "\\1", trimmedFastaAnnot)
    trimmedFastaDesc = gsub("(^.*\\|.*\\|)(.*) OS=.*$", "\\2",
trimmedFastaAnnot)
    trimmedFastaDesc = gsub("^[A-Z|0-9]*_[A-Z|0-9]* ", "",
trimmedFastaDesc)
    #trimmedFastaDesc = gsub(" ", "_", trimmedFastaDesc)
    finalIdAndDescriptionDF = data.frame(idList, trimmedFastaDesc)
    names(finalIdAndDescriptionDF) = c("Identifier_Input", "description")
    return(finalIdAndDescriptionDF)
}

#--------------------- END FUNCTIONS ---------------------#
#-------------------------------------------------------#
```

```
if(.Platform$OS.type == "windows"){
      designPath = tk_choose.files(default =
"C:/Users/paul_/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall",
caption = "Select the heatmap design",

             multi = FALSE, filters = NULL, index = 1)
} else {
      designPath = tk_choose.files(default =
"~/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall", caption =
"Select the heatmap design",

             multi = FALSE, filters = NULL, index = 1)
}

if(.Platform$OS.type == "windows"){
      dataTablePath = tk_choose.files(default =
"C:/Users/paul_/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall",
caption = "Select the table of values for heatmap",

                        multi = FALSE, filters = NULL, index = 1)
} else {
      dataTablePath = tk_choose.files(default =
"~/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall", caption =
"Select the table of values for heatmap",

                        multi = FALSE, filters = NULL, index = 1)
}

if(.Platform$OS.type == "windows"){
      fastaDatabaseFile = tk_choose.files(default =
"C:/Users/paul_/Google_Drive/PhD/DataBases/Protein/Databases_used_in_Mascot
_Search", caption = "FASTA database path",

                        multi = FALSE, filters = NULL, index = 1)
} else {
      fastaDatabaseFile = tk_choose.files(default =
"~/Google_Drive/PhD/DataBases/Protein/Databases_used_in_Mascot_Search",
caption = "FASTA database path",

                        multi = FALSE, filters = NULL, index = 1)
}

if(.Platform$OS.type == "windows"){
      SlimsSummaryAllFile = tk_choose.files(default =
"C:/Users/paul_/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall/Sl
ims_Summary_Outputfiles", caption = "Full Slims table (fullSlimsTable.csv)",

                                    multi = FALSE, filters = NULL, index = 1)
} else {
      SlimsSummaryAllFile = tk_choose.files(default =
"~/Google_Drive/PhD/TMT_Results/TMTSummaries/ResultsOverall/Slims_Summary_O
utputfiles", caption = "Full Slims table (fullSlimsTable.csv)",

                                    multi = FALSE, filters = NULL, index = 1)
}

studyDirPath = dirname(dataTablePath)
studyName = gsub(".*/(.*)$", "\\1", studyDirPath)
outDir = paste0(studyDirPath, "/Pheatmap_Outputfiles")
if (!file.exists(outDir)) {
      dir.create(outDir)
```

```
} else {
      print("File Exists!")
}


designFile = gsub("^.*\\/(.*)\\.xlsx$", "\\1", designPath)
dataTableFile = gsub("^.*\\/(.*)\\.xlsx$", "\\1", dataTablePath)


TMTDesign = read_excel(path = designPath, sheet = "design", range =
cell_cols("A:C"))
TMTDesign = TMTDesign[colSums(!is.na(TMTDesign)) > 0]



dataTable = read_excel(path = dataTablePath, sheet = "AllData", range =
cell_cols("A:O"))
dataTable = dataTable[order(dataTable$Clusters),]
# dataTable = na.omit(dataTable)
names(dataTable)[1] = "Identifier_Input"






#### Determine whether the identifiers are Wheat or Barley and start the
initial tidy of data
prepPro_AllIds = dataTable
if (any(grepl("^MLOC_|^AK[0-9]*", prepPro_AllIds[[1]]))) {
      species = "Barley"
      proteinFastaFile = fastaDatabaseFile
} else {
      species = "Wheat"
      proteinFastaFile = fastaDatabaseFile
}
print(species)
#prepPro_AllIds = prepPro_AllIds[order(prepPro_AllIds$Clusters),]
########## Make a dataframe of IDs only ##########
AllIdsColumn = prepPro_AllIds[1]
names(AllIdsColumn) = "Identifier_Input"
# Original Identifiers and then add uniprot Ids plus descriptions
start.time <- Sys.time()
if (species == "Barley") {
      SlimsSummaryAllsub = read.csv(SlimsSummaryAllFile, stringsAsFactors =
FALSE)
      SlimsSummaryAllsub = SlimsSummaryAllsub[ , c("Identifier_Input",
"uniprot", "description")]
      Identifier_Input = AllIdsColumn
      Identifier_Input$Identifier_Input = gsub("(^AK.*)\\..$", "\\1",
Identifier_Input$Identifier_Input)
      idsTable = left_join(Identifier_Input , SlimsSummaryAllsub, by =
"Identifier_Input")
} else if (species == "Wheat") {
      uniprot = AllIdsColumn
      names(uniprot) = "uniprot"
      idsTable = cbind(AllIdsColumn, uniprot)
      Identifier_Input = idsTable$Identifier_Input
      descriptionTable = getWheatDescription(Identifier_Input,
proteinFastaFile)
      idsTable$description =
descriptionTable$description[match(idsTable$Identifier_Input,
descriptionTable$Identifier_Input)]
}
end.time <- Sys.time()
time.taken <- end.time - start.time
time.taken



sampleNames = names(dataTable)[grep("^R1.X(.*$)", names(dataTable))]
comparisonLabel = unique(gsub("^.*\\.(.*$)", "\\1", sampleNames))
# Save the row to be deleted if needed
TMTDesignLabelRow = TMTDesign[grep(comparisonLabel, TMTDesign$Label), ]
TMTDesign = TMTDesign[grep(comparisonLabel, TMTDesign$Label, invert = TRUE),
]


sampleDF = data.frame(Sample = sampleNames, stringsAsFactors = FALSE)
sampleDF$Label = gsub("^R1.X(.*$)", "\\1", sampleDF$Sample)
sampleDF$Label = gsub("(^.*)\\..*$", "\\1", sampleDF$Label)

fullDesign = TMTDesign %>% full_join(sampleDF, by = "Label")
fullDesign = fullDesign[ ,c(ncol(fullDesign), 1:3)]
fullDesign[ , "Sample"] = gsub("^R1\\.X(.*)$", "\\1", fullDesign$Sample)


                                              # #This is the full design to
aid in making other plots
                                              #      #such as violin plots
                                              # fullDesign = read.csv(file =
designPath, header = TRUE, stringsAsFactors = FALSE)
                                              #    names(fullDesign)[1]    =
"Sample"
                                              #  fullDesign[  ,  "Sample"]  =
gsub("^R1\\.X(.*)$", "\\1", fullDesign$Sample)



#The design to be used for making a matrix for the heatmap
Design = fullDesign
rowNamesDesign = Design$Sample
                  rowNamesDesign = gsub("^R1.X(.*$)", "\\1", rowNamesDesign)
                  Design[,"Sample"] = NULL
                  rownames(Design) = rowNamesDesign


#Merge datatable and dataTable to get identifiers
dataTable$Identifier_Input      =       gsub("(^AK.*)\\..$",       "\\1",
dataTable$Identifier_Input)
dataTable = dataTable %>% full_join(idsTable, by = "Identifier_Input")
idsAndDescriptions = dataTable[ ,c(1,ncol(dataTable))]
                  names(dataTable)     =     gsub("^R1.X(.*$)",     "\\1",
names(dataTable))
namesWanted = names(dataTable[c(1,ncol(dataTable))])
dataTable = dataTable[ ,c(namesWanted, rowNamesDesign)]
dataTable = unfactorize(dataTable)
dataTable = na.omit(dataTable)
dataTableNums = dataTable[,rowNamesDesign]
#dataTable = cbind(idsAndDescriptions, dataTableNums)
```

```
write.csv(dataTable, file = paste0(outDir, "/All_", dataTableFile,
"_dataTable.csv"), row.names = FALSE)
dataTableWriteLog = cbind(dataTable[ , c(1,2)], log10(dataTable[ ,
c(3:11)]))
write.csv(dataTableWriteLog, file = paste0(outDir, "/All_", dataTableFile,
"_dataTable_log10_All.csv"), row.names = FALSE)

# dataTable = read.csv(file = dataTablePath, row.names = 1, header = TRUE,
stringsAsFactors = FALSE)
dataTable.log.ids = log10(dataTable[ ,rowNamesDesign])
row.names(dataTable.log.ids) = dataTable$Identifier_Input

### Annotate our heatmap (optional)
annotation <- data.frame(Group = Design[,"Group"],
row.names=row.names(Design))

# Reorder Density levels
annotation$Group = factor(annotation$Group, levels =
unique(annotation$Group))
choiceColours = c("red", "green", "blue", "yellow", "orange", "purple")

AnnCol = choiceColours[1:(length(levels(annotation$Group)))]
names(AnnCol) <- levels(annotation$Group)


AnnColour <- list(
      Group = AnnCol)



      # Clustering Distance options = "euclidean", "maximum", "manhattan",
"canberra",
      # "binary" or "minkowski"
      # CLustering Method options = "ward.D", "ward.D2", "single",
"complete", "average",
      # "mcquitty", "median", "centroid"
sigProteins_heatmap = pheatmap(dataTable.log.ids, color =
colorRampPalette(c("lawngreen", "black", "firebrick1"))(100),
                        cluster_rows = TRUE, show_rownames=TRUE, annotation=
annotation, annotation_colors = AnnColour, border_color = NA,
                        width = 10, height = 8, fontsize = 10, fontsize_row
= 6, scale = "row", annotation_names_col = FALSE,
                        clustering_distance_cols = "manhattan",
clustering_method = "complete",
                        legend_breaks = c(-2,-1,0,1,2,2.1), main = "",
legend_labels = c("-2", "-1", "0", "1", "2", "      \n\n(log10)"))


# Draw grobs to improve the look of the graph
fillerRectangle = grid.rect(width = 0.5, height = 0.5, gp = gpar(fill =
"white", col = "white", alpha = 0.8))
grid.arrange(fillerRectangle, sigProteins_heatmap[[4]], fillerRectangle,
nrow=1, widths = c(1,20,1))

# Save the grobs (including the pheatmap heatmap) to disk
png(paste0(outDir, "/All_", dataTableFile, "_ids.png"), width = 10, height
= 8, res = 600, units = "in") # Open a new pdf file
grid.arrange(fillerRectangle, sigProteins_heatmap[[4]], fillerRectangle,
nrow=1, widths = c(1,20,1))
dev.off()
```

```
# With protein descriptions in the Y-axis instead of identifiers
dataTable.log.desc = log10(dataTable[ ,rowNamesDesign])
rowNamesHeatmapDesc = make.names(gsub(" ", "_", dataTable$description),
unique = TRUE)
row.names(dataTable.log.desc) = rowNamesHeatmapDesc

# The below graph is too messy with the descriptions
            # sigProteins_heatmap_with_Desc = pheatmap(dataTable.log.desc,
color = colorRampPalette(c("lawngreen", "black", "firebrick1"))(100),
            #
                              cluster_rows = TRUE, show_rownames=TRUE,
annotation= annotation, annotation_colors = AnnColour, border_color = NA,
            #
                              width = 10, height = 8, fontsize = 10,
fontsize_row = 6, scale = "row", annotation_names_col = FALSE,
            #
                              clustering_distance_cols = "manhattan",
clustering_method = "complete",
            #
                              legend_breaks = c(-2,-1,0,1,2,2.1), main = "",
legend_labels = c("-2", "-1", "0", "1", "2", "      \n\n(log10)"))
            #
            #
            # # Draw grobs to improve the look of the graph
            # fillerRectangle = grid.rect(width = 0.5, height = 0.5, gp =
gpar(fill = "white", col = "white", alpha = 0.8))
            #    grid.arrange(fillerRectangle,    sigProteins_heatmap[[4]],
fillerRectangle, nrow=1, widths = c(1,20,1))
            #
            # # Save the grobs (including the pheatmap heatmap) to disk
            # png(paste0(outDir, "/All_", dataTableFile, "_desc.png"), width
= 10, height = 8, res = 600, units = "in") # Open a new pdf file
            #                                  grid.arrange(fillerRectangle,
sigProteins_heatmap_with_Desc[[4]],  fillerRectangle,  nrow=1,  widths  =
c(1,20,1))
            # dev.off()


#--------------------- PCA ---------------------#
dataTable2 = dataTable[c(1,3:11)]
dataTableTranspose = t(dataTable2)
dfNames = dataTableTranspose[1,]
dataTableTranspose  =  data.frame(dataTableTranspose,  stringsAsFactors  =
FALSE)
dataTableTranspose = dataTableTranspose[c(2:nrow(dataTableTranspose)), ]
names(dataTableTranspose) = dfNames
dataTableTranspose$TMT = row.names(dataTableTranspose)
dataTableTranspose     =     dataTableTranspose[c(ncol(dataTableTranspose),
1:(ncol(dataTableTranspose)-1))]
row.names(dataTableTranspose) = NULL
dataTableTranspose[        ,        c(2:ncol(dataTableTranspose))]        =
as.data.frame(lapply(dataTableTranspose[,c(2:ncol(dataTableTranspose))],
as.numeric))


dataTableTranspose["Group"] = NA

for (i in 1:nrow(dataTableTranspose)) {
      dataTableTranspose$Group[i]                                         =
fullDesign$Group[match(dataTableTranspose$TMT[i], fullDesign$Sample)]
}
```

```
dataTableTranspose  =  dataTableTranspose[c(1,  c(ncol(dataTableTranspose),
2:(ncol(dataTableTranspose)-1)))]
# Generate PCA
dataTablePCA_Result        <-        prcomp(dataTableTranspose[        ,
c(3:ncol(dataTableTranspose))],
                                                   center = TRUE,
                                                   scale. = TRUE)
# Prepare graph
ir.species = dataTableTranspose[ ,"Group"]

library(ggbiplot)
PCA_Plot  <-  ggbiplot(dataTablePCA_Result,  obs.scale  =  1,  var.scale  =  1,
var.axes = FALSE,
                                        groups = ir.species, ellipse = TRUE,
                                        circle = TRUE)
PCA_Plot <- PCA_Plot + scale_color_discrete(name = '')
PCA_Plot <- PCA_Plot + theme(legend.direction = 'horizontal',
                                         legend.position = 'top')

print(PCA_Plot)
PCA_Plot_Figure  =  ggsave(file  =  paste0(outDir,  "/All_",  dataTableFile,
"_PCA_plot.png"), plot = PCA_Plot, h=6, w=6, units="in",dpi=600)

#-------------------- END PCA --------------------#
#------------------------------------------------#




#################### VIOLIN AND BOX PLOTS ####################
# prepare data for Violin Plot
#New dataframe name
dataTable.log.df = dataTable.log.ids
#Add column
dataTable.log.df$Protein = row.names(dataTable.log.ids)
# Re-arrange columns to improve table
dataTable.log.df     =     dataTable.log.df[     ,c(ncol(dataTable.log.df),
1:(ncol(dataTable.log.df)-1))]
# Remove row names
row.names(dataTable.log.df) = NULL
# Make a wide table into a long table
dataTable.log.df_long = dataTable.log.df %>%
      gather(TMT, Log_10_value, 2:10)
# Add a new column called "Group" and fill it with empty values
dataTable.log.df_long["Group"] = NA


# Refer to the design table to replace "TMT" with the appropriate "Group"
name
for (i in 1:nrow(dataTable.log.df_long)) {
      dataTable.log.df_long$Group[i]                                      =
fullDesign$Group[match(dataTable.log.df_long$TMT[i], fullDesign$Sample)]
}

# Get order of x-axis from pheatmap
col.order = sigProteins_heatmap$tree_col$order

# Turn the "TMT" character variable into a factor and define levels
dataTable.log.df_long$TMT  =  factor(dataTable.log.df_long$TMT,  levels  =
unique(dataTable.log.df_long$TMT))
                       # # Change the "TMT" levels to be the same as the x-
```

```
axis from the "pheatmap"
                        #          dataTable.log.df_long$TMT          =
factor(dataTable.log.df_long$TMT,
levels(dataTable.log.df_long$TMT)[col.order])

# Draw a violin plot
violin_plot = ggplot(dataTable.log.df_long, aes(x = TMT, y = Log_10_value,
fill = Group)) +
      geom_violin()  +  ylab("Expression  (Log10)")  +  theme(axis.text.x  =
element_text(angle = 90))
violin_plot
ggsave(paste0(outDir, "/All_", dataTableFile, "_violin_plot.png"))

#Draw a boxplot with points of data underlaying it
box_plot = ggplot(dataTable.log.df_long, aes(x = TMT, y = Log_10_value, fill
= Group)) +
geom_boxplot(outlier.shape   =   NA)   +   ylab("Expression   (Log10)")   +
theme(axis.text.x = element_text(angle = 270))
box_plot = box_plot + geom_point(colour = "black", size = 4, alpha = 0.2)
box_plot
ggsave(paste0(outDir, "/All_", dataTableFile, "_box_plot.png"), dpi = 600,
units = "cm", width = 29.7, height = 21)

#Draw a violin plot with points of data underlaying it
dot_violin_plot  =  ggplot(dataTable.log.df_long,  aes(x  =  TMT,  y  =
Log_10_value, fill = Group)) +
geom_violin() +
ylab("Expression (Log10)") + theme(axis.text.x = element_text(angle = 90))
dot_violin_plot = dot_violin_plot + geom_point(colour = "black", size = 4,
alpha = 0.2)
dot_violin_plot
ggsave(paste0(outDir, "/All_", dataTableFile, "_dot_violoin_plot.png"), dpi
= 600, units = "cm", width = 29.7, height = 21)
```

Appendix D.

D.1 Scatterplots

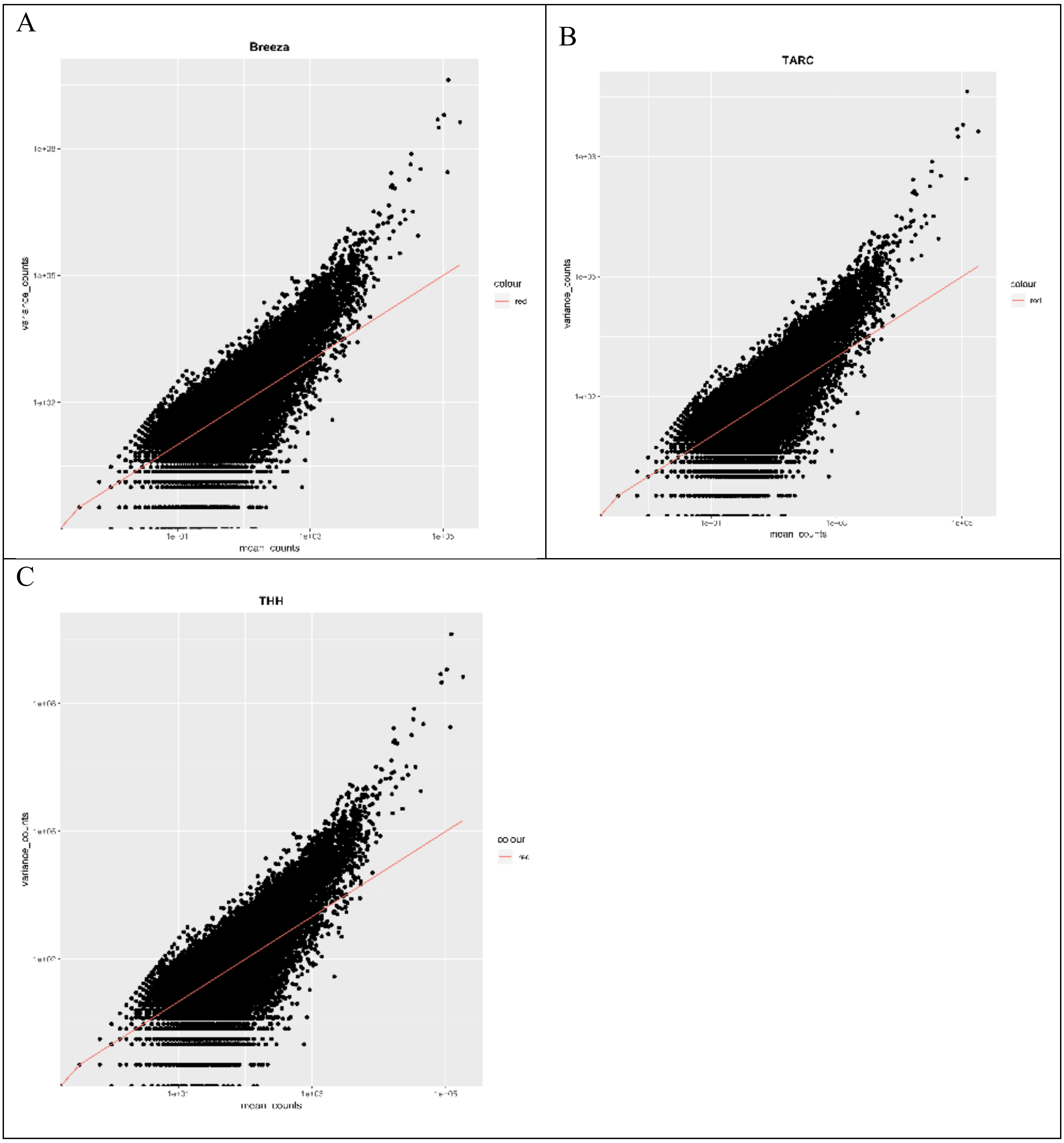


Figure D.1. Scatterplot of counts, mean versus variance for the three sample groups of Spitfire wheat grain, grown at Breeza, TARC, and THH farm. Data for both axes is log10 scale.

D.2 Heatmaps of sample-transcript expression

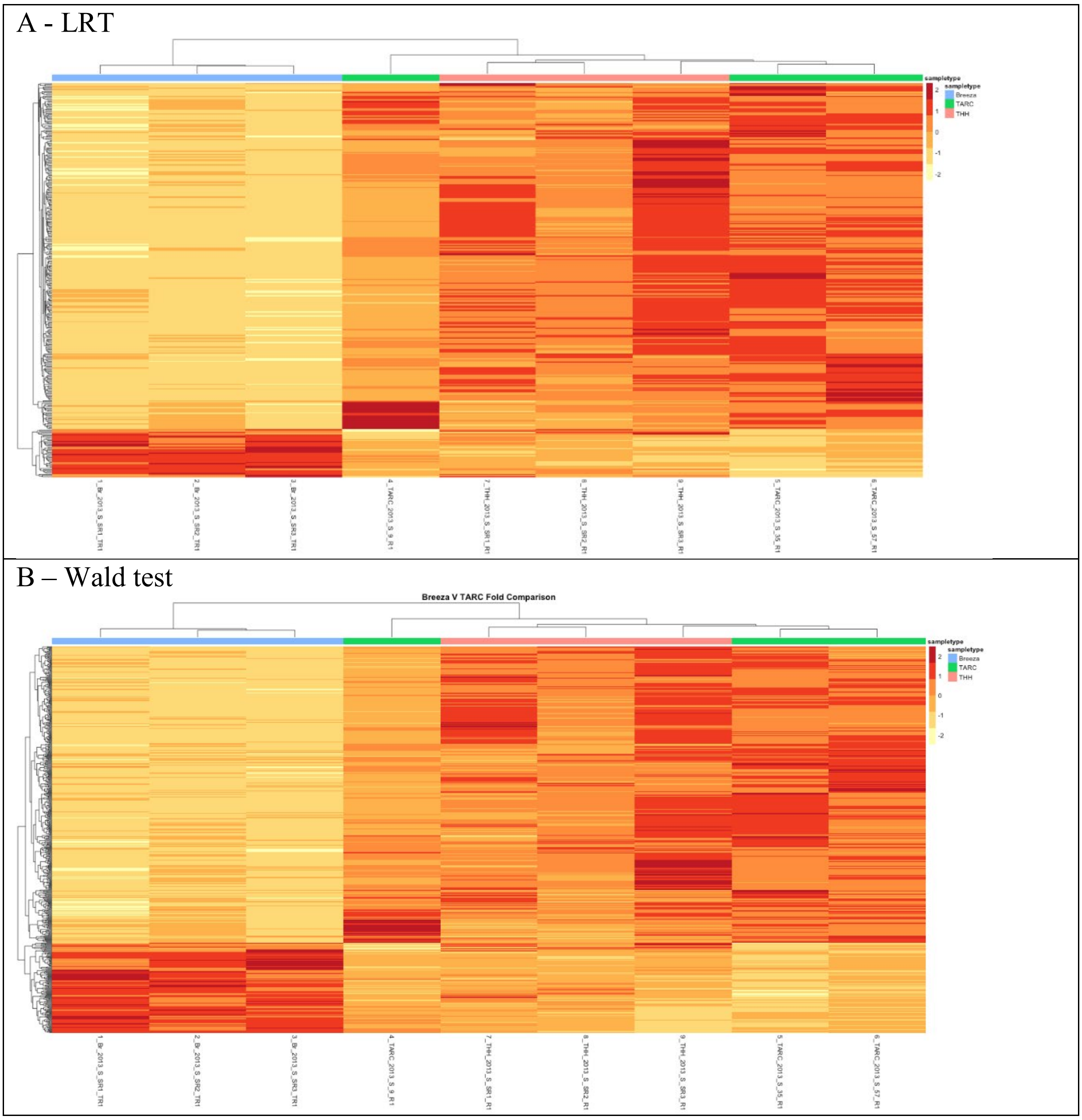


Figure D.2. Messenger RNA transcript expression from Spitfire wheat grain that was grown at Breeza, TARC, and THH farms. Results are after a Breeza and TARC comparison of expression data, with testing for significance using LRT (A) or the Wald test (B) to discover transcripts with statistically significant differential expression.

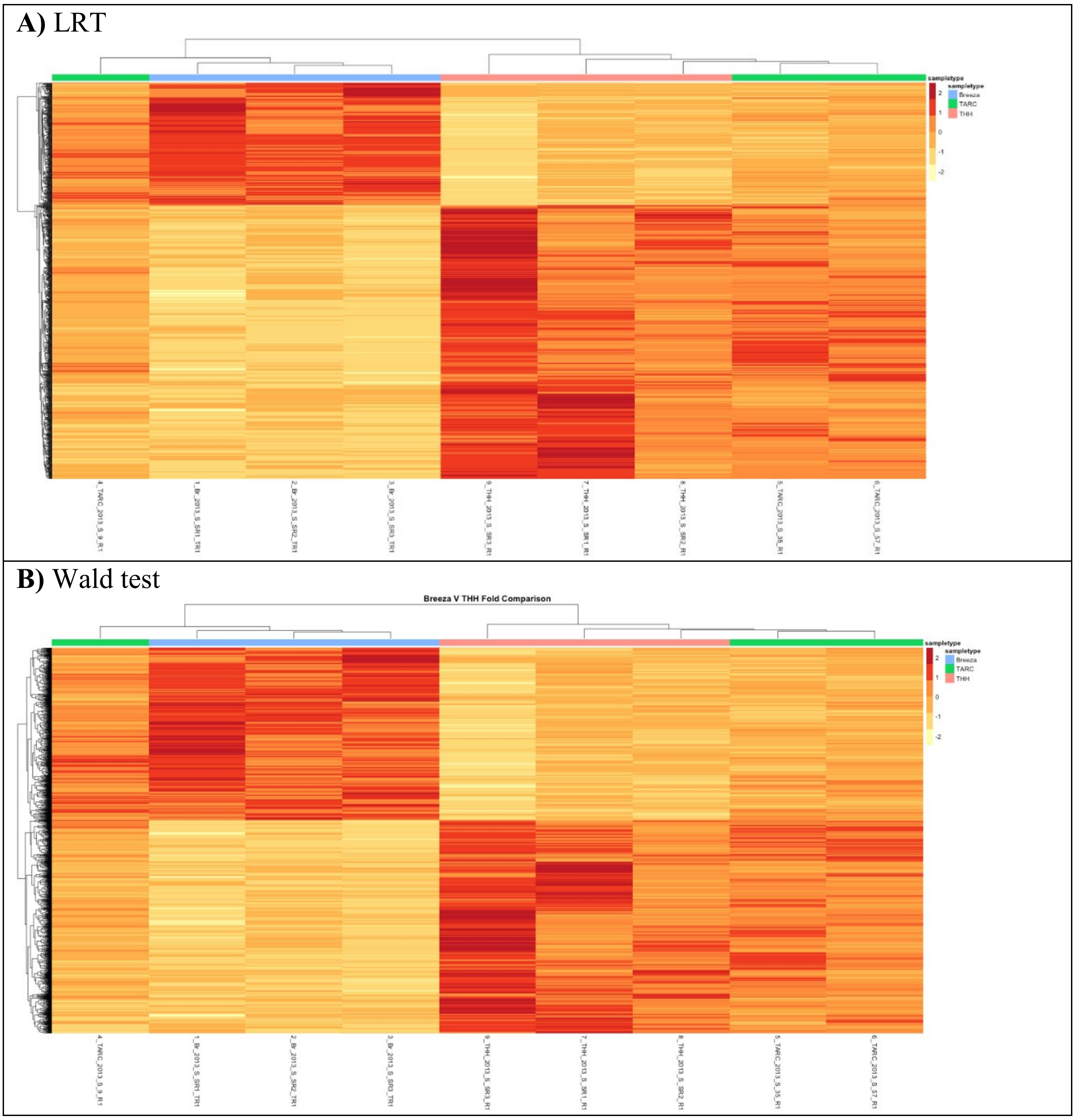


Figure D.3. Same as for Figure D.2, except the comparison was between Breeza and THH transcript expression data. LRT (A) or the Wald test (B) were also used to discover transcripts with statistically significant differential expression.

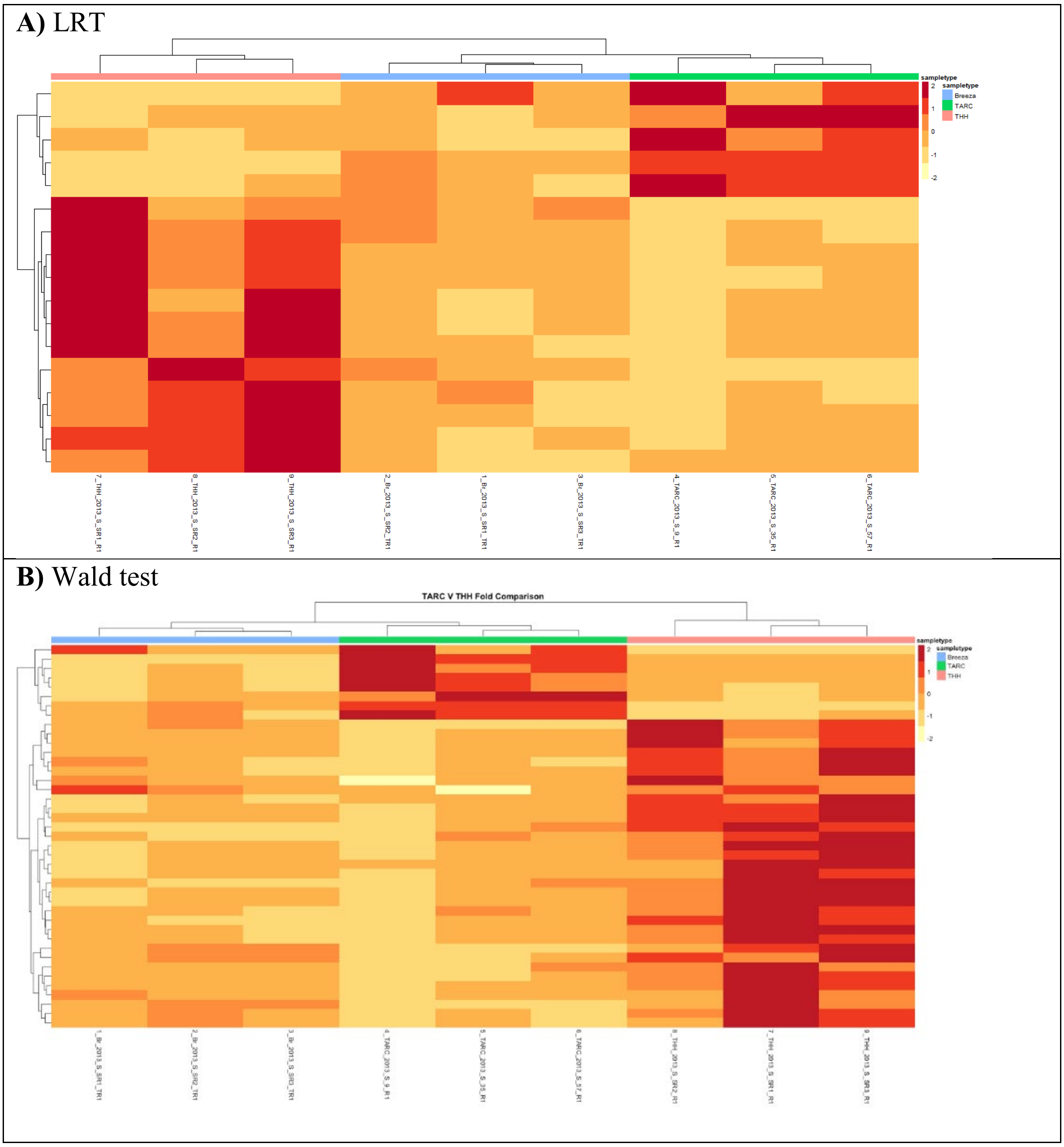


Figure D.4. Same as for Figure D.2, except the comparison was between TARC and THH transcript expression data. LRT (A) or the Wald test (B) were also used to discover transcripts with statistically significant differential expression.

D.3 Volcano plots of sample-transcript expression

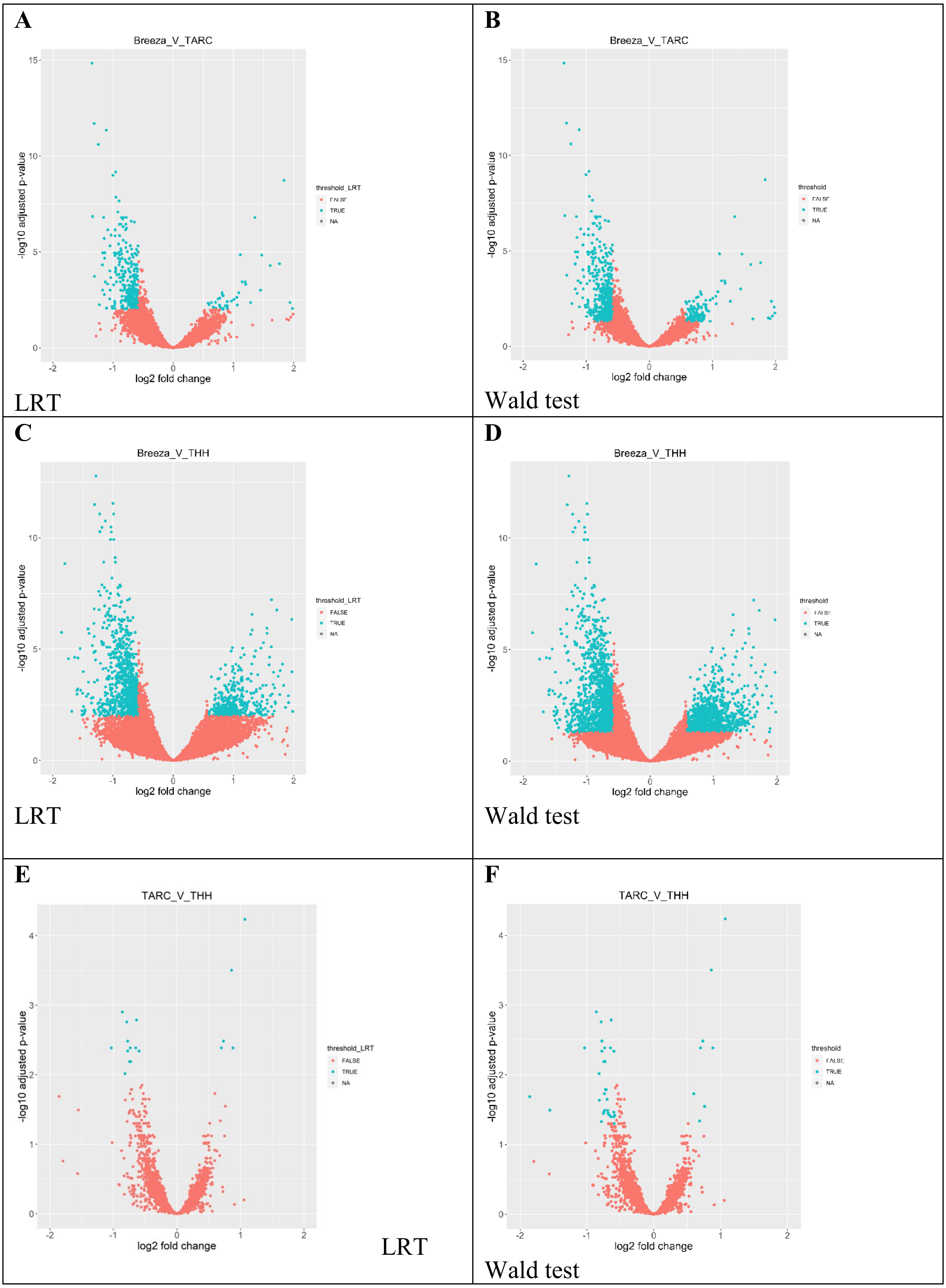
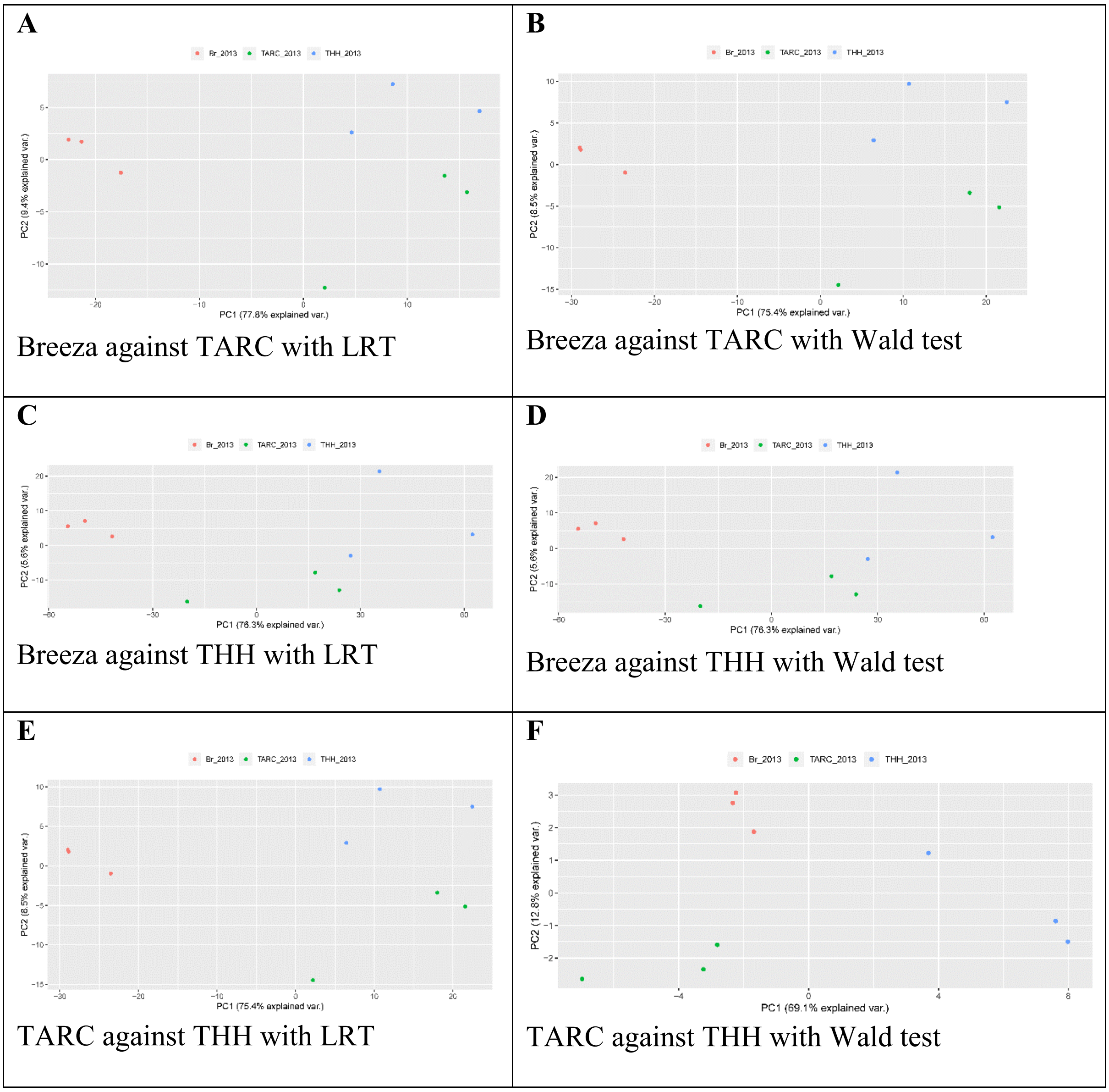


Figure D.5. Volcano plots of data after calculating differential expression of RNA transcripts.

D.4 PCAs of sample-transcript expression



Breeza against TARC with LRT

Breeza against TARC with Wald test

Breeza against THH with LRT

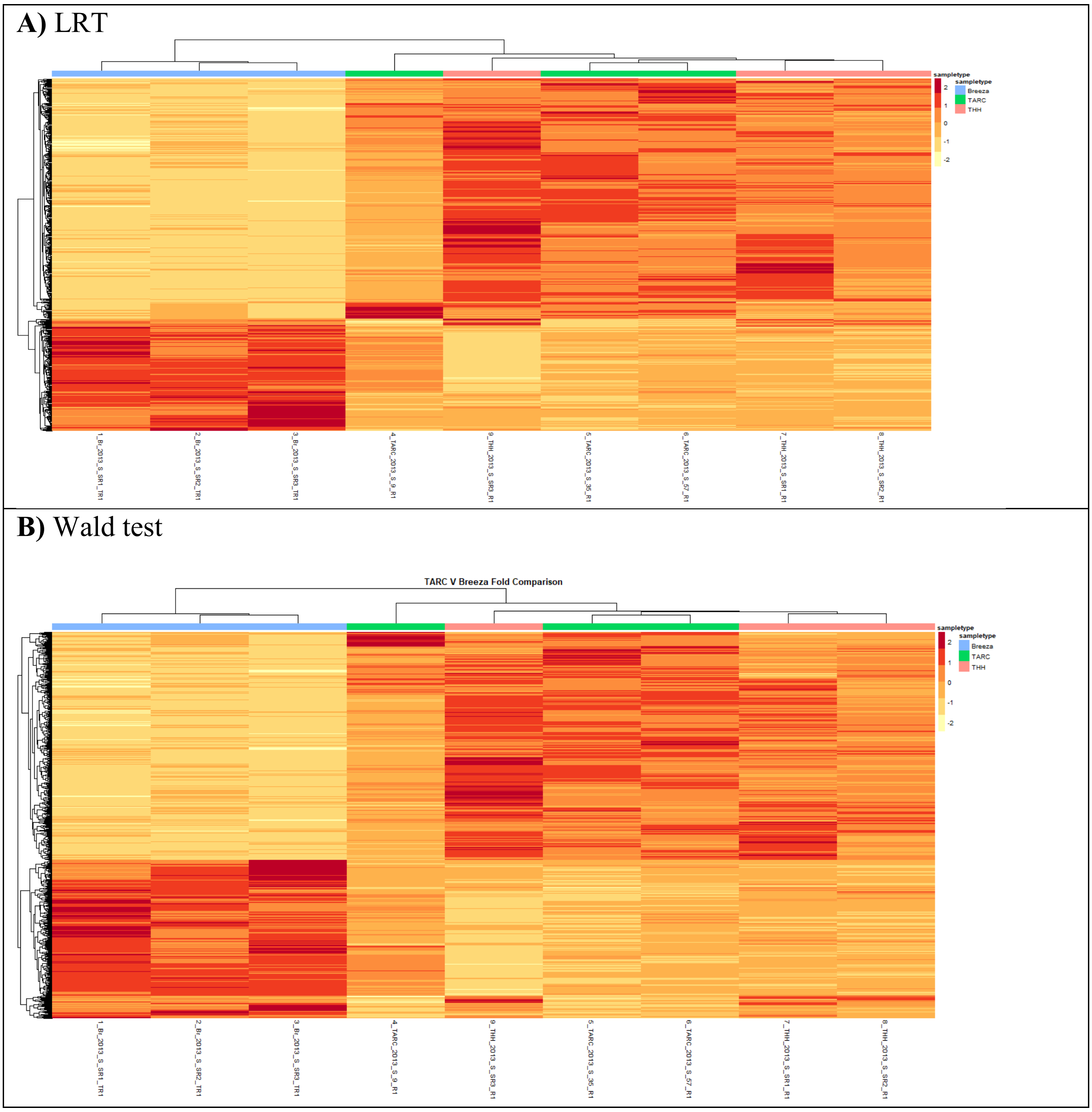
Breeza against THH with Wald test

TARC against THH with LRT

TARC against THH with Wald test

Figure D.6. PCA plots of sample data, with data from each sample representing differentially expressed transcripts detected after comparisons described above (A to F), followed by tests for significance using either LRT or Wald test (listed in A to F).

D.5 Heatmaps of sample-transcript expression – opposite comparison order



A) LRT

B) Wald test

Figure D.7. Messenger RNA transcript expression from Spitfire wheat grain that was grown at Breeza, TARC, and THH farms. Results are after a TARC and Breeza comparison of expression data, with testing for significance using LRT (A) or the Wald test (B) to discover transcripts with statistically significant differential expression.

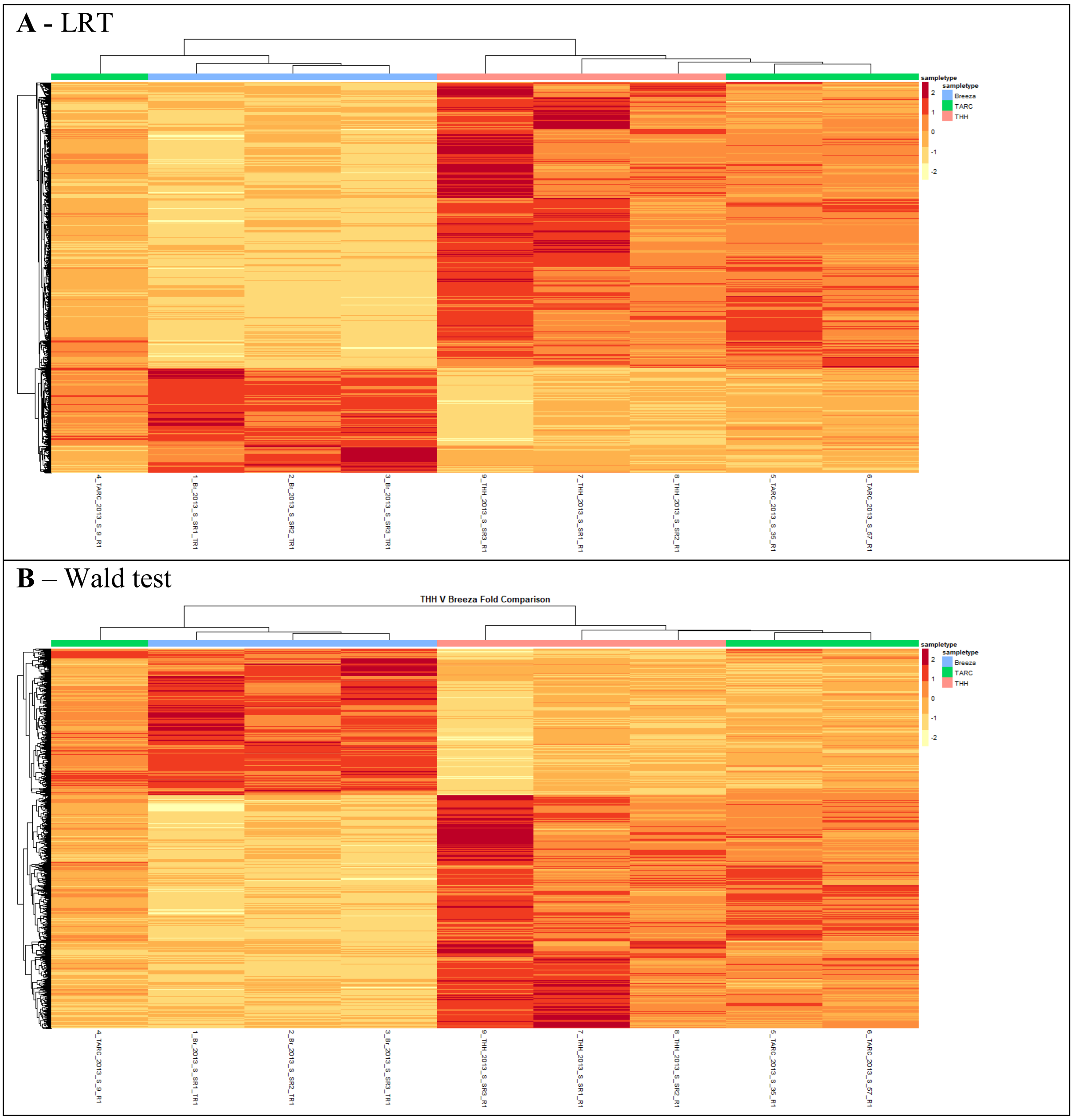


Figure D.8. THH compared with Breeza using LRT (A) or the Wald test (B) to discover transcripts with statistically significant differential expression.

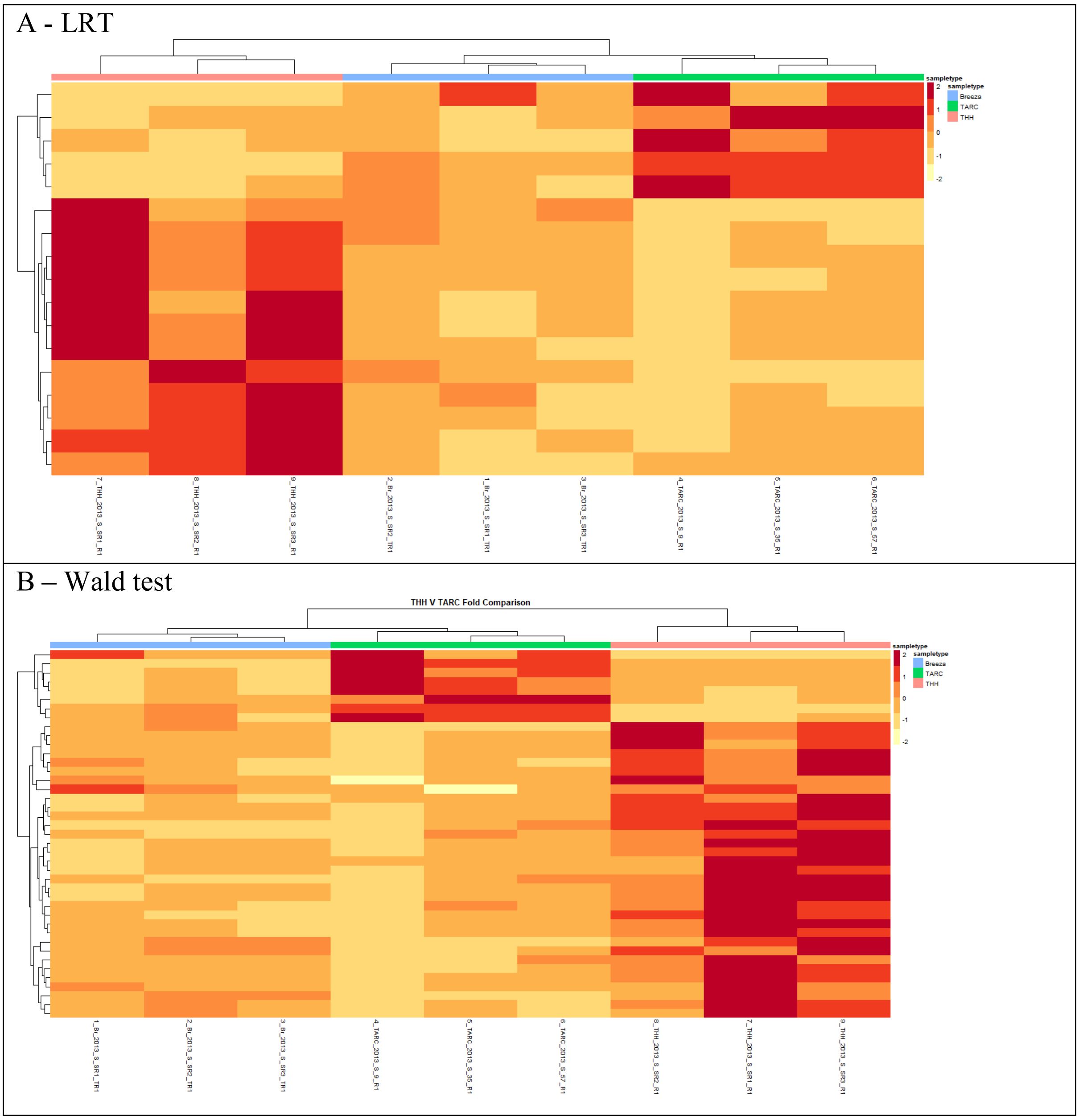


Figure D.9. THH compared with TARC using LRT (A) or the Wald test (B) to discover transcripts with statistically significant differential expression.

D.6 Volcano plots of sample-transcript expression - opposite comparison order

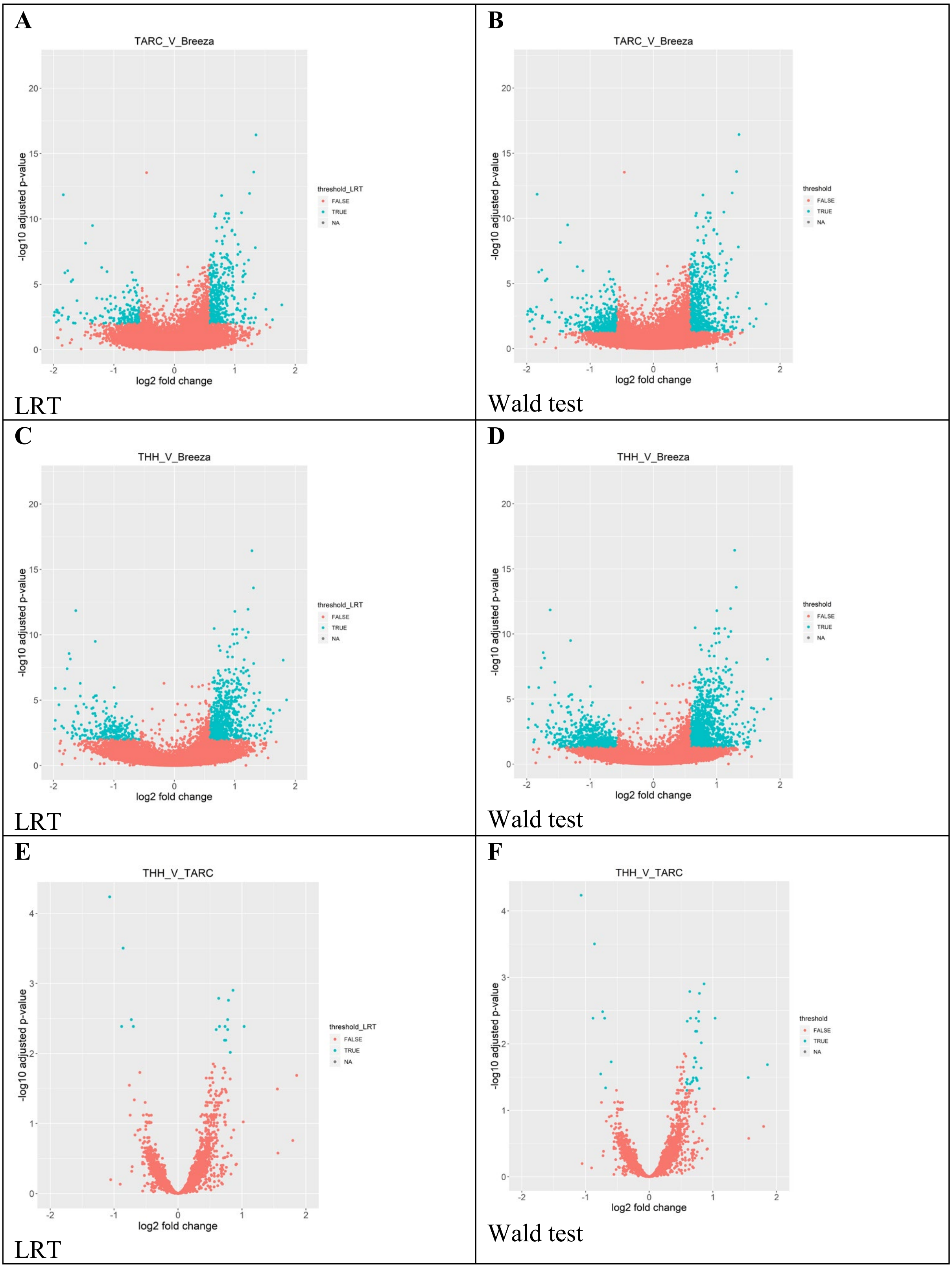
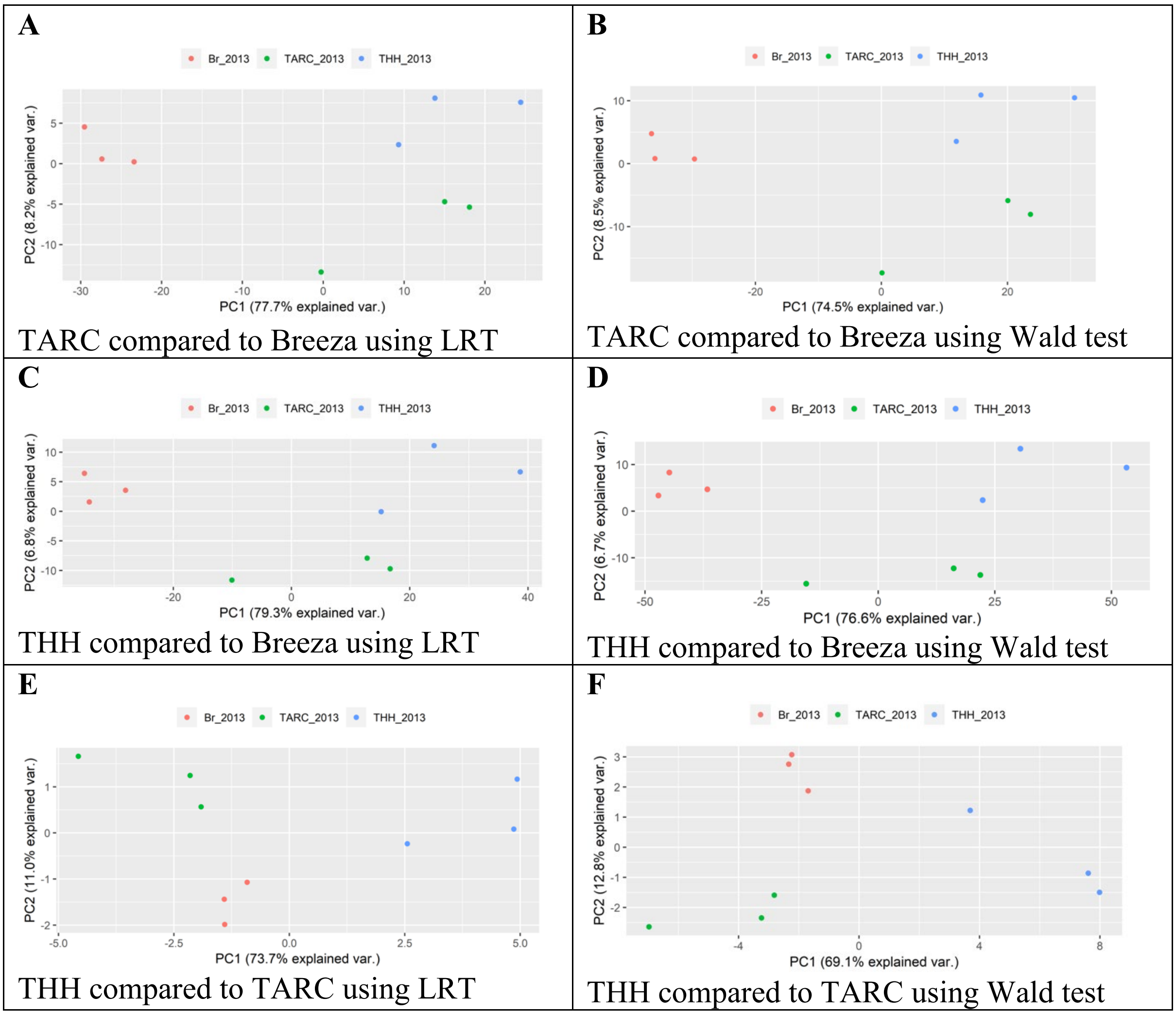


Figure D.10. Volcano plots displaying significant and non-significant RNA transcript data around an axis of negative and positive (log2) fold change. Comparison of farms sites used is at the top of each graph (A to F).

D.7 PCAs of sample-transcript expression – opposite comparison order



TARC compared to Breeza using LRT

TARC compared to Breeza using Wald test

THH compared to Breeza using LRT

THH compared to Breeza using Wald test

THH compared to TARC using LRT

THH compared to TARC using Wald test

Figure D.11. PCA plots of sample data. Data from each sample represented differentially expressed transcripts detected after comparisons described above (A to F), followed by tests for significance using either LRT or Wald test (listed in A to F).

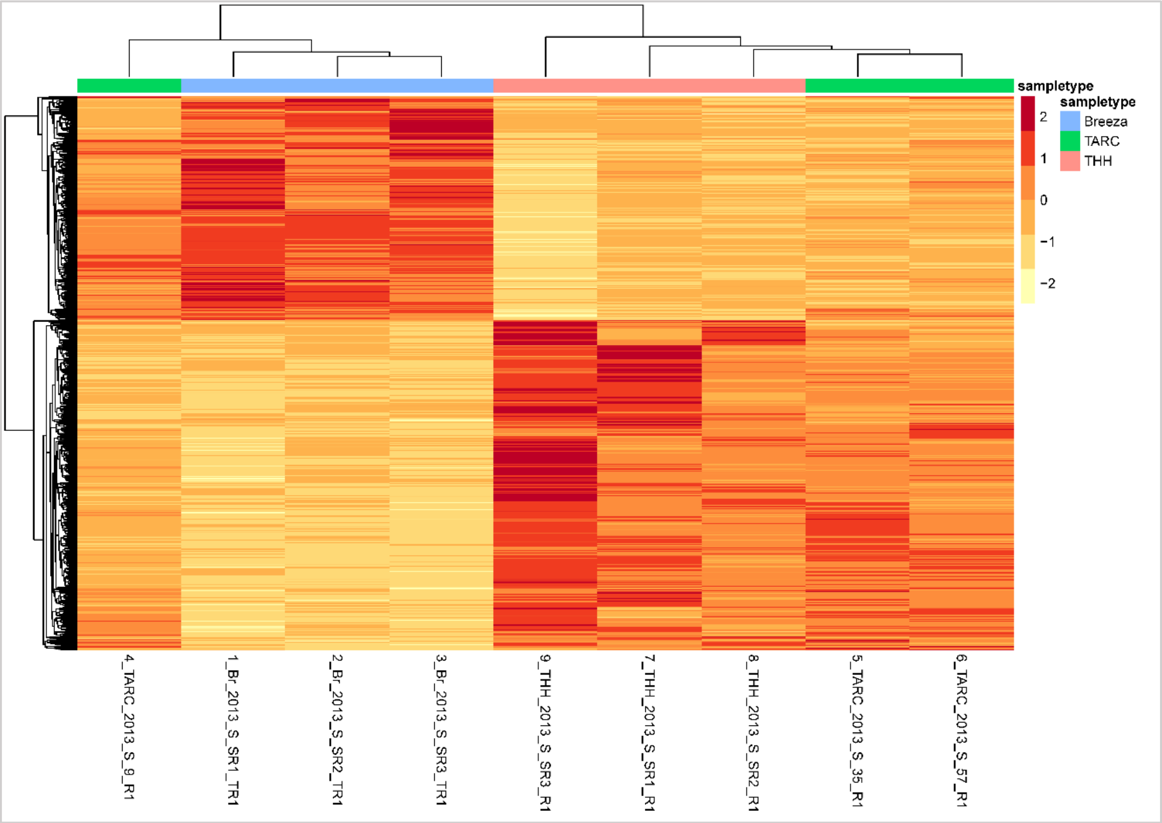


Figure D.12. Heatmap of expression levels of transcripts for determined after an ANOVA-like comparison across all three sample groups and significance calculation via LRT. The heatmap contains 1880 rows of transcript data.

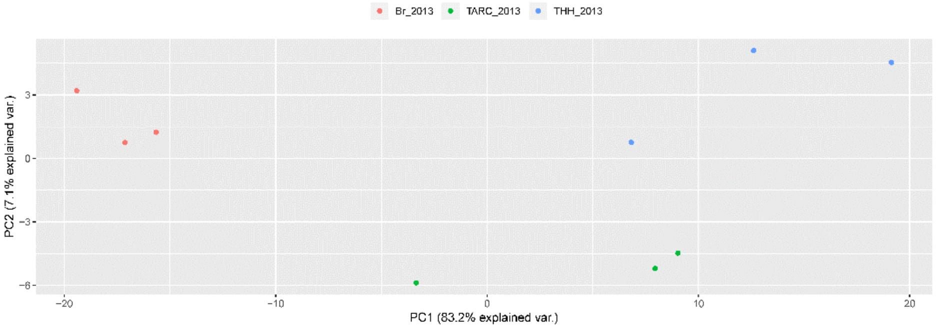


Figure D.13. PCA plots calculated from transcript data of wheat grain samples after an ANOVA-like comparison and significance calculation via LRT (p-adjusted value of 0.0001). The sample groups are spitfire wheat grain harvested from three different farms: Breeza (Br), TARC, and Terry Hie Hie (THH).

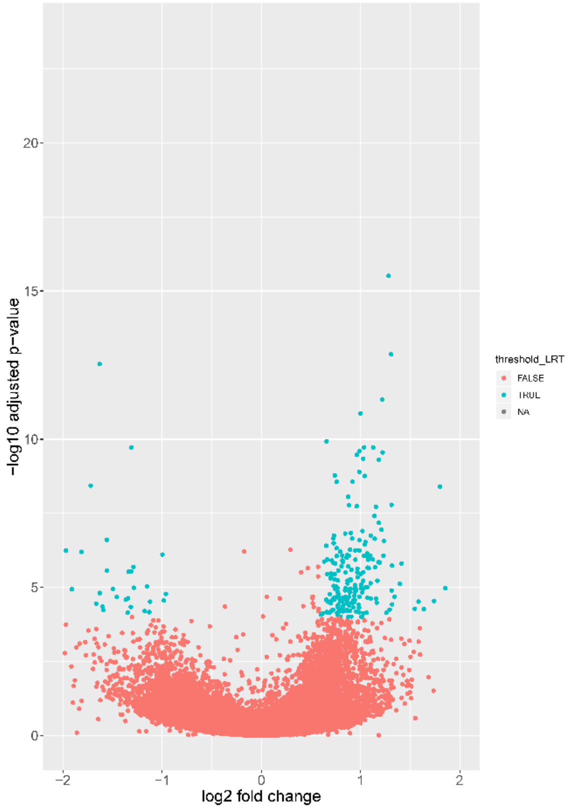


Figure D.14. Volcano plots of significant and non-significant transcript data from wheat grain samples after an ANOVA-like comparison and significance calculation via LRT (p-adjusted value of 0.0001).

D.8 R-scripts mentioned in Chapter 5

D.8.1 R-script: 'tximportAfterKallisto'

```
# if (!requireNamespace("BiocManager", quietly = TRUE))
#   install.packages("BiocManager")
# BiocManager::install("rtracklayer", version = "3.8")
# BiocManager::install("tximport", version = "3.8")
# BiocManager::install("DESeq2", version = "3.8")
# BiocManager::install("AnnotationDbi", version = "3.8")
# BiocManager::install("rhdf5", version = "3.8")


# # Windows Computer
# if(.Platform$OS.type == "windows"){
#   Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jdk-11.0.1')
# }
if(.Platform$OS.type == "windows"){
  Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jdk1.8.0_191')
}
library("tcltk")
library("rJava")
```

```
detach("package:rJava", unload=TRUE)
library("rChoiceDialogs")
library("tximport")
library("AnnotationDbi")
library("readr")
library("rtracklayer")
library("dplyr")
library("DESeq2")
library("rhdf5")

#Name the base directory and its path
##### Input paths for conditional statement below #####
macPathExHD = "/Volumes/Seagate_Backup_Plus_Drive/"
macPath = "/Users/Paul/"
desktopPath = "/Users/43533698/"



# Point to the protein or RNA expression file
if(.Platform$OS.type == "windows"){
  base_dir              =              tk_choose.dir(default              =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/RNASeq_Data/RNASeq_Wheat_Folder
_5-12-16",
                                   caption = "Select the folder with Kalisto
data ('D_Kallisto/Kallisto_Output')")
} else {
  base_dir              =              tk_choose.dir(default              =
"/Users/Paul/Google_Drive/PhD/DEanalysis/RNASeq_Data/RNASeq_Wheat_Folder_5-
12-16",
                             caption = "Select the folder with Kalisto data
('D_Kallisto/Kallisto_Output')")
}


# Check what is in the directory
list.files(base_dir)
#Get the sample identifiers (In this case directory names)
sample_id = list.dirs(base_dir, full.names = FALSE, recursive = FALSE)
#Construct full paths to each sample directory
kal_dirs <- sapply(sample_id, function(x) file.path(base_dir, x))

#Load in table of experimental information
raw_text = "
sample Cultivar Harvest_Year Biological_Replicate condition
1_Br_2013_S_SR1_TR1 Spitfire 2013 1 Breeza
2_Br_2013_S_SR2_TR1 Spitfire 2013 2 Breeza
3_Br_2013_S_SR3_TR1 Spitfire 2013 3 Breeza
4_TARC_2013_S_9_R1 Spitfire 2013 1 TARC
5_TARC_2013_S_35_R1 Spitfire 2013 2 TARC
6_TARC_2013_S_57_R1 Spitfire 2013 3 TARC
7_THH_2013_S_SR1_R1 Spitfire 2013 1 THH
8_THH_2013_S_SR2_R1 Spitfire 2013 2 THH
9_THH_2013_S_SR3_R1 Spitfire 2013 3 THH
"
summaryTable  <-  read.table(header=TRUE,  text=raw_text,  stringsAsFactors  =
FALSE)

##################### ".tsv files from Kallisto #####################
##### Import ".tsv" files from Kallisto
# Vector of file paths to the "abundance.tsv" file
files <- file.path(kal_dirs, "abundance.tsv")
```

```
names(files) <- summaryTable$sample
all(file.exists(files))

    # ##################### OR use ".h5" files from Kallisto
#####################
    # ######### Import ".h5" files from kallisto output ##########
    # #Read in a vector of ".h5" files with sample data
    # files <- file.path(kal_dirs, "abundance.h5")
    # #Add names to the "files" vector (Make it a named vector)
    # names(files) <- summaryTable$sample
    # #Check that the paths and files are correct
    # all(file.exists(files))
    # ######### End Import ".h5" files ##########

##### Get the transcript and gene identifiers from the ".gtf" file for wheat
  #using "import" from the "rtracklayer" package

if(.Platform$OS.type == "windows"){
  gtfData =
import("C:/Users/paul_/Google_Drive/PhD/DataBases/RNASeq/Triticum_aestivum.
TGACv1.35.gtf")
} else {
  gtfData =
import("~/Google_Drive/PhD/DataBases/RNASeq/Triticum_aestivum.TGACv1.35.gtf
")
}


geneData = gtfData@elementMetadata
# Select only rows with "transcript" in the "type" column of the "geneData"
dataframe
geneData = geneData[grep("transcript", geneData$type),]
# Select only the "transcript_id" and "gene_id" columns
geneData = as.data.frame(geneData[,c("transcript_id", "gene_id")])
# Save memory and remove the "gtfData" object
# rm(gtfData)

# Copy and rename the dataframe and rename one of the columns
tx2gene = geneData
# Again, save memory and remove "geneData" data frame
# rm(geneData)

############# Gene estimates ############
# Import Gene and transcript level estimates
  # uses "readr" to load data faster (reader = read_tsv)
txi <- tximport(files = files, type = "kallisto", tx2gene = tx2gene)

# Check that the import worked
check = as.data.frame(txi[["counts"]])
check = cbind(row.names(check), check)
rownames(check) = NULL
names(check)[1] = "id"
############# End Gene estimates ############


################### Alternative where gene-level summation is avoided
###################
      # # Avoid gene-level summarization
      # # Use the original transcript level estimates as a list of matrices.
      # txi.tx <- tximport(files, type = "kallisto", txOut = TRUE, tx2gene
= tx2gene,
      #                    reader = read_tsv)
      # # These matrices can then be summarized afterwards
      #   # using the function "summarizeToGene".
      # txi.sum <- summarizeToGene(txi.tx, tx2gene)
      # all.equal(txi$counts, txi.sum$counts)



##################################################
    # # Get sample names from the "txi" object (from tximport)
    #   # Turn the column names into row names to
    # rownames(summaryTable) <- colnames(txi$counts)

# Creating a DESeqDataSet for use with DESeq2:
# Turn the condition column into factors
summaryTable$condition = as.factor(summaryTable$condition)
# Keep "sample" column as text
summaryTable$sample = as.character(summaryTable$sample)
# Get sample names from the "txi" object (from tximport)
# Turn the column names into row names to
rownames(summaryTable) = summaryTable$sample
# Delete the "sample" column as this is now "row.names"
summaryTable$sample = NULL
ddso <- DESeqDataSetFromTximport(txi, summaryTable, ~condition)

#Raw Expression Data
head(counts(ddso))
#Experimental Data
ddso@colData

# Prepare raw counts for saving to ."csv" file
rawRNACountsViaTximport = data.frame(counts(ddso))
rawRNACountsViaTximport$ids = row.names(rawRNACountsViaTximport)
rawRNACountsViaTximport                                                    =
rawRNACountsViaTximport[,c(length(rawRNACountsViaTximport),
1:(length(rawRNACountsViaTximport)-1))]
row.names(rawRNACountsViaTximport) = NULL

# Check directory exists and write it if necessary
if (dir.exists("~/Google_Drive/PhD/DEanalysis/preDESeq2Data/")) {
  print("Directory exists!")
} else  {dir.create("~/Google_Drive/PhD/DEanalysis/preDESeq2Data/")}

# Save raw counts to ".csv"
write.csv(rawRNACountsViaTximport,                  file                   =
"~/Google_Drive/PhD/DEanalysis/preDESeq2Data/rawRNACountsViaTximport.csv",
row.names = FALSE)
write.csv(summaryTable,                        file                        =
"~/Google_Drive/PhD/DEanalysis/preDESeq2Data/summaryTable.csv", row.names =
FALSE)
```

D.8.2 R-script: ‘DESeq2_Functions.R’

```
# if (!requireNamespace("BiocManager", quietly = TRUE))
#   install.packages("BiocManager")
# BiocManager::install("rtracklayer", version = "3.8")
# BiocManager::install("tximport", version = "3.8")
# BiocManager::install("DESeq2", version = "3.8")
# BiocManager::install("AnnotationDbi", version = "3.8")
# BiocManager::install("ReportingTools", version = "3.8")

## Gene-level differential expression analysis using DESeq2

# Load in the necessary R packages
# if (!requireNamespace("BiocManager", quietly = TRUE))
#   install.packages("BiocManager")
# BiocManager::install("DESeq2", version = "3.8")
library("ggplot2")
library("RColorBrewer")
library("DESeq2")
library("pheatmap")
library("ReportingTools")

    #
    #        rawRNACountsViaTximport        =        read.table(file        =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/preDESeq2Data/rawRNACountsViaTx
import.csv", row.names = 1, stringsAsFactors = FALSE, sep = ",", header =
TRUE)
    # rawRNACountsViaTximport = as.matrix(rawRNACountsViaTximport)
    #
    #           summaryTable           =           read.table(file           =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/preDESeq2Data/summaryTable.csv"
, stringsAsFactors = FALSE, sep = ",", header = TRUE)
    # summaryTable = as.matrix(summaryTable)


####################  SAVE HEATMAP Function ####################
#Save pheatmap function
save_pheatmap <- function(x, filename, width=1500, height=800) {
  stopifnot(!missing(x))
  stopifnot(!missing(filename))
  png(filename = filename, width = width, height=height)
  grid::grid.newpage()
  grid::grid.draw(x$gtable)
  dev.off()
}
#################### END FUNCTION ####################

# Clear all plots from memory
graphics.off()
# Can also use the following but you will get an error if nothing is open:
# dev.off(dev.list()["RStudioGD"])


##### tximport via "tximportAfterKallisto" script has created
# DESeq2 object

#Experimental Data from tximport object
summaryTable = data.frame(ddso@colData)
# Turn the condition column into factors
# if they are not already
summaryTable$condition = as.factor(summaryTable$condition)
```

```
# # Complex designs - last factor should be the condition of interest
#   # Example:
# design <- ~ sex + age + treatment
#
# # Interactions can also be added like so:
#   # Example:
# design <- ~ sex + age + treatment + sex:treatment

##### Check Raw counts and experimental data #####
#Raw Expression Data
tximportDESeq2RawCounts = as.data.frame(counts(ddso))
head(tximportDESeq2RawCounts)
#Experimental design information
  #Needed to construct DESeq2 object using "DESeqDataSetFromMatrix"
tximportDESeq2Experiment = as.data.frame(ddso@colData)

### Check that sample names match in both files
all(colnames(tximportDESeq2RawCounts) %in% rownames(summaryTable))
all(colnames(tximportDESeq2RawCounts) == rownames(summaryTable))

  # Check raw counts
  # tximportDESeq2RawCounts



        # barplot(colSums(tximportDESeq2RawCounts)/1000000,
        #         main="Total number of reads per sample (million)",
        #         #col=c("red","green", "blue"),
        #         col=summaryTable$condition,
        #         #        names.arg = "",
        #         las=1,  horiz=TRUE,
        #         ylab="Samples", cex.names=0.5,
        #         xlab="Million counts")
        #
        #
        # epsilon = 1
        # ## Boxplots
        #       boxplot(log2(tximportDESeq2RawCounts       +       epsilon),
col=summaryTable$condition, names.arg = summaryTable$condition, pch=".",
        #         horizontal=TRUE, cex.axis=0.5,
        #         las=1, ylab="Samples", xlab="log2(Counts +1)")


## Density
## We will require one function from the affy package
if(!require("affy")){
  source("http://bioconductor.org/biocLite.R")
  biocLite("affy")
}
library(affy)
plotDensity(log2(tximportDESeq2RawCounts + 1), lty=1, xlab="log2(RNA Seq Raw
Counts)", col=summaryTable$condition, lwd=2)
grid()
col.strain <- c("Breeza"="green","TARC"="orange", "THH"="red") # Choose one
color per strain
legend("topright", legend=names(col.strain), col=col.strain, lwd=2)
```

```
dds <- DESeqDataSetFromMatrix(countData = tximportDESeq2RawCounts, colData
= tximportDESeq2Experiment, design = ~ condition)

#Remove any rows with no counts
dds <- dds[ rowSums(counts(dds)) > 10, ]

# Check dimensions
dim(dds)


#################### Normalisation of counts ####################
##Run analysis
dds <- DESeq(dds)
    # dds <- DESeq(dds, test="LRT", reduced=~1)

res <- results(dds)
############################################################


# Plot the distribution of RNA-Seq counts (Histogram)
RNASeqDist = ggplot(tximportDESeq2RawCounts) +
  geom_histogram(aes(x = tximportDESeq2RawCounts[,1]), stat = "bin", bins =
200) +
  xlab("Raw expression counts") +
  ylab("Number of genes")
print(RNASeqDist)

if(.Platform$OS.type == "windows"){
  ggsave("RNASeqDist.png", plot = RNASeqDist, path =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
} else {
  ggsave("RNASeqDist.png", plot = RNASeqDist, path =
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
}

# Same as above but zooming in on the region -5 to 500 counts
RNASeqDistzoom = ggplot(tximportDESeq2RawCounts) +
  geom_histogram(aes(x = tximportDESeq2RawCounts[,1]), stat = "bin", bins =
200) +
  xlim(-5, 500) + ylim(0, 500) +
  xlab("Raw expression counts (if y > 500, data removed") +
  ylab("Number of genes")
print(RNASeqDistzoom)

if(.Platform$OS.type == "windows"){
  ggsave("RNASeqDistzoom.png", plot = RNASeqDistzoom, path =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
} else {
  ggsave("RNASeqDistzoom.png", plot = RNASeqDistzoom, path =
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
}
```

```
#####  The  RNASeq  data  should  be  modeled  around  the  negative  bionomial
distribution
#Check with a scatter plot of mean verses variance of data
# Preparation code to plot mean versus variance

# ----- BREEZA -----
mean_counts <- apply(tximportDESeq2RawCounts[, 1:3], 1, mean)
variance_counts <- apply(tximportDESeq2RawCounts[, 1:3], 1, var)
df <- data.frame(mean_counts, variance_counts)

#Check if data fits negative binomial distribution
# Plot variance against the mean for -- Breeza --
checkBreezaData = ggplot(df) +
  geom_point(aes(x=mean_counts, y=variance_counts)) +
  geom_line(aes(x=mean_counts, y=mean_counts, color="red")) +
  scale_y_log10() +
  scale_x_log10() +
  ggtitle("Breeza") +
  theme(plot.title = element_text(face="bold", hjust = 0.5))
print(checkBreezaData)
# Check directory exists and write it if necessary
if(.Platform$OS.type == "windows"){
  if
(dir.exists("C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_A
LL")) {
    print("Directory exists!")
  }                                                                        else
{dir.create("C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_A
LL")}
} else {
  if (dir.exists("~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL")) {
    print("Directory exists!")
  }                                                                        else
{dir.create("~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL")}
}


if(.Platform$OS.type == "windows"){
  ggsave("checkBreezaData.png",   plot   =   checkBreezaData,   path   =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
} else {
  ggsave("checkBreezaData.png",   plot   =   checkBreezaData,   path   =
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
}



# ----- TARC ----
mean_counts2 <- apply(tximportDESeq2RawCounts[, 4:6], 1, mean)
variance_counts2 <- apply(tximportDESeq2RawCounts[, 4:6], 1, var)
df2 <- data.frame(mean_counts2, variance_counts2)

# Plot variance against the mean for -- TARC --
checkTARCData = ggplot(df2) +
  geom_point(aes(x=mean_counts, y=variance_counts)) +
  geom_line(aes(x=mean_counts, y=mean_counts, color="red")) +
  scale_y_log10() +
  scale_x_log10() +
  ggtitle("TARC") +
  theme(plot.title = element_text(face="bold", hjust = 0.5))
print(checkTARCData)

if(.Platform$OS.type == "windows"){
  ggsave("checkTARCData.png",    plot    =    checkTARCData,    path    =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
} else {
  ggsave("checkTARCData.png",    plot    =    checkTARCData,    path    =
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
}


# ----- THH ----
mean_counts3 <- apply(tximportDESeq2RawCounts[, 7:9], 1, mean)
variance_counts3 <- apply(tximportDESeq2RawCounts[, 7:9], 1, var)
df3 <- data.frame(mean_counts3, variance_counts3)

# Plot variance against the mean for -- THH --
checkTHHData = ggplot(df3) +
  geom_point(aes(x=mean_counts, y=variance_counts)) +
  geom_line(aes(x=mean_counts, y=mean_counts, color="red")) +
  scale_y_log10() +
  scale_x_log10() +
  ggtitle("THH") +
  theme(plot.title = element_text(face="bold", hjust = 0.5))
print(checkTHHData)


if(.Platform$OS.type == "windows"){
  ggsave("checkTHHData.png",    plot    =    checkTHHData,    path    =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
} else {
  ggsave("checkTHHData.png",    plot    =    checkTHHData,    path    =
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
}


##### Mapping and quantitation of transcripts
# 1. Splice aware mapping to genome (STAR, etc.)
# OR
# 2. Pseudoalignment (Sailfish, Salmon, Kallisto)
# Then:
# Quantitation:
# 1. Count unique reads mapped to genes
# 2. Normalisation of counts
# a. Account for Gene level and sample level ?????
#3. Differential exression (DE) analysis


#################### Normalisation of counts ####################
```

```
# #*** Generate size factors, separately to "DESeq" function
# dds <- estimateSizeFactors(dds)

#Check the normalisation for size --- SIZE
sizefactorNormalised = sizeFactors(dds)

#*** Retrieve normalized counts from matrix --- NORMALISED COUNTS
normalized_counts = counts(dds, normalized=TRUE)
normalized_counts_df = data.frame(normalized_counts)
normalized_counts_df$ids = row.names(normalized_counts_df)
normalized_counts_df =
normalized_counts_df[,c(length(normalized_counts_df),
1:(length(normalized_counts_df)-1))]
row.names(normalized_counts_df) = NULL



if(.Platform$OS.type == "windows"){
  #Save normalised counts in a tab-delimited table --- SAVE NORMALISED COUNTS
  write.csv(normalized_counts_df,
file="C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL/DESe
q2Normalized_counts_df.csv", row.names = FALSE)
} else {
  #Save normalised counts in a tab-delimited table --- SAVE NORMALISED COUNTS
  write.csv(normalized_counts_df,
file="~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL/DESeq2Normalized_
counts_df.csv", row.names = FALSE)
}

##### Sample level QC
# Log2 transformed normalised counts are used to check similarity between
samples
# Including PCA and hierachical clustering

###***t Transform counts for data visualization
rld <- rlog(dds, blind=TRUE)
# To check the transformed data use "assay"
head(assay(rld))

### Plot PCA
# Need an rlog object ("rld" as above)
# Need the intgroup (column of metadata of interest)
# Use DESeq2 PCA
plotPCA(rld, intgroup="condition", ntop = 50000)

# Plot PCA via ggplot2
ggplotPCA = plotPCA(rld, intgroup="condition", ntop = 50000,
returnData=TRUE)
percentVar <- round(100 * attr(ggplotPCA, "percentVar"))
PCAplot = ggplot(ggplotPCA, aes(PC1, PC2, color=condition)) +
geom_point(size=3) +
  xlab(paste0("PC1: ",percentVar[1],"% variance")) +
  ylab(paste0("PC2: ",percentVar[2],"% variance")) +
  theme(legend.title=element_blank())
print(PCAplot)

if(.Platform$OS.type == "windows"){
  ggsave("PCAplot.png", plot = PCAplot, path =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
} else {
  ggsave("PCAplot.png", plot = PCAplot, path =
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
}

### Extract the rlog matrix from the object
rld_mat <- assay(rld) ## assay() is function from the
"SummarizedExperiment" package that was loaded when you loaded DESeq2

### Compute pairwise corrrelation values
rld_cor <- cor(rld_mat)    ## cor() is a base R function

## check the output of cor(), make note of the rownames and colnames
head(rld_cor)

### Plot heatmap
pairCorrHeatMap = pheatmap(rld_cor)

if(.Platform$OS.type == "windows"){
  save_pheatmap(pairCorrHeatMap,
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL/pairCorrH
eatMap.png")
} else {
  save_pheatmap(pairCorrHeatMap,
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL/pairCorrHeatMap.png")
}


# Heatmap with different colours
heat.colors <- brewer.pal(6, "Blues")
pairCorrHeatMap2 = pheatmap(rld_cor, color = heat.colors, border_color=NA,
fontsize = 10,
                           fontsize_row = 10, height=20)
if(.Platform$OS.type == "windows"){
  save_pheatmap(pairCorrHeatMap2,
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL/pairCorrH
eatMap2.png")
} else {
  save_pheatmap(pairCorrHeatMap2,
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL/pairCorrHeatMap2.png")
}


#################### General summary data ####################
# Get size factors
sizeFactors(dds)
# Total number of raw reads for each sample
countsPerSample = colSums(counts(dds))
# Total number of reads per sample after normalisation
normalisationPerSample = colSums(counts(dds, normalized=TRUE))
#NOTE: Other gene specific normalization factors can be applied to data such
as GC content

#################### Data Frame of raw and normalised counts
####################
dfRawNormal = data.frame(countsPerSample, normalisationPerSample)
#New column from row names
dfRawNormal$sample = row.names(dfRawNormal)
```

```
#Delete row names
row.names(dfRawNormal) = NULL
#Rearrange columns
dfRawNormal = dfRawNormal[,c(3,1,2)]
# Write the raw mean of normalised counts per sample
if(.Platform$OS.type == "windows"){
  write.csv(dfRawNormal,                         file                    =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL/rawNormal
isedValuesSum.csv", row.names = FALSE)
} else {
  write.csv(dfRawNormal,                         file                    =
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL/rawNormalisedValuesSum
.csv", row.names = FALSE)
}

#Reshape the columns for graphing with ggplot2
library(reshape2)
dfRawNormalMelt = melt(dfRawNormal, variable.name = "name",
                       value.names = "value", id.vars = c("sample"))

# Plot to compare raw counts verses normalised counts
rawVnormalized   =   ggplot(dfRawNormalMelt,   aes(x   =   sample,   y=value,
fill=name)) + geom_bar(stat = "identity", position = position_dodge()) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
print(rawVnormalized)

if(.Platform$OS.type == "windows"){
  ggsave("rawVnormalized.png",    plot    =    rawVnormalized,    path    =
"C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
} else {
  ggsave("rawVnormalized.png",    plot    =    rawVnormalized,    path    =
"~/Google_Drive/PhD/DEanalysis/results/DESeq2_QC_ALL",
         scale = 1, width = 20, height = 20, units = "cm",
         dpi = 600)
}



############################################################
#################### GENE WIDE dispersions ####################
#2 Estimating gene-wise dispersions
#  Dispersion  is  a  measure  of  spread  or  variability  in  the  data,  such  as
Variance, standard deviation, IQR, etc.
# dispersion (α), Var = μ + α*μ^2

##### Plot dispersion estimates #####
# Genes with surrounding blue dots around black are not shrunken
# as they probably do not follow modelling assumptions
# Shrunken log2 foldchanges (LFC)
# DESeq2 shrinks the LFC estimates toward zero when the
# information for a gene is low (Low counts, High dispersion values)
plotDispEsts(dds)


############################################################
#################### WALD Test Function ####################
waldTestFunction = function (dds, contrastColumn, numerator, denominator) {
    saveString = paste0(numerator, "_V_", denominator)
    if(.Platform$OS.type == "windows"){
      specificResultsDir =
paste0("C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/", saveString,
"_Wald/")
    } else {
      specificResultsDir = paste0("~/Google_Drive/PhD/DEanalysis/results/",
saveString, "_Wald/")
    }

    if (dir.exists(specificResultsDir)) {
      print("Directory exists!")
    } else  {dir.create(specificResultsDir)}

    # Set contrasts and extract results table
    res_Wald = results(dds, contrast=c(contrastColumn, numerator,
denominator))
    #Examine Data
    head(res_Wald)
    #Examine the object
    class(res_Wald)
    # Extract information about each column (of "res_Wald")
    mcols(res_Wald, use.names=TRUE)
    #####
    ## **** Summarize results **** ##
    summary(res_Wald, alpha = 0.05)
    #####
    ## ****Add a fold change threshold **** ##
    # This is needed because stringency needs to be increased which is
    # done by filtering for values above a certain fold change threshold
    ### Set threshold variables
    padj.cutoff <- 0.05
    lfc.cutoff <- 0.58 #Equivalent to 1.5 fold change
    #####
    #Create a logical vector whose length is equal to the total number of
genes in the dataset
    threshold <- res_Wald$padj < padj.cutoff & abs(res_Wald$log2FoldChange)
> lfc.cutoff
    #How many of the above vector are true
    #How many genes are differentially expressed according to our criteria
    length(which(threshold == TRUE))
    #Add "threshold" values to the "res_Wald" object
    res_Wald$threshold <- threshold
    #Subset the values (rows) that pass the threshold
    subset(res_Wald, threshold == TRUE)
    #Check data
    head(res_Wald, n = 10)

    # Plot expression for single gene
    #plotCounts(dds, gene="TRIAE_CS42_1AL_TGACv1_000002_AA0000030",
intgroup="condition")

    # Create dataframe for plotting
    df_res_Wald = data.frame(res_Wald)

    # Volcano plot
    volcano = ggplot(df_res_Wald) +
      geom_point(aes(x=log2FoldChange, y=-log10(padj), colour=threshold)) +
      xlim(c(-2,2)) +
      ggtitle(paste0(numerator, "_V_", denominator)) +
   xlab("log2 fold change") +
   ylab("-log10 adjusted p-value") +
   theme(#legend.position = "none",
     plot.title = element_text(size = rel(1.5), hjust = 0.5),
     axis.title = element_text(size = rel(1.5)),
     axis.text = element_text(size = rel(1.25)))
  # Draw plot
  volcano
  # Save plot

  ggsave(paste0(saveString, "_volcano.png"), plot = volcano, path =
specificResultsDir,
             scale = 1, width = 20, height = 20, units = "cm",
             dpi = 600)

  # Sort the results tables
  res_Wald_ordered <- res_Wald[order(res_Wald$padj), ]
  # Get significant genes from object
  sig_Wald_genes                                                   <-
row.names(res_Wald_ordered)[which(res_Wald_ordered$threshold)]
  # Normalised counts into a new object
  normWald_sig = normalized_counts[sig_Wald_genes,]

  # Make a data frame of values
  sig_Wald_genes_df = data.frame(res_Wald_ordered[sig_Wald_genes, ])
  sig_Wald_genes_df$Ids = row.names(sig_Wald_genes_df)
  sig_Wald_genes_df = sig_Wald_genes_df[c(length(sig_Wald_genes_df),
1:length(sig_Wald_genes_df)-1)]
  row.names(sig_Wald_genes_df) = NULL
  # Make a dataframe of ids and their values then save it
  write.csv(sig_Wald_genes_df, file = paste0(specificResultsDir,
saveString, "_Wald_values.csv"), row.names = FALSE)
  # Save the text file of identifiers
  write.table(sig_Wald_genes_df$Ids,
               file = paste0(specificResultsDir, saveString,
"_Wald_Sig_Ids.txt"),
               row.names = FALSE, sep = "\t", col.names = FALSE)
  # Also save a copy in the
"~/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/" folder for the blastx
workflow


  if(.Platform$OS.type == "windows"){
   write.table(sig_Wald_genes_df$Ids,
                file                                                   =
paste0("C:/Users/paul_/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/",
saveString, "_Wald_Sig_Ids.txt"),
                row.names = FALSE, sep = "\t", col.names = FALSE)
  } else {
   write.table(sig_Wald_genes_df$Ids,
                file                                                   =
paste0("~/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/", saveString,
"_Wald_Sig_Ids.txt"),
                row.names = FALSE, sep = "\t", col.names = FALSE)
  }


  ### Annotate our heatmap (optional)
  annotation <- data.frame(sampletype=summaryTable[,'condition'],
                             row.names=row.names(summaryTable))
```

```
    ### Set a color palette
    heat.colors <- brewer.pal(6, "YlOrRd")

    ### 1. Run pheatmap "norm_sig"
    wald_heatmap = pheatmap(normWald_sig, color = heat.colors, cluster_rows
= T, clustering_distance_cols = "manhattan",
                         show_rownames=F,    annotation=    annotation,
border_color=NA, fontsize = 10, scale="row",
                         fontsize_row   =   10,   height=20,   main   =
paste(numerator, "V", denominator, "Fold Comparison"))
    # Save pheatmap:
    save_pheatmap(wald_heatmap,   paste0(specificResultsDir,   saveString,
"_Wald_heatmap.png"))

    #norm_LRTsig = res_tableLRT_sorted[sigLRT,]
    normWald_sig_DataFrame = data.frame(normWald_sig)
    normWald_sig_DataFrame$ids = row.names(normWald_sig_DataFrame)
    normWald_sig_DataFrame                                               =
normWald_sig_DataFrame[,c(length(normWald_sig_DataFrame),
1:(length(normWald_sig_DataFrame)-1))]
    row.names(normWald_sig_DataFrame) = NULL
    write.csv(normWald_sig_DataFrame,  file  =  paste0(specificResultsDir,
saveString, "_Wald_Exp.csv"), row.names = FALSE)

    #################### HTML Report ####################
    ##### Write a HTML report
    # First re-set working directory for R, then continue
    setwd(specificResultsDir)
    htmlRep <- HTMLReport(shortName = "Report",
                        title = "Differential expression analysis",
                        reportDirectory      =      paste0(saveString,
"_HTMLreport"))
    publish(normWald_sig_DataFrame, htmlRep)
    url <- finish(htmlRep)
    #browseURL(url)
    # Return R working directory to normal
    setwd(specificResultsDir)
    #################### END HTML Report ####################
}
#################### END Function ####################


######################################################################
#################### Run Wald Test Function ####################
# Samples (locations):
      # Breeza
      # TARC
      # THH

waldTestFunction(dds, "Breeza", "THH", "TARC")
waldTestFunction(dds, "condition", "Breeza", "THH")
waldTestFunction(dds, "condition", "TARC", "THH")
waldTestFunction(dds, "condition", "Breeza", "TARC")

waldTestFunction(dds, "condition", "TARC", "Breeza")
waldTestFunction(dds, "condition", "THH", "Breeza")
waldTestFunction(dds, "condition", "THH", "TARC")


#################### END Wald Test ####################
############################################################
```

```
############################################################
#################### LRT: LIKELIHOOD RATIO TEST ####################

##################################################
##### First step in the likelihood ratio test #####
dds1 <- DESeq(dds, test="LRT", full = ~ condition, reduced = ~ 1)


#################### FUNCTION: Likelihood ratio test and Fold Change
####################
lrtFunction = function (dds_lrt, contrastColumn, numerator, denominator) {
    saveString = paste0(numerator, "_V_", denominator)
          #saveString = "_AnovaVolco_"
    if(.Platform$OS.type == "windows"){
      specificResultsDir =
paste0("C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/", saveString,
"_LRT/")
    } else {
      specificResultsDir = paste0("~/Google_Drive/PhD/DEanalysis/results/",
saveString, "_LRT/")
    }

    if (dir.exists(specificResultsDir)) {
      print("Directory exists!")
    } else  {dir.create(specificResultsDir)}

    # Set contrasts and extract results table
    dds_lrt = results(dds, contrast=c(contrastColumn, numerator,
denominator))
          #dds_lrt = results(dds)

    #Set cutoff
    LRT.padj.cutoff = 0.0001
    #Define a threshold (FDR)
    LRT.lfc.cutoff <- 0.58 #(1.5 fold)

    threshold_LRT = dds_lrt$padj < LRT.padj.cutoff &
abs(dds_lrt$log2FoldChange) > LRT.lfc.cutoff
    #threshold_LRT = dds_lrt$padj < padjLRT.cutoff
    # Add a column of significant genes
    dds_lrt$threshold_LRT <- threshold_LRT
    #Count the values (rows) that pass the threshold
    numberThatPassed = subset(dds_lrt, threshold_LRT == TRUE)
    length(numberThatPassed$threshold_LRT)

    # Sort the results tables
    dds_lrt_sorted <- dds_lrt[order(dds_lrt$padj), ]
    # Get significant genes from object
    sigLRT <-
row.names(dds_lrt_sorted)[which(dds_lrt_sorted$threshold_LRT)]

    # Plot expression for single gene
    #plotCounts(dds, gene="TRIAE_CS42_1AL_TGACv1_000002_AA0000030",
intgroup="condition")

    # Create dataframe for plotting
    df_res_LRT = data.frame(dds_lrt)

    # Volcano plot
    volcano = ggplot(df_res_LRT) +
      geom_point(aes(x=log2FoldChange,y=-log10(padj),
colour=threshold_LRT)) +
      xlim(c(-2,2)) +
      #ggtitle(paste0(numerator, "_V_", denominator)) +
      xlab("log2 fold change") +
      ylab("-log10 adjusted p-value") +
      theme(#legend.position = "none",
        plot.title = element_text(size = rel(1.5), hjust = 0.5),
        axis.title = element_text(size = rel(1.5)),
        axis.text = element_text(size = rel(1.25)))
    # Draw plot
    volcano
    # Save plot
    ggsave(paste0(saveString, "_volcano.png"), plot = volcano, path =
specificResultsDir,
           scale = 1, width = 20, height = 20, units = "cm",
           dpi = 600)

    # Normalised counts into a new object
    norm_LRTsig <- normalized_counts[sigLRT,]
    # Write identifiers of interest
    norm_LRTsig_genes = data.frame(row.names(norm_LRTsig))
    names(norm_LRTsig_genes) = "Ids"



    if(.Platform$OS.type == "windows"){
      if
(dir.exists("C:/Users/paul_/Google_Drive/PhD/DataBases/RNASeq/sigIdentifier
s/")) {
        print("Directory exists!")
      } else
{dir.create("C:/Users/paul_/Google_Drive/PhD/DataBases/RNASeq/sigIdentifier
s/")}
    } else {
      if
(dir.exists("~/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/")) {
        print("Directory exists!")
      } else
{dir.create("~/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/")}
    }


    # Save the text file of identifiers
    write.table(norm_LRTsig_genes$Ids,
                file = paste0(specificResultsDir, saveString,
"_LRT_Sig_Ids.txt"),
                row.names = FALSE, sep = "\t", col.names = FALSE)
    # Also save a copy in the
"~/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/" folder for the blastx
workflow

    if(.Platform$OS.type == "windows"){
        write.table(norm_LRTsig_genes$Ids,
                  file =
paste0("C:/Users/paul_/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/",
saveString, "_LRT_Sig_Ids.txt"),
                  row.names = FALSE, sep = "\t", col.names = FALSE)
    } else {
        write.table(norm_LRTsig_genes$Ids,
                  file =
paste0("~/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/", saveString,
"_LRT_Sig_Ids.txt"),
                  row.names = FALSE, sep = "\t", col.names = FALSE)
    }

    # Make a dataframe of significant ids and their values then save it
    df_lrt_values = data.frame(dds_lrt_sorted)
    sig_lrt_values = df_lrt_values[sigLRT,]
    sig_lrt_values$Ids = row.names(sig_lrt_values)
    sig_lrt_values = sig_lrt_values[c(length(sig_lrt_values),
1:length(sig_lrt_values)-1)]
    sig_lrt_values = data.frame(sig_lrt_values, row.names = NULL)
    write.csv(sig_lrt_values, file = paste0(specificResultsDir, saveString,
"_LRT_values.csv"), row.names = FALSE)


    ### Annotate our heatmap (optional)
    annotation <- data.frame(sampletype=summaryTable[,'condition'],
                             row.names=row.names(summaryTable))

    ### Set a color palette
    heat.colors <- brewer.pal(6, "YlOrRd")

    ### Run pheatmap
    LRT_heatmap = pheatmap(norm_LRTsig, color = heat.colors, cluster_rows =
T, clustering_distance_cols = "manhattan",
                           clustering_method = 'average', show_rownames=F,
annotation= annotation, border_color=NA,
                           fontsize = 10, scale="row", fontsize_row = 10,
height=20)
    save_pheatmap(LRT_heatmap, paste0(specificResultsDir, saveString,
"_LRT_heatmap.png"))

    norm_LRTsig_DataFrame = data.frame(norm_LRTsig)
    norm_LRTsig_DataFrame$ids = row.names(norm_LRTsig_DataFrame)
    norm_LRTsig_DataFrame =
norm_LRTsig_DataFrame[,c(length(norm_LRTsig_DataFrame),
1:(length(norm_LRTsig_DataFrame)-1))]
    row.names(norm_LRTsig_DataFrame) = NULL
    write.csv(norm_LRTsig_DataFrame, file = paste0(specificResultsDir,
saveString, "_LRT_Exp.csv"), row.names = FALSE)


    #################### HTML Report ####################
    ##### Write a HTML report
    # First re-set working directory for R, then continue
    options(scipen=999)
    setwd(specificResultsDir)
    htmlRep <- HTMLReport(shortName = "Report",
                          title = paste("Normalised LRT differential
expression for", saveString),
                          reportDirectory = paste0(saveString,
"_HTMLreport"))
    publish(norm_LRTsig_DataFrame, htmlRep)
    url <- finish(htmlRep)
    #browseURL(url)
    # Return R working directory to normal
    setwd(specificResultsDir)
    #################### END HTML Report ####################
}
```

```
######################################################################
####################     Run     Likelihood     Ratio     Test     Functions
####################
lrtFunction(dds_lrt, "condition", "Breeza", "THH")
lrtFunction(dds_lrt, "condition", "TARC", "THH")
lrtFunction(dds_lrt, "condition", "Breeza", "TARC")

lrtFunction(dds_lrt, "condition", "TARC", "Breeza")
lrtFunction(dds_lrt, "condition", "THH", "Breeza")
lrtFunction(dds_lrt, "condition", "THH", "TARC")

#################### END LRT Function ####################
############################################################
```

D.8.3 R-script: ‘gtfToFastaThenBlastx.R’

```
##################################################
#### Before starting this script, do the following: #####
#*****Decide whether you need transcripts or genes and perform part 1 based
on that information

#1  Run  tximport  and  DESeq2  to  get  the  list  of  genes,  or  the  list  of
transcripts, and save as:
# ~/Google_Drive/PhD/DataBases/RNASeq/RNASeqSignificantIds.txt
########################################


########## START ##########
#2 Use the list of transcripts or genes from above (1) to subset a GTF file
# through the grep (gnugrep) command:


###########################################################################
###############
##########  COPY  THE  NECESSARY  TEXT  FILE  OF  SIGNIFICANT  IDENTIFIERS,  TO:
###############
##########   "~/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/BLASTtemp"
#############
###########################################################################
###############

#NOTE: If the script fails try "dos2unix" on the file of 'Identifiers'

# Load in the necessary packages:
library(seqinr)
library(dplyr)


##### Input paths for conditional statement below #####
macPath = "/Users/Paul/"
desktopPath = "/Users/43533698/"
externalHD = "/Volumes/Seagate_Backup_Plus_Drive/"
```

```
##### Check which computer is being used #####
##### Add Paths ######
if (file.exists(macPath)){
      if (file.exists(externalHD)){
      #Paths in
      sourceGTF                                                            =
"/Volumes/Seagate_Backup_Plus_Drive/DataBases/RNASeq/Triticum_aestivum.TGAC
v1.35.gtf"
      sourceGenomeFASTA                                                    =
"/Volumes/Seagate_Backup_Plus_Drive/DataBases/RNASeq/Triticum_aestivum.TGAC
v1.dna.toplevel.fa"
      } else
        {print("Connect the external HD to the laptop!")}
} else if (file.exists(desktopPath)){
  #Paths in
  sourceGTF                                                                =
"~/BioInformatics/DataBases/RNASeq/Triticum_aestivum.TGACv1.35.gtf"
  sourceGenomeFASTA                                                        =
"~/BioInformatics/DataBases/RNASeq/Triticum_aestivum.TGACv1.dna.toplevel.fa
"
} else {print("Check Paths!")}


###################################################
##### Use "gnugrep" called by R (system) to subset GTF #####

# Construct the query string
# Using gnugrep as it is many times faster than default
idsInputDir                                                                =
"~/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/BLASTtemp/"
getBlastFile = list.files(idsInputDir)
inputFile = grep("Sig_Ids.txt", getBlastFile, value = TRUE)
idString = gsub("(^.*_.*_.*_.*)_.*_.*.txt$", "\\1", inputFile)
outputDirBase = "~/Google_Drive/PhD/DataBases/RNASeq/Blastx_Results/"

outputDirResults = paste0(outputDirBase, idString, "_BlastxResults/")
if (dir.exists(outputDirResults)) {
  print("Directory exists!")
} else  {dir.create(outputDirResults)}

grepSystemScript = paste("gnugrep -F -f",
                         paste0(idsInputDir, inputFile),
                         sourceGTF,
                         paste0(">    ",    outputDirResults,    idString,
"_GTF_Subset.gtf"))

## Print the command to screen to check syntax ##
print(grepSystemScript)
# Send out a system call for terminal to process grep subsetting
system(grepSystemScript)


###################################################
##### Subset the GTF file #####
#3  Use  the  subsetted  gtf  file  of  significant  transcripts  or  genes
("TritAv_RNASeqSignifGenes.gtf")
##### Step 2 #####
gffreadSystemScript = paste("gffread",
                            "-w",    paste0(outputDirResults,    idString,
"_FASTA_subset.fa"),
                           "-M -g", sourceGenomeFASTA,
                           paste0(outputDirResults, idString,
"_GTF_subset.gtf"))

## Print the command to screen to check syntax ##
print(gffreadSystemScript)
# Tell R to run the command in the terminal
system(gffreadSystemScript)

# Load in the fasta file "RNASeqSignificantIdsFASTA.fa" that was generated
by the above script
  # which is a set of significantly expressed genes (or genes and transcripts)
in fasta format
    # Gel all the transcripts, the set of genes can be subset later
subset_IdsOfInterest_FASTA = read.fasta(file = paste0(outputDirResults,
idString, "_FASTA_subset.fa"), seqtype = "DNA", as.string = TRUE,
set.attributes = TRUE)
#################### STOP HERE IF LIST OF ALL SIGNIFICANT TRANSCRIPTS IS
ENOUGH ####################
########## Continue if the subset of genes is needed ##########
# (One sequence per gene instead of one sequence per transcript) #

#################### GENES ONLY ####################
##### Use script below to subset genes from all transcripts #####
# Subset of transcripts of significant interest are in
"subset_IdsOfInterest_FASTA"

# Get the list of significantly expressed genes from
"RNASeqSignificantGenes.txt"
genesOfInterest = read.table(file = paste0(idsInputDir, inputFile),
stringsAsFactors = FALSE, col.names = "ids")
# Turn the table into a character vector
genesOfInterest = genesOfInterest$ids

# Get a subset of genes from the transcripts
  # If there are multiple transcripts per gene
    # only the first transcript will be chosen per gene
subset_Genes_FASTA =
subset_IdsOfInterest_FASTA[match(paste0(genesOfInterest, ".1"),
names(subset_IdsOfInterest_FASTA))]
# Check that the number of "genesOfInterest" equals the
  # Number in the "subset_Genes_FASTA" list
length(genesOfInterest) == length(names(subset_Genes_FASTA))
# Get annotations to use as names when saving fasta file
sub_Genes_fa_ANNOT = getAnnot(subset_Genes_FASTA)
#Remove the ">" to avoid ">>" when saving the fasta file
  #write.fasta automatically adds a ">" character to the name
  #so if one exists there will be two written to file
  #hence the ">>" characters that must be removed
sub_Genes_fa_ANNOT = lapply(sub_Genes_fa_ANNOT, function(x) gsub(">","",x))

#################### Save FASTA file of genes of interest
####################
# Save FASTA file of genes using the annotations as names
  # This retains the coding region information (eg. "CDS = 1, 1000")
    # The final file is: "subset_Genes_FASTA.fasta"
write.fasta(sequences = subset_Genes_FASTA, names = sub_Genes_fa_ANNOT,
            nbchar = 60, file.out = paste0(outputDirResults, idString,
"_subset_Genes.fa"))

#################### Save FASTA file of all transcripts ####################
```

```
# Get annotations to use as names when saving fasta file
subset_IdsOfInterest_fa_ANNOT = getAnnot(subset_IdsOfInterest_FASTA)
#Remove the ">" to avoid ">>" when saving the fasta file
#write.fasta automatically adds a ">" character to the name
#so if one exists there will be two written to file
#hence the ">>" characters that must be removed
subset_IdsOfInterest_fa_ANNOT    =    lapply(subset_IdsOfInterest_fa_ANNOT,
function(x) gsub(">","",x))

write.fasta(sequences     =     subset_IdsOfInterest_FASTA,     names     =
subset_IdsOfInterest_fa_ANNOT,
            nbchar  =  60,  file.out  =  paste0(outputDirResults,  idString,
"_subset_All_Ids_FASTA.fa"))

#  The previous "_FASTA_subset.fa" file (around line 90) is equivalent to
"_subset_All_Ids_FASTA.fa"
  # Therefore delete this file
rmString     =     paste("rm",     paste0(outputDirResults,     idString,
"_FASTA_subset.fa"))
system(rmString)

###################################################
########## Make BLAST Database ##########
##### The above script will result in a FASTA file for both the genes of
interest
  # and all the transcripts (isoforms) of the genes of interest
# Gene = TRIAE_CS42_1AL_TGACv1_000756_AA0018500
#   Transcripts   of   gene   =   TRIAE_CS42_1AL_TGACv1_000756_AA0018500.1,
TRIAE_CS42_1AL_TGACv1_000756_AA0018500.2
                      #            TRIAE_CS42_1AL_TGACv1_000756_AA0018500.3,
TRIAE_CS42_1AL_TGACv1_000756_AA0018500.4

#################### First run "MAKEBLASTDB" through R ###################
# This will first make a blastdb database and then run a blastx query

if
(!file.exists("~/Google_Drive/PhD/DataBases/BLASTproteinDB/wheatBLASTdb/whe
atBLASTdb.psd")) {
  # Run makeBLASTdb
  #4a   Make   a   BLAST   database   from   the   current   wheat   genome
("uniprot_wheat_2017_plusIsoforms_Download.fasta")
  makeDbCommandString = paste("makeblastdb",
                              "-in",
"~/Google_Drive/PhD/DataBases/BLASTproteinDB/wheatBLASTdb_plusIsoforms/unip
rot_wheat_2017_plusIsoforms_Download.fasta",
                              "-dbtype", "prot", "-parse_seqids",
                              "-out",
"~/Google_Drive/PhD/DataBases/BLASTproteinDB/wheatBLASTdb_plusIsoforms/whea
tBLASTdb_plusIsoforms")
  # Print the script to screen for checking
  print(makeDbCommandString)
  # Run the makeblastdb command
  system(makeDbCommandString)

} else  {print("Blastdb exists!")}


############################## BLASTX ##############################
#4b Run a blastx command to transcribe DNA to amino acid sequence to search
a protein database
```

```
########## Function to unfactorise all factored columns ##########
unfactorize <- function(df){
  for(i  in  which(sapply(df,  class)  ==  "factor"))  df[[i]]  =
as.character(df[[i]])
  return(df)
}
#################### END Function ####################


blastxCommandString  =  paste("blastx",  "-query",  paste0(outputDirResults,
idString, "_subset_Genes.fa"),
                            "-db
~/Google_Drive/PhD/DataBases/BLASTproteinDB/wheatBLASTdb_plusIsoforms/wheat
BLASTdb_plusIsoforms",
                            "-outfmt  '10  qseqid  sseqid  bitscore  pident
evalue'",
                            "-evalue .001 -max_target_seqs 1",
                            "-out",   paste0(outputDirResults,   idString,
"_BlastxResult.csv"))

# Print the script to screen for checking
print(blastxCommandString)
# Run the blastx command
system(blastxCommandString)

# Load in blastx results
IdsAfterBlastx   =   read.csv(file   =   paste0(outputDirResults,   idString,
"_BlastxResult.csv"), header = FALSE)
names(IdsAfterBlastx) = c("queryId", "fullUniprotIds", "bitscore", "pident",
"evalue")
IdsAfterBlastx$uniprotIds      =      gsub("^.*\\|(.*)\\|.*$",      "\\1",
IdsAfterBlastx$fullUniprotIds)
IdsAfterBlastx                    =                    IdsAfterBlastx[,c(1:2,
length(IdsAfterBlastx),4:length(IdsAfterBlastx)-1)]

    # source("https://bioconductor.org/biocLite.R")
    # biocLite("UniProt.ws")
#Load the "UniProt.ws" package into R
library("UniProt.ws")
# Need to detach "dplyr" or the "Uniprot.ws" package will not work
detach("package:dplyr", unload = TRUE)

###########Connect  to  UniProt  and  get  UniProt  Identifiers  from  Barley
Identifiers########
#Set the Taxon number for Barley (Wheat = 4565)
speciesId = UniProt.ws(taxId=4565)
#Key Type or Database for the input Identifiers
inputKeyType = "UNIPROTKB"
#Data columns that will be output
outputColumn = c("PROTEIN-NAMES", "SEQUENCE")
#The command to retrive UniProt information
identifierList = IdsAfterBlastx$uniprotIds

# Get the amino acid sequence attached to the uniprot identifier
retrievedIdsSequence  =  select(speciesId,  identifierList,  outputColumn,
inputKeyType)
#Change the name of the first column in the akRetrieve data.frame
names(retrievedIdsSequence)[1] = "uniprotIds"
detach("package:UniProt.ws", unload = TRUE)
library("dplyr")
```

```
# Match "uniprotIds" column between "retrievedIdsSequence" and
"IdsAfterBlastx" tables and add matched data to "IdsAfterBlastx"
IdsAfterBlastx$Protein_Names = retrievedIdsSequence$`PROTEIN-
NAMES`[match(retrievedIdsSequence$uniprotIds, IdsAfterBlastx$uniprotIds)]
IdsAfterBlastx$AA_Sequence =
retrievedIdsSequence$SEQUENCE[match(retrievedIdsSequence$uniprotIds,
IdsAfterBlastx$uniprotIds)]
# Turn any factorised columns to character type
IdsAfterBlastx = unfactorize(IdsAfterBlastx)
# Save the full data frame to file
write.csv(IdsAfterBlastx, file = paste0(outputDirResults, idString,
"_BlastxAndSequence.csv"), row.names = FALSE)
# Save the Identifiers to file for use in GO Retriever or other Gene Ontology
packages
write.table(IdsAfterBlastx$uniprotIds, file = paste0(outputDirResults,
idString, "_Ids_For_GO.txt"), sep = "\t", row.names = FALSE, col.names =
FALSE)

########################################################################
#################################
#################### Use "GORetriever" to get gene ontology Identifiers from
protein identifiers ####################

# Get GO Identifiers using "_Ids_For_GO.txt" file

###############################################
########## Get paste0(idString, "_All_GO.csv") - GO Retriever ##########
########## Get paste0(idString, "_GO_Slim.txt") - GO Slims Viewer ##########
########## Get "goSlims" from "goslimviewer" ##########
###############################################
```

D.8.4 R-script: ‘extraGOScript.R’

```
options(scipen = 0)

library("GO.db") # Also loads in "AnnotationDbi" for using command "Term"
below
                    # This command gets GO Terms from GO Ids

if(.Platform$OS.type == "windows"){
  # Set input directory
  idsInputDir =
"C:/Users/paul_/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/BLASTtemp/
"
} else {
  # Set input directory
  idsInputDir =
"~/Google_Drive/PhD/DataBases/RNASeq/sigIdentifiers/BLASTtemp/"
}

# Check filenames in the directory
getBlastFile = list.files(idsInputDir)
# Capture the file of interest
```

```
inputFile = grep("Sig_Ids.txt", getBlastFile, value = TRUE)
# Get the information string from the file of interest
idString = gsub("(^.*_.*_.*)_.*_.*.txt$", "\\1", inputFile)

if(.Platform$OS.type == "windows"){
  # Set an output base directory
  outputDirBase                                                        =
"C:/Users/paul_/Google_Drive/PhD/DataBases/RNASeq/Blastx_Results/"
} else {
  # Set an output base directory
  outputDirBase = "~/Google_Drive/PhD/DataBases/RNASeq/Blastx_Results/"
}


# Set the full output directory
outputDirResults = paste0(outputDirBase, idString, "_BlastxResults/")

##################################################
########## Get paste0(idString, "_All_GO.csv") ##########
########## Get paste0(idString, "_GO_Slim.txt") ##########
##################################################


# Load in the file ending in "_All_GO.csv" after getting go:ids and go:Slims
from agbase website
idsTable = read.csv(file = paste0(outputDirResults, idString,
"_All_GO.csv"), stringsAsFactors = FALSE)
# Select the columns needed
idsTable = idsTable[, c("Input_Accession", "Input_GOID", "Input_GO_Name",
"GO_Type")]
# Replace spaces or commas with "_"
idsTable$Input_GO_Name = gsub(" ", "_", idsTable$Input_GO_Name)
idsTable$Input_GO_Name = gsub(",", "_", idsTable$Input_GO_Name)
# Aggregate the columns based on the protein identifier (similar to "wego"
format)
idsTable = aggregate(cbind(Input_GOID, Input_GO_Name, GO_Type) ~
Input_Accession, data = idsTable, paste, collapse = " ", na.action = na.pass)
names(idsTable)[1] = "Ids"


# Load in the results from the BLASTX
blastxTable = read.csv(file = paste0(outputDirResults, idString,
"_BlastxAndSequence.csv"), stringsAsFactors = FALSE, header = TRUE)
# Copy the blastxTable data frame and use "match" to construct more columns
blastxPlusFullInfo = blastxTable
blastxPlusFullInfo$GO_ID                                               =
idsTable$GO_ID[match(blastxPlusFullInfo$uniprotIds, idsTable$uniprotIds)]
blastxPlusFullInfo$GO_Term_Name                                        =
idsTable$GO_Term_Name[match(blastxPlusFullInfo$uniprotIds,
idsTable$uniprotIds)]
blastxPlusFullInfo$Aspect                                              =
idsTable$Aspect[match(blastxPlusFullInfo$uniprotIds, idsTable$uniprotIds)]

# Load in the "GO_Slim" results saved as ".csv"
goSlimTable = read.table(file = paste0(outputDirResults, idString,
"_GO_Slim.txt"), stringsAsFactors = FALSE, header = FALSE, sep = "\t")
titles = c("uniprotIds", "goSlim", "goSlim_Aspect")
names(goSlimTable) = titles
# Get the GO Term from the GO Ids
go_Terms = Term(goSlimTable$goSlim)
goSlimTable$goSlim_Terms = go_Terms
```

```
# Rearrange the table
goSlimTable = goSlimTable[c(1:2, length(goSlimTable), length(goSlimTable)-
1)]
# Replace spaces or commas with "_"
goSlimTable$goSlim_Terms = gsub(" ", "_", goSlimTable$goSlim_Terms)
goSlimTable$goSlim_Terms = gsub(",", "_", goSlimTable$goSlim_Terms)
# Aggregate so the data can be matched with blastx results
goSlimTable  =  aggregate(cbind(goSlim,  goSlim_Terms,  goSlim_Aspect)  ~
uniprotIds, data = goSlimTable, paste, collapse = " ", na.action = na.pass)
# Grow the blastx table using the "match" command
blastxPlusFullInfo$goSlim                                              =
goSlimTable$goSlim[match(blastxPlusFullInfo$uniprotIds,
goSlimTable$uniprotIds)]
blastxPlusFullInfo$goSlim_Terms                                        =
goSlimTable$goSlim_Terms[match(blastxPlusFullInfo$uniprotIds,
goSlimTable$uniprotIds)]
blastxPlusFullInfo$goSlim_Aspect                                       =
goSlimTable$goSlim_Aspect[match(blastxPlusFullInfo$uniprotIds,
goSlimTable$uniprotIds)]
blastxPlusFullInfo[is.na(blastxPlusFullInfo)] = "Uncharacterized protein"
names(blastxPlusFullInfo)[1] = "Ids"
# Write the table to disc
write.csv(blastxPlusFullInfo,  file  =  paste0(outputDirResults,  idString,
"_blastxPlusFullInfo.csv"), row.names = FALSE)
# rm(list = ls(pattern="[^blastxPlusFullInfo]"))

###################################################################
########### Make a GO Slims summary and save it######################
library("dplyr")
#get the "path" to the "biological process" file for go Slims
bpPath = list.files(path = paste0(outputDirResults, "goSlims/"), pattern =
".*\\.bp\\.txt", all.files = FALSE, full.names = TRUE)
#get the "path" to the "cellular component" file for go Slims
ccPath = list.files(path = paste0(outputDirResults, "goSlims/"), pattern =
".*\\.cc\\.txt", all.files = FALSE, full.names = TRUE)
#get the "path" to the "molecular function" file for go Slims
mfPath =  list.files(path = paste0(outputDirResults, "goSlims/"), pattern =
".*\\.mf\\.txt", all.files = FALSE, full.names = TRUE)

#Read the tab delimited file from GOSlimViewer and turn it into a data.frame
biologicalProcess = read.table(bpPath, header = FALSE, stringsAsFactors =
FALSE, sep = "\t")
#Write in the column names (vectors)
names(biologicalProcess) = c("Slims_GO_ID", "go_Term", "Count")
#Add a new column and fill it with the "Biological Process" description
biologicalProcess$Process = "Biological Process"

#Read the tab delimited file from GOSlimViewer and turn it into a data.frame
cellularComponent = read.delim(ccPath, header = FALSE, stringsAsFactors =
FALSE)
#Write in the column names (vectors)
names(cellularComponent) = c("Slims_GO_ID", "go_Term", "Count")
#Add a new column and fill it with the "Biological Process" description
cellularComponent$Process = "Cellular Component"

#Read the tab delimited file from GOSlimViewer and turn it into a data.frame
molecularFunction = read.delim(mfPath, header = FALSE, stringsAsFactors =
FALSE)
#Write in the column names (vectors)
names(molecularFunction) = c("Slims_GO_ID", "go_Term", "Count")
#Add a new column and fill it with the "Biological Process" description
```

```
molecularFunction$Process = "Molecular Function"

#Bind the "biologicalProcess", "cellularComponent", "molecularFunction"
data.frames together
allSlims = bind_rows(biologicalProcess, cellularComponent,
molecularFunction)
#Remove any rows with "_" in the go_Term column (these are level one terms)
allSlims = allSlims %>% filter(!grepl(".*_.*", go_Term))

# Write the GO Slims summary to disc
write.csv(allSlims, file = paste0(outputDirResults, idString,
"_allSlimsList.csv"), row.names = FALSE)



if (any(grepl("(^.*)\\.[0-9]*$", blastxPlusFullInfo$Ids))){
# Do the following for genes - not transcripts
  blastxPlusFullInfo$Ids = gsub("(^.*)\\.[0-9]*$", "\\1",
blastxPlusFullInfo$Ids)
print("Changed!")
}


if(.Platform$OS.type == "windows"){
  getValuesPathBase = "C:/Users/paul_/Google_Drive/PhD/DEanalysis/results/"
} else {
  getValuesPathBase = "~/Google_Drive/PhD/DEanalysis/results/"
}

getFullValuesPath = paste0(getValuesPathBase, idString, "/")
studyId = gsub("^(.*_.*_.*)_.*$", "\\1", idString)

studyId = "All_anova_like_LRT"
listFiles = list.files(getFullValuesPath, pattern = "values.csv")
Values = read.csv(file = paste0(getFullValuesPath,listFiles),
stringsAsFactors = FALSE)

ValuesJoinBlastx = left_join(Values, blastxPlusFullInfo, by = "Ids")
ValuesJoinBlastx = unique(ValuesJoinBlastx)
ValuesJoinBlastx =
ValuesJoinBlastx[!is.na(ValuesJoinBlastx$fullUniprotIds), ]

write.csv(ValuesJoinBlastx, file = paste0(outputDirResults, idString,
"_ValuesJoinBlastx.csv"), row.names = FALSE)
```

D.8.5 R-script: ‘makeFastaFromTable.R’

```
# If prior steps which require connection to Unipro.ws database do not work,
  # get the protein identifiers and manually do a Uniparc search
  # with a conversion to UniprotKB


# Use seqinr to make a fasta file
```

```
library("seqinr")
options(scipen=999)
if(.Platform$OS.type == "windows"){
  Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jdk1.8.0_191')
}
library("rJava")
detach("package:rJava", unload=TRUE)
library("tcltk")
library("rChoiceDialogs")


unfactorize <- function(df){
  for(i in which(sapply(df, class) == "factor")) df[[i]] =
as.character(df[[i]])
  return(df)
}


if(.Platform$OS.type == "windows"){
  rnaseqAndBlastxTablePath = tk_choose.files(default =
"C:/Users/paul_/Google_Drive/PhD/DataBases/RNASeq/Blastx_Results", caption
= "Select the table ('..._blastxPlusFullInfo.csv') with amino acid sequence
and names",
                               multi = FALSE, filters = NULL, index = 1)
} else {
  rnaseqAndBlastxTablePath = tk_choose.files(default =
"~/Google_Drive/PhD/DataBases/RNASeq/Blastx_Results", caption = "Select the
table ('..._blastxPlusFullInfo.csv') with amino acid sequence and names",
                               multi = FALSE, filters = NULL, index = 1)
}

rnaseqAndBlastxTableFile = paste0(gsub( "^(.*)\\.csv", "\\1",
rnaseqAndBlastxTablePath), "_FASTA.faa")


rnaseqAndBlastxTable = read.csv(rnaseqAndBlastxTablePath, stringsAsFactors =
FALSE)
rnaseqAndBlastxTableSeqAndNames = rnaseqAndBlastxTable[ ,c("AA_Sequence",
"uniprotIds")]


write.fasta(sequences =
as.list(rnaseqAndBlastxTableSeqAndNames$AA_Sequence), names =
rnaseqAndBlastxTableSeqAndNames$uniprotIds, file.out =
rnaseqAndBlastxTableFile, as.string = TRUE)
```

Appendix E.

Pages 283-298 of this thesis have been removed as they contain published material under copyright. Removed contents published as:

Mirzaei M., Wu Y., Worden P., Jerkovic A., Atwell B.J. (2016) How Proteomics Contributes to Our Understanding of Drought Tolerance. In: Salekdeh G. (eds) *Agricultural Proteomics Volume 2*. Springer, Cham. https://doi.org/10.1007/978-3-319-43278-6_6

Appendix F.

Pages 300-310 of this thesis have been removed as they contain published material under copyright. Removed contents published as:

Robert D. Willows, Paul Worden, Mehdi Mirzaei, (2017) Barley Grain Proteomics, in Colgrave, M. L. (Ed.), *Proteomics in Food Science*, (pp. 75-88) Academic Press. https://doi.org/10.1016/B978-0-12-804007-2.00005-9

References

Abbo, S., A. Gopher, Z. Peleg, Y. Saranga, T. Fahima, F. Salamini, et al. 2006. The ripples of "The Big (agricultural) Bang": the spread of early wheat cultivation. Genome 49: 861-863. doi:10.1139/g06-049.

Aliscioni, A., H.L. Bell, G. Besnard, P.A. Christin, J.T. Columbus, M.R. Duvall, et al. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytologist 193: 304-312. doi:10.1111/j.1469-8137.2011.03972.x.

Altenbach, S.B., W.H. Vensel and F.M. DuPont. 2010. Analysis of expressed sequence tags from a single wheat cultivar facilitates interpretation of tandem mass spectrometry data and discrimination of gamma gliadin proteins that may play different functional roles in flour. BMC Plant Biology 10.

Amalraj, R.S., N. Selvaraj, G.K. Veluswamy, R.P. Ramanujan, R. Muthurajan, M. Palaniyandi, et al. 2010. Sugarcane proteomics: Establishment of a protein extraction method for 2-DE in stalk tissues and initiation of sugarcane proteome reference map. Electrophoresis 31: 1959-1974. doi:10.1002/elps.200900779.

Anwar, M.R., D.L. Liu, R. Farquharson, I. Macadam, A. Abadi, J. Finlayson, et al. 2015. Climate change impacts on phenology and yields of five broadacre crops at four climatologically distinct locations in Australia. Agricultural Systems 132: 133-144. doi:https://doi.org/10.1016/j.agsy.2014.09.010.

Ashoub, A., T. Berberich, T. Beckhaus and W. Brüggemann. 2011. A competent extraction method of plant proteins for 2-D gel electrophoresis. Electrophoresis 32: 2975-2978. doi:10.1002/elps.201100150.

Balter, M. 2007. Seeking agriculture's ancient roots. Science (New York, N.Y.) 316: 1830-1835.

Barak, S., D. Mudgil and B.S. Khatkar. 2015. Biochemical and Functional Properties of Wheat Gliadins: A Review. Critical Reviews in Food Science and Nutrition 55: 357-368. doi:10.1080/10408398.2012.654863.

Barros, E., S. Lezar, M.J. Anttonen, J.P. Van Dijk, R.M. Röhlig, E.J. Kok, et al. 2010. Comparison of two GM maize varieties with a near-isogenic non-GM variety using transcriptomics, proteomics and metabolomics. Plant Biotechnology Journal 8: 436-451. doi:doi:10.1111/j.1467-7652.2009.00487.x.

Batey, I.L.G., R. B. and MacRitchie. F. . 1991. Use of Size-Exclusion High-Performance Liquid Chromatography in the Study of Wheat Flour Proteins: An Improved Chromatographic Procedure. Cereal Chemistry 68: 207-209.

Bathgate, G.N. and G.H. Palmer. 1972. A Reassessment of the Chemical Structure of Barley and Wheat Starch Granules. Starch - Stärke 24: 336-341. doi:10.1002/star.19720241004.

Bazargani, M.M., E. Sarhadi, A.-A.S. Bushehri, A. Matros, H.-P. Mock, M.-R. Naghavi, et al. 2011. A proteomics view on the role of drought-induced senescence and oxidative stress defense in enhanced stem reserves remobilization in wheat. Journal of Proteomics 74: 1959-1973. doi:http://dx.doi.org/10.1016/j.jprot.2011.05.015.

Blochet, J.-E.K., A.; Compoint, J. P. and Marion, D. 1991. Gluten Proteins 1990
Chapter: "Amphiphilic proteins from wheat flour: specific extraction, structure and lipid binding properties. "AACC, St. Paul, MN, USA.

Bottomley, R.C., H.F. Kearns and J.D. Schofield. 1982. Characterisation of wheat flour and gluten proteins using buffers containing sodium dodecyl sulphate. Journal of the science of food and agriculture 33: 481-491. doi:10.1002/jsfa.2740330514.

Branlard, G. and M. Dardevet. 1994. A Null Gli-D1 Allele with a Positive Effect on Bread Wheat Quality. Journal of Cereal Science 20: 235-244. doi:http://dx.doi.org/10.1006/jcrs.1994.1063.

Bray, N.L., H. Pimentel, P. Melsted and L. Pachter. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotech 34: 525-527. doi:10.1038/nbt.3519 http://www.nature.com/nbt/journal/v34/n5/abs/nbt.3519.html#supplementary-information.

Brown, J.W.S. and R.B. Flavell. 1981. Fractionation of wheat gliadin and glutenin subunits by two-dimensional electrophoresis and the role of group 6 and group 2 chromosomes in gliadin synthesis. Theoretical and Applied Genetics 59: 349-359. doi:10.1007/bf00276448.

Budak, H., B.A. Akpinar, T. Unver and M. Turktas. 2013. Proteome changes in wild and modern wheat leaves upon drought stress by two-dimensional electrophoresis and nanoLC-ESI-MS/MS. Plant Mol Biol 83: 89-103. doi:10.1007/s11103-013-0024-5.

Capriotti, A.L., G.M. Borrelli, V. Colapicchioni, R. Papa, S. Piovesana, R. Samperi, et al. 2014. Proteomic study of a tolerant genotype of durum wheat under salt-stress conditions. Anal Bioanal Chem 406: 1423-1435. doi:10.1007/s00216-013-7549-y.

Carlson, M. 2016. GO.db: A set of annotation maps describing the entire Gene Ontology.

Caruso, G., C. Cavaliere, P. Foglia, R. Gubbiotti, R. Samperi and A. Laganà. 2009. Analysis of drought responsive proteins in wheat (Triticum durum) by 2D-PAGE and MALDI-TOF mass spectrometry. Plant Science 177: 570-576. doi:http://dx.doi.org/10.1016/j.plantsci.2009.08.007.

Charif, D. and J.R. Lobry. 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: U. Bastolla, M. Porto, H. E. Roman and M. Vendruscolo, editors, Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Springer Berlin Heidelberg, Berlin, Heidelberg. p. 207-232.

Cremer, F. and C. Van de Walle. 1985. Method for extraction of proteins from green plant tissues for two-dimensional polyacrylamide gel electrophoresis. Analytical Biochemistry 147: 22-26. doi:http://dx.doi.org/10.1016/0003-2697(85)90004-1.

Dachkevitch, T. and J. Autran. 1989. PREDICTION OF BAKING QUALITY OF BREAD WHEATS IN BREEDING PROGRAMS BY SIZE-EXCLUSION HIGH-PERFORMANCE LIQUID-CHROMATOGRAPHY. Cereal Chem. 66: 448-456.

Damerval, C., D. De Vienne, M. Zivy and H. Thiellement. 1986. Technical improvements in two-dimensional electrophoresis increase the level of genetic variation detected in wheat-seedling proteins. Electrophoresis 7: 52-54. doi:10.1002/elps.1150070108.

Danno, G.-I., K. Kanazawa and M. Natake. 1974. Extraction of Wheat Flour Proteins with Sodium Dodecyl Sulfate and Their Molecular Weight Distribution. Agricultural and Biological Chemistry 38: 1947-1953. doi:10.1080/00021369.1974.10861440.

Darragh, A.J., D.J. Garrick, P.J. Moughan and W.H. Hendriks. 1996. Correction for Amino Acid Loss during Acid Hydrolysis of a Purified Protein. Analytical Biochemistry 236: 199-207. doi:http://dx.doi.org/10.1006/abio.1996.0157.

Domżalska, L., A. Mikuła and J.J. Rybczyński. 2016. Protein extraction from Ca-alginate encapsulated plant material for comparative proteomic analysis. Protein Expression and Purification 126: 55-61. doi:http://dx.doi.org/10.1016/j.pep.2016.05.014.

Dreccer, M.F., J. Fainges, J. Whish, F.C. Ogbonnaya and V.O. Sadras. 2018. Comparison of sensitive stages of wheat, barley, canola, chickpea and field pea to temperature and water stress across Australia. Agricultural and Forest Meteorology 248: 275-294. doi:https://doi.org/10.1016/j.agrformet.2017.10.006.

Dubcovsky, J. and J. Dvorak. 2007. Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. Science 316: 1862.

Dunn, M.J. 1995. Editorial. Electrophoresis 16: 1077-1078. doi:10.1002/elps.11501601182.

Elton, G.A. and J.A. Ewart. 1960. Starch-gel electrophoresis of wheat proteins. Nature 187: 600.

FAO. 2005. Grasslands of the worldFOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS,, Rome.

Farooq, M., M. Hussain and K.H.M. Siddique. 2014. Drought Stress in Wheat during Flowering and Grain-filling Periods. Critical Reviews in Plant Sciences 33: 331-349. doi:10.1080/07352689.2014.875291.

Feldman, M. and M.E. Kislev. 2007. Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. Israel journal of plant sciences 55: 207-221. doi:10.1560/IJPS.55.3-4.207.

Franco, O.L., D.J. Rigden, F.R. Melo and M.F. Grossi-de-Sá. 2002. Plant α-amylase inhibitors and their interaction with insect α-amylases. European Journal of Biochemistry 269: 397-412. doi:10.1046/j.0014-2956.2001.02656.x.

Fuertes-Mendizábal, T., J. González-Torralba, L.M. Arregui, C. González-Murua, M.B. González-Moro and J.M. Estavillo. 2013. Ammonium as sole N source improves grain quality in wheat. Journal of the science of food and agriculture 93: 2162-2171. doi:10.1002/jsfa.6022.

Gao, L. and W. Bushuk. 1992. Solubilization of glutenin in urea/SDS solutions at elevated temperature. Journal of Cereal Science 16: 81-89. doi:10.1016/S0733-5210(09)80081-7.

Garcia-Seco, D., M. Chiapello, M. Bracale, C. Pesce, P. Bagnaresi, E. Dubois, et al. 2017. Transcriptome and proteome analysis reveal new insight into proximal and distal responses of wheat to foliar infection by Xanthomonas translucens. Scientific Reports 7: 10157. doi:10.1038/s41598-017-10568-8.

Gaut, B.S. 2002. Tansley Review No. 132. Evolutionary Dynamics of Grass Genomes. The New Phytologist 154: 15-28.

Ge, P., C. Ma, S. Wang, L. Gao, X. Li, G. Guo, et al. 2012. Comparative proteomic analysis of grain development in two spring wheat varieties under drought stress. Anal Bioanal Chem 402: 1297-1313. doi:10.1007/s00216-011-5532-z.

Giles, R.J. and T.A. Brown. 2006. GluDy allele variations in Aegilops tauschii and Triticum aestivum: implications for the origins of hexaploid wheats. Theor Appl Genet 112: 1563-1572. doi:10.1007/s00122-006-0259-5.

Gómez-Vidal, S., M. Tena, L.V. Lopez-Llorca and J. Salinas. 2008. Protein extraction from Phoenix dactylifera L. leaves, a recalcitrant material, for two-dimensional electrophoresis. Electrophoresis 29: 448-456. doi:10.1002/elps.200700380.

Graham, J.S.D. 1963. Starch-Gel Electrophoresis of Wheat Flour Proteins. Australian Journal of Biological Sciences 16: 342-349. doi:http://dx.doi.org/10.1071/BI9630342.

Graveland, A., P. Bongers and P. Bosveld. 1979. Extraction and fractionation of wheat flour proteins. Journal of the science of food and agriculture 30: 71-84. doi:10.1002/jsfa.2740300112.

Guess, H.A. 1900. The gluten constituents of wheat and flour and their relationship to breadmaking properties. Journal of the American Chemical Society 22: 263-268.

Gupta, P.K., R.R. Mir, A. Mohan and J. Kumar. 2008. Wheat Genomics: Present Status and Future Prospects. International Journal of Plant Genomics 2008: 896451. doi:10.1155/2008/896451.

Gupta, R.B. and F. MacRitchie. 1994. Allelic variation at glutenin subunit and gliadin loci, Glu-1, Glu-3 and Gli-1 of common wheats. II. Biochemical basis of the allelic effects on dough properties. Allelic variation at glutenin subunit and gliadin loci, Glu-1, Glu-3 and Gli-1 of common wheats. II. Biochemical basis of the allelic effects on dough properties: 19-29.

H. Samarah, N. 2005. Effects of drought stress on growth and yield of barley. Agronomy for Sustainable Development 25: 145-149.

Hajheidari, M., A. Eivazi, B.B. Buchanan, J.H. Wong, I. Majidi and G.H. Salekdeh. 2007. Proteomics uncovers a role for redox in drought tolerance in wheat. Journal of proteome research 6: 1451-1460. doi:10.1021/pr060570j.

Halton, P. 1924. The chemistry of the strength of wheat flour. J. Agric. Sci. 14: 587-599. doi:10.1017/S0021859600004007.

Hari, V. 1981. A method for the two-dimensional electrophoresis of leaf proteins. Analytical Biochemistry 113: 332-335. doi:http://dx.doi.org/10.1016/0003-2697(81)90085-3.

He, J., S. Penson, S.J. Powers, C. Hawes, P.R. Shewry and P. Tosi. 2013. Spatial Patterns of Gluten Protein and Polymer Distribution in Wheat Grain. Journal of Agricultural and Food Chemistry 61: 6207-6215. doi:10.1021/jf401623d.

He, Z., L. Liu, X. Xia and J. Liu. 2005. Composition of HMW and LMW Glutenin Subunits and Their Effects on Dough Properties, Pan Bread, and Noodle Quality of Chinese Bread Wheats. Cereal Chemistry 82: 345-350.

Helbaek, H. 1959. Domestication of Food Plants in the Old World. Science 130: 365-372.

Hillman, G.C. and M.S. Davies. 1990. 6. Domestication rates in wild-type wheats and barley under primitive cultivation. Biological Journal of the Linnean Society 39: 39-78. doi:10.1111/j.1095-8312.1990.tb01611.x.

Holding, D., M. Otegui, B. Li, R. Meeley, T. Dam, B. Hunter, et al. 2007. The Maize Floury1 Gene Encodes a Novel Endoplasmic Reticulum Protein Involved in Zein Protein Body Formation(W). Plant Cell 19: 2569-2582. doi:10.1105/tpc.107.053538.

Hopf, M. 1983. Jericho plant remains. Excavations at Jericho 5: 576-621.

Hoseney, R.C., K.F. Finney, Y. Pomeranz and M.D. Shogren. 1969. Functional (breadmaking) and biochemical properties of wheat flour components. IV. Gluten protein fractionation by solubilizing in 70 percent ethyl alcohol and in dilute lactic acid. Cereal Chemistry 46: 495-502.

Huebner, F.R. and J.S. Wall. 1976. Fractionation and Quantitative Differences of Glutenin from Wheat Varieties Varying in Baking Quality. Cereal Chemistry 53: 258 - 269.

Huntley, M.A., J.L. Larson, C. Chaivorapol, G. Becker, M. Lawrence, J.A. Hackney, et al. 2013. ReportingTools: an automated result processing and presentation toolkit for high-throughput genomic analyses. Bioinformatics (Oxford, England) 29: 3220-3221. doi:10.1093/bioinformatics/btt551.

Hurkman, W.J. and C.K. Tanaka. 1986. Solubilization of plant membrane proteins for analysis by two-dimensional gel electrophoresis. Plant Physiol 81: 802-806.

Hurkman, W.J., C.K. Tanaka, W.H. Vensel, R. Thilmony and S.B. Altenbach. 2013. Comparative proteomic analysis of the effect of temperature and fertilizer on gliadin and glutenin accumulation in the developing endosperm and flour from Triticum aestivum L. cv. Butte 86. Proteome science 11: 8. doi:10.1186/1477-5956-11-8.

Hurkman, W.J., W.H. Vensel, C.K. Tanaka, L. Whitehand and S.B. Altenbach. 2009. Effect of high temperature on albumin and globulin accumulation in the endosperm proteome of the developing wheat grain. Journal of Cereal Science 49: 12-23. doi:http://dx.doi.org/10.1016/j.jcs.2008.06.014.

Jerkovic, A. 2011. Biochemical and physiological studies of abscisic acid treated wheat (Triticum aestivum) grain Macquarie University.

Jerkovic, A., A.M. Kriegel, J.R. Bradner, B.J. Atwell, T.H. Roberts and R.D. Willows. 2010. Strategic Distribution of Protective Proteins within Bran Layers of Wheat Protects the Nutrient-Rich Endosperm. Plant Physiology 152: 1459.

Ji, Q., X. Xu and K. Wang. 2013. Genetic transformation of major cereal crops.

Jiang, S.-S., X.-N. Liang, X. Li, S.-L. Wang, D.-W. Lv, C.-Y. Ma, et al. 2012. Wheat Drought-Responsive Grain Proteome Analysis by Linear and Nonlinear 2-DE and MALDI-TOF Mass Spectrometry. International Journal of Molecular Sciences 13: 16065-16083. doi:10.3390/ijms131216065.

Kasarda, D.D., H.P. Tao, P.K. Evans, A.E. Adalsteins and S.W. Yuen. 1988. Sequencing of protein from a single spot of a 2-D gel pattern: N-terminal sequence of a major wheat LMW-Glutenin subunit. Sequencing of protein from a single spot of a 2-D gel pattern: N-terminal sequence of a major wheat LMW-Glutenin subunit 39: 899-906.

Katz, S. and M. Voigt. 1986. Bread and Beer. Expedition 28: 23.

Kerber, E.R. 1964. Wheat: Reconstitution of the Tetraploid Component (AABB) of Hexaploids. Science 143: 253-255.

Kew, R. 2016. The State of the World's Plants Report – 2016. Royal Botanic Gardens, Kew; Sfumato Foundation, Royal Botanic Gardens, Kew.

Kislev, M. 1980. Triticum parvicoccum sp. Nov., the oldest naked wheat.

Kislev, M.E. 1984. Emergence of Wheat Agriculture. Paléorient 10: 61-70. doi:10.3406/paleo.1984.940.

Kolde, R. 2015. pheatmap: Pretty Heatmaps.

Kosová, K., P. Vítámvás, I.T. Prášil and J. Renaut. 2011. Plant proteome changes under abiotic stress — Contribution of proteomics studies to understanding plant stress response. Journal of Proteomics 74: 1301-1322. doi:http://dx.doi.org/10.1016/j.jprot.2011.02.006.

Kurowska, E. and W. Bushuk. 1988. Solubility of flour and gluten protein in a solvent of acetic acid, urea, and cetyltrimethylammonium bromide, and its relationship to dough strength. Solubility of flour and gluten protein in a solvent of acetic acid, urea, and cetyltrimethylammonium bromide, and its relationship to dough strength: 156-158.

Laemmli, U.K. 1970. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. Nature 227: 680-685. doi:10.1038/227680a0.

Lev-Yadun, S., A. Gopher and S. Abbo. 2000. The Cradle of Agriculture. Science 288: 1602-1603. doi:10.1126/science.288.5471.1602.

Li, Z. and H. Trick. 2005. Rapid method for high-quality RNA isolation from seed endosperm containing high levels of starch. Biotechniques 38: 872-+. doi:10.2144/05386BM05.

Lomax, J. 2005. Get ready to GO! A biologist's guide to the Gene Ontology. Briefings in Bioinformatics 6: 298-304. doi:10.1093/bib/6.3.298.

Love, M.I., W. Huber and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15: 550. doi:10.1186/s13059-014-0550-8.

Love, M.I., W. Huber and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 15: 550-550. doi:10.1186/s13059-014-0550-8.

Lowe, R., N. Shirley, M. Bleackley, S. Dolan and T. Shafee. 2017. Transcriptomics technologies. PLOS Computational Biology 13: e1005457. doi:10.1371/journal.pcbi.1005457.

Lugg, J.W.H. 1946. Problems associated with the acid hydrolysis of an impure protein preparation. Biochemical Journal 40: 88-96.

Ma, D., X. Huang, J. Hou, Y. Ma, Q. Han, G. Hou, et al. 2018. Quantitative analysis of the grain amyloplast proteome reveals differences in metabolism between two wheat cultivars at two stages of grain development.(Report). BMC Genomics 19. doi:10.1186/s12864-018-5174-z.

MacRitchie, F. 2016. Seventy years of research into breadmaking quality. Journal of Cereal Science 70: 123-131. doi:http://dx.doi.org/10.1016/j.jcs.2016.05.020.

MacRitchie, F., D.D. Kasarda and D.D. Kuzmicky. 1991. Characterization of wheat protein fractions differing in contributions to breadmaking quality. Cereal Chemistry 68: 122-130.

Madani, A., A.S. Rad, A. Pazoki, G. Nourmohammadi and R. Zarghami. 2010. Wheat (Triticum aestivum L.) grain filling and dry matter partitioning responses to source:sink modifications under postanthesis water and nitrogen deficiency. Acta Scientiarum. Agronomy 32: 145-151. doi:10.4025/actasciagron.v32i1.6273.

Maldonado, A.M., S. Echevarria-Zomeno, S. Jean-Baptiste, M. Hernandez and J.V. Jorrin-Novo. 2008. Evaluation of three different protocols of protein extraction for Arabidopsis thaliana leaf proteome analysis by two-dimensional electrophoresis. J Proteomics 71: 461-472. doi:10.1016/j.jprot.2008.06.012.

Matsuoka, Y. and S. Nasuda. 2004. Durum wheat as a candidate for the unknown female progenitor of bread wheat: an empirical study with a highly fertile F1 hybrid with Aegilops tauschii Coss. Theor Appl Genet 109: 1710-1717. doi:10.1007/s00122-004-1806-6.

McCarthy, F.M., N. Wang, G.B. Magee, B. Nanduri, M.L. Lawrence, E.B. Camon, et al. 2006. AgBase: a functional genomics resource for agriculture. BMC Genomics 7: 229. doi:10.1186/1471-2164-7-229.

McCorriston, J. 2000. Barley. In: K. F. Kiple and K. C. Ornelas, editors, The Cambridge World History of Food. Cambridge University Press, Cambridge. p. 81-89.

Meredith, O.B. and J.J. Wren. 1966. Determination of Molecular-Weight Distribution in Wheat-Flour Proteins by Extraction and Gel Filtration in a Dissociating Medium. Cereal Chemistry 43: 169 - 186.

Metakovsky, E.V., P. Annicchiarico, G. Boggini and N.E. Pogna. 1997. Relationship Between Gliadin Alleles and Dough Strength in Italian Bread Wheat Cultivars. Journal of Cereal Science 25: 229-236. doi:10.1006/jcrs.1996.0088.

Metakovsky, E.V., I. Felix and G. Branlard. 1997. Association Between Dough Quality (W value) and Certain Gliadin Alleles in French Common Wheat Cultivars. Journal of Cereal Science 26: 371-373. doi:http://dx.doi.org/10.1006/jcrs.1997.0130.

Mirzaei, M., D. Pascovici, J.X. Wu, J. Chick, Y. Wu, B. Cooke, et al. 2017. TMT One-Stop Shop: From Reliable Sample Preparation to Computational Analysis Platform. Methods Mol Biol 1549: 45-66. doi:10.1007/978-1-4939-6740-7_5.

Monaghan, J.M., J.W. Snape, A.J.S. Chojecki and P.S. Kettlewell. 2001. The use of grain protein deviation for identifying wheat cultivars with high grain protein concentration and yield. Euphytica 122: 309-317. doi:10.1023/a:1012961703208.

Montealegre, C., M.L. Marina and C. Garcia-Ruiz. 2010. Separation of olive proteins combining a simple extraction method and a selective capillary electrophoresis (CE) approach: application to raw and table olive samples. J Agric Food Chem 58: 11808-11813. doi:10.1021/jf1026313.

Nadel, D., D.R. Piperno, I. Holst, A. Snir and E. Weiss. 2015. New evidence for the processing of wild cereal grains at Ohalo II, a 23 000-year-old campsite on the shore of the Sea of Galilee, Israel. Antiquity 86: 990-1003. doi:10.1017/S0003598X00048201.

Nesbitt, M. 2001. Wheat evolution: integrating archaeological and biological evidence. The Linnean 3: 37-59.

Nesbitt, M. 2002. When and where did domesticated cereals first occur in southwest Asia.

Nesbitt, M. and D. Samuel. 1996. From staple crop to extinction? The archaeology and history of the hulled wheats.

Neuwirth, E. 2014. RColorBrewer: ColorBrewer Palettes.

O'Farrell, P.H. 1975. High Resolution Two-Dimensional Electrophoresis of Proteins. The Journal of biological chemistry 250: 4007-4021.

Osborne, T. 1907. Proteins of the wheat kernel, by Thomas B. Osborne / Osborne, Thomas B. (Thomas Burr)S.l. : s.n., S.l.].

Osborne, T.B. and C.G. Voorhees. 1893. The Proteids of the Wheat Kernel. Amer. Chem. Journal 15: 392-471.

Østergaard, O., S. Melchior, P. Roepstorff and B. Svensson. 2002. Initial proteome analysis of mature barley seeds and malt. Proteomics 2: 733-739. doi:10.1002/1615-9861(200206)2:6<733::AID-PROT733>3.0.CO;2-E.

Pagès, H., M. Carlson, S. Falcon and N. Li. 2017. AnnotationDbi: Annotation Database Interface.

Payne, P.I. 1987. Genetics of Wheat Storage Proteins and the Effect of Allelic Variation on Bread-Making Quality. Annu. Rev. Plant Physiol. p. 141-153.

Peleg, Z. 2011. Genetic analysis of wheat domestication and evolution under domestication. Journal of experimental botany 62: 5051-5061. doi:10.1093/jxb/err206.

Peng, J. 2003. Domestication quantitative trait loci in Triticum dicoccoides, the progenitor of wheat. Proceedings of the National Academy of Sciences - PNAS 100: 2489-2494. doi:10.1073/pnas.252763199.

Peng, J.H. and J. Peng. 2011. Domestication evolution, genetics and genomics in wheat. Molecular breeding 28: 281-301. doi:10.1007/s11032-011-9608-4.

Peng, Z., M. Wang, F. Li, H. Lv, C. Li and G. Xia. 2009. A Proteomic Study of the Response to Salinity and Drought Stress in an Introgression Strain of Bread Wheat. Molecular & Cellular Proteomics 8: 2676-2686. doi:10.1074/mcp.M900052-MCP200.

Piperno, D.R., E. Weiss, I. Holst and D. Nadel. 2004. Processing of wild cereal grains in the Upper Palaeolithic revealed by starch grain analysis. Nature 430: 670. doi:10.1038/nature02734.

Plessis, A. 2013. Association study of wheat grain protein composition reveals that gliadin and glutenin composition are trans-regulated by different chromosome regions. Journal of experimental botany 64: 3627-3644. doi:10.1093/jxb/ert188.

Pomeranz, Y. 1965. Dispersibility of wheat proteins in aqueous urea solutions—a new parameter to evaluate bread-making potentialities of wheat flours. Journal of the science of food and agriculture 16: 586-593. doi:10.1002/jsfa.2740161002.

Pompa, M., M.M. Giuliani, C. Palermo, F. Agriesti, D. Centonze and Z. Flagella. 2013. Comparative Analysis of Gluten Proteins in Three Durum Wheat Cultivars by a Proteomic Approach. Journal of Agricultural and Food Chemistry 61: 2606-2617. doi:10.1021/jf304566d.

Pradhan, G.P., P.V.V. Prasad, A.K. Fritz, M.B. Kirkham and B.S. Gill. 2012. Effects of drought and high temperature stress on synthetic hexaploid wheat. Functional Plant Biology 39: 190-198. doi:http://dx.doi.org/10.1071/FP11245.

Ramankutty, N. and J.A. Foley. 1999. Estimating historical changes in global land cover: Croplands from 1700 to 1992. Global Biogeochemical Cycles 13: 997-1027. doi:10.1029/1999GB900046.

Rangan, P., A. Furtado and R.J. Henry. 2017. The transcriptome of the developing grain: a resource for understanding seed development and the molecular control of the functional and nutritional properties of wheat. BMC genomics 18: 766-766. doi:10.1186/s12864-017-4154-z.

Rathmell, W. 2000. Proteome Approach to the Characterisation of Protein Composition in the Developing and Mature Wheat-grain Endosperm. Journal of Cereal Science 32: 169-188.

Raymond, S. and L. Weintraub. 1959. Acrylamide Gel as a Supporting Medium for Zone Electrophoresis. Science 130: 711-711.

Reyes, F.C., M.S. Otegui, T. Chung, R. Vierstra, D. Holding and R. Jung. 2011. Delivery of prolamins to the protein storage vacuole in maize aleurone cells. Plant Cell 23: 769-784. doi:10.1105/tpc.110.082156.

Rollins, J.A., E. Habte, S.E. Templer, T. Colby, J. Schmidt and M. von Korff. 2013. Leaf proteome alterations in the context of physiological and morphological responses to drought and heat stress in barley (Hordeum vulgare L.). Journal of Experimental Botany 64: 3201-3212. doi:10.1093/jxb/ert158.

Salamini, F., H. Ozkan, A. Brandolini, R. Schafer-Pregl and W. Martin. 2002. Genetics and geography of wild cereal domestication in the near east. Nat Rev Genet 3: 429-441.

Saravanan, R.S. and J.K.C. Rose. 2004. A critical evaluation of sample extraction techniques for enhanced proteomic analysis of recalcitrant plant tissues. Proteomics 4: 2522-2532. doi:10.1002/pmic.200300789.

Schuster, A.M. and E. Davies. 1983. Ribonucleic Acid and Protein Metabolism in Pea Epicotyls: I. The Aging Process. Plant Physiology 73: 809-816. doi:10.1104/pp.73.3.809.

Sheoran, I.S., A.R.S. Ross, D.J.H. Olson and V.K. Sawhney. 2009. Compatibility of plant protein extraction methods with mass spectrometry for proteome analysis. Plant Science 176: 99-104. doi:http://dx.doi.org/10.1016/j.plantsci.2008.09.015.

Shevchenko, A., H. Tomas, J. Havli, J.V. Olsen and M. Mann. 2007. In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nature Protocols 1: 2856. doi:10.1038/nprot.2006.468.

Shewry, P.R., N.G. Halford and A.S. Tatham. 1992. High molecular weight subunits of wheat glutenin. Journal of Cereal Science 15: 105-120. doi:http://dx.doi.org/10.1016/S0733-5210(09)80062-3.

Singh, N.K., G.R. Donovan, I.L. Batey and F. MacRitchie. 1990. Use of Sonication and Size-Exclusion High-Performance Liquid Chromatography in the Study of Wheat Flour Proteins. I. Dissolution of Total Proteins in the Absence of Reducing Agents. Cereal Chemistry 67: 150-161.

Singh, N.K. and F. MacRitchie. 1989. Controlled degradation as a tool for probing wheat protein structure. Proceedings from the International Cereal Chemistry 1989 Symposium. Lahti, Finland. p. 321-326.

Skylas, D.J., L. Copeland, W.G. Rathmell and C.W. Wrigley. 2001. The wheat-grain proteome as a basis for more efficient cultivar identification. Proteomics 1: 1542-1546. doi:10.1002/1615-9861(200111)1:12<1542::AID-PROT1542>3.0.CO;2-K.

Skylas, D.J., J.A. Mackintosh, S.J. Cordwell, D.J. Basseal, B.J. Walsh, J. Harry, et al. 2000. Proteome Approach to the Characterisation of Protein Composition in the Developing and Mature Wheat-grain Endosperm. Journal of Cereal Science 32: 169-188. doi:http://dx.doi.org/10.1006/jcrs.2000.0321.

Song, Y., H. Zhang, G. Wang and Z. Shen. 2012. DMSO, an Organic Cleanup Solvent for TCA/Acetone-Precipitated Proteins, Improves 2-DE Protein Analysis of Rice Roots. Plant Molecular Biology Reporter 30: 1204-1209. doi:10.1007/s11105-012-0442-6.

Soreng, R.J., P.M. Peterson, K. Romaschenko, G. Davidse, F.O. Zuloaga, E.J. Judziewicz, et al. 2015. A worldwide phylogenetic classification of the Poaceae (Gramineae). Journal of Systematics and Evolution 53: 117-137. doi:10.1111/jse.12150.

Stotz, H.U., B. Spence and Y. Wang. 2009. A defensin from tomato with dual function in defense and development. Plant Mol Biol 71: 131-143. doi:10.1007/s11103-009-9512-z.

Sultan, A., B. Andersen, B. Svensson and C. Finnie. 2016. Exploring the Plant–Microbe Interface by Profiling the Surface-Associated Proteins of Barley Grains. Journal of proteome research 15: 1151-1167. doi:10.1021/acs.jproteome.5b01042.

Summers, D.F., J.V. Maizel and J.E. Darnell. 1965. Evidence for Virus-Specific Noncapsid Proteins in Poliovirus-Infected HeLa Cells. Proceedings of the National Academy of Sciences of the United States of America 54: 505-513.

Sun, D., N. Wang and L. Li. 2014. In-Gel Microwave-Assisted Acid Hydrolysis of Proteins Combined with Liquid Chromatography Tandem Mass Spectrometry for Mapping Protein Sequences. Analytical Chemistry 86: 600-607. doi:10.1021/ac402802a.

Taddei, G. 1819. Ricerche sul glutine del frumento. Giornale di Fisica, Chimica e Storia Naturale, Brugnatelli 2: 360-361.

Tanno, K. and G. Willcox. 2006. How Fast Was Wild Wheat Domesticated? Science (New York, N.Y.) 311: 1886-1886. doi:10.1126/science.1124635.

Trapnell, C., B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 28: 511-515.

Trümper, C., K. Paffenholz, I. Smit, P. Kössler, P. Karlovsky, H.-P. Braun, et al. 2016. Identification of regulated proteins in naked barley grains (Hordeum vulgare nudum) after

Fusarium graminearum infection at different grain ripening stages. Journal of Proteomics 133: 86-92. doi:10.1016/j.jprot.2015.11.015.

Uthayakumaran, S., P.W. Gras, F.L. Stoddard and F. Bekes. 1999. Effect of Varying Protein Content and Glutenin-to-Gliadin Ratio on the Functional Properties of Wheat Dough. Cereal Chemistry Journal 76: 389-394. doi:10.1094/CCHEM.1999.76.3.389.

Vensel, W.H., L. Harden, C.K. Tanaka, W.J. Hurkman and W.F. Haddon. 2002. Identification of wheat endosperm proteins by MALDI mass spectrometry and LC-MS/MS. Journal of biomolecular techniques : JBT 13: 95.

Vincent, D., M.D. Wheatley and G.R. Cramer. 2006. Optimization of protein extraction and solubilization for mature grape berry clusters. Electrophoresis 27: 1853-1865. doi:10.1002/elps.200500698.

Wang, N., X. Wu, L. Ku, Y. Chen and W. Wang. 2016. Evaluation of Three Protein-Extraction Methods for Proteome Analysis of Maize Leaf Midrib, a Compound Tissue Rich in Sclerenchyma Cells. Frontiers in Plant Science 7: 856. doi:10.3389/fpls.2016.00856.

Wang, W., M. Scali, R. Vignani, A. Spadafora, E. Sensi, S. Mazzuca, et al. 2003. Protein extraction for two-dimensional electrophoresis from olive leaf, a plant tissue containing high levels of interfering compounds. Electrophoresis 24: 2369-2375. doi:10.1002/elps.200305500.

Wang, W., F. Tai and S. Chen. 2008. Optimizing protein extraction from plant tissues for enhanced proteomics analysis. Journal of Separation Science 31: 2032-2039. doi:10.1002/jssc.200800087.

Wang, W., R. Vignani, M. Scali and M. Cresti. 2006. A universal and rapid protocol for protein extraction from recalcitrant plant tissues for proteomic analysis. Electrophoresis 27: 2782-2786. doi:10.1002/elps.200500722.

Wang, X., Z.-F. Yuan, J. Fan, K.R. Karch, L.E. Ball, J.M. Denu, et al. 2016. A Novel Quantitative Mass Spectrometry Platform for Determining Protein O-GlcNAcylation Dynamics. Molecular & Cellular Proteomics 15: 2462-2475. doi:10.1074/mcp.O115.049627.

Weiss, E. and D. Zohary. 2011. The Neolithic Southwest Asian Founder Crops
Their Biology and Archaeobotany. Current Anthropology 52: S237-S254. doi:10.1086/658367.

Wellman, C.H., P.L. Osterloff and U. Mohiuddin. 2003. Fragments of the earliest land plants. Nature 425: 282. doi:10.1038/nature01884.

Wendelboe-Nelson, C. and P.C. Morris. 2012. Proteins linked to drought tolerance revealed by DIGE analysis of drought resistant and susceptible barley varieties. Proteomics 12: 3374-3385. doi:10.1002/pmic.201200154.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data AnalysisSpringer-Verlag New York.

Wickham, H., R. François, L. Henry and K. Müller. 2017. dplyr: A Grammar of Data Manipulation.

Wickham, H., J. Hester and R. Francois. 2017. readr: Read Rectangular Text Data.

Wieser, H., W. Bushuk and F. MacRitchie. 2006. Chapter 7 The Polymeric Glutenins. Gliadin and Glutenin: The Unique Balance of Wheat Quality. AACC International, Inc. p. 213-240.

Willcox, G. and G. Willcox. 2008. Early Holocene cultivation before domestication in northern Syria. Vegetation history and archaeobotany 17: 313-325. doi:10.1007/s00334-007-0121-y.

Woychik, J.H., J.A. Boundy and R.J. Dimler. 1961. Starch gel electrophoresis of wheat gluten proteins with concentrated urea. Archives of Biochemistry and Biophysics 94: 477-482. doi:http://dx.doi.org/10.1016/0003-9861(61)90075-3.

Wrigley, C.W.B., F.; and Bashuk, W. 2006. Chapter 1. Gluten: A Balance of Gliadin and Glutenin. In: C. B. Wrigley, W., editor Gliadin and Glutenin: The Unique Balance of Wheat Quality. AACC International University of Manitoba, Winnipeg, MB, Canada. p. 3-32.

Wu, Y., M. Mirzaei, D. Pascovici, J.M. Chick, B.J. Atwell and P.A. Haynes. 2016. Quantitative proteomic analysis of two different rice varieties reveals that drought tolerance is correlated with reduced abundance of photosynthetic machinery and increased abundance of ClpD1 protease. J Proteomics 143: 73-82. doi:10.1016/j.jprot.2016.05.014.

Xiang, X., S. Ning, X. Jiang, X. Gong, R. Zhu, L. Zhu, et al. 2010. Protein extraction from rice (Oryza sativa L.) root for two-dimensional electrophresis. Frontiers of Agriculture in China 4: 416-421. doi:10.1007/s11703-010-1031-9.

You, L. and S. Wood. 2005. Assessing the spatial distribution of crop areas using a cross-entropy method. International Journal of Applied Earth Observation and Geoinformation 7: 310-323. doi:https://doi.org/10.1016/j.jag.2005.06.010.

Zhang, S., L.-L. Zhang, K.-K. Zhou, Y.-J. Liu and Z. Zhao. 2015. Evaluation of three types of protein extraction methods for tetraploid black locust (Robinia pseudoacacia L.) phloem tissue proteome analysis by two-dimensional electrophoresis. Analytical Methods 7: 1008-1017. doi:10.1039/C4AY02038C.

Zhang, Y., B.R. Fonslow, B. Shan, M.-C. Baek and J.R. Yates. 2013. Protein Analysis by Shotgun/Bottom-up Proteomics. Chemical reviews 113: 2343-2394. doi:10.1021/cr3003533.

Zhen, Y. and J. Shi. 2011. Evaluation of sample extraction methods for proteomic analysis of coniferous seeds. Acta Physiol Plant 33: 1623-1630. doi:10.1007/s11738-010-0697-1.

Zhou, J., C. Ma, S. Zhen, M. Cao, F.J. Zeller, S.L.K. Hsam, et al. 2016. Identification of drought stress related proteins from 1S(l)(1B) chromosome substitution line of wheat variety Chinese Spring. Botanical studies 57: 20-20. doi:10.1186/s40529-016-0134-x.

Zhu, K. and K. Khan. 2002. Quantitative variation of HMW glutenin subunits from hard red spring wheats grown in different environments. Cereal Chemistry 79: 783-786.

Žilić, S., M. Barać, M. Pešić, D. Dodig and D. Ignjatović-Micić. 2011. Characterization of Proteins from Grain of Different Bread and Durum Wheat Genotypes. International Journal of Molecular Sciences 12: 5878-5894. doi:10.3390/ijms12095878.