VALUE ADDED WHEAT THROUGH APPLIED GENOMIC PREDICTION: A GENOMIC APPROACH FOR BREEDING LOW GLUTEN EPITOPE WHEAT

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Plant Breeding and Genetics

at Massey University, Palmerston North,

New Zealand

Jamie Macalister

2023

Abstract

Gluten epitopes are known to trigger coeliac disease (CD) in affected consumers and are believed to be linked to some cases of gluten intolerance. Research suggests that if consumers were exposed to wheat with reduced concentrations of gluten epitopes, the incidence of CD and gluten intolerance may be reduced. Methods have recently been developed allowing researchers to measure gluten epitope concentrations in wheat. This offers wheat breeders the potential to select towards varieties with lower epitope concentrations than existing cultivars. However, the methods for measuring epitope concentrations remain costly and time consuming. Therefore, it is proposed that a genomic based approach for breeding low epitope wheat lines is a more practical method than traditional phenotype-based selections. The genetic factors associated with epitope concentrations remain poorly understood. In this thesis, heritability estimates of between 0.37-0.93 are reported for concentrations of 6 distinct gluten epitopes. The associations between epitope concentrations and baking quality are also assessed and are shown to range from being near zero for some epitopes to strong positive correlations between other epitopes and particular baking quality characteristics. A Genome Wide Association Study and a model for genomic prediction are employed to determine the genetic factors associated with epitope concentrations. In these analyses, 3 significant genomic windows are identified as being associated with concentrations of 3 particular epitopes. Empirical prediction accuracies of between 0.16-0.53 are observed for predictions of epitope concentrations in a breeding population. Additionally, accuracies of between 0.37-0.67 are achieved by adjusting the population structure to represent the ideal circumstances that breeders would aim to achieve in their training and target populations. These results demonstrate that genomic selection (GS) will be an effective method for breeding low gluten epitope wheat. The outcome of this thesis will allow implementation of GS in the New Zealand Institute for Plant & Food Research wheat breeding program where epitope concentrations will be established as a new breeding target. This is expected to lead to the release of niche, low epitope cultivars with a value-add component that benefits growers, industry and consumers.

Acknowledgements

This research would not have been possible without the support of a number of incredible organisations. Thank you to The New Zealand Institute for Plant & Food Research (PFR) for their hands on support of my studies over the past three years and for enabling me to focus full time on my research. Thanks also to The Agricultural and Marketing Research and Development Trust (AGMARDT), who made a substantial contribution to resourcing this project and whose support of this research has been essential in enabling industry collaboration. This work has also received the backing of a number of industry partners, including the Baking Industry Research Trust (BIRT), the Foundation for Arable Research (FAR) and the Milling Industry Research Trust, I am deeply grateful for the support of each of these organisations and believe that these collaborations will continue to be essential to making significant improvements across the wheat growing, milling and baking industries in the future.

Thank you to my supervisory panel, consisting of Prof. Dorian Garrick, Dr Rhiannon Handcock and Dr Paul Johnston. Thank you all for your guidance and regular constructive discussions, I have learnt a considerable amount over the past three years and have greatly appreciated your ongoing support which has been essential to my research. Your patience as I come to terms with new concepts or am painstakingly troubleshooting code has been greatly appreciated. Your comments and suggestions on each of my chapters have also been incredibly valuable and have ensured that my research is always of highest possible quality.

Thank you to my fellow colleagues at PFR. There has been an endless list of fellow researchers, technicians and team leaders who have assisted in all areas of my PhD, from preparing the field for my trials to assisting in the lab and even helping write regular reports and funding applications. So, thank you again to, Doug Bowron, Ben & Janelle Browne, Samantha Baldwin, Antonia Miller, Zac Hanley, Andy Hay, Vanessa Winchester, Samantha Gibbs, Sarah Roberts, Nigel Joyce, Cath Wade and many more. None of this work would have been possible without the help of each and every one of you.

I would also like to acknowledge the governance group who have helped oversee this work, consisting of Tania Watson, Ralph Thorogood, Ivan Lawrie, Andy Worrill, Alison Stewart, Zac Hanley, Antonia Miller and Paul Johnston. Your support of this project has been greatly appreciated and has ensured that my work remains relevant and as practically useful as possible to the wheat growing, milling and baking industries.

To my family and friends who have continued to support me throughout my studies, Mum, Dad, Scott, Andrew, Nikki, I could never put into words how valuable your support has been and always will be. Thank you for always being there, through the good times and the bad and thank you for reminding me that sometimes, there is more to life than study.

Table of Contents

Chapter 1

General introduction

1.1 Introduction

Wheat is one of the most widely consumed food products globally and in many countries is considered to be the primary source of carbohydrates in the average diet (Shewry & Hey, 2015). Wheat is also a major source of protein, fibre and other essential vitamins and minerals (Shewry & Hey, 2015). Due to its versatility as a crop, wheat can be grown in many different climates making it one of the most widely grown crops globally (Igrejas & Branlard, 2020). Possibly the most unique characteristic of wheat, is the presence of gluten proteins. Gluten is most widely known for providing strong, elastic dough for bread making but is also now commonly used as a flavour enhancer and stabilising agent in many processed foods (Biesiekierski, 2017).

Gluten epitopes are small, specific peptides that primarily exist within the gliadin fraction of wheat gluten (Ciccocioppo et al., 2005). Highly enriched in glutamine and proline amino acids the gluten epitopes resist digestion by human proteases and persist into the intestinal tract where they can trigger coeliac disease (CD) in affected consumers and may also be linked to symptoms of gluten intolerance in non-coeliacs (King et al., 2020). Coeliac disease is an autoimmune disorder that occurs in genetically susceptible individuals who carry at least one copy of either the DQ2 or DQ8 haplotypes for Human Leukocyte Antigens(Romanos et al., 2009). In coeliac consumers, gluten epitopes are identified as antigens and are presented to gluten-sensitive T-cell lymphocytes (van den Broeck et al., 2010). This triggers gut inflammation and degradation of villi in the small intestine leading to symptoms such as diarrhoea, fatigue, weight loss, bloating and anaemia (van den Broeck et al., 2010). Currently, the only treatment for CD is a lifelong gluten free diet. However, evidence suggests that reduced exposure to gluten epitopes in the diet may reduce the incidence of CD and severity of gluten intolerance (Molberg et al., 2005).

Wheat has a relatively large (c.17 Gb) and complex genome (Uauy, 2017). Modern bread wheat (*T. aestivum*) is an allohexaploid species, consisting of three sub genomes (A, B and D) that originated from different ancestral species (Uauy, 2017). The wheat genome also contains many repetitive, overlapping sequences and over 80% of it is made up of transposable elements (Wicker et al., 2018). These characteristics mean that the wheat genome remains poorly mapped compared to many other valuable plant and animal species. However, as technology progresses, the genetic make-up of wheat is becoming better understood. This combined with the increasing efficiencies being made with regards to high-throughput genotyping technologies, the application of genomic tools in wheat breeding is becoming increasingly practical and effective. Due to the vital role wheat plays in the modern diet, any advancements in wheat breeding capabilities offer substantial benefits to the health and wellbeing of consumers around the world.

This study aims to utilise genotype information from Single Nucleotide Polymorphism (SNP) data and apply Bayesian techniques to understand the genetic regulation of 6 specific gluten epitopes. Using these results, the objective of this research is to develop a model for genomic prediction that is effective in selecting towards wheat lines with low epitope phenotypes. It is hoped that this will act as an additional tool for breeders to reliably breed for low epitope cultivars in commercial breeding programs whilst continuing to breed for other important traits such as agronomic performance and baking quality. Low epitope wheat cultivars are expected to provide benefits across the entire supply chain. These range from farmers who can expect to receive a premium for growing low epitope cultivars, to the bakeries who will be able to develop niche food products. Benefits are also expected for the final consumer with the potential for delaying or preventing the onset of CD in susceptible consumer and reduced symptoms of gluten intolerance in others.

1.2 Thesis objectives and outline

The main objective of this thesis was to develop a genomic prediction model for identifying and breeding wheat lines with reduced concentrations of particular gluten epitopes. A literature review is presented in Chapter 2 to give an overview of coeliac disease and gluten intolerance as well as discuss the functions of gluten and the unique genetic characteristics of wheat. Chapter 2 concludes with a discussion around different potential strategies for developing low gluten epitope wheat products and provides justification for a genomic based breeding approach. Little is understood about the complex genetic factors that regulate gluten epitope concentrations in wheat and therefore many of the chapters in this thesis are dedicated to better understanding these. Chapter 3 investigates the heritability and genotype x environment characteristics of gluten epitope regulation to determine the likely effectiveness of a breeding approach in reducing epitope concentrations in wheat. This is followed up in Chapter 4 with an analysis of the associations between gluten epitope concentrations and a number of baking quality traits in wheat. This chapter aims to determine whether it is possible to breed for low gluten epitope phenotypes whilst maintaining acceptable baking quality characteristics in their cultivars. Chapter 5 consists of a Genome Wide Association Study to identify significant loci that are associated with gluten epitope concentrations. This is then followed up in Chapter 6 with an analysis of the prediction accuracies of the genomic model to predict epitope concentrations in validation populations under three different scenarios. In this chapter, a number of scenarios are tested to determine the criteria needed to obtain useful prediction accuracies. The practical implications of our results, along with the limitations, are also discussed. Chapter 7 is written to bring together all of the findings presented in the preceding chapters. This chapter provides a general discussion and considers how the results of this study may be applicable in advancing the ability of wheat breeders to accurately select for complex quantitative traits.

1.3 PhD supervisors

Professor Dorian Garrick

Dr Rhiannon Handcock

Dr Paul Johnston

1.4 References

Biesiekierski, J. R. (2017). What is gluten? *Journal of Gastroenterology and Hepatology*, *32 Suppl 1*, 78-81. https://doi.org/10.1111/jgh.13703

Ciccocioppo, R., Di Sabatino, A., & Corazza, G. R. (2005). The immune recognition of gluten in coeliac disease. *Clinical & Experimental Immunology*, *140*(3), 408-416. https://doi.org/https://doi.org/10.1111/j.1365-2249.2005.02783.x

Igrejas, G., & Branlard, G. (2020). The Importance of Wheat. In G. Igrejas, T. M. Ikeda, & C. Guzmán (Eds.), *Wheat Quality For Improving Processing And Human Health* (pp. 1-7). Springer International Publishing. https://doi.org/10.1007/978-3-030-34163-3_1

King, J. A., Jeong, J., Underwood, F. E., Quan, J., Panaccione, N., Windsor, J. W., . . . Kaplan, G. G. (2020). Incidence of Celiac Disease Is Increasing Over Time: A Systematic Review and Meta-analysis. *Official journal of the American College of Gastroenterology | ACG*, *115*(4). https://journals.lww.com/ajg/Fulltext/2020/04000/Incidence_of_Celiac_Disease_Is_Increasing_Over.9.aspx

Molberg, Ø., Uhlen, A. K., Jensen, T., Flæte, N. S., Fleckenstein, B., Arentz–Hansen, H., . . . Sollid, L. M. (2005). Mapping of gluten T-cell epitopes in the bread wheat ancestors: Implications for celiac disease. *Gastroenterology*, *128*(2), 393-401. https://doi.org/https://doi.org/10.1053/j.gastro.2004.11.003

Romanos, J., van Diemen, C. C., Nolte, I. M., Trynka, G., Zhernakova, A., Fu, J., . . . Wijmenga, C. (2009). Analysis of HLA and Non-HLA Alleles Can Identify Individuals at High Risk for Celiac Disease. *Gastroenterology*, *137*(3), 834-840.e833. https://doi.org/https://doi.org/10.1053/j.gastro.2009.05.040

Shewry, P. R., & Hey, S. J. (2015). The contribution of wheat to human diet and health [https://doi.org/10.1002/fes3.64]. *Food and Energy Security*, *4*(3), 178-202. https://doi.org/https://doi.org/10.1002/fes3.64

Uauy, C. (2017). Wheat genomics comes of age. *Current Opinion in Plant Biology*, *36*, 142-148. https://doi.org/https://doi.org/10.1016/j.pbi.2017.01.007

van den Broeck, H. C., de Jong, H. C., Salentijn, E. M. J., Dekking, L., Bosch, D., Hamer, R. J., . . . Smulders, M. J. M. (2010). Presence of celiac disease epitopes in modern and old hexaploid wheat varieties: wheat breeding may have contributed to increased prevalence of celiac disease. *Theoretical and Applied Genetics*, *121*(8), 1527-1539. https://doi.org/10.1007/s00122-010-1408-4

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R. H., . . . International Wheat Genome Sequencing, C. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology*, *19*(1), 103. https://doi.org/10.1186/s13059-018-1479-0

Chapter 2

Gluten epitopes, coeliac disease and the role of wheat breeding: A review

2.1 Introduction

Wheat is a major staple of the human diet worldwide and based on production data, is the third most important crop globally, behind rice and maize (Shewry & Hey, 2015). Wheat has long been an essential crop for humans. It can be cultivated in a vast range of environments (Feldman et al., 1995) and is extremely versatile in its uses, due in particular to the functional properties of gluten (Shewry & Hey, 2015). Furthermore, wheat provides a substantial proportion of many vital nutrients to the human diet, including protein, vitamins and dietary fibre.

As important as wheat is to the general population, it does cause significant health issues for some. Coeliac disease is an autoimmune disease caused by gluten consumption where affected individuals develop a permanent intolerance to dietary gluten. The disease affects around 1% of the general population, making it the most common food sensitive enteropathy in humans (Vader et al., 2002). There is also strong evidence that the incidence of coeliac disease is increasing globally which appears to be linked to increased gluten consumption per capita (Scherf et al., 2020).

This review considers a number of questions that relate to understanding the biochemical compounds that cause coeliac disease, the manner in which these compounds influence bread baking quality traits and the genomic controls within wheat that regulate the production of these compounds. This leads to a description of tools that could be applied to reduce or eliminate these compounds from wheat or from wheat-based products. The review sets the scene for investigating strategies to develop new wheat varieties that reduce health impacts for susceptible consumers without compromising baking quality.

2.2 Triggers of coeliac disease & wheat sensitivity

Coeliac disease is a condition where affected individuals present an immunogenic response to dietary gluten. In response to consuming gluten, patients with coeliac disease will become very sick. The onset of coeliac disease can occur at any age and will often be linked to an environmental factor such as a period of unrelated sickness, a surgical procedure or even pregnancy (Scherf et al., 2020). Typical symptoms of the disease include pain in the abdomen, diarrhoea, indigestion, nausea, vomiting, fatigue and skin rash. There is no known cure for coeliac disease and the only treatment is a lifelong gluten free diet. These autoimmune responses are triggered in coeliac patients when certain gluten peptides are absorbed into the intestine and identified by specific dendritic cells. These cells contain HLA-DQ surface receptor proteins which can sometimes comprise two known HLA-DQ haplotypes (HLA-DQ2 & HLA-DQ8) that will identify gluten peptides as antigens. At least one of these two serotypes appear to be carried by all coeliac patients and can therefore be used to identify consumers with a susceptibility to the disease (Nasr et al., 2016). Once identified as antigens, the gluten peptides are presented to CD4+ T-cells which in turn initiate an immune

response. This immunogenic reaction will then result in significant damage to the intestinal walls, significantly reducing the patient's ability to absorb essential nutrients (Bonini, 2012; Peña & Wijmenga, 2001).

Numerous research groups have investigated the structure and composition of gluten and it has been found that there are a number of specific peptides within gluten that appear to cause coeliac disease, these have been termed gluten epitopes (Martucci & Corazza, 2002). Gluten comprises two main fractions, these being glutenins and gliadins. Originally, these two fractions were classified as either soluble (gliadins) or insoluble (glutenins) in aqueous alcohol (Howdle et al., 1984). However, the two fractions are now more accurately classified by their quaternary amino acid structure as either monomeric (gliadins) or covalent aggregates (glutenins); (Howdle, 2006). It was originally believed that gliadins were solely responsible for coeliac disease while glutenins were non-toxic and more important for baking quality (Van De Kamer et al., 1953). However, further studies have determined that this is not the case. Gliadins do appear to play a more significant role in causing coeliac disease but glutenins have also been shown to be toxic to susceptible consumers (Vader et al., 2002). Both gliadins and glutenins can be broken down into a number of smaller subunits. Gliadin subunits are described as either α-gliadins, γ-gliadins or ω-gliadins; while glutenins are typically divided into low molecular weight (LWM) or high molecular weight (HMW) subunits. Further, gluten subunits can be categorized into S-rich, S-poor or HMW proteins. The ω-gliadins are described as S-poor and the 3 subunits, α-gliadins, γ-gliadins and LMW glutenins are described as S-rich (Figure 2.1); (Balakireva & Zamyatnin, 2016; Shewry et al., 1986). Of particular interest, is the structure of the S-rich proteins which often consist of many repetitive sequences. These proteins are particularly rich in glutamine and proline and contain a high degree of homology causing many of them to be identified as antigens in affected consumers (Howdle, 2006).

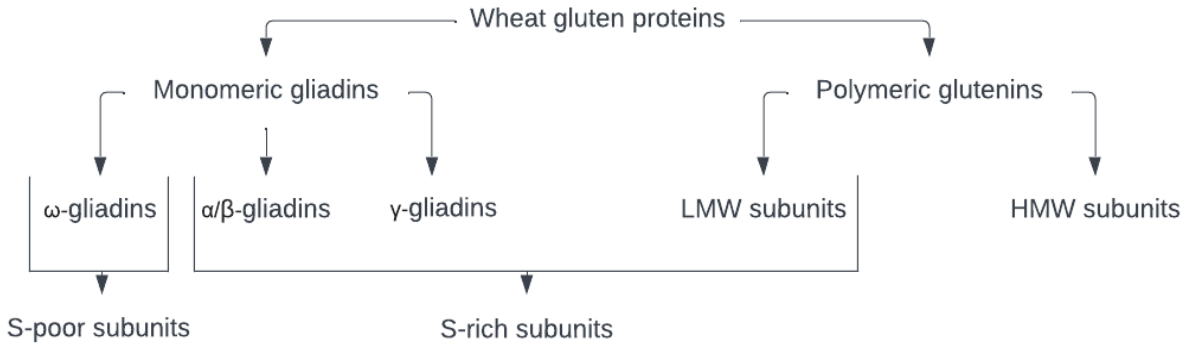

Figure 2.1: Classification of gluten proteins (LMW = Low Molecular Weight, HMW = High Molecular weight)

Once absorbed into the intestine, gluten peptides will be recognised as antigens in patients with either the HLA-DQ2 or HLA-DQ8 haplotypes triggering an immune response which in turn leads to tissue damage in

the gut (Figure 2.2); (Sollid, 2002). This damage prompts the release of transglutaminase (tTG) enzymes which modify the gluten peptides and increase their affinity to antigen presenting cells. This in turn creates a positive feedback loop causing greater intestinal damage. Transglutaminase enzymes can also be released in response to an unrelated gut infection or even in response to a recent surgery and therefore, such events can be a trigger for the onset of coeliac disease in susceptible consumers (Sollid, 2002). The preferred substrate for tTG modification is characterised by repetitive glutamine and proline residues, such as those often found in the S-rich gluten subunits. The tTG modification proceeds as either deamidation or transamidation and it was originally believed that this modification was a requirement for an immunogenic response. However, more recently, certain gluten peptides have been identified that will trigger an immune response without tTG modification (Vader et al., 2002).

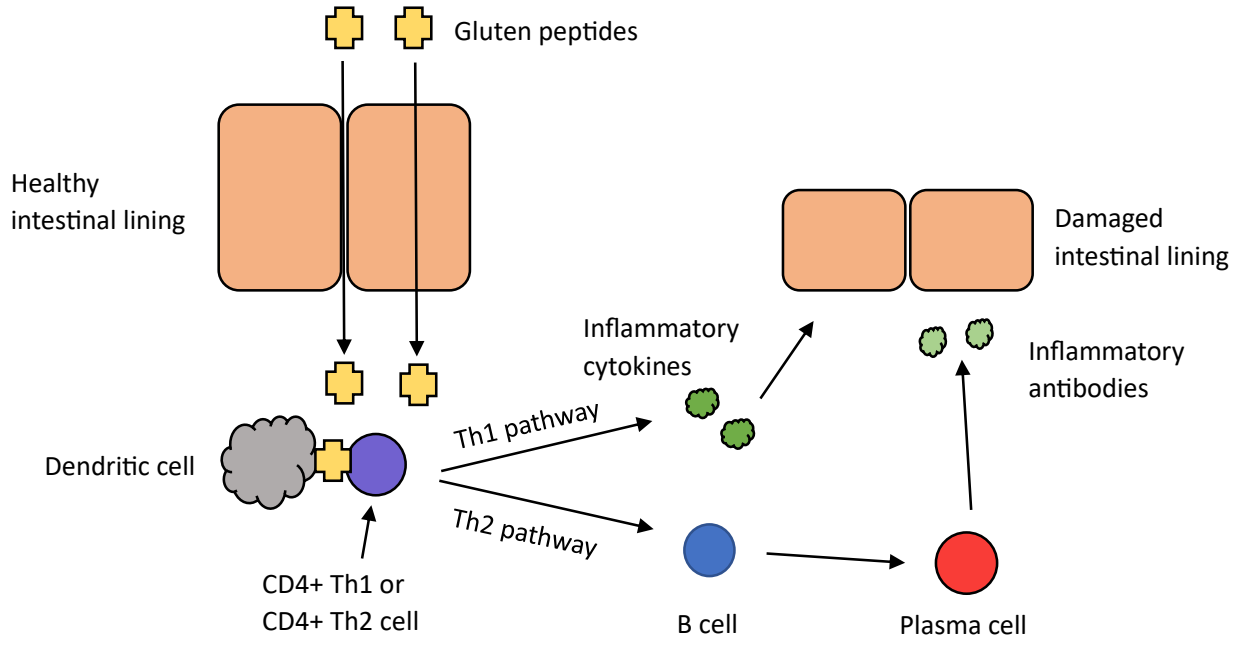

Figure 2.2: Inflammatory pathway of coeliac peptides in the intestinal lining. Gluten peptides are absorbed through the cell wall. Peptides are identified by dendritic cells and presented to either CD4+ Th1 or CD4+ Th2 cells. CD4+ Th1 cells trigger the production of inflammatory cytokines which damage the intestinal lining. CD4+ Th2 cells trigger the activation of B cells which become plasma cells producing inflammatory antibodies which also damage the intestinal lining. Adapted from "Coeliac disease: dissecting a complex inflammatory disorder", by L. M. Sollid, 2002, Nature reviews. Immunology, 2, 647-655.

Coeliac disease is not the sole cause of dietary complications resulting from wheat consumption. Wheat allergy is another condition that can occur in response to consuming wheat based products or in some cases can be due to inhaling wheat flour ("baker's asthma"); (Baur et al., 1998). The mechanisms of many wheat allergens are poorly understood and can be highly dependent on genetic and environmental factors. Symptoms of wheat allergy include, hives, anaphylaxis and irritation of the mouth and throat. It is believed

that there are a range of compounds in wheat products that can cause allergies, however one specific gliadin protein (ω-5 gliadin) has been identified as a significant allergen, particularly in children (Inomata, 2009). It is however not clear whether ω-5 gliadin is an important peptide involved in coeliac disease, although, just like with coeliac disease, it is believed that the S-rich gliadin and glutenin subunits are all implicated in wheat allergy (Nasr et al., 2016). Therefore, any work to develop low gluten epitope wheat may also have the added benefit of reducing wheat allergens.

In addition to coeliac disease and wheat allergy, many non-coeliac consumers have also reported symptoms such as bloating, fatigue and nausea in response to consuming wheat based products (Czaja-Bulsa, 2015). This condition is generally termed non-coeliac gluten sensitivity (NCGS). It is broadly diagnosed in non-coeliac consumers who remove gluten from their diet and in response, report that the symptoms have disappeared. It is not clear however that gluten is the actual cause of NCGS and therefore, the condition is now more frequently termed, non-coeliac wheat sensitivity (NCWS). It has been regularly disputed that NCWS should not actually be considered a real medical condition because no reliable biomarkers exist and the exact cause remains unknown (Molina-Infante et al., 2015).

There are however some compounds found in wheat that could be potential triggers for NCWS, these being Amylase Trypsin Inhibitors (ATI's); (Fasano et al., 2015) and fermentable oligo-, di-, mono-saccharides and polyols (FODMAPs); (Barrett & Gibson, 2012). ATI's are a family of proteins found in wheat which comprise around 2-4% of total wheat protein (Dupont et al., 2011). They are important agronomically as they form part of a plants natural defence system against parasites and insects and therefore have the potential to both reduce crop losses and reduce the amount of pesticide/insecticide that needs to be applied to a crop (Priya et al., 2013). Breeders have for many years selected cultivars that demonstrate increased pest resistance and it is believed that for this reason, the presence of ATI's in modern wheats has increased significantly above levels seen in older varieties (Junker et al., 2012). While this may have important benefits for growers, many studies have also shown that these compounds have the ability to trigger inflammation of the gut in human consumers and therefore are likely to be a cause of the symptoms seen in NCWS (Zevallos et al., 2017). The substantial increase in levels of ATI's found in modern wheats may also explain the apparent increase of NCWS seen in the global wheat consuming population. However, because ATIs are not found in gluten, it is also likely that there is no link between ATI's and coeliac disease.

In addition to ATI's, FODMAPs in wheat are also a potential cause of NCWS. These are a group of short-chain carbohydrates that are incompletely absorbed in the human gut. Numerous studies have identified FODMAPs as being a likely trigger for irritable bowel syndrome along with other common symptoms of NCWS (Barrett & Gibson, 2012). FODMAPs are present in many food products but interestingly the levels of FODMAPs in modern wheats tends to be relatively low when compared to a number of other foods. For

example, work from Varney et al. (2017), measuring FODMAP levels in different products shows split peas as having FODMAP levels many multiples higher than those found in white bread and this appears to also be the case for a range of other products that do not contain any wheat. It is therefore believed that the reason FODMAPs in wheat are likely to be an issue for many consumers is because the average daily intake of wheat-based products around the world is relatively high and therefore the overall exposure to FODMAPs through wheat is also high. Being carbohydrates, it is also highly unlikely that there is any link between FODMAPs and coeliac disease.

2.3 Links between gluten epitopes and baking quality characteristics in wheat

Gluten is a unique family of proteins found in wheat and other cereals and is responsible for providing the particular characteristics of dough essential for producing a variety of baked products. The two fractions of wheat gluten (gliadins & glutenins) are each considered responsible for different aspects of dough quality. Gliadins tend to provide the viscosity and extensibility characteristics while glutenins are considered more important for dough strength and elasticity, therefore making the balance of the two components essential to the final characteristics of the dough (Biesiekierski, 2017; Wieser, 2007). A key characteristic of both gliadins and glutenins is their high concentration of proline and glutamine. Glutamine concentrations can range from 26-53% of the total amino acid composition of different gluten proteins while proline concentrations range from 10-29% (Table 2.1); (Scherf et al., 2016). There is evidence that the levels of glutamine and proline in wheat gluten play a significant role in determining the baking and dough characteristics of flour Fermin et al. (2005). However, it has proven difficult to quantify the effects of different glutamine and proline concentrations due to the large range of other genetic and environmental factors that affect baking quality.

Fermin et al. (2005), provides insight into the roles of glutamine and proline by comparing the quality characteristics of two different wheat varieties under a range of flour treatments (no added proline or glutamine, added proline, added glutamine, added proline and glutamine). The two varieties used in the study (‘Sisson 36’ & ‘Karl 92’) varied significantly for proline and glutamine levels (with ‘Karl 92’ being higher in both respects) yet still had similar total protein concentrations. Initial quality work clearly showed that ‘Karl 92’ produced a much stronger, more extensible and resistant dough along with a more cohesive and springy bread with greater loaf volume. However, with the addition of proline and glutamine to the mix, almost all of the dough and bread properties mentioned above improved in ‘Sisson 36’ to match those values seen in ‘Karl 92’ indicating the value of high proline and glutamine levels to the functional characteristics of wheat flour for baking. Furthermore, it is important to note that the addition of only one of either proline or glutamine resulted in very little improvement to dough functionality, and in some cases

a reduction in functionality, suggesting that it is the complementarity of proline and glutamine that is important as opposed to one amino acid having advantages over the other.

Table 2.1: Range of glutamine and proline concentrations of gluten proteins based on their molecular weight (Scherf et al., 2016)

	High Molecular Weight	Medium Molecular Weight	Low Molecular Weight
Glutamine	26-36%	37-53%	28-36%
Proline	10-15%	20-29%	11-22%

Another unique characteristic of gluten proteins is the significant presence of repetitive amino acid sequences. Typically, repetitive domains account for around 30-50% of the protein sequence in S-rich gliadins and LMW subunits and 75-85% of the protein sequence in HMW subunits (Shewry, 2019). There appears to be greater diversity in repetitive sequences across the other gluten classes, however the one defining characteristic of all the repetitive domains in gluten is that they are rich in proline and glutamine (Shewry, 2019). In a review by Belton (1999), it is suggested that it is because of these repetitive sequences that gluten derives its unique solubility characteristics, essential to dough functionality, as it allows protein:protein hydrogen bonds to form resulting in the insolubility of HMW subunits in water. The exact role of different repetitive domains in gluten appears to depend greatly on other characteristics of the differing peptides within gluten, however due to the extent that proline and glutamine repetitive sequences exist, it is likely that they are an essential component for gluten functionality (Wieser, 2007).

Along with being important to the functionality of wheat flour, the high proline and glutamine repetitive sequences of wheat gluten proteins also appear to be a critical component of wheat gluten epitopes. Sollid et al. (2012), compiled a list of all of the known gluten epitopes showing that every epitope existed within a repetitive domain. The high glutamine and proline content within the repetitive domains of gluten means that gastrointestinal proteases are unable to break them down. The result of this is that many gluten proteins remain intact when absorbed in the gut and therefore, as discussed earlier, they remain in a form that can be identified as antigens by HLA-DQ2 or HLA-DQ8 receptors (Janssen et al., 2015). Due to the significant role that glutamine and proline rich repetitive domains play in gluten functionality while also being resistant to intestinal degradation, it is probable that a strong relationship exists between the immunogenicity of gluten epitopes and their value in baking quality. This therefore presents a significant challenge to producing functional wheat-based products with reduced immunogenicity.

2.4 Genetic variation in gluten epitope profiles in modern and ancient wheats

Common bread wheat (*T. aestivum*) is an allohexaploid species consisting of 3 distinct genomes (AABBDD), each with 7 pairs of homologous chromosomes. The A and B genomes come from the tetraploid species, *T. turgidum* (AABB), which itself was formed through intergenic hybridization of *T. urartu* (einkorn) and *Ae. speltoides*, while the D genome of common wheat is derived from Ae*. tauschii* or wild goat grass (Figure 2.3); (Dvořák, 2001; Molberg et al., 2005). When compared to its ancestors, common hexaploid wheat is unique primarily due to its valuable baking properties. It is believed that the D genome of common wheat is the key contributor to these valuable characteristics (Kerber & Tipples, 1969). The D genome contributes numerous significant loci that encode multiple common genes for HMW glutenins, LMW glutenins as well as a number of different gliadin subunits (Kucek et al., 2015). However, due to the apparent linkage between baking quality traits and gluten epitope profiles, the D genome also appears to be a significant contributor of T-cell stimulatory epitopes in common bread wheat. Numerous studies have repeatedly described one specific epitope, termed the "33-mer", as being the most significant at stimulating coeliac disease in affected patients (Comino et al., 2012; Tye-Din et al., 2010). The reason the 33-mer peptide is considered so important is because it contains three overlapping T-cell epitopes making it highly immunogenic to susceptible consumers. Furthermore, the 33-mer contains a high number of proline and glutamine residues also making it extremely resistant to breakdown by intestinal peptidases (Camarca et al., 2009; Lu Shan et al., 2002). The 33-mer peptide is also encoded by the gliadin subunit, α2-gliadin which is believed to exist only within the D genome. This is supported by work from Schalk et al. (2017),who tested 38 different bread wheat varieties as well as 2 spelt cultivars (which also carry the D genome), and detected 33-mer in the flour of all these varieties in concentrations ranging from 91-603ug/g of flour whereas no 33-mer was detected in any durum, emmer or einkorn species (all which lack the D genome).

Although the D genome does appear to be the key contributor to bread making quality and T-cell stimulatory gluten epitopes in common wheat, ancient wheats consisting of the A and B genomes do still contribute to the epitope profiles of common wheat. Work from Molberg et al. (2005), used polyclonal T-cell lines from intestinal biopsy samples of coeliac patients to test the immunogenicity of flour from ancient wheat species (einkorn, emmer and durum) and compared these results to the immunogenicity of Ae*. tauschii* (D genome donor). All of the gluten digests from Ae*. tauschii* were recognised by the T-cell lines whereas the gluten digests from einkorn, emmer and durum were only recognised by 25-38% of the T-cell lines. Work from both Vader et al. (2002) and Vincentini et al. (2009), measuring the T-cell responses in young coeliac patients to gluten from ancient wheats further supports this data but also provides more complexities. In both studies, all of the coeliac patients responded to the gluten profiles of ancient wheats,

however in both cases each patient responded differently to different epitopes and also some only appeared to respond to specific combinations of epitopes. Although there is clearly a significant reduction in gluten immunogenicity of ancient wheat species lacking the D genome, it is important to note that no coeliac patient can consume any gluten from any type of wheat without becoming seriously ill, meaning that even with the significantly reduced epitope profiles of ancient wheats, consuming these is still not a viable option.

Due to the inherent link between gluten epitopes and baking quality in modern wheat, it is fair to assume that over the thousands of years that humans have been selectively breeding bread wheat for superior bread making characteristics, people would also be unintentionally selecting for lines with higher epitope levels. Furthermore, there is increasing evidence that, at least over the past decade, the global incidence and/or awareness of coeliac disease and other gluten related disorders, is increasing. As only a proportion of those people genetically predisposed to coeliac disease ever actually become coeliac (Gujral et al., 2012), it seems possible that the increasing incidence of the disease could result from consumers being exposed to higher levels of epitopes in wheat based products and the increased use of gluten as an additive in non-wheat based products. However, based on existing literature, it is not conclusive whether recent breeding efforts have resulted in the unintentional development of more immunogenic wheats. One study from van den Broeck et al. (2010), measured the prevalence of two different gluten epitopes (Glia-α9 and Glia-α20) in 36 modern commercial European cultivars compared to 50 older landraces. Overall, the presence of the more significant epitope (Glia-α9) was higher in the modern cultivars, while Glia-α20 was lower in the modern varieties. There was no conclusive evidence that modern breeding programs are consistently selecting higher epitope varieties, yet the improvement in baking quality characteristics in modern wheats compared to older landraces has been consistent and substantial, suggesting that it may well be possible to develop lower epitope wheats without compromising gluten functionality. Furthermore, van den Broeck et al. (2010) only looked at two of the many known T-cell stimulatory epitopes and so most notably is lacking data for the most significant epitope, 33-mer (Glia-α2). A study from Schalk et al. (2017), however, does compare the levels of 33-mer in modern wheats with those seen in older varieties (cultivars registered prior to 1950). This work used varieties from a greater range of geographic locations with a total of 23 modern cultivars and 15 old cultivars. The results of that work clearly showed significant variations in levels of 33-mer within both the modern cultivars and within the old cultivars. The line with the lowest levels of 33-mer, by a significant margin, was an old cultivar, however the cultivar with the highest levels of 33-mer per g of α-gliadin, was also one of the old varieties. Because of the substantial variation within groups, there was no identifiable trend to suggest that one group (either modern or old cultivars) were significantly different to the other.

It is also important to consider environmental influences on epitope levels and it should be noted that not all of the tested lines used for experimental comparison have always been grown in the same conditions. As noted by Schalk et al. (2017), in a small trial with 5 separate varieties grown over 3 different seasons, it was found that harvest year had a greater influence on 33-mer content than the genotype, indicating that environmental factors are indeed a crucial consideration when comparing epitope profiles. It appears that little work has been done to quantify the effect of specific environmental variables on epitope profiles, however Dubois et al. (2018),used different varieties of spelt wheat (*T. aestivum* ssp*. spelta*) to provide further evidence that the role of the environment on epitope levels is likely to be both dynamic and complex. It was shown that the effect of location, year and nitrogen fertilisation strategy can all have variable effects on epitope levels (Dubois et al., 2018). Although little work has been done looking specifically at the relationship between epitopes and environment, a greater amount of work has been done to study the relationship between gliadin levels and environment in modern wheat, but again these effects have proven difficult to quantify. It does however seem clear that environmental conditions are a significant factor in determining gliadin levels (Wieser, 2007). In order to gain a comprehensive understanding of how environmental variables influence gluten epitope profiles, more trials of a much greater scale than have been undertaken in the past will be required.

2.5 Potential strategies for developing low epitope wheat products

Over 30 specific T-cell stimulatory gluten epitopes have been identified and are known to trigger coeliac disease in affected patients (Sollid et al., 2012). Typically, gluten epitopes are just part of a larger gluten peptide and like all gliadin and glutenin proteins, are controlled by multiple gene families at complex loci (Payne, 1987). This is further complicated by that fact that common bread wheat (*T. aestivum*) is also an allohexaploid species and so carries 3 genetically distinct sets of chromosomes, all of which carry genes encoding gluten proteins, meaning that understanding or manipulating the regulation of gluten epitopes are highly complicated tasks.

Research groups from around the world have been working on a variety of strategies to develop no or low epitope wheats. Most of these strategies have focused on applying molecular techniques to either down-regulate or delete specific epitope genes. Additional to this, rather than trying to influence wheat genomes, some groups have been looking at ways to reduce epitope levels in standard wheat-based products by utilizing specific enzymes in the baking process or specific processing techniques to breakdown epitopes. In all of these areas, substantial progress has indeed been made, however the link between epitope levels and baking quality has proven to be the biggest challenge as significant reductions in epitope levels appears to almost always be linked with reduced baking quality characteristics.

Work from van den Broeck et al. (2009), used deletion lines of the wheat 'Chinese Spring' to determine the effects of certain gliadin and glutenin loci on both epitope levels and baking properties. Removing the α-gliadin locus from chromosome 6-D showed significant reductions in total epitope levels but also resulted in significant losses in many baking quality traits. This was as expected considering that the α-gliadin genes from the D genome are believed to be the main source of gluten epitopes in wheat. Furthermore, the loss of a major gliadin loci was sure to alter the glutenin:gliadin ratio in the flour which is known to be a crucial determining factor for its baking quality. In contrast to this however, van den Broeck et al. (2009) also showed that in deletion lines individually lacking ω-gliadin, γ-gliadin and LMW glutenin loci from chromosome 1-D in 'Chinese Spring' did not suffer significant losses in baking quality characteristics despite showing decreases in epitope levels. This indicates that some epitopes may well be less important to baking quality traits than others. It is also important to note that 'Chinese Spring' may not necessarily be the best variety to study baking quality characteristics in as it tends to produce a less favourable bread than most modern varieties in the western world. However, the genome of 'Chinese Spring' has been sequenced and it is therefore an extremely useful variety for studying genetic controls in wheat.

A range of studies have also applied transgenic techniques in an attempt to produce low epitope wheat. This has generally been done through the down regulation of gliadins. Using this method, Wieser et al. (2007) and Becker et al. (2012),were both able to reduce α-gliadins by over 60% in their transgenic lines. Although the exact effect on epitope levels were not measured, it is expected that there would be a significant decrease in the presence of epitopes seeing as α-gliadins are the main source of these. The effect of reduced α-gliadins on baking quality characteristics however, appeared to be rather variable. Gluten strength in the transgenic lines was increased while bread volume was decreased and no difference was seen in either crumb structure or appearance, suggesting that overall baking quality was not affected too severely. Gil-Humanes, et al. (2014a,b),applied similar transgenic techniques to achieve reductions of 60-88% in total gliadins in a set of transgenic lines which also resulted in lower loaf volume with little effect on crumb structure or texture but did lead to reduced dough strength. These results suggest that lower epitope levels could potentially be achieved while retaining acceptable technical properties, however without measuring the effect of down-regulating gliadins has on the presence of specific epitopes, it is still unclear what benefit would be achieved. The application of the recently developed CRISPR-Cas9 gene editing technique has also been proposed as a promising method for reducing gluten epitopes in wheat while having minimal effect on gluten functionality. CRISPR-Cas9 can cause very specific changes to target genes and could theoretically be used to shorten epitope sequences or even cause small mutations within epitopes so they are no longer antigenic while having little effect on the overall protein structure and function. To demonstrate the potential for CRISPR-Cas9, Jouanin et al. (2018), applied these techniques to edit multiple gliadin genes and have since also developed methods to identify and characterise the

mutations that were achieved. Furthermore, Sánchez-León et al. (2018), were able to develop multiple mutant lines using CRISPR-Cas9 with up to 85% reduced immunogenicity from the 33-mer epitope. These efforts support the potential of CRISPR-Cas9 in achieving coeliac-safe wheat, although it has not yet determined what effect these changes have on baking quality. Further complications for CPRISPR-Cas9 are the ethical and legal concerns in some countries arising from the engineering that is required as an intermediate step in the process, even though the final resulting lines do not contain any foreign DNA.

Less attention has been given to exploiting existing natural variation for low epitope levels and applying conventional breeding techniques to identify and develop low epitope cultivars. Work from van den Broeck et al. (2010) and Schalk et al. (2017), has shown that there is substantial variation in epitope profiles for current commercial wheat cultivars and that low epitope lines with acceptable baking quality characteristics can be identified. However, there has been little work done to identify the genomic alleles that determine such characteristics. One study analysed a large range of α-gliadin genomic sequences in 11 different bread wheats and used bioinformatic techniques to identify a number of naturally occurring variants causing non-functional antigenic properties while appearing to have minimal effect on overall protein structure and function (Mitea et al., 2010). Furthermore, work from van den Broeck et al. (2009), suggests that cultivars lacking γ-gliadin epitopes are still likely to retain acceptable baking quality traits, while Spaenij–Dekking et al. (2005), provides evidence that epitopes in γ-gliadin's significantly contribute to a wheats overall epitope profile. This suggests that opportunities may exist to exploit naturally occurring variants for the purpose of developing low epitope wheats with retained functional qualities.

A number of strategies can be employed to exploit naturally occurring variants for low gluten epitope phenotypes, ranging from conventional phenotype-based breeding to more modern genotype-based selections. Genomic selection utilizes genotype data to predict an individual's genetic merit for specific traits, therefore allowing these data to be used to make breeding selections. Genomic selection can provide numerous advantages for rapid genetic improvement and would be well suited to breeding for low gluten epitope wheat varieties. Currently, measuring epitope concentrations in wheat is a costly and time-consuming exercise and it is unlikely to be practical in most breeding programs. Therefore, a genomic selection approach to breeding can significantly reduce the requirements for phenotyping epitope traits. Furthermore, genomic selection can accelerate breeding by allowing selection of superior individuals at earlier generations and is particularly effective for selecting towards complex, quantitative traits. To develop the tools for genomic selection, breeders can undertake a Genome Wide Association Study (GWAS) to identify genetic markers and loci that are significantly associated with their trait of interest. These data can then be utilized for the selection of superior lines through genomic selection and also Marker Assisted Selection (MAS). The latter being particularly useful for selecting towards or away from a limited number of strong effect genes. Naturally occurring genetic variants can also be exploited to achieve low epitope

wheats through the development of synthetic hexaploid lines. Synthetic hexaploid lines are developed by recreating the hybridisation process that led to the formation of common hexaploid wheats. In doing this, breeders are able to utilize the vast genetic diversity that exists in the D genome among wild *Ae. tauschii* accessions and integrate this diversity into common wheat lines. This strategy offers great potential considering the fact that the existing gene pool for common bread wheat has been narrowing consistently for many years due to the universal focus on a relatively small number of production and processing traits (van den Broeck et al., 2010). There has however been very little work looking into the variation in epitope profiles of available *Ae. tauschii* accessions or any work to develop synthetic hexploid lines for this purpose. More effort is necessary before this strategy can be properly assessed. Furthermore, the process of developing and assessing synthetic hexaploid lines, then incorporating promising lines into high performing modern cultivars would be a very long-term endeavour.

Rather than focusing on developing low epitope wheat, some researchers have instead focused on specific processing strategies for using standard wheat flour to produce low epitope products. One such example is the use of sourdough fermentation to catabolise gluten epitopes during processing. A number of studies have already demonstrated significant reductions in gluten epitopes through sourdough fermentation with *Lactobacillus* bacteria (Jouanin et al., 2018). Greco et al. (2011), used *Lactobacillus* and fungal proteases to breakdown gluten epitopes sufficiently to produce a wheat-based product suitable for coeliac patients. Other processing strategies that have shown promise include the use of malted grains or the addition of certain peptidases to degrade gluten epitopes in the raw products, however, as with sourdough fermentation, these processes would result in a different product when compared to standard wheat-based breads (Jouanin et al., 2018). Near total separation of gliadins from whole gluten and substitution of these with oat avenins to retain baking quality characteristics has also been achieved at lab scale (Jouanin et al., 2018). This method has also been proposed as a way of processing flour to be used in coeliac safe products. However, this process would present significant challenges in order to be scaled up to industrial levels. Furthermore, oat avenins are not actually considered to be 100% coeliac safe. Most coeliac patients can tolerate the consumption of oats but there are at least two immunogenic epitopes that have been identified (Gilissen et al., 2016). This means that oats are significantly safer than wheat-based products for coeliac consumers but can still illicit an immune response. Overall, these processes appear to show great promise in regard to developing coeliac-safe wheat-based products and combining them with novel low epitope wheat varieties may well be the best opportunity for achieving this without compromising baking quality.

2.6 Conclusion

It is clear that there are many promising strategies for developing low epitope wheats, each strategy has its own unique challenges, but it does seem likely that in the future low epitope or low gluten wheat products will be available to consumers. One important and currently unanswered question however, is at what point does a 'low epitope' wheat become beneficial to the health of consumers? Until a wheat variety is developed with no detectable epitopes, it will not be safe for coeliac consumers, but is there any benefit of a low epitope wheat to non-coeliac consumers? Perhaps the most widely accepted theory is that low epitope wheats have the potential to turn the tide on the increasing incidence of coeliac disease worldwide. Only a small portion of those genetically pre-disposed to coeliac disease ever become coeliac, yet a significant portion of the population are at risk and it is believed that the increasing consumption of gluten worldwide is causing the increasing incidence of the disease. Therefore, by producing low epitope wheat, it may be possible to reduce the exposure of potentially harmful epitopes to susceptible consumers and so reduce their risk of coeliac disease without the need for reduced gluten intake. Although a number of promising strategies have been discussed to reduce epitope levels in wheat-based products (Figure 2.4), there still remains a number of significant challenges. Due to the highly complex nature of gluten proteins themselves, the complex genetic controls of gluten proteins and the inherent link between gluten epitopes and bread baking quality, it is clear that a significant amount of work is still required to achieve an added value, low epitope wheat.

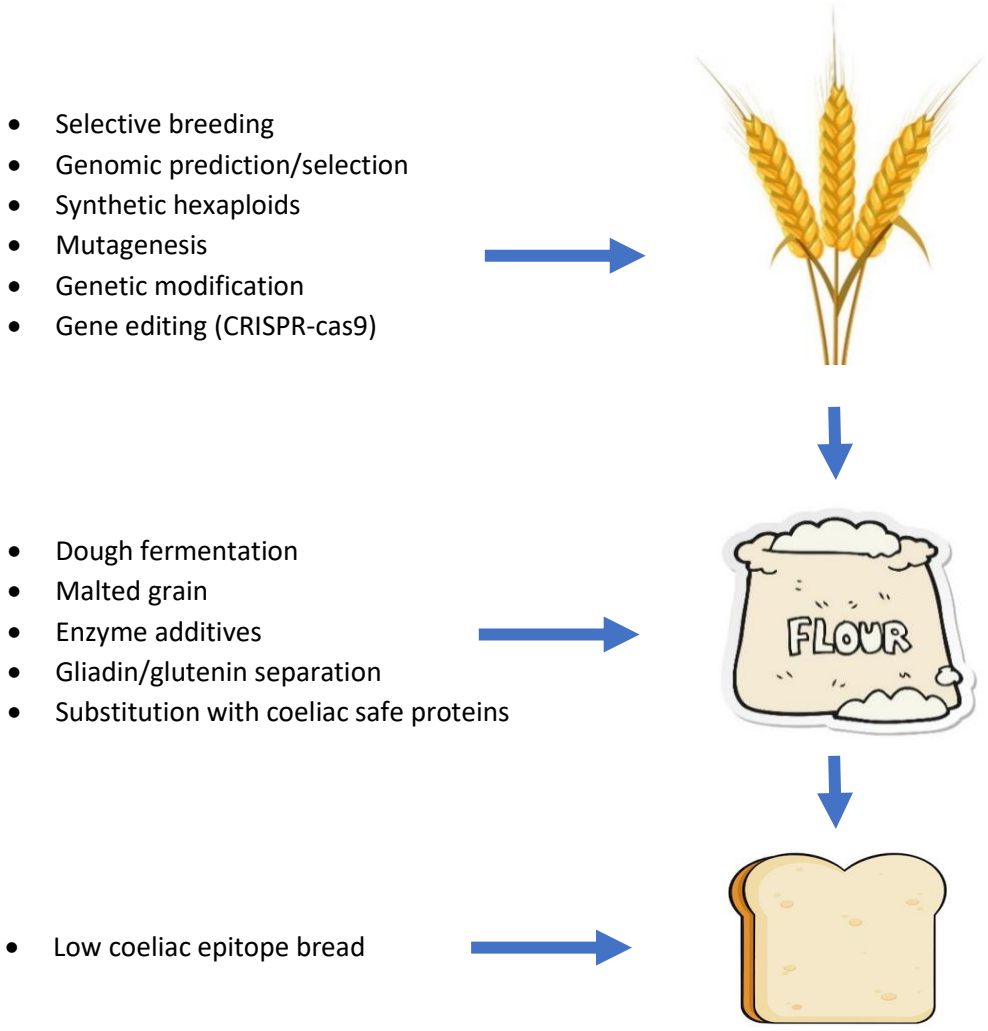

Figure 2.3: Potential strategies for developing low coeliac epitope wheat and the stages of the production cycle that they can be applied.

2.7 References

Balakireva, A. V., & Zamyatnin, A. A. (2016). Properties of Gluten Intolerance: Gluten Structure, Evolution, Pathogenicity and Detoxification Capabilities. *Nutrients*, *8*(10).

Barrett, J. S., & Gibson, P. R. (2012). Fermentable oligosaccharides, disaccharides, monosaccharides and polyols (FODMAPs) and nonallergic food intolerance: FODMAPs or food chemicals? *Therapeutic Advances in Gastroenterology*, *5*(4), 261-268. https://doi.org/10.1177/1756283X11436241

Baur, X., Degens, P. O., & Sander, I. (1998). Baker's asthma: Still among the most frequent occupational respiratory disorders. *Journal of Allergy and Clinical Immunology*, *102*(6), 984-997. https://doi.org/https://doi.org/10.1016/S0091-6749(98)70337-9

Becker, D., Wieser, H., Koehler, P., Folck, A., Muhling, K. H., & Zorb, C. (2012). Protein composition and techno-functional properties of transgenic wheat with reduced a-gliadin content obtained by RNA interference. *Journal of Applied Botany and Food Quality*, *85*, 23-33.

Belton, P. S. (1999). Mini Review: On the Elasticity of Wheat Gluten. *Journal of Cereal Science*, *29*(2), 103-107. https://doi.org/https://doi.org/10.1006/jcrs.1998.0227

Biesiekierski, J. R. (2017). What is gluten? *Journal of Gastroenterology and Hepatology*, *32*, 78-81. https://doi.org/10.1111/jgh.13703

Bonini, S. (2012). Immunology IV: Clinical Applications in Health and Disease. *The World Allergy Organization Journal*, *5*(8). https://doi.org/https://doi.org/10.1097/WOX.0b013e3182641db0

Camarca, A., Anderson, R., Mamone, G., Fierro, O., Facchiano, A., Costantini, S., . . . Troncone, R. (2009). Intestinal T Cell Responses to Gluten Peptides Are Largely Heterogeneous: Implications for a Peptide-Based Therapy in Celiac Disease. *Journal of immunology*, *182*, 4158-4166. https://doi.org/10.4049/jimmunol.0803181

Comino, I., Real, A., Vivas, S., Síglez, M. Á., Caminero, A., Nistal, E., . . . Sousa, C. (2012). Monitoring of gluten-free diet compliance in celiac patients by assessment of gliadin 33-mer equivalent epitopes in feces. *The American Journal of Clinical Nutrition*, *95*(3), 670-677. https://doi.org/https://doi.org/10.3945/ajcn.111.026708

Czaja-Bulsa, G. (2015). Non coeliac gluten sensitivity – A new disease with gluten intolerance. *Clinical Nutrition*, *34*(2), 189-194. https://doi.org/https://doi.org/10.1016/j.clnu.2014.08.012

Dubois, B., Bertin, P., Hautier, L., Muhovski, Y., Escarnot, E., & Mingeot, D. (2018). Genetic and environmental factors affecting the expression of α-gliadin canonical epitopes involved in celiac disease in a wide collection of spelt (Triticum aestivum ssp. spelta) cultivars and landraces. *BMC Plant Biology*, *18*(1), 262. https://doi.org/10.1186/s12870-018-1487-y

Dupont, F. M., Vensel, W. H., Tanaka, C. K., Hurkman, W. J., & Altenbach, S. B. (2011). Deciphering the complexities of the wheat flour proteome using quantitative two-dimensional electrophoresis, three proteases and tandem mass spectrometry. *Proteome Science*, *9*(1), 10. https://doi.org/10.1186/1477-5956-9-10

Dvořák, J. (2001). Triticum Species (Wheat). In S. Brenner & J. H. Miller (Eds.), *Encyclopedia of Genetics* (pp. 2060-2068). Academic Press. https://doi.org/https://doi.org/10.1006/rwgn.2001.1672

Fasano, A., Sapone, A., Zevallos, V., & Schuppan, D. (2015). Nonceliac gluten sensitivity. *Gastroenterology*, *148*(6), 1195–1204.

Feldman, M., Lupton, F. G. H., & Miller, T. E. (1995). *Evolution of crop plants* Longman Scientific & technical.

Fermin, B., Hahm, T. S., Radinsky, J., Kratochvil, R., Hall, J., & Lo, M. (2005). Effect of Proline and Glutamine on the Functional Properties of Wheat Dough in Winter Wheat Varieties. *Journal of Food Science*, *70*, 273-278. https://doi.org/10.1111/j.1365-2621.2005.tb07183.x

Gil-Humanes, J., Pistón, F., Altamirano-Fortoul, R., Real, A., Comino, I., Sousa, C., . . . Barro, F. (2014). Reduced-Gliadin Wheat Bread: An Alternative to the Gluten-Free Diet for Consumers Suffering Gluten-Related Pathologies. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0090898

Gil-Humanes, J., Pistón, F., Barro, F., & Rosell, C. M. (2014). The Shutdown of Celiac Disease-Related Gliadin Epitopes in Bread Wheat by RNAi Provides Flours with Increased Stability and Better Tolerance to Over-Mixing. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0091931

Gilissen, L. J. W. J., van der Meer, I. M., & Smulders, M. J. M. (2016). Why Oats Are Safe and Healthy for Celiac Disease Patients.*Medical sciences (Basel, Switzerland),* 4(4), 21. https://doi.org/10.3390/medsci4040021

Greco, L., Gobbetti, M., Auricchio, R., Di Mase, R., Landolfo, F., Paparo, F., . . . Auricchio, S. (2011). Safety for Patients With Celiac Disease of Baked Goods Made of Wheat Flour Hydrolyzed During Food Processing. *Clinical Gastroenterology and Hepatology*, *9*(1), 24-29. https://doi.org/https://doi.org/10.1016/j.cgh.2010.09.025

Gujral, N., Freeman H, J., & Thomson, A. B. (2012). Celiac disease: prevalence, diagnosis, pathogenesis and treatment. *World Journal of Gastroenterology*, *18*(42), 6036-6059. https://doi.org/10.3748/wjg.v18.i42.6036

Howdle, P. D. (2006). Gliadin, glutenin or both? The search for the Holy Grail in coeliac disease. *European Journal of Gastroenterology & Hepatology*, *18*(7).

Howdle, P. F., Ciclitira, P. J., Simpson, F. G., & Losowsky, M. S. (1984). Are all gliadins toxic in coeliac disease? An in vitro study of alpha, beta, gamma, and w gliadins. *Scandinavian journal of gastroenterology*, *19*(1), 41-47.

Inomata, N. (2009). Wheat allergy. *Current Opinion in Allergy and Clinical Immunology*, *9*(3).

Janssen, G., Christis, C., Kooy-Winkelaar, Y., Edens, L., Smith, D., van Veelen, P., & Koning, F. (2015). Ineffective Degradation of Immunogenic Gluten Epitopes by Currently Available Digestive Enzyme Supplements. *PLoS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0128065

Jouanin, A., Gilissen, L. J. W. J., Boyd, L. A., Cockram, J., Leigh, F. J., Wallington, E. J., . . . Smulders, M. J. M. (2018). Food processing and breeding strategies for coeliac-safe and healthy wheat products. *Food Research International*, *110*, 11-21. https://doi.org/https://doi.org/10.1016/j.foodres.2017.04.025

Junker, Y., Zeissig, S., Kim, S.-J., Barisani, D., Wieser, H., Leffler, D. A., . . . Schuppan, D. (2012). Wheat amylase trypsin inhibitors drive intestinal inflammation via activation of toll-like receptor 4. *Journal of Experimental Medicine*, *209*(13), 2395-2408. https://doi.org/10.1084/jem.20102660

Kerber, E. R., & Tipples, K. H. (1969). Effects of the D genome on milling and baking properties of wheat. *Canadian Journal of Plant Science*, *49*(3), 255-263. https://doi.org/10.4141/cjps69-046

Kucek, L. K., Veenstra, L. D., Amnuaycheewa, P., & Sorrells, M. E. (2015). A Grounded Guide to Gluten: How Modern Genotypes and Processing Impact Wheat Sensitivity. *Comprehensive Reviews in Food Science and Food Safety*, *14*(3), 285-302. https://doi.org/https://doi.org/10.1111/1541-4337.12129

Martucci, S., & Corazza, G. R. (2002). Spreading and focusing of gluten epitopes in celiac disease. *Gastroenterology*, *122*(7), 2072-2075. https://doi.org/https://doi.org/10.1053/gast.2002.34102

Mitea, C., Salentijn, E. M. J., van Veelen, P., Goryunova, S. V., van der Meer, I. M., van den Broeck, H. C., . . . Smulders, M. J. M. (2010). A Universal Approach to Eliminate Antigenic Properties of Alpha-Gliadin Peptides in Celiac Disease. *PLoS ONE*, *5*(12). https://doi.org/10.1371/journal.pone.0015637

Molberg, Ø., Uhlen, A. K., Jensen, T., Flæte, N. S., Fleckenstein, B., Arentz–Hansen, H., . . . Sollid, L. M. (2005). Mapping of gluten T-cell epitopes in the bread wheat ancestors: Implications for celiac disease. *Gastroenterology*, *128*(2), 393-401. https://doi.org/https://doi.org/10.1053/j.gastro.2004.11.003

Molina-Infante, J., Santolaria S Fau - Sanders, D. S., Sanders Ds Fau - Fernández-Bañares, F., & Fernández-Bañares, F. (2015). Systematic review: noncoeliac gluten sensitivity. (1365-2036).

Nasr, I., Shekeili, L. A. I., Wahshi, H. A. A., Nasr, M. H., & Ciclitira, P. J. (2016). Celiac disease, wheat allergy and non celiac gluten sensitivity. *Integrative Food, Nutrition and Metabolism*, *3*(4), 336-340. https://doi.org/doi:10.15761/IFNM.1000155

Payne, P. I. (1987). Genetics of Wheat Storage Proteins and the Effect of Allelic Variation on Bread-Making Quality. *Annual Review of Plant Physiology*, *38*(1), 141-153. https://doi.org/10.1146/annurev.pp.38.060187.001041

Peña, A. S., & Wijmenga, C. (2001). Genetic factors underlying gluten-sensitive enteropathy. (1529-7322).

Priya, S., Kumar, S., Kaur, N., & Gupta, A. K. (2013). Specificity of α-amylase and trypsin inhibitor proteins in wheat against insect pests. *New Zealand Journal of Crop and Horticultural Science*, *41*(1), 49-56. https://doi.org/10.1080/01140671.2012.722112

Sánchez-León, S., Gil-Humanes, J., Ozuna, C. V., Giménez, M. J., Sousa, C., Voytas, D. F., & Barro, F. (2018). Low-gluten, nontransgenic wheat engineered with CRISPR/Cas9. *Plant Biotechnology Journal*, *16*(4), 902-910. https://doi.org/https://doi.org/10.1111/pbi.12837

Schalk, K., Lang, C., Wieser, H., Koehler, P., & Scherf, K. A. (2017). Quantitation of the immunodominant 33-mer peptide from alpha-gliadin in wheat flours by liquid chromatography tandem mass spectrometry. *Sci Rep*, *7*, 45092. https://doi.org/10.1038/srep45092

Scherf, K. A., Catassi, C., Chirdo, F., Ciclitira, P. J., Feighery, C., Gianfrani, C., . . . Koehler, P. (2020). Recent Progress and Recommendations on Celiac Disease From the Working Group on Prolamin Analysis and Toxicity. *Frontiers in Nutrition*, *7*, 29.

Scherf, K. A., Koehler, P., & Wieser, H. (2016). Gluten and wheat sensitivities – An overview. *Journal of Cereal Science*, *67*, 2-11. https://doi.org/10.1016/j.jcs.2015.07.008

Shan, L., Molberg, Ø., Parrot, I., Hausch, F., Filiz, F., Gray, G. M., . . . Khosla, C. (2002). Structural Basis for Gluten Intolerance in Celiac Sprue. *Science*, *297*(5590), 2275-2279. https://doi.org/10.1126/science.1074129

Shewry, P. (2019). What Is Gluten-Why Is It Special? *Front Nutr*, *6*, 101. https://doi.org/10.3389/fnut.2019.00101

Shewry, P. R., & Hey, S. J. (2015). The contribution of wheat to human diet and health. *Food and Energy Security*, *4*(3), 178-202. https://doi.org/https://doi.org/10.1002/fes3.64

Shewry, P. R., Tatham, A. S., Forde, J., Kreis, M., & Miflin, B. J. (1986). The classification and nomenclature of wheat gluten proteins: A reassessment. *Journal of Cereal Science*, *4*(2), 97-106. https://doi.org/https://doi.org/10.1016/S0733-5210(86)80012-1

Sollid, L. (2002). Coeliac disease: dissecting a complex inflammatory disorder. *Nature reviews. Immunology*, *2*, 647-655. https://doi.org/10.1038/nri885

Sollid, L. M., Qiao, S.-W., Anderson, R. P., Gianfrani, C., & Koning, F. (2012). Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. *Immunogenetics*, *64*(6), 455-460. https://doi.org/10.1007/s00251-012-0599-z

Spaenij–Dekking, L., Kooy–Winkelaar, Y., van Veelen, P., Wouter Drijfhout, J., Jonker, H., van Soest, L., . . . Koning, F. (2005). Natural Variation in Toxicity of Wheat: Potential for Selection of Nontoxic Varieties for Celiac Disease Patients. *Gastroenterology*, *129*(3), 797-806. https://doi.org/https://doi.org/10.1053/j.gastro.2005.06.017

Tye-Din, J. A., Stewart, J. A., Dromey, J. A., Beissbarth, T., van Heel, D. A., Tatham, A., . . . Anderson, R. P. (2010). Comprehensive, Quantitative Mapping of T Cell Epitopes in Gluten in Celiac Disease. *Science Translational Medicine*, *2*(41). https://doi.org/10.1126/scitranslmed.3001012

Vader, W., Kooy, Y., van Veelen, P., de Ru, A., Harris, D., Benckhuijsen, W., . . . Koning, F. (2002). The gluten response in children with celiac disease is directed toward multiple gliadin and glutenin peptides. *Gastroenterology*, *122*(7), 1729-1737. https://doi.org/https://doi.org/10.1053/gast.2002.33606

Van De Kamer, J. H., Weijers, H. A., & Dicke, W. K. (1953). Coeliac disease. IV. An investigation into the injurious constituents of wheat in connection with their action on patients with coeliac disease. *Acta paediatrica*, *42*(3), 223-231. https://doi.org/https://doi.org/10.1111/j.1651-2227.1953.tb05586.x

van den Broeck, H., de Jong, H., Salentijn, E., Dekking, L., Bosch, D., Hamer, R., . . . Smulders, M. J. M. (2010). Presence of celiac disease epitopes in modern and old hexaploid wheat varieties: Wheat breeding may have contributed to increased prevalence of celiac disease. *Theoretical and Applied Genetics*, *121*, 1527-1539. https://doi.org/10.1007/s00122-010-1408-4

van den Broeck, H. C., van Herpen, T. W., Schuit, C., Salentijn, E. M., Dekking, L., Bosch, D., . . . van der Meer, I. M. (2009). Removing celiac disease-related gluten proteins from bread wheat while retaining technological properties: a study with Chinese Spring deletion lines. *BMC Plant Biol*, *9*, 41. https://doi.org/10.1186/1471-2229-9-41

Varney, J., Barrett, J., Scarlata, K., Catsos, P., Gibson, P. R., & Muir, J. G. (2017). FODMAPs: food composition, defining cutoff values and international application. *Journal of Gastroenterology and Hepatology*, *32*(S1), 53-61. https://doi.org/https://doi.org/10.1111/jgh.13698

Vincentini, O., Borrelli, O., Silano, M., Gazza, L., Pogna, N., Luchetti, R., & De Vincenzi, M. (2009). T-cell response to different cultivars of farro wheat, Triticum turgidum ssp. dicoccum, in celiac disease patients. *Clinical Nutrition*, *28*(3), 272-277. https://doi.org/https://doi.org/10.1016/j.clnu.2009.03.013

Wieser, H. (2007). Chemistry of gluten proteins. *Food Microbiology*, *24*(2), 115-119. https://doi.org/https://doi.org/10.1016/j.fm.2006.07.004

Wieser, H., Koehler, P., Folck, A., & Becker, D. (2007). Characterization of wheat with strongly reduced α-gliadin content. *Gluten proteins*, 13-16.

Zevallos, V. F., Raker, V., Tenzer, S., Jimenez-Calvente, C., Ashfaq-Khan, M., Rüssel, N., . . . Schuppan, D. (2017). Nutritional Wheat Amylase-Trypsin Inhibitors Promote Intestinal Inflammation via Activation of Myeloid Cells. *Gastroenterology*, *152*(5), 1100-1113. https://doi.org/https://doi.org/10.1053/j.gastro.2016.12.006

Chapter 3

Heritability analysis of 6 coeliac disease related gluten peptides from the α-gliadin fraction of wheat flour

3.1 Abstract

Concentrations of 6 gluten epitopes (P1-P6) were measured in 471 wheat lines across 3 growing seasons in Lincoln, New Zealand. Phenotypic correlations were calculated using Pearson's correlation coefficient (r). Positive correlations were observed between years for each epitope ranging from r=0.41 for P1 up to r=0.92 for P3. Total epitope concentrations (P1-P6) were also positively correlated between years ranging from r=0.40 to r=0.65. Positive genetic correlations were observed for epitope concentrations between years. These ranged from 0.62 for P4 to 0.97 for P3 while genetic correlations for total epitope concentrations (P1-P6) ranged from 0.72-0.89 when compared between different seasons. Heritability estimates (h^2) and variance components were calculated using the BayesC method. In the analysis of epitope concentrations across all three growing seasons, h^2 estimates ranged from as low as 0.44 for P1 up to 0.75 for P5 and h^2 estimates for total epitope concentrations (P1-P6) were 0.52. These data indicate that each of the measured epitopes are at least moderately to highly heritable with high genetic correlations across environments. This suggests that selective breeding is likely to be an effective tool for developing new wheat cultivars with significantly reduced concentrations of gluten epitope. However, few significant genetic associations appear to exist between individual epitopes suggesting that breeders will need to target individual epitopes rather than selecting only on total epitope concentrations.

3.2 Introduction

Wheat (*Triticum aestivum*) has been domesticated for over 10,000 years (Charmet, 2011). It is believed that over that time farmers have been selecting desirable plants, produced through natural outcrossing, to produce their following crop, making this the first step from natural selection towards artificial breeding. Wheat breeding began to evolve further in the early 19th century with the development of hybridisation techniques and an understanding of Mendelian genetics(Lupton, 1987). Since that time, the understanding of the genetics underlying breeding outcomes has progressed rapidly, along with the development of new technologies allowing this knowledge to be applied through modern day breeding programs. As technology advanced, breeders began to phenotype novel traits and determine the expected response to selection on those traits by measuring variance, standard deviation and heritability in their populations(Balkan, 2018). In more recent years, genotyping allowed new approaches to be investigated such as Genome Wide Association Studies and Marker Assisted Selection. These techniques were originally developed in livestock breeding and are increasingly being adapted for and applied to plant breeding programs. These new approaches are allowing plant breeders to both increase their selection accuracy and to target more quantitative and difficult to measure traits.

For decades now, wheat breeding programs have targeted a number of key traits that are valued by growers and by the milling and baking industries. Such traits include total grain yield, resistance to prominent diseases, non-shattering seed, high total protein and large seed size (Fischer, 2022). There has been less focus on consumer targeted traits. However, with an increasing incidence of coeliac disease, non-coeliac gluten sensitivity (NCGS) and general gluten avoidance globally(Scherf et al., 2020), it seems prudent for modern day breeders to shift some of their focus to the nutritional quality of wheat and wheat products. Specific epitopes have been identified in wheat gluten as the trigger for coeliac disease and may also contribute to the development of NCGS in some consumers (Sharma et al., 2020). Most gluten epitopes exist within the α-gliadin fraction (Ruiz-Carnicer et al., 2019) and 6 of these epitopes are assessed in this study (P1-P6). These epitopes have all been shown to be immunogenic in coeliac patients, with the most potent epitope being P5, also commonly referred to as 33-mer as it is made up of a sequence of 33 amino acids (Cebolla et al., 2018). A number of research groups around the world are working to eliminate some of these epitopes from their cultivars through mutagenesis, gene editing using CRISPR-Cas9 or other methods (Shewry & Tatham, 2016). Progress has been hindered by the complex nature of gluten coding genes, protein regulation and linkage to other important breeding traits (Becker et al., 2012; Jouanin et al., 2019; Sanchez-Leon et al., 2018; van den Broeck et al., 2009; Wieser, 2006). Far less work has been done on the use of conventional breeding methods to reduce epitope concentrations. This study aims to address that gap in knowledge by assessing genetic variation and covariation, phenotypic variation and heritability of different gluten epitopes in a breeding population to determine the breeding potential for different epitope traits.

3.3 Materials & methods

3.3.1 Field trials

A population of 471 different wheat varieties was grown across three seasons (2019-2020, 2020-2021 and 2021-2022) at the New Zealand Institute of Plant & Food Research (PFR) in Lincoln, Canterbury. These varieties were sown in mid-May each season as separate two row plots (1m x 40cm) and harvested in late January each year. Varieties were sown in order of anticipated flowering date. No replicates were sown in the field trial, replication was only achieved when phenotyping traits by taking multiple subsamples from the harvested grain. The population consisted of a range of diverse material acquired from the PFR genebank. This material included advanced New Zealand milling wheat breeding lines that were developed from as early as 1986 to advanced selections developed from as late as 2018. The population included current and ex-commercial cultivars from New Zealand, Australia, the United Kingdom and Canada, as well as lines developed through targeted low epitope crosses using double haploids. All of the conventionally bred lines were at the F7 stage or later and therefore genetic heterogeneity within lines was expected to be

minimal. Approximately 75% of the population consisted of lines bred in New Zealand, primarily through the PFR breeding program, with the remainder of lines being sourced from international germplasm. Lines specifically bred for milling and bread baking made up 78% of the population while the remaining 22% of lines were considered low quality varieties, bred for biscuit production or animal feed. A full list of the named varieties grown for this study is included in the Appendix (Table 8.1). The bread wheat varieties Reliance, Viceroy, Hanson, Duchess and Conquest were utilized as checks for phenotype results. These varieties have each been grown and phenotyped at the Lincoln site every year for at least the previous 5 years prior to the 2019-2020 season and therefore their phenotypic characteristics are well documented. Grain quality data (Thousand Grain Weights, proportion of undersize grain and hectolitre weights) were also collected on all lines along with disease susceptibility scores for septoria and leaf rust.

The 2019-2020 season (Year 1) was defined by a cooler, mostly overcast spring (Sep-Oct) with generally consistent rainfall, ending in a dry summer (Table 3.1; Figure 3.1). That season also produced less Growing Degree Days (GDD) than the following two seasons (Table 3.2). The 2020-2021 growing season (Year 2) was characterised by a relatively dry winter and early spring (Aug-Oct) then finished with a warm but wet summer (Nov-Jan); (Table 3.1; Figure 3.1). In the 2021-2022 growing season (Year 3), the winter and spring periods received consistently mild weather with high levels of rainfall throughout the year, particularly in December when the level of solar radiation was recorded to be well below that of the previous two seasons (Table 3.1; Figure 3.1).

Table 3.1: Monthly weather data for each growing season recorded in Lincoln, New Zealand

	Mean AirTemp Max (°C)			**Mean AirTemp Min (°C)**			**Total Solar (Mj/m²/day)**			**Rain Total (mm)**		
	Year 1	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**
May	16.1	16	15.7	5.4	5.2	5	187.3	178.1	209.5	11	22.4	108
Jun	11.7	12.2	12.3	1.6	4.3	4.6	138.1	108.1	116.6	59	87.6	99.8
Jul	13.1	11.3	12.2	4.5	2.4	2.5	136.7	158.6	185.3	68	54.2	21.2
Aug	12.9	13.7	13	2.5	3.9	4.8	248.4	259.9	238.1	50.4	6.4	76.8
Sep	15	17	15.1	4.7	4.5	4.9	362.8	444.8	417.1	31.6	31.4	27.6
Oct	15.7	18.1	16.7	6.1	8.2	7.9	495.3	560.4	523.1	54.2	12.4	51.2
Nov	20.8	19.3	18.8	10.1	9.1	10.4	615.5	606.5	645.3	57.8	64.6	29.2
Dec	20.8	20.9	19.8	10.3	10	12.2	706.7	672.6	547.2	37.6	52.2	110.4
Jan	22.2	22.5	21.7	11.7	12.2	12.7	719.4	692.8	772.5	6.6	32.6	48.2

Table 3.2: Growing Degree Days (GDD) recorded between sowing and harvest dates for each growing season

Season	**GDD**
Year 1	2,720
Year 2	2,819
Year 3	2,834

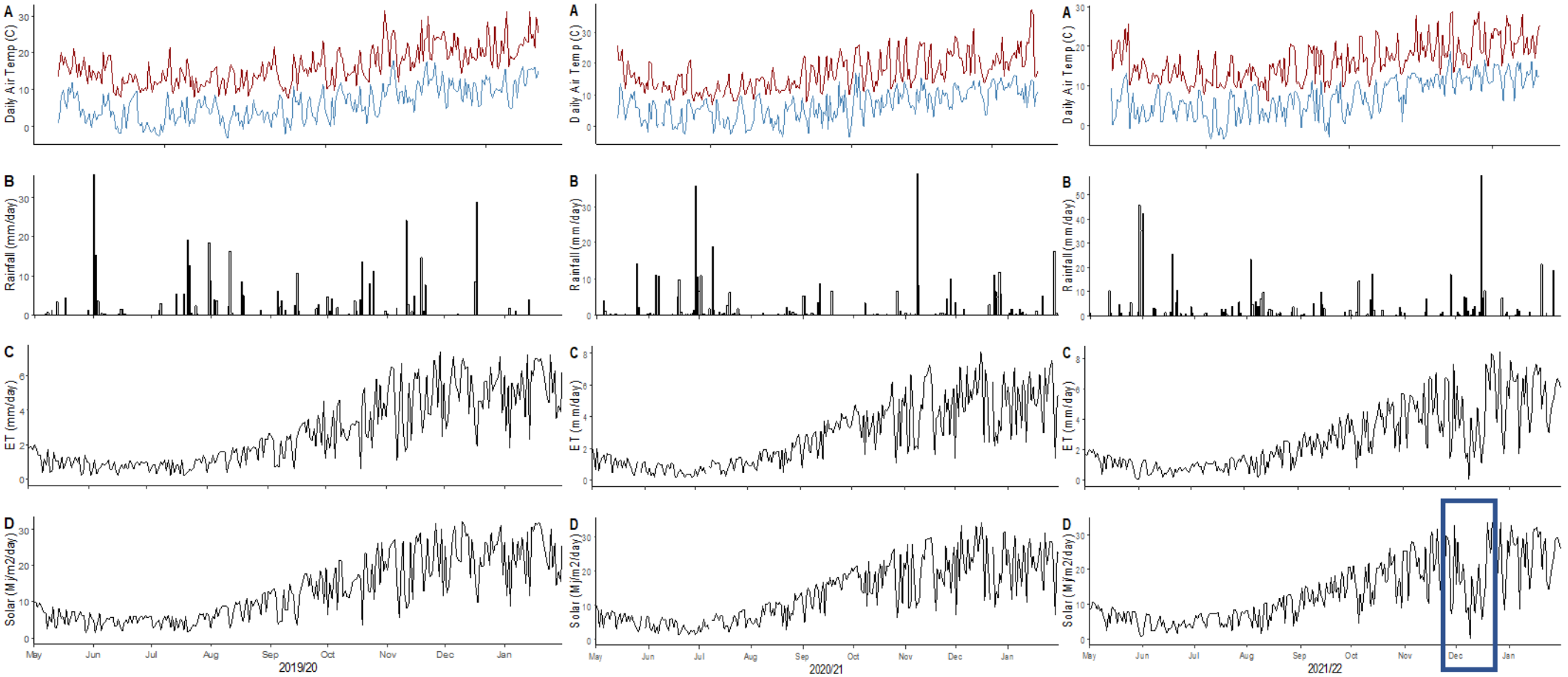

Figure 3.1: Weather data recorded in Lincoln, New Zealand between May and February for each of the three growing seasons. A: Daily minimum (blue line) and maximum (red line) temperatures; B: Daily Rainfall; C: Daily Evapotranspiration Rate (ET); D: Daily Solar radiation. A notable reduction in Solar radiation is highlighted in December 2021

3.3.2 Flour protein content (FPC)

The FPC was determined using near-infrared spectroscopy (NIR). Whole-grain flour samples were milled through a 0.5mm screen on a Foss CT293 Cyclotec™ (Foss Analytical Co., Ltd., Suzhou, China) then measurements were recorded using a NIR Technicon Infralyzer 450 (Technicon Ireland Ltd., Swords, Co., Dublin). The NIR was calibrated annually to meet industry criteria set by the New Zealand Flour Millers Association.

3.3.3 Epitope quantification

Epitope quantification was undertaken using a method based on the techniques described by Ogilvie et al. (2021). For each year of testing, a 20 g grain sample was taken from every variety and ground into a whole-grain flour using a Foss CT293 Cyclotec™ (Foss Analytical Co., Ltd., Suzhou, China). Three replicate 100mg (+/- 2mg) flour samples were then weighed out for each variety and placed in a labelled 1.5ml Eppendorf tube and the flour weights were recorded. All samples were then stored in the dark at 4°C until needed.

Gliadin was extracted by addition of 1 mL of 70% (v/v) aqueous ethanol with continuous mixing using a rotation wheel for 60 min at room temperature, followed by centrifugation at 10,000 rpm for 10 min at room temperature. The residue was re-extracted with 70% aqueous ethanol for 30 min at room temperature with continuous rotating, followed again by centrifugation at 10,000 rpm for 10 min at room temperature. The two supernatants were combined and stored in the dark at 4°C until needed.

Gliadin extracts were warmed to room temperature followed by centrifugation at 10,000 rpm for 10 min, prior to digestion. Aliquots of 80 µL were taken and four-fold diluted with 100 mM Tris-HCL (pH 8.0)/10 mM CaCl2, spiked with heavy isotope labelled standard P1 (P1H; 1 µg/mL). Proteins were digested with chymotrypsin using an approximate 1:5 ratio (protease:protein) at 30°C for 24 hours. Digestion was quenched with the addition of trifluoroacetic acid (final concentration 0.5%), followed by centrifugation at 10,000 rpm for 10 min. Aliquots were removed for Liquid Chromatography-Mass Spectrometry(LC-MS) analysis. Samples were extracted and analysed in batches of 70 samples at a time.

The LC-MS system consisted of a Thermo Scientific™ (San Jose, CA, USA) Q Exactive™ Plus Orbitrap (HR/AM) LC-MS/MS coupled with a Vanquish™ UHPLC system (Binary Pump H, Split Sampler HT, DAD HL, Dual Oven).

For the LC-MS analysis, six model immunogenic peptides from the gliadin fraction of gluten (P1-P6) were synthesised and purchased from New England Peptide (USA) (Table 3.3). For each of the peptides, a stock solution of 50 mM was prepared in 100% methanol and stored at 20°C. Calibration standards were prepared by diluting the individual stock solutions using 10% methanol. Calibration standard curves were measured using 10, 5, 4, 3, 2, 1 and 0.5 ppm.

A 2-µL aliquot of each standard and prepared extract was separated with a mobile phase consisting of 0.1% formic acid in water and 0.1% formic acid in acetonitrile by reversed phase chromatography (Aeris™ 1.7 µm PEPTIDE XB-C18 100 Å, LC Column 100 x 2.1 mm and guard, Part Number:00D-4506-AN, Phenomenex, Torrance, CA, USA) maintained at 30°C with a flow rate of 300 µL/min.

Table 3.3: Synthetic peptide standards

Mass (m/z)	Formula (M)	Start (min)	End(min)	Peptide
813.90480 (2+)	RPQQPYPQPQPQY	2.40	4.50	P6
784.92721 (2+)	LQLQPFPQPQLPY	4.50	6.40	P1
1032.54291 (3+)	LQLQPFPQPQLPYPQPHLPYPQPQPF	6.20	6.80	P4
755.06852 (3+)	LQLQPFPQPQLPYPQPQPF	6.30	6.95	P2
1029.54280 (3+)	LQLQPFPQPQLPYPQPQLPYPQPQPF	6.75	7.40	P3
1304.01708 (3+)	LQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPF	6.90	8.20	P5
978.26463 (4+)	LQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPF	6.90	8.20	P5

The peptide concentration (ppm) was calculated using the slope coefficient obtained from the linear regression of the synthetic standard peptide (P1-P6) curves. Data were processed with the aid of Xcalibur®4 and Proteome Discoverer v1.4.1.14 (Thermo Electron Corporation).

3.3.4 Genotyping

To prepare DNA samples for genotyping, 6 SSR markers (previously selected for their high number of unique alleles) were analysed from DNA samples of 3 plants of each variety in both the training and validation populations. This process was carried out to ensure the DNA being genotyped was 'true to type' and not from an inadvertent outcross or seed contamination.

Tissue samples for DNA extractions were cut from 2-month-old plants and placed in separate wells of a 96-well plate. Before freezing at –20°C, 2 tungsten carbide beads (Qiagen) were added to each well. Samples were freeze dried for 48 hours immediately prior to DNA extractions. Tissue samples were disrupted on the Tissuelyser II (Qiagen) at 20Hz for 30sec, the plates were rotated after each 30sec, and the process repeated 3 more times. DNA extractions were then carried out using the protocol described by Pallotta et al. (2003).

SSR marker analysis was carried out using 3 different multiple PCR runs on each DNA sample. Each PCR consisted of a combination of two SSR markers each with a common fluorescently labelled primer (20uM working stock). The combinations are outlined in Table 3.4.

Table 3.4: SSR marker and primer combinations for 'True to type' testing

	SSR Marker 1	SSR Marker 2	Fluorescent Primer
PCR1	wmc422	gwm583	Rhod
PCR2	wmc285	wmc420	FAM
PCR3	wmc144	wmc14	Hex

The PCRs were run using the conditions outlined in Table 3.5.The PCR products were separated for 9 mins at 280V in agarose gel (2.2g agarose powder with 110ml of 1xTBE (Tris, Boric Acid and EDTA)) and 5ul of SYBRsafe (Thermo Fisher Scientific). A gel image was then obtained under trans-UV light using a Bio-Rad Gel Doc XR+ Imaging System to determine if the PCR was successful.

Table 3.5: PCR conditions

	Temperature (°C)	Time	Number of cycles
Start	94	2 mins	
Denature	94	30 sec	
Annealing	55	30 sec	40
Elongation	72	30 sec	
End	72	5 mins	

PCR products were analysed using an ABI3130xI Genetic Analyser (Applied Biosystems). First 1ul of PCR product was added to 9ul of ddH_2O and 3ul of EDTA (125mM). Then 30ul of absolute ethanol was added before samples were mixed by inversion 10 times and stored in the dark for 15 mins. The samples were then centrifuged at 3,000 rcf at 4°C for 30 mins. The supernatant was then removed, and the pellet dried at 65°C for 10 mins. 10ul of a Formamide/Genescan Liz600 ladder mix (1,000:7.5) was added to each sample, centrifuged at 1,000 rpm for 1 min and then denatured at 94°C for 5 mins before being cooled on ice for an additional 5 mins. The samples were then processed using the Genetic Analyser.

The PCR results were analysed using GeneMarkerv1.3 (Softgenetics, State College, PA). Each of the 6 SSR markers were compared across the three plants from each wheat variety. Plants were considered 'true to type' if the marker results matched those of at least one of the other plants of that variety. One 'true to type' plant was kept for each variety with the remaining plants being discarded. A new DNA sample was then extracted from each of the 'true to type' plants using the same process as described above but using single Eppendorf tubes instead of 96 well plates. The quantity of DNA in the solutions was then determined using PicoGreen fluorescent dye (Quant-iTPicoGreen DS DNA assay, Invitrogen P7589) following the

procedures outlined by Invitrogen (2022). Finally, 3ug of DNA from each sample were sent to GenomNZ for SNP chip genotyping.

All varieties were genotyped by GenomNZ using a 90K Illumina SNP bead chip (Infinium HD iSelect-24 Kit, 48 samples, 3072-90K) through the Infinium HD workflow. A consensus genetic map for wheat (Wang et al., 2014) was used to define the chromosomal location and base pair positions of the SNPs. A total of 42,755 of 81,587 SNPs were able to be mapped. Clustering of SNPs was determined using Genome Studio(v2.0.5). The DBSCAN algorithm was used to set cluster distance to 0.07 and the minimum points per cluster was set to 20. Markers with a call rate of below 0.95 were removed.A total of 0.6% of calls were heterozygous, as expected for an inbred population. The genotype calls were converted to an allele dosage of 0, 1 or 2. For our analysis, the small proportion of missing calls were replaced with the population mean dosage for that locus from called samples while monomorphic markers and those with a minor allele frequency less than 0.05 were removed from subsequent analyses. The remaining 26,386 markers after these edits were included in all calculations.

3.3.5 Statistical analysis

Variance components and heritability for the concentrations of gluten epitopes in different wheat varieties were estimated using the BayesC method (Habier et al., 2011) in the "Julia for Whole-genome Analyses Software" (JWAS) package (Cheng et al., 2018) run in a Julia computing environment. In the single trait analysis of gluten epitope concentrations, variance components were estimated by fitting the model equation:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Ma} + \boldsymbol{e}$$

Where $\boldsymbol{y}$ was the vector of observed phenotypes, $\boldsymbol{b}$ was the vector of fixed effects, $\boldsymbol{a}$ was the vector of random additive SNP effects, $\boldsymbol{e}$ was the vector of residuals and $\boldsymbol{X}$ was the design matrix corresponding to $\boldsymbol{b}$ and $\boldsymbol{M}$ was the matrix of additive dosage covariates representing 0, 1 or 2 copies of the Illumina B allele. Fixed effects were the batch in which phenotypes were measured, the resulting batch effect was used to adjust gluten epitope concentration for reporting the phenotypic correlations. Epitope quantification was done in batches of 140 samples at a time, alongside a bulk standard. The batch adjustment accounts for technical variability between LCMS batches, within years, as determined by variation in the standard sample between batches. In the multi-year analyses, the year in which samples were grown is also fitted as a fixed effect to account for the seasonal variability between years. Gluten epitope concentrations are measured as ug of epitope per g of flour but are reported as deviations representing batch adjusted values. Phenotypic correlations were calculated using Pearson's correlation coefficient in RStudio (version 1.1.456) using the ggpubr package (Kassambara, 2023).

Bivariate analyses were conducted using the same model equation as the single trait model. These analyses were fit pairwise and applied to different growing seasons to estimate genetic correlations for gluten epitopes across seasons. Priors for the multiple trait genetic (G_0) and residual (R_0) variances between traits were determined using the posterior means from the single trait models to define the diagonal elements and an assumed genetic correlation of 0.8 and residual correlation of 0 to define the off-diagonal elements. Analyses were also run fitting an assumed genetic correlation of 0.5, however the choice of prior had little effect on posterior mean genetic correlations for epitope concentrations. This provides confidence that the phenotypic measurements themselves were providing most of the information for estimation. A Markov chain Monte Carlo (MCMC) approach was used to determine posterior mean estimates of heritability and other variance components. All analyses were run with a chain length of 100,000 samples after a burn-in of 1,000 samples that were discarded. All other samples were retained. Convergence of the MCMC chain was assessed by visual inspection. The MCMC data were assumed to have converged if the mean in the first portion of the chain was not different to the mean at the end of the chain and no prolonged variations in the mean were obvious. In all cases, MCMC chains exhibited evidence of convergence.

3.4 Results

3.4.1 Seasonal changes in epitope concentrations

The phenotypic variances for epitope concentrations are presented in Table 3.6. The distribution of epitope concentrations as deviations from the batch means (P1-P6) for each of the three growing seasons are presented as violin plots in Figure 3.2. The range in epitope concentrations appears to remain relatively consistent across years for each of the epitopes, although a slight reduction in phenotypic variance can be observed across consecutive seasons for the P2, P3 and P4 epitopes (Table 3.6). The most notable variation in epitope concentrations across seasons is the reduction in variance in the concentrations of P6 in Year 2 (Figure 3.2, Table 3.6). In most cases, epitope concentrations appear to follow a normal distribution, however notable outlier groups exist for concentrations of low values of P2 and high values of P3 epitopes (Figure 3.2).

Table 3.6: Phenotypic variance for batch adjusted gluten epitope concentrations (P1-P6) across the three growing seasons

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
Year 1	593	280	1,698	327	2929	22,747
Year 2	647	206	1,516	268	2620	5,823
Year 3	478	170	743	174	3074	17,648

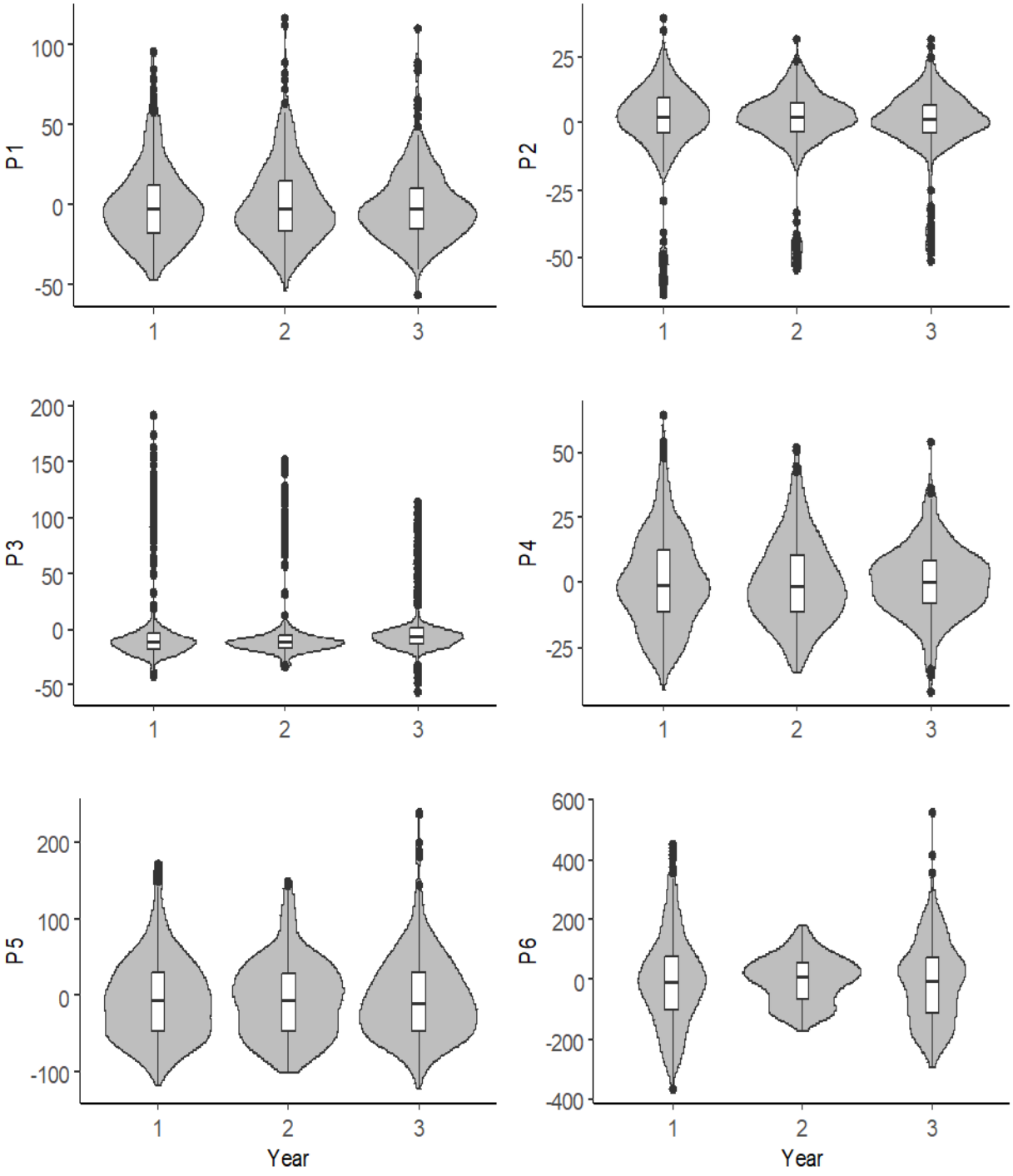

Figure 3.2: Violin plots illustrating the distribution in batch adjusted epitope concentrations (P1-P6) for each of the 3 growing seasons (Year)

Positive phenotypic correlations of between 0.4-0.65 are observed for the combined concentrations of P1-P6 in each of the three growing seasons (P<0.001; Figure 3.3) Concentrations of each of the individually tested gluten epitopes (P1-P6) appear to be significantly positively correlated between growing seasons with correlation coefficients (r) ranging between 0.41-0.92 (P<0.001; Figure 3.4). Phenotypic correlations were always lower between Year 1 and Year 3 than either between Year 1 and Year 2 or between Year 2 and Year 3 (Figure 3.3; Figure 3.4).

Significant positive correlations can be observed when comparing the sum total concentration of P1-P6 epitopes between years (Figure 3.3), however when this is broken down to comparing concentrations of individual epitopes against each other, the relationships are less clear. For example, near zero phenotypic correlations can be observed between the concentrations of P1 versus either P2 or P3, between P2 and P6, between P3 and P4 and between P5 and P6 (Figure 3.5). Moderate correlations, some positive and some negative, can be observed between the concentrations of some epitopes (e.g. P1 vs P6, P1 vs P4, P2 vs P5, P3 vs P5 and P4 vs P6; Figure 3.5) however some of these appear to be affected by outlier groups.

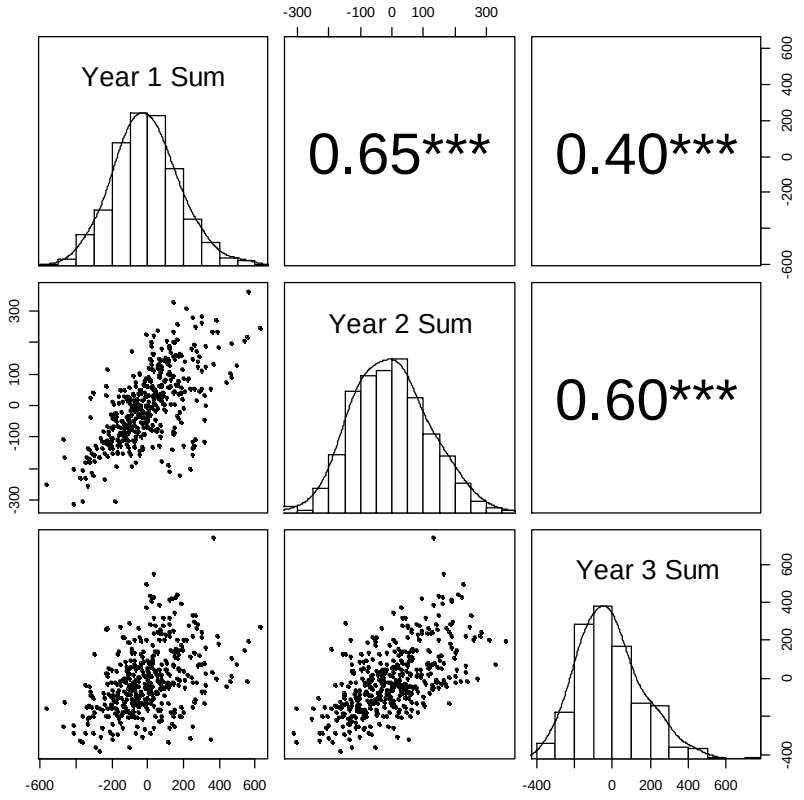

*Figure 3.3: Scatter matrix for phenotypic correlations of total gluten epitope concentrations (Sum = total P1-P6 concentration deviations) between each of the three growing seasons. Pearson's correlation coefficient (r) is presented in upper right panels with significance level (*** = P<0.001)*

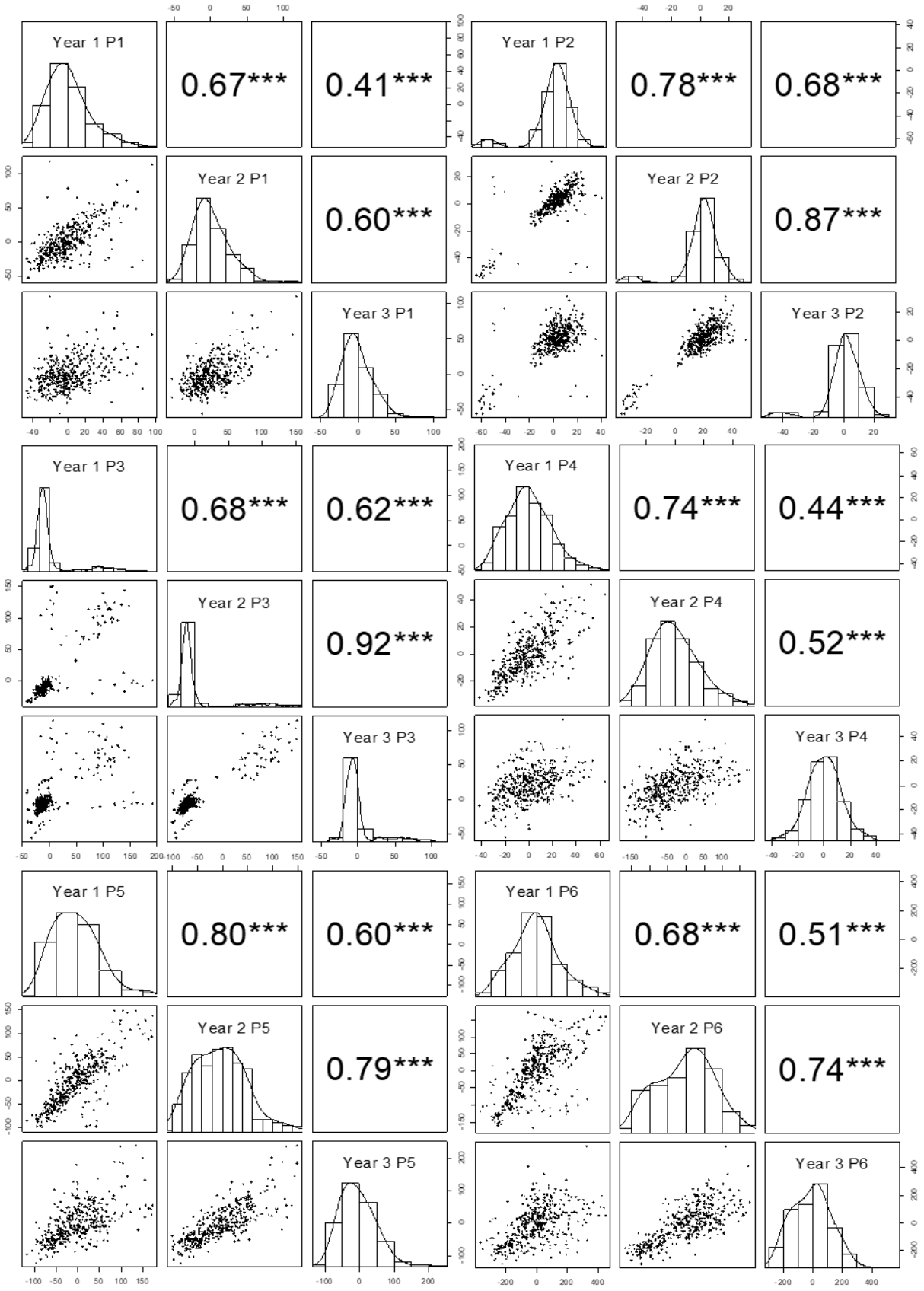

*Figure 3.4: Scatter matrix for batch adjusted epitope concentrations across 3 growing seasons. Pearson's correlation coefficient (r) is presented in upper right panels with significance level (*** = P<0.001).*

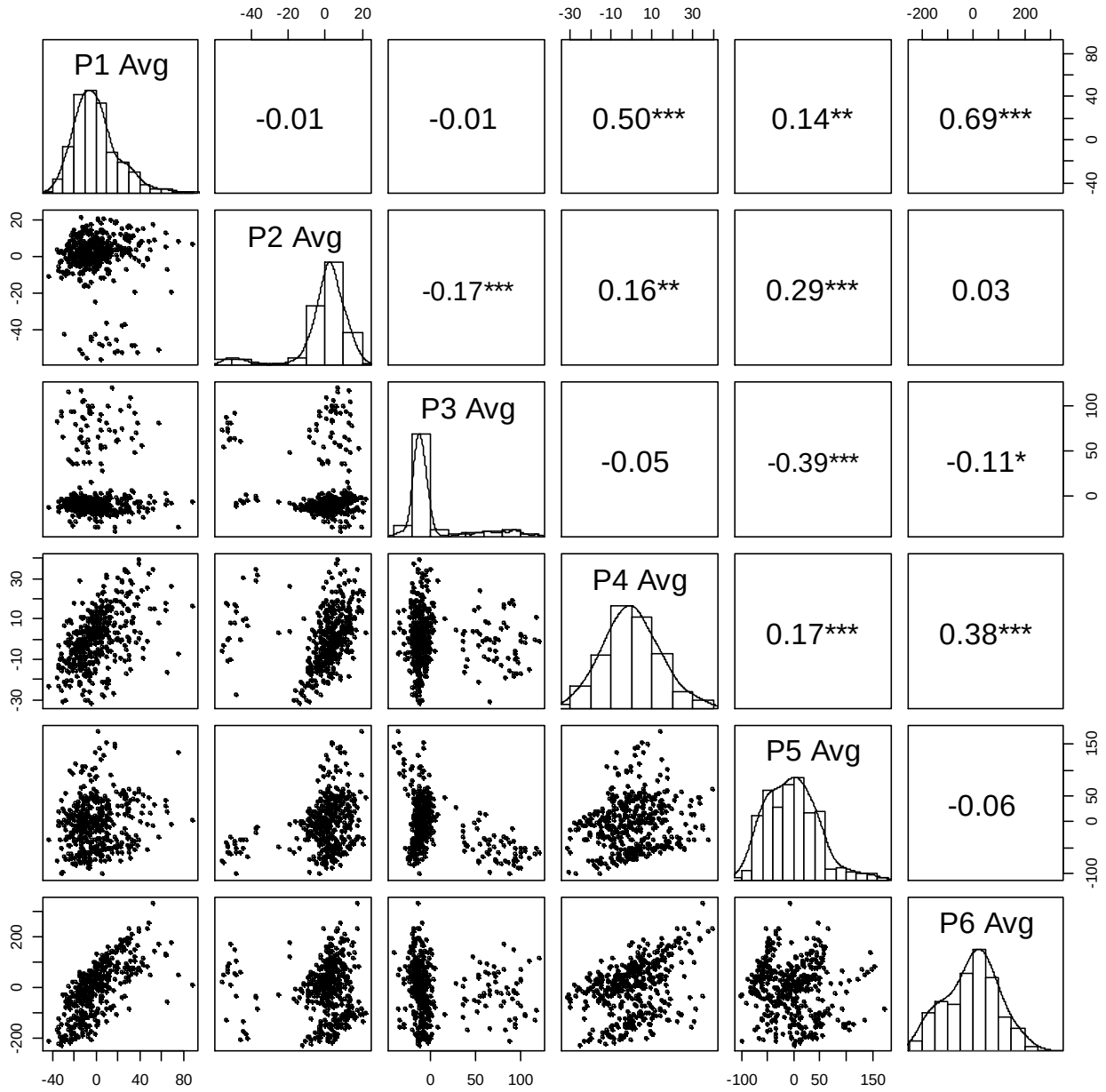

Figure 3.5: Scatter matrix for batch adjusted epitope concentrations averaged over three years (P1 Avg-P6 Avg). Correlation coefficients (r) between concentrations of different epitopes are presented in upper right panel with significant level (= P<0.05; ** = P<0.01; *** = P<0.001)*

3.4.2 Heritability analysis

Heritability estimates (h^2) were calculated for each of the gluten epitopes (P1-P6) along with the total P1-P6 concentrations for each of the individual growing seasons as well as by combining all 3 seasons. The mean h^2 estimates are presented alongside the 95% credibility intervals in Table 3.7.

In the single year analyses, the width in 95% credibility intervals ranged from as low as 0.08 for the P2 epitope in Year 2, up to 0.25 for the P6 epitope in Year 3. However, in the combined analysis, mean h^2 estimates became more similar for all epitopes while credibility intervals narrowed (Table 3.7). In the

combined analysis, mean h^2results ranged from a low of 0.44 for the P1 epitope, up to 0.75 for P5 while the h^2 estimate for total epitope concentration was calculated at 0.52 (Table 3.7).

Table 3.7: Mean and Credibility Intervals (CI) for heritability estimates for each epitope (P1-P6) and total epitope concentration (Sum) for each individual growing season (Year 1, Year2, Year 3) and combined seasons (3 Years)

Mean	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Sum**
Year 1	0.38	0.74	0.48	0.61	0.63	0.58	0.58
Year 2	0.53	0.93	0.91	0.61	0.81	0.77	0.67
Year 3	0.35	0.72	0.64	0.41	0.56	0.48	0.37
3 Years	0.44	0.74	0.68	0.54	0.75	0.53	0.52
CI upper (95%)							
Year 1	0.48	0.8	0.56	0.69	0.74	0.66	0.69
Year 2	0.62	0.96	0.95	0.68	0.85	0.85	0.76
Year 3	0.45	0.8	0.73	0.52	0.65	0.63	0.50
3 Years	0.48	0.77	0.71	0.58	0.77	0.59	0.58
CI lower (95%)							
Year 1	0.28	0.66	0.39	0.53	0.53	0.48	0.45
Year 2	0.44	0.88	0.86	0.53	0.75	0.7	0.56
Year 3	0.25	0.64	0.54	0.31	0.47	0.38	0.24
3 Years	0.39	0.72	0.65	0.5	0.72	0.48	0.45

3.4.3 Genetic correlations

Mean genetic correlations were estimated at 0.62 or above for all gluten epitopes (P1-P6) in all of the combined years analysis (Table 3.8). Correlations were lowest when comparing Year 1 data with Year 3 data and apart from the results for the P4 epitope, were highest when comparing the Year 2 data with the Year 3 data (Table 3.8). The range in mean correlation estimates between each of the bivariate analyses, for each epitope were as low as 0.08 and 0.09 for the P2 and P5 epitopes respectively and as high as 0.19 and 0.22 for the P4 and P3 epitopes respectively. When the total epitope concentrations were obtained for each year, mean genetic correlations between years remained within the range of mean correlations calculated for each individual epitope in each of the multi-year analyses (Table 3.8). Genetic correlations between different gluten epitopes ranged from -0.39 to 0.7 (Table 3.9), generally in line with the phenotypic correlations reported in Figure 3.4.

Table 3.8: Mean and credibility Intervals (CI) for estimates of genetic correlation between years for each epitope (P1-P6) and total combined epitope concentrations (Sum)

Mean	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Sum**
yr1_yr2	0.72	0.82	0.78	0.81	0.92	0.84	0.89
yr1_yr3	0.66	0.82	0.75	0.62	0.84	0.76	0.72
yr2_yr3	0.80	0.90	0.97	0.72	0.93	0.93	0.84
CI upper (95%)							
yr1_yr2	0.80	0.86	0.83	0.87	0.96	0.90	0.94
yr1_yr3	0.79	0.88	0.82	0.76	0.91	0.84	0.83
yr2_yr3	0.87	0.92	0.98	0.80	0.96	0.95	0.91
CI lower (95%)							
yr1_yr2	0.65	0.77	0.73	0.73	0.88	0.77	0.84
yr1_yr3	0.52	0.76	0.67	0.48	0.76	0.66	0.59
yr2_yr3	0.72	0.88	0.96	0.60	0.90	0.90	0.75

Table 3.9: Mean genetic correlations between concentrations of different gluten epitopes (P1-P6).

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
P1	1	0.03	0.03	0.51	0.09	0.70
P2		1	-0.14	0.17	0.28	0.03
P3			1	-0.05	-0.39	-0.11
P4				1	0.10	0.37
P5					1	-0.04
P6						1

Data were collected for Flour Protein Content (FPC) and Thousand Grain Weight (TGW) across the three growing seasons. The phenotypic correlations and spread of results are presented in Figure 3.6. Significant positive correlations can be observed in all cases.

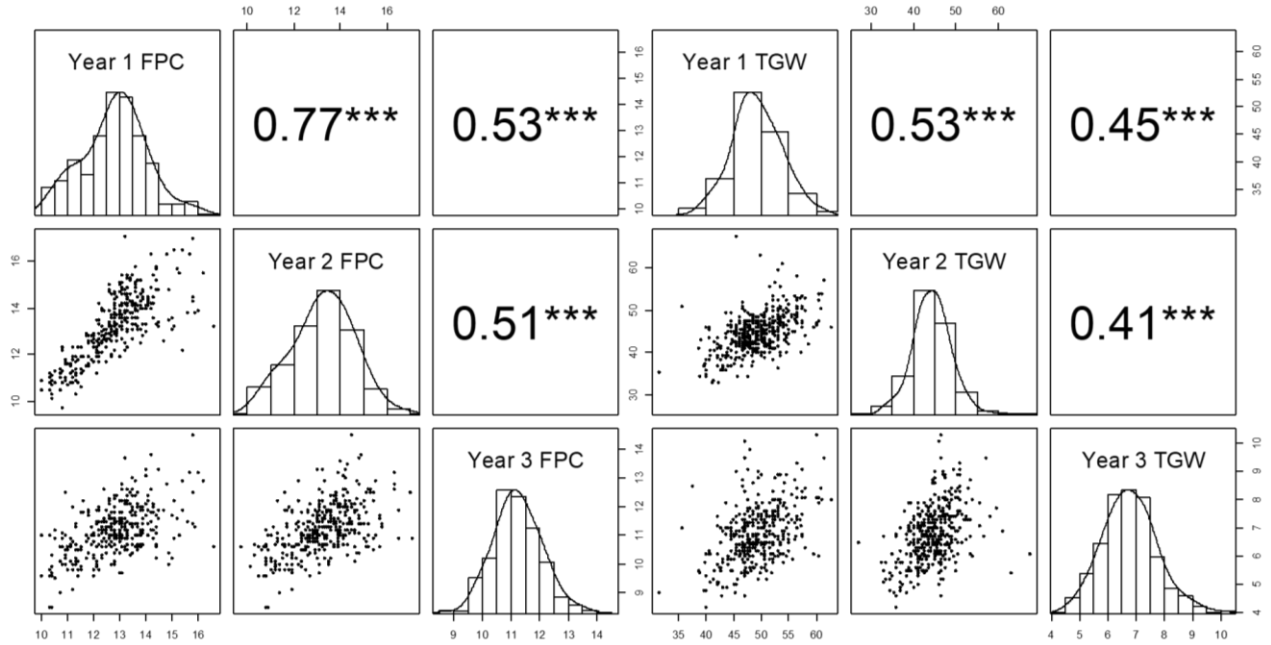

*Figure 3.6: Scatter matrices for Flour Protein Content (FPC %) and Thousand Grain Weight (TGW g) across the three growing seasons. Pearson's correlation coefficient (r) is presented in upper right panels with significance level (*** = P<0.001)*

3.5 Discussion

In order to successfully breed to change the mean of a complex quantitative trait, the phenotype needs to be determined to a significant extent by genotypic factors (reflected by heritability) and to be relatively stable across different environments (reflected by genetic correlations). The rate of genetic gain in breeding can be influenced by a number of factors such as selection intensity over time, selection accuracy, heritability and phenotype variability (Kelly, 2011; Richards et al., 2010). The heritability of a trait and its naturally occurring variability are outside of a breeder's control but may be influenced by genotype x environment (GxE) interactions. To determine the breeding potential of a new trait, it is important to understand the role that genotypic and environmental factors can have on the phenotype. In this study, the concentrations of different gluten epitopes (P1-P6) were measured in 471 wheat lines over three consecutive growing seasons to determine the heritability of epitope concentrations as well as the stability of epitope concentrations across environments. The concentrations of each of the tested gluten epitopes appear to be significantly correlated between growing seasons (Figure 3.4), although the strength of these correlations does vary depending on which two seasons are being compared. Of particular note, is that phenotypic correlations between Years 1 and 3 were always less than those recorded between both Years 1 and 2 and Years 2 and 3. This is an unexpected trend which may simply be a result of variations in environmental conditions across years but may also indicate possible degradation in sample quality across the three years potentially through genetic outcrossing or differences in sample preparation. Additionally, the reduction in variation across years for some epitopes (Table 3.6) may also point to potential sample degradation. Some individual samples also produced unexpected results where their recorded epitope concentrations were high in one year but low in another year. Eleven of these samples were re-tested to determine if there was a sampling error. This was done by re-extracting the gliadin fraction from the original flour samples for analysis, however there was no appreciable differences in the re-tested samples. Wheat lines with highly variable phenotypes such as these are undesirable in breeding programs. It is important that breeders establish cultivars with reliable phenotypes to ensure that growers know what to expect when they plant a particular variety. In the case of gluten epitope concentrations, this is also important for the baking industry if they are aiming to produce specialty, low gluten epitope wheat products.

Flour Protein Content (FPC) and Thousand Grain Weight (TGW) were measured in all samples across the three years and are included to assist in determining whether or not there has been degradation in the quality of epitope phenotyping. Correlations for these traits between years were significant but were always highest when comparing data between Year 1 and Year 2 than when comparing either Year 1 or 2 with Year 3 (Figure 3.4). This trend may be more easily explained by environmental conditions due to the abnormal spring/summer period in Year 3 where solar radiation was reduced due to regular rainfall and overcast skies (Figure 3.1), during a critical period for grain development. These results suggest that issues such as sample

contamination or outcrossing in the field are less likely and if degradation in sampling has occurred over the three years, it may be limited only to epitope quantification. If degradation was the cause, the most likely explanation may be linked to the degradation of reagents in the lab across time, however this remains inconclusive. Even when considering the unexpected trend in phenotypic correlations, moderate-strong positive correlations can be observed for all of the tested epitopes between different years. These data indicate that by at least maintaining similar crop management practices across seasons, breeders may be able to reliably select for epitope concentrations each year, across multiple years. This is further supported by the estimates for genetic correlations, where data for each epitope in each year were compared between years. Genetic correlationsgreater than 0.8 for a particular trait measured in different environments are generally considered high enough for breeders to select for this trait in a single environment. This is because a high genetic correlation across environments indicates that a trait is influenced by the same set of genes in a similar manner, regardless of environmental variables (Hill, 2001; Neyhart et al., 2019). Mean genetic correlation estimates for the P2 and P5 epitopes were above 0.8 in all of the multi-year analyses, while the average correlation across all of the analyses was also above 0.8 for the P3 and P6 epitopes (Table 3.8). For the P1 and P4 epitopes however, the mean estimate for genetic correlation across all of the multi-year analyses were 0.73 and 0.72 respectively, suggesting that the concentrations of these epitopes, although still highly correlated and primarily determined by a specific set of genotypic factors, are possibly more influenced by genotype-environmental interactions than the other tested epitopes. This is also in line with the phenotypic correlations where the P1 and P4 epitopes generally displayed the weakest correlations across years than the other tested epitopes. The h^2 estimates were also calculated to further determine GxE interactions for gluten epitope concentrations. Additive genetic effects were included in the model calculations for h^2 and due to the nature of SNP models, linkage effects were also captured which would account for some or all of the non-additive effects. This means that h^2 calculations in this study are not clearly broad sense or narrow sense but include variables for both (Furlotte et al., 2014). In wheat breeding programs, traits with an estimated h^2 of greater than 0.7 are generally considered to be highly heritable while h^2 values below 0.4 are considered to be low (Plavsin et al., 2021). In line with the estimates for genetic correlations, in the combined 3 Year analysis of h^2, the mean h^2 estimates for the P2 and P5 epitopes were greater than 0.7 while the lowest mean estimate for h^2 was 0.44 for the P1 epitope (Table 3.7). These results indicate that all of the gluten epitopes tested in this study are at least moderately to highly heritable, further supporting the conclusion that gluten epitope concentrations can be accurately selected for in different environments.

At the population level, it appears that concentrations of each of the tested epitopes are determined by a specific set of genotypic factors that are not highly influenced by environmental variables that differed across the three seasons. This is encouraging from the perspective of breeding towards lower epitope

cultivars as it indicates that breeding selections based on either genotype information or phenotypic measurements are likely to be effective. Furthermore, the range in epitope concentrations between cultivars is high. In all cases, the varieties with the highest concentration of each epitope are many times higher than the varieties with the lowest concentrations of each epitope. This suggests that the potential rate of genetic gain towards reduced epitope concentrations in a breeding population is also high. Furthermore, when comparing the range of epitope concentrations in the entire testing population against those lines that are commercially available to farmers and industry, low epitope germplasm is already available to breeders but yet to be exploited. For example, there are over 70 wheat lines in the testing population with lower total levels of epitope (P1-P6) than the lowest commercially available wheat line in New Zealand as of 2022, which itself has 62% higher levels of total average epitope than the lowest tested line. Further observations of the data also indicate the possible existence of large effect but rare, outlier alleles for some epitopes. Of particular interest are the small outlier groups seen in the concentrations of the P2 and P3 epitopes. In all three seasons, 22 wheat lines from a range of pedigrees, showed significantly lower levels of the P2 epitope than the rest of the population (Figure 3.4). The consistently low levels of P2 measured along with the magnitude of difference from the remainder of the population may indicate the presence of a small number of large effect alleles that if introgressed into the breeding population, could rapidly and significantly reduce the concentration of P2 epitope in that population. In contrast, a similar outlier group of 42 wheat lines presented significantly higher concentrations of the P3 epitope than the remainder of the population, once again indicating the possible presence of a small number of large effect alleles, however, in this case, a breeder would select away from such alleles in order to reduce the concentrations of P3 in the breeding population. Interestingly however, 12 of the 22 low P2 lines are also 12 of the 42 high P3 lines which may lead to negative correlations between the two phenotypes and challenges in simultaneously reducing the concentrations of both P2 and P3.

Total concentrations of P1-P6 have shown significant positive phenotypic correlations across years (Figure 3.3), suggesting that the wheat lines with high concentrations of one particular epitope are also likely to contain high concentrations of other epitopes and vice versa. This is further supported by the high genetic correlations seen in total epitope concentrations between years (Table 3.8). However, when breaking down the analysis to correlations between individual epitopes, it appears that this is not necessarily the case as correlations between individual epitopes ranged from -0.39 to 0.65 (Figure 3.5) and were rarely significant. In some cases, for example, when comparing concentrations of P3 against concentrations of P5, the significant negative correlation observed in the data becomes near zero when the outlier group of high P3 lines are removed. This is again the case when removing the outlier groups seen in the correlations between P2 and P5, P2 and P3, P3 and P6. In the analysis of genetic correlations between the concentrations of different epitopes, the results are in many ways similar to the phenotypic correlations observed in Figure

3.4, the strongest of which being between P1 and P6 at a mean of 0.7 with all other mean correlations being well below this (Table 3.9). These data suggest that the results for total epitope concentrations may be skewed based on the strong positive correlations seen between some epitopes, such as between P1 and P6 or P1 and P4. Additionally, the data could also be skewed towards the more abundant epitopes, particularly P6, which in many cases is found to be many multiple times higher than the concentrations of epitopes such as P1, P2 and P4. Along with this, P6 has shown to have high genetic correlations between years and so varieties with high levels of P6 in one year are also likely to have high levels of P6 in another year. Therefore, it is likely to be important that when selecting towards wheat varieties with lower levels of gluten epitope, breeders are best to consider the concentrations of individual epitopes across years. This is particularly the case for the P5 epitope for example which is believed to be more immunodominant than each of the other tested epitopes (Camarca et al., 2009; Shan et al., 2002), meaning that wheats with a higher concentration of total epitope but lower P5 may well be considered less immunogenic than the alternative. However, in order to be able to accurately predict the estimated immunogenicity of different wheat varieties, more work is certainly required for researchers to be able to quantify the toxicity of different epitopes in susceptible consumers.

Based on the generally high heritability's and positive phenotypic and genetic correlations of the different gluten epitopes, it appears that variations in epitope concentrations are primarily determined by genotype, rather than environmental variables. This is promising for breeders wishing to select towards lower epitope wheat cultivars as it suggests that either phenotypic or genomic based selection criteria are likely to be effective breeding tools. Although the data points to genetic factors being the dominant variable in determining epitope concentrations, the mode of action is still unclear. Research by Taranto et al. (2020), looking at SNP chip data in Durum wheat suggests that the genetic controls for some gluten epitopes are primarily regulatory rather than structural due to the lack of marker trait associations observed at specific gliadin loci. This suggests that the regulation of some epitopes may well be affected by other external variables, particularly at the individual level. Possible evidence of this can be observed when looking at the concentrations of the P2 and P3 epitopes in this study, a small number of wheat lines were recorded as having epitope concentrations significantly above the rest of the population in one year, yet the following year they were recorded as being significantly below the rest of the population. These data were double checked by re-testing the flour samples and the results remained consistent. This suggests that there may have been certain differences in environmental conditions between the two growing seasons that particularly affected the epitope concentrations of these lines. In this study, environmental variables were kept as consistent as possible across the three years, using a similar location (soil and micro-climate) and similar inputs each year. Therefore, our results do not consider a number of variables that would exist when growing wheat varieties on different farms in different locations. Such variables would include growing

degree days, nutrient availability, biotic and abiotic stresses and chemical inputs, all of which have been shown to affect gliadin content in wheat (Ronga et al., 2020). Therefore, more robust multi-environment data would be useful to better understand the role of environmental factors on epitope levels. Additionally, in order to establish a new wheat variety as a low epitope cultivar, breeders would need to ensure they have comprehensively tested their variety across environments to ensure epitope concentrations are stable.

3.6 References

Balkan, A. (2018). Genetic variability, heritability and genetic advance for yield and quality traits in m2-4 generations of bread wheat (*triticum aestivum l*.) genotypes. *Turkish Journal Of Field Crops*, 173-179. https://doi.org/10.17557/tjfc.485605

Becker, D., Wieser, H., Koehler, P., Folck, A., Muhling, K. H., Zorb, C. (2012). Protein composition and techno-functional properties of transgenic wheat with reduced a-gliadin content obtained by RNA interference. *Journal of Applied Botany and Food Quality*, *85*, 23-33.

Camarca, A., Anderson, R. P., Mamone, G., Fierro, O., Facchiano, A., Costantini, S., . . . Gianfrani, C. (2009). Intestinal T cell responses to gluten peptides are largely heterogeneous: implications for a peptide-based therapy in celiac disease. *J Immunol*, *182*(7), 4158-4166. https://doi.org/10.4049/jimmunol.0803181

Cebolla, A., Moreno, M. L., Coto, L., & Sousa, C. (2018). Gluten Immunogenic Peptides as Standard for the Evaluation of Potential Harmful Prolamin Content in Food and Human Specimen. *Nutrients*, *10*(12). https://doi.org/10.3390/nu10121927

Charmet, G. (2011). Wheat domestication: lessons for the future. *C R Biol*, *334*(3), 212-220. https://doi.org/10.1016/j.crvi.2010.12.013

Cheng, H., Fernando, R. L., & Garrick, D. J. (2018). JWAS: Julia implementation of whole-genome analysis software. Proceedings of the World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand.

Furlotte, N. A., Heckerman, D., & Lippert, C. (2014). Quantifying the uncertainty in heritability. *Journal of Human Genetics*, *59*(5), 269-275. https://doi.org/10.1038/jhg.2014.15

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*(1), 186. https://doi.org/10.1186/1471-2105-12-186

Hill, W. G. (2001). Genetic Correlation. In S. Brenner & J. H. Miller (Eds.), *Encyclopedia of Genetics* (pp. 823-825). Academic Press. https://doi.org/https://doi.org/10.1006/rwgn.2001.1423

Invitrogen. (2022). Quant-iT™ PicoGreen™ dsDNA Reagent and Kit. In T. F. S. Inc. (Ed.), *User Guide*.

Jouanin, A., Schaart, J. G., Boyd, L. A., Cockram, J., Leigh, F. J., Bates, R., . . . Smulders, M. J. M. (2019). Outlook for coeliac disease patients: towards bread wheat with hypoimmunogenic gluten by gene editing of alpha- and gamma-gliadin gene families. *BMC Plant Biol*, *19*(1), 333. https://doi.org/10.1186/s12870-019-1889-5

Kassambara, A. (2023). ggpubr: 'ggplot2' Based Publication Ready Plots. In.

Kelly, J. K. (2011). The breeder's equation. *Nature Education Knowledge*, *4*(5).

Lupton, F. G. H. (1987). History of wheat breeding. In F. G. H. Lupton (Ed.), *Wheat Breeding: Its scientific basis* (pp. 51-70). Springer Netherlands. https://doi.org/10.1007/978-94-009-3131-2_3

Neyhart, J. L., Lorenz, A. J., & Smith, K. P. (2019). Multi-trait Improvement by Predicting Genetic Correlations in Breeding Crosses. *G3 (Bethesda)*, *9*(10), 3153-3165. https://doi.org/10.1534/g3.119.400406

Ogilvie, O. J., Gerrard, J. A., Roberts, S., Sutton, K. H., Larsen, N., & Domigan, L. J. (2021). A Case Study of the Response of Immunogenic Gluten Peptides to Sourdough Proteolysis. *Nutrients*, *13*(6). https://doi.org/10.3390/nu13061906

Pallotta, M., Warner, P., Fox, R., Kuchel, H., Jefferies, S., & Langridge, P. (2003). Marker assisted wheat breeding in the southern region of Australia. Proceedings of the Tenth International Wheat Genetics Symposium,

Plavsin, I., Gunjaca, J., Satovic, Z., Sarcevic, H., Ivic, M., Dvojkovic, K., & Novoselovic, D. (2021). An Overview of Key Factors Affecting Genomic Selection for Wheat Quality Traits. *Plants (Basel)*, *10*(4). https://doi.org/10.3390/plants10040745

Richards, R. A., Rebetzke, G. J., Watt, M., Condon, A. G., Spielmeyer, W., & Dolferus, R. (2010). Breeding for improved water productivity in temperate cereals: phenotyping, quantitative trait loci, markers and the selection environment. *Functional Plant Biology*, *37*(2), 85-97. https://doi.org/https://doi.org/10.1071/FP09219

Ronga, D., Laviano, L., Catellani, M., Milc, J., Prandi, B., Boukid, F., . . . Francia, E. (2020). Influence of environmental and genetic factors on content of toxic and immunogenic wheat gluten peptides. *European Journal of Agronomy*, *118*. https://doi.org/10.1016/j.eja.2020.126091

Ruiz-Carnicer, A., Comino, I., Segura, V., Ozuna, C. V., Moreno, M. L., Lopez-Casado, M. A., . . . Sousa, C. (2019). Celiac Immunogenic Potential of alpha-Gliadin Epitope Variants from Triticum and Aegilops Species. *Nutrients*, *11*(2). https://doi.org/10.3390/nu11020220

Sanchez-Leon, S., Gil-Humanes, J., Ozuna, C. V., Gimenez, M. J., Sousa, C., Voytas, D. F., & Barro, F. (2018). Low-gluten, nontransgenic wheat engineered with CRISPR/Cas9. *Plant Biotechnol J*, *16*(4), 902-910. https://doi.org/10.1111/pbi.12837

Scherf, K. A., Catassi, C., Chirdo, F., Ciclitira, P. J., Feighery, C., Gianfrani, C., . . . Koehler, P. (2020). Recent Progress and Recommendations on Celiac Disease From the Working Group on Prolamin Analysis and Toxicity. *Front Nutr*, *7*, 29. https://doi.org/10.3389/fnut.2020.00029

Shan, L., Molberg Ø Fau - Parrot, I., Parrot I Fau - Hausch, F., Hausch F Fau - Filiz, F., Filiz F Fau - Gray, G. M., Gray Gm Fau - Sollid, L. M., . . . Khosla, C. (2002). Structural basis for gluten intolerance in celiac sprue. *Science*, *297*(5590), 2275-2279.

Sharma, N., Bhatia, S., Chunduri, V., Kaur, S., Sharma, S., Kapoor, P., . . . Garg, M. (2020). Pathogenesis of Celiac Disease and Other Gluten Related Disorders in Wheat and Strategies for Mitigating Them. *Front Nutr*, *7*, 6. https://doi.org/10.3389/fnut.2020.00006

Shewry, P. R., & Tatham, A. S. (2016). Improving wheat to remove coeliac epitopes but retain functionality. *Journal of Cereal Science*, *67*, 12-21. https://doi.org/https://doi.org/10.1016/j.jcs.2015.06.005

Taranto, F., D'Agostino, N., Catellani, M., Laviano, L., Ronga, D., Milc, J., . . . Francia, E. (2020). Characterization of Celiac Disease-Related Epitopes and Gluten Fractions, and Identification of Associated Loci in Durum Wheat. *Agronomy*, *10*(9). https://doi.org/10.3390/agronomy10091231

Tony Fischer, R. A. (2022). History of Wheat Breeding: A Personal View. In M. P. Reynolds & H.-J. Braun (Eds.), *Wheat Improvement: Food Security in a Changing Climate* (pp. 17-30). Springer International Publishing. https://doi.org/10.1007/978-3-030-90673-3_2

van den Broeck, H. C., van Herpen, T. W., Schuit, C., Salentijn, E. M., Dekking, L., Bosch, D., . . . van der Meer, I. M. (2009). Removing celiac disease-related gluten proteins from bread wheat while retaining technological properties: a study with Chinese Spring deletion lines. *BMC Plant Biol*, *9*, 41. https://doi.org/10.1186/1471-2229-9-41

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., . . . Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnol J*, *12*(6), 787-796. https://doi.org/10.1111/pbi.12183

Wieser, H., Koehler, P., Folck, A., Becker, D. (2006). *Characterization of wheat with strongly reduced a-gliadin content.* 9th Annual Gluten Workshop, San Fransisco, California, USA.

Chapter 4

Associations between gluten epitope concentrations and baking quality characteristics in bread wheat (*T. aestivum*)

4.1 Abstract

Wheat varieties with reduced concentrations of gluten epitopes are becoming an area of interest for wheat breeders around the world. However, it is thought that epitope concentrations may be associated with the dough and bread quality characteristics of wheat. In this study, estimated phenotypic and genetic correlations were calculated between the concentrations of 6 different gluten epitopes (P1-P6) and a range of baking quality traits to determine the presence and strength of any associations. The quality traits tested include Flour Protein Content (FPC), Mixograph quality parameters, Farinograph Water Absorption (WA) and Development Time (DT), bread loaf volume, and bread loaf texture. The FPC and mixograph data were derived from a population of 471 different wheat lines while the farinograph and bread quality data were derived from 148 of those lines. Genotypic information was obtained using a 90K Illumina wheat SNP chip. Moderate-strong positive phenotypic and genetic correlations were observed between FPC and each of the measured quality traits, with the exception of FPC and the mixograph parameter Midline Peak Time (MPT) where only weak to near zero positive correlations were observed. Estimates for mean genetic correlations were highest between FPC and the two farinograph traits, WA (0.78) and DT (0.75) as well as the baking quality traits Loaf Volume (0.76) and Loaf Texture (0.89). In most cases, weak-moderate positive correlations were observed between epitope concentrations and baking quality traits, however these were always lower than the correlations observed between FPC and baking quality. Estimates for genetic correlations between epitope concentrations and baking quality traits were almost always below 0.7. A partial relationship analysis was used to estimate mean genetic correlations between epitope concentrations andbaking quality characteristics whilst controlling for the genetic effects of FPC. In almost all cases, mean genetic correlations between epitope concentrations and baking quality were reduced when the effects of FPC were excluded from analysis. Concentrations of the P4 epitope generally had the greatest associations with baking quality traits with mean genetic correlations ranging from 0.48-0.80 for most baking quality metrics and mean phenotypic correlations ranging from 0.24-0.77. The P3 and P5 epitopes generally showed the weakest associations with baking quality. The highest mean phenotypic correlation observed between P3 and any of the baking quality metrics was 0.14, whilst the highest for P5 was 0.27. Additionally, in the partial relationship analysis of genetic correlations excluding the effects of FPC, mean genetic correlations for both P3 and P5 were consistently below 0.35. Due to the consistently low mean estimates for genetic and phenotypic correlations between concentrations of most gluten epitopes and the different baking quality metrics, these results indicate that breeders may be able to use a targeted selective breeding approach to breed new wheat cultivars with reduced concentrations of specific epitopes while maintaining high levels of baking quality. An effective metric for such selections may be a low epitope to total protein ratio.

4.2 Introduction

Wheat (*Triticum aestivum*) is one of the most widely grown crops in the world and is a major source of carbohydrates and protein in the modern diet, contributing more calories than any other cereal crop (Kumar et al., 2011; Sharma et al., 2020). What makes wheat unique is the presence of gluten proteins found in the grain. Gluten is a plant storage protein and its purpose, from a biological perspective, is to provide the seed with the energy and biochemical resources required to germinate and establish its root system (Shewry, 2019). However, from a human perspective, gluten provides wheat flour with highly unique viscoelastic and mechanical properties that make it incredibly useful for a vast swathe of food products (Biesiekierski, 2017; Shewry, 2019). Gluten proteins are made up of two primary subunits, gliadins and glutenins which can be further characterised into α/β, γ and δ gliadins, or low molecular weight (LMW) and high molecular weight (HMW) glutenins (Biesiekierski, 2017; Tosi et al., 2009). The subunits each have unique properties and form together to create what is known as the gluten matrix. That matrix of hundreds of different proteins results in gluten that varies significantly between different wheat cultivars and from different growing conditions (Wieser, 2007).

While gluten is vital to produce a significant range of products, such as bread, pasta, cakes, pastries, noodles and so on, it also contains compounds that trigger serious health conditions in some consumers. Gluten epitopes are specific peptide sequences that exist primarily within the gliadin fraction of gluten (Biesiekierski, 2017). Those peptides are recognised in the gut of coeliac patients as antigens, triggering an adaptive immune response which leads to gut inflammation. This inflammation causes tissue damage, releasing human transglutaminase, an enzyme which itself further modifies these gluten epitopes, amplifying their immunogenicity (Dewar et al., 2004; Kivelä et al., 2021).Immune responses to gluten include symptoms such as vomiting, nausea and diarrhoea. Gluten epitopes may also be responsible for triggering similar symptoms in non-coeliac patients who suffer from gluten intolerance (Scherf et al., 2016). Some evidence suggests that reducing exposure to gluten epitopes in children susceptible to developing coeliac disease may reduce the risk of them ever developing the disease (Ivarsson, 2005; Ivarsson et al., 2002). Reduced exposure to gluten epitopes in patients already diagnosed with coeliac disease, is likely to prevent further irreversible damage to the gut lining. It is also possible that exposure to gluten epitopes is a trigger for symptoms of non-coeliac gluten sensitivity (NCGS), however the mechanisms of that condition remain poorly understood (Sharma et al., 2020). Work by Anderson et al. (2013); on the genetic controls for γ-gliadins in the wheat 'Chinese Spring' demonstrated how expressed sequence tags can vary between different cultivars and so could be used as a criterion for selecting specific alleles with differing epitope abundance. For these reasons, plant breeding efforts to produce new wheat cultivars with reduced levels of gluten epitope may well offer significant health benefits to consumers who are either at risk of developing coeliac disease or are currently suffering from NCGS.

Gliadins are known to be an important determinant of baking quality in wheat and therefore associations may also exist between gluten epitope concentrations and baking quality (Xue et al., 2019). Existing research shows that wheat lines with significant reductions in α-gliadins in 'Chinese Spring' deletion lines, have reduced baking quality properties compared to naturally occurring lines with higher levels of α-gliadins (van den Broeck et al., 2009). However, less research has been done to investigate the importance of specific gluten epitopes on baking quality. Gluten epitopes appear to only make up a fraction of the total gliadin content in wheat, for example, of more than 890 published amino acid sequences of α-gliadins, only 19 sequences contain the P5 epitope, one of the most immunodominant epitopes (Schalk et al., 2017). Therefore, it is thought that by specifically selecting away from target epitopes while maintaining acceptable levels of total gluten, breeders may be able to develop novel wheat cultivars which maintain desirable levels of baking quality but with reduced immunogenicity in affected consumers. The objective of this study was to characterise naturally occurring variation of 6 different gluten epitopes in a range of wheat cultivars and determine if the epitope composition contributes to specific baking quality parameters, such as mixing tolerance, dough strength, loaf texture, weight and volume, water absorption, work input and development time.

4.3 Materials & methods

4.3.1 Field trials

Experiments were conducted using 471 wheat lines grown from some of the germplasm resources at the New Zealand Institute of Plant & Food Research. These varieties were grown across two seasons (2020-2021 = Year 2; 2021-2022 = Year 3) in Lincoln, New Zealand. Trials were sown in mid-May each year as two row plots and harvested in late January each year.

4.3.2 Flour protein content (FPC)

The FPC was determined using near-infrared spectroscopy (NIR). Whole-grain flour samples were milled through a 0.5mm screen on a Foss CT293 CyclotecTM (Foss Analytical Co., Ltd., Suzhou, China)then measurements were recorded using a NIR Technicon Infralyzer 450 (Technicon Ireland Ltd., Swords, Co., Dublin). The NIR was calibrated annually to meet industry criteria set by the New Zealand Flour Millers Association.

4.3.3 Mixograph testing

A 10g bowl mixograph was used to obtain three key metrics for dough quality,these being Midline Peak Time (MPT), Midline Peak Value (MPV) and Midline Peak Width (MPW). MPT being the time (mins) to peak dough development, MPV (%) being dough strength at peak development and MPW (%) being reflective of dough tolerance to overmixing. Mixogram data were calculated using MixSmart software

(National Manufacturing Division, TMCO, Lincoln, NE). The mixograph was operated at 25°C to ensure consistency throughout the day. A fixed water absorption of 65% was maintained across all samples, meaning for each sample, 10 g of flour were mixed in the mixograph bowl with 6.5 mls of water for a total of 5 minutes. Flour samples were prepared using a BrabenderQuadrumat Junior Experimental Mill (Brabender OHG, Duisberg, Germany). Mixograph testing was carried out on samples from the Year 2 field trial.

4.3.4 Bake testing

Bake testing was carried out on samples grown in the Year 3 field trial. 148 lines were selected for bake testing from 471 lines that were genotyped and assessed for gluten epitopes. Testing numbers were limited due to the extensive resource requirements for bake testing. The lines chosen for further testing were selected to include a diverse range of material. This was achieved by creating 4 separate subsets. Each subset being for lines cross-classified for: high or low total P1-P6 epitope concentrations; and high or low MPV. Subsets were of equal sizes, each containing 37 lines.

Flour for bake testing was prepared using a Buhler MLU-202 Automatic Laboratory Mill. Samples were tempered at 27°C to a moisture content of 16-16.5%, 18-24 hours prior to milling. The bran and pollard fractions of the milled samples were cleaned using a Buhler Laboratory Impact Finisher (MLU-302) and the white flour portion of these fractions were combined with the break and reduction flours from the Buhler mill with the remaining bran and pollard discarded. All of the white flour fractions were combined by mixing for 15 mins in a Hobart Mixer to make up the final flour sample for bake testing. Samples were kept at room temperature for 7 days post milling before being stored at -4°C until further testing.

Optimum water absorption (WA) and development time (DT) for these samples was determined using a Brabender Farinograph with a 50g mixing bowl(Brabender OHG, Duisberg, Germany) operated at 30°C. The WA was determined as the percentage of water required to achieve a midline mixing peak of 500 Brabender Units (BU). The DT was determined as the time (mins) from beginning mixing to the midline peak. This procedure is based on that described in the AACC International Method 54-21.01 (AACC, 2000).

Bake tests were conducted using 200 g of flour per sample combined with 3 g of salt, 1 g of softener, 1 g of improver and 6 g of fresh compressed yeast. These ingredients were mixed with water in a 400 g paddle mixer built at the New Zealand Institute for Plant & Food Research. The amount of water added was calculated using the WA values obtained from the farinograph test. The mixer temperature was maintained at 22°C for all samples. The mixer was run at 200rpm for the first 20 seconds of mixing before being increased to 400rpm. All samples were mixed to a work input of 10.5 wH, in line with industry standards. Once mixed, the dough was proofed for 10 mins at 32°C and then moulded using a Mono Universal Dough Moulder (Mono Equipment Ltd, Swansea, UK). The moulded dough is then proofed again at 40°C and 80% humidity for 40 mins before being baked at 200°C for 20 mins. Loaf volume is

determined using a Volscan Profiler volume analyser (Stable Mycrosystems, Surrey, UK). Loaf Texture is a subjective score based on cell size and structure of the bread. Scores are between 1-10 with 1 being a bread that hasn't risen and has little to no cell structure and 10 being a well risen bread with large well distributed cells. Scores were determined against photographed standards to ensure consistency. Two technical replicates were baked for each milled sample and the loaf volume and loaf texture scores presented are the average of the two replicates.

4.3.5 Epitope quantification and genotyping

The concentrations of 6 gluten epitopes (P1-P6) were measured each season as described in Chapter 3. A 90K Illumina SNP bead chip was used to conduct genotyping on all lines. Those processes are described in Chapter 3.

4.3.6 Statistical analysis

Variance components were estimated for individual traits using the BayesC method (Habier et al., 2011) and a Markov chain Monte Carlo (MCMC) approach in the Julia Whole-genome Analyses Software (JWAS) package (Cheng et al., 2018) run in a Julia computing environment. The MCMC chain length was 100,000 samples after a burn-in of 1,000 samples that were discarded. Every 10th sample was retained. Variance components were estimated by fitting the model equation:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Ma} + \boldsymbol{e}$$

Where **y** was the vector of observed phenotypes, **b** was the vector of fixed effects, **a** was the vector of random additive SNP effects, ***e*** was the vector of residuals, **X** was the design matrix corresponding to **b** and **M** was the matrix of additive dosage covariates representing 0, 1 or 2 copies of the Illumina B allele. In the case of gluten epitope concentrations, fixed effects were the batch in which epitope concentrations were measured and the resulting batch effect was used to adjust gluten epitope concentration for reporting the data. This adjustment accounts for technical variability between LCMS batches, within years. Gluten epitope concentrations are measured as ug of epitope per g of flour but are reported as deviations representing batch adjusted values.

Bivariate analyses were conducted using the same model equation as the single trait model. These analyses were fit pairwise to multiple traits to estimate phenotypic, genetic and residual correlations between traits. The posterior mean used for estimating correlations was derived from the mean of the MCMC chain. Priors for the multiple trait genetic (G0) and residual (R0) var-covariances were determined using the posterior means from the single trait models to define the diagonal elements and assumed genetic correlations of 0.8 and residual correlations of 0 to define the off-diagonal elements. Convergence of the MCMC chain was assessed by visual inspection. The MCMC data were assumed to have converged if the mean in the first

portion of the chain was not different to the mean at the end of the chain and no prolonged variations in the mean were obvious. In all cases, MCMC chains exhibited evidence of convergence.

Partial relationship analyses were conducted to estimate the genetic correlation between individual baking quality traits and gluten epitope concentrations whilst controlling for the genetic effects of FPC. The variance components from the three traits in each analysis were obtained from the respective bivariate models. These data were used to create a 3x3 matrix where the diagonal elements contained the genetic variance of each trait, and the off-diagonal elements were the co-variance estimates between two traits. This was converted to a 2x2 matrix using the methods described by Lu et al. (2015) to exclude the effects of FPC on the genetic variance of the remaining traits. Estimates for genetic correlations were then calculated and compared to the original genetic correlations to determine the effects of epitope concentrations on baking quality traits independent of FPC.

4.4 Results

4.4.1 Associations between epitope concentrations, FPC and mixograph quality parameters

The distribution of results for FPC, MPT, MPV and MPW are shown in Figure 4.1. Each of these traits appear to be normally distributed around the mean, although a small group of outliers with high values for MPT and a small group of outliers with low values for MPV do exist (Figure 4.1). Epitope concentrations are plotted against FPC, MPT, MPV and MPW in the Appendix (Figures 8.1-8.6).

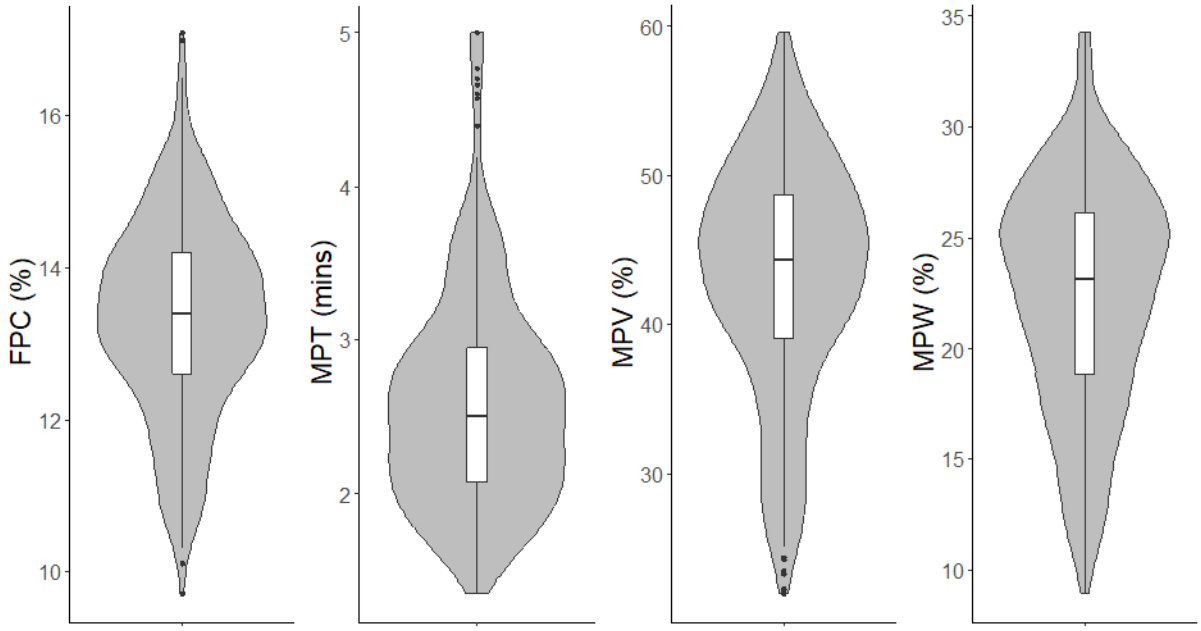

Figure 4.1: Violin plots illustrating the distribution of data for Flour Protein Content (FPC) and mixograph quality parameters in all 471 lines (MPT=Midline Peak Time, MPV=Midline Peak Value, MPW=Midline Peak Width)

Estimates for phenotypic and genetic correlations between mixograph quality parameters and gluten epitope concentrations are in Tables 4.1-4.2. The standard deviations for phenotypic and genetic correlation estimates are consistently low (Table 4.1, Table 4.2). Mean estimates for phenotypic and genetic correlations are generally very similar (Table 4.1, Table 4.2). The main exceptions being between FPC and mixograph quality where the mean estimates for genetic correlations were higher than the phenotypic correlations. The mean estimates for genetic correlations between P6 concentrations and MPT as well as between total epitope concentrations (P1-P6) and MPT were also higher than the corresponding phenotypic correlations.

Mean estimates for genetic correlations between FPC and epitope concentrations ranged between 0-0.8 while the mean estimates for phenotypic correlations ranged from 0.02-0.77. In both cases, the lowest mean correlation estimates were observed between P3 concentrations and FPC while the highest correlations were observed between P4 and FPC (Table 4.1; Table 4.2). Mean genetic and phenotypic correlations between P3 concentrations and each of the mixograph quality traits were also consistently near zero. Genetic correlations between FPC and the traits MPV and MPW were always higher than those observed between epitope concentrations and either MPV or MPW except in the case of P4 concentrations where genetic correlations were similar to those observed between FPC and both MPV and MPW. However, these correlations were still consistently below 0.8. Concentrations of the P4 epitope also appear to show the strongest mean phenotypic correlations with both MPV (0.61) and MPW (0.41) than any of the other epitopes and in this case were higher than the mean phenotypic correlations observed between FPC and both MPV and MPW. Estimates for mean genetic and phenotypic correlations between epitope concentrations and MPT were consistently low in all cases.

Table 4.1: Phenotypic correlations between FPC, gluten epitope concentrations (P1-P6; Sum = Total concentrations of P1-P6) and mixograph quality parameters (MPT = Midline Peak Time; MPV = Midline Peak Value; MPW = Midline Peak Width) measured in all 471 lines; (95% CI = 95% Credibility Intervals; Std = Sample Standard Deviation)

	FPC			**MPT**			**MPV**			**MPW**		
	Mean	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**
FPC	-	-	-	0.05	-0.01,0.11	0.04	0.37	0.31,0.42	0.04	0.30	0.24,0.36	0.04
P1	0.57	0.51,0.62	0.03	-0.03	-0.11,0.04	0.04	0.37	0.31,0.44	0.04	0.20	0.12,0.28	0.04
P2	0.17	0.10,0.24	0.04	-0.04	-0.11,0.03	0.04	0.19	0.13,0.25	0.04	0.08	0.02,0.15	0.04
P3	0.02	-0.03,0.08	0.04	0.02	-0.03,0.08	0.04	0.13	0.07,0.19	0.04	0.05	-0.01,0.11	0.04
P4	0.77	0.74,0.80	0.02	-0.09	-0.17,-0.02	0.04	0.61	0.56,0.65	0.03	0.41	0.35,0.46	0.04
P5	0.26	0.19,0.33	0.04	-0.08	-0.15,-0.01	0.04	0.11	0.04,0.18	0.04	0.12	0.06,0.19	0.04
P6	0.51	0.46,0.55	0.03	0.04	-0.03,0.10	0.04	0.37	0.32,0.43	0.03	0.24	0.19,0.29	0.03
Sum	0.66	0.62,0.70	0.02	-0.03	-0.09,0.05	0.04	0.51	0.45,0.56	0.03	0.31	0.25,0.37	0.04

Table 4.2: Genetic correlations between FPC, gluten epitope concentrations (P1-P6; Sum = Total concentrations of P1-P6) and mixograph quality parameters (MPT = Midline Peak Time; MPV = Midline Peak Value; MPW = Midline Peak Width) measured in all 471 lines; (95% CI = 95% Credibility Intervals; Std = Sample Standard Deviation)

	FPC			MPT			MPV			MPW		
	Mean	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**
FPC	-	-	-	0.22	0.10,0.35	0.08	0.65	0.57,0.72	0.05	0.57	0.49,0.67	0.05
P1	0.59	0.49,0.65	0.05	0.19	0.05,0.32	0.08	0.37	0.26,0.48	0.07	0.20	0.07,0.34	0.08
P2	0.18	0.09,0.27	0.06	-0.03	-0.13,0.06	0.06	0.19	0.11,0.27	0.05	0.09	-0.01,0.18	0.06
P3	0.00	-0.07,0.07	0.04	0.01	-0.07,0.09	0.05	0.12	0.04,0.20	0.05	0.03	-0.06,0.12	0.05
P4	0.80	0.77,0.86	0.03	0.02	-0.11,0.15	0.08	0.66	0.59,0.73	0.04	0.48	0.39,0.58	0.06
P5	0.24	0.14,0.34	0.06	-0.02	-0.13,0.08	0.06	0.10	0.01,0.19	0.06	0.14	0.05,0.24	0.06
P6	0.48	0.41,0.55	0.04	0.16	0.07,0.27	0.06	0.34	0.26,0.42	0.05	0.25	0.18,0.32	0.04
Sum	0.68	0.61,0.73	0.04	0.21	0.09,0.32	0.07	0.52	0.44,0.61	0.05	0.37	0.27,0.47	0.06

4.4.2 Associations between epitope concentrations and bake testing parameters

The distribution of data for each of the baking quality parameters tested in this study are presented in Figure 4.2. In most cases, the data appear to be normally distributed. Some outliers exist with high development times and loaf volumes. Data for loaf texture appears to be bimodal, with many average scores of either 8.0 or 9.0. Epitope concentrations are plotted against WA, DT, Loaf Volume and Loaf Texture in the Appendix (Figures 8.7-8.12).

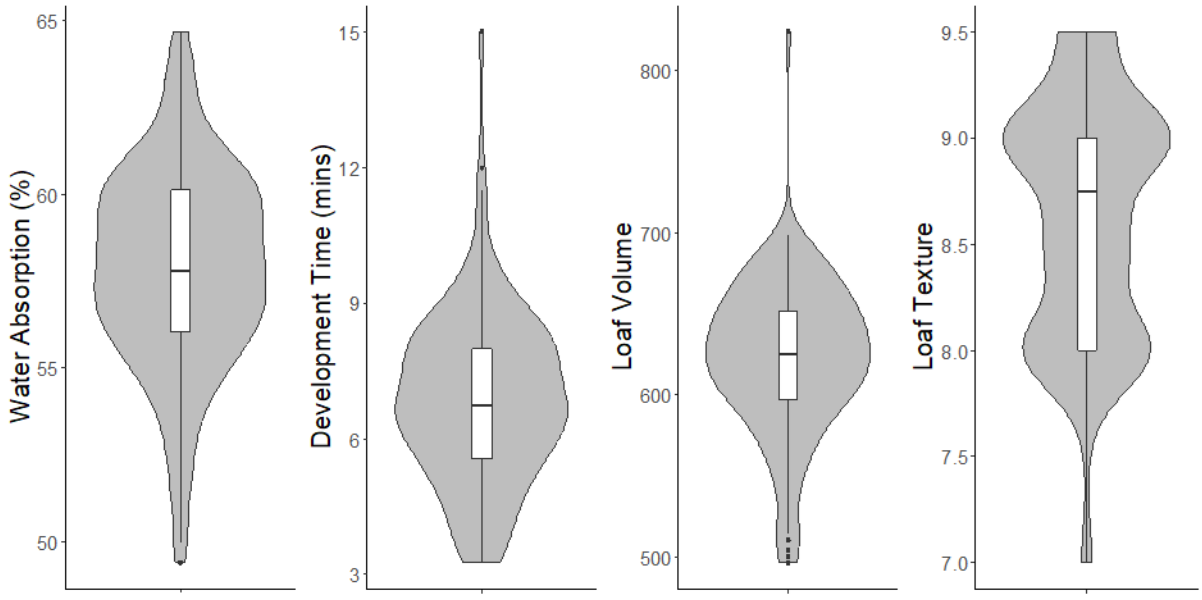

Figure 4.2: Violin plots illustrating the distribution of data for different baking quality metrics in the 148 baked lines.

Estimates for phenotypic and genetic correlations between baking quality parameters and gluten epitope concentrations are in Tables 4.3-4.4. Standard deviations for estimates of genetic correlations are relatively

high (Table 4.4) while the standard deviations for phenotypic correlations are consistently low (Table 4.3). Genetic correlations between FPC and baking quality parameters are consistently higher than those observed between epitope concentrations and baking quality. The P4 epitope appears to have the strongest genetic correlations with WA (0.57), DT (0.64) and loaf volume (0.66) than any of the other tested epitopes. A similar mean estimate for the genetic correlation between P4 concentrations and loaf texture (0.61) can also be observed (Table 4.4). However, the mean estimates for genetic correlations between loaf texture and both the P6 epitope (0.75) and total epitope concentrations (P1-P6); (0.82) are higher. A similar trend can be observed for mean estimates of phenotypic correlations between epitope concentrations and baking quality traits; however, these estimates are consistently lower than those for the corresponding genetic correlations.

Table 4.3: Phenotypic correlations between FPC, gluten epitope concentrations (P1-P6; Sum = Total concentrations of P1-P6) and baking quality parameters (WA = Water Absorption; DT = Development Time; Volume = Loaf Volume; Texture = Loaf Texture) measured in the 148 baked lines; (95% CI = 95% Credibility Intervals; Std = Sample Standard Deviation)

	WA			**DT**			**Volume**			**Texture**		
	Mean	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**
FPC	0.52	0.42,0.61	0.06	0.58	0.49,0.67	0.05	0.52	0.42,0.61	0.06	0.59	0.50,0.67	0.05
P1	0.21	0.07,0.35	0.09	0.15	0.01,0.30	0.08	0.21	0.08,0.33	0.08	0.20	0.07,0.33	0.08
P2	0.07	-0.07,0.21	0.08	0.18	0.05,0.32	0.08	0.20	0.06,0.33	0.08	0.25	0.12,0.38	0.08
P3	0.14	-0.01,0.29	0.09	0.09	-0.04,0.23	0.08	0.11	-0.02,0.24	0.08	0.09	-0.05,0.22	0.08
P4	0.34	0.21,0.45	0.07	0.45	0.34,0.55	0.06	0.26	0.13,0.38	0.08	0.24	0.12,0.36	0.07
P5	0.16	0.04,0.29	0.08	0.27	0.14,0.39	0.07	0.28	0.16,0.40	0.07	0.17	0.04,0.30	0.08
P6	0.26	0.12,0.38	0.08	0.21	0.07,0.34	0.08	0.27	0.15,0.38	0.07	0.30	0.17,0.42	0.08
Sum	0.29	0.16,0.41	0.08	0.30	0.18,0.43	0.08	0.36	0.24,0.47	0.07	0.34	0.21,0.46	0.07

Table 4.4: Genetic correlations between FPC, gluten epitope concentrations (P1-P6; Sum = Total concentrations of P1-P6) and baking quality parameters (WA = Water Absorption; DT = Development Time; Volume = Loaf Volume; Texture = Loaf Texture) measured in the 148 baked lines; (95% CI = 95% Credibility Intervals; Std = Sample Standard Deviation)

	WA			DT			Volume			Texture		
	Mean	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**	**Mean**	**95% CI**	**Std**
FPC	0.78	0.69,0.88	0.06	0.75	0.63,0.87	0.07	0.76	0.63,0.89	0.08	0.89	0.69,0.92	0.07
P1	0.38	0.07,0.60	0.16	0.34	0.10,0.55	0.15	0.38	0.09,0.63	0.16	0.53	0.19,0.79	0.18
P2	0.43	0.18,0.65	0.15	0.40	0.09,0.64	0.17	0.38	0.05,0.69	0.19	0.42	0.07,0.69	0.19
P3	0.23	0.01,0.51	0.15	0.18	-0.02,0.43	0.14	0.25	-0.01,0.53	0.16	0.39	0.11,0.67	0.17
P4	0.57	0.38,0.74	0.11	0.64	0.46,0.83	0.11	0.66	0.43,0.86	0.13	0.61	0.45,0.84	0.12
P5	0.27	0.07,0.56	0.15	0.43	0.23,0.61	0.12	0.56	0.35,0.75	0.12	0.58	0.30,0.75	0.14
P6	0.37	0.18,0.56	0.12	0.22	-0.02,0.42	0.13	0.39	0.14,0.65	0.15	0.75	0.42,0.84	0.13
Sum	0.46	0.26,0.69	0.13	0.39	0.17,0.62	0.14	0.57	0.33,0.75	0.13	0.82	0.59,0.89	0.09

4.4.3 Estimates for genetic correlations independent of FPC

In almost all cases, the mean estimates for genetic correlations between gluten epitope concentrations and each of the tested quality parameters is reduced when the genetic effects of FPC are excluded from analysis (Table 4.5). With the effects of FPC excluded, the P4 epitope continues to show the strongest mean estimate for genetic correlations between MPV (0.6) and MPW (0.25) than any of the other tested epitopes (Table 4.5). In contrast, the mean estimates for genetic correlations between epitope concentrations and WA are all reduced to below 0.24 when the genetic effects of FPC are excluded from analysis. Mean genetic correlations are also reduced between epitope concentrations and both DT and loaf volume, the most notable being for P6 concentrations where the genetic correlation is reduced to -0.18 against DT and -0.12 against loaf volume. In this case, the highest estimates for genetic correlations with DT are for P4 (0.36) and P5 (0.31) concentrations while the highest estimate with loaf volume if for P5 concentrations (0.34).

Table 4.5: Genetic correlations (rg) between gluten epitope concentrations (P1-P6; Sum = total concentration of P1-P6) and both mixograph quality traits in all 471 lines and baking quality traits in the 148 baked lines (MPV = Midline Peak Value; MPW = Midline Peak Width; WA = Water Absorption; DT = Development Time; Volume = Loaf Volume) excluding the genetic effects of Flour Protein Content (rg/FPC). (Change in rg = change in genetic correlations when excluding the effects of FPC vs when including the effects of FPC).

	MPV		**MPW**		**WA**		**DT**		**Volume**	
	rg/FPC	**Change in rg**	**rg/FPC**	**Change in rg**	**rg/FPC**	**Change in rg**	**rg/FPC**	**Change in rg**	**rg/FPC**	**Change in rg**
P1	0.11	-0.37	-0.11	-0.31	0.23	-0.15	0.13	-0.21	0.14	-0.24
P2	0.12	-0.15	0.02	-0.07	0.20	-0.23	0.18	-0.22	0.26	-0.12
P3	0.14	-0.06	0.04	0.01	0.11	-0.12	0.05	-0.13	0.11	-0.14
P4	0.60	-0.13	0.25	-0.23	0.17	-0.40	0.36	-0.28	0.15	-0.51
P5	-0.02	-0.21	0.04	-0.10	-0.05	-0.32	0.31	-0.12	0.34	-0.22
P6	0.12	-0.30	0.05	-0.20	0.06	-0.31	-0.18	-0.40	-0.12	-0.51
Sum	0.28	-0.33	0.08	-0.29	0.12	-0.34	0.02	-0.37	0.14	-0.43

4.5 Discussion

Baking quality in wheat is a highly complex trait, there are many different variables that determine the overall quality of a cultivar and in some cases, overall quality can be highly subjective (Kumar et al., 2019). In this study, a range of quality metrics have been assessed using industry accepted methods. These data were then compared against the concentrations of 6 different gluten epitopes (P1-P6) to identify whether any associations exist between epitope concentrations and baking quality. Phenotypic correlations between gluten epitope concentrations and FPC varied for different epitopes. Weak positive or near zero correlations were observed between P2, P3 and P5 concentrations and FPC while moderate to strong positive correlations could be observed between the remaining epitopes and FPC. The strongest being between P4 concentrations and FPC (0.77). Estimates for genetic correlations between FPC and epitope concentrations were also variable but in line with the observed phenotypic correlations. The genetic correlation between P4 and FPC was estimated to be 0.8, suggesting a strong association with FPC. However, genetic correlations between the concentrations of all of the other measured epitopes were below 0.6, suggesting weaker associations. These results are not unexpected as gluten content has been shown to be highly correlated with FPC in flour (Cato & Mullan, 2020). Therefore, it is expected that high protein wheats may also contain high concentrations of certain gluten epitopes when compared to varieties with lower protein. However, excluding the correlations observed between P4 concentrations and FPC, these correlations are not particularly strong, suggesting that targeted selections may still allow breeders to select for individuals with a low ratio of epitope to FPC, particularly for the P2, P3 and P5 epitopes where genetic correlations with FPC were all below 0.25. In the case of P5, this is particularly encouraging as the P5 epitope is considered to be the most immunodominant of all the previously described gluten epitopes (Camarca et al., 2009; Shan et al., 2002).

Protein content is often used in the milling and baking industries as an easy to measure proxy for baking quality and while this can provide an indication of potential dough strength, it certainly lacks the nuances of other baking quality metrics (Xue et al., 2019). This is likely due to the complex interactions that exist between gliadins and glutenins which are often more important in determining overall baking quality than total gluten content (Biesiekierski, 2017). For example, when comparing FPC with different mixograph quality parameters, phenotypic correlations tend to be relatively weak (Table 4.1). This is particularly the case when comparing FPC with MPT. The MPT is a measure of the time it takes for a flour sample to reach peak dough development, this is when a dough has its maximum consistency and optimum gas retention (Chung et al., 2001). For industry purposes, it is desirable to have a dough that achieves peak dough development as quickly as possible to enable efficient, cost-effective mixing (Cauvain, 2012). However, MPT has little association with dough strength and is likely determined more by the gliadin:glutenin ratio in the flour than the protein content, meaning that FPC is a poor predicter of MPT (Barak et al., 2013). The traits

MPV and MPW however, have previously been shown to have a greater association with dough strength than MPT (Labuschagne et al., 2016; Sun et al., 2015) which may explain the weak-moderate positive phenotypic correlations that can be observed between FPC and both MPV and MPW. A similar trend can be observed when comparing the concentrations of different gluten epitopes with the mixograph data. For example, it appears that a near zero phenotypic correlation exists between MPT and the concentrations of each epitope (Table 4.1). Additionally, the estimates for genetic correlations between MPT and epitope concentrations are also low with the highest being only 0.21 when comparing total epitope concentrations (P1-P6) with MPT results (Table 4.2). These data indicate that neither FPC nor epitope concentrations have any strong associations with MPT. In contrast to MPT, a number of moderate positive phenotypic correlations were observed when comparing MPV and MPW with concentrations of each of the tested gluten epitopes (Table 4.1). In particular, mean estimates for the phenotypic and genetic correlations between P4 concentrations and MPV were 0.61 and 0.66 respectively. These data suggest that moderate positive associations may exist between P4 concentrations and MPV. Furthermore, when the genetic effects of FPC are excluded from the analysis, the mean genetic correlation between P4 concentrations and MPV is only reduced to 0.60 (Table 4.5). This suggests that the P4 concentrations may contribute to MPV, irrespective of FPC. In contrast, with the effects of FPC removed, the genetic correlations observed between each of the other tested epitopes and MPV are reduced to between -0.02-0.14, suggesting that no major associations exist between each of these epitopes and MPV. Similar results can be seen in the partial relationship analysis between epitope concentrations and MPW. However, in this case, the associations between P4 concentrations and MPW are reduced to a greater extent than for MPV, suggesting that no major associations exist between any of the tested gluten epitopes and MPW.

A subset of the wheat varieties grown in this study received further quality testing, including Farinograph tests and bake tests. Data derived from the farinograph tests included DT and WA, with DT being measured in minutes and WA the percentage of water to flour required to achieve optimum mixing conditions. The DT is a similar metric to MPT, however the key differences are in the testing methods. The MPT is measured with a pin mixer at a fixed water absorption of 65% while the farinograph has a paddle mixer and the amount of water added is calculated for each sample so that peak mixing occurs at 500 Brabender Units. This ensures that the optimal amount of water is being added to each flour sample for bake testing. Estimates for genetic and phenotypic correlations between epitope concentrations and DT were generally higher than those observed between epitope concentrations and MPT. Again, P4 concentrations appear to have the strongest associations with DT than the other tested epitopes, although both the phenotypic and genetic correlations are still lower than those observed between FPC and DT (Table 4.3; Table 4.4). Mean genetic correlations between epitope concentrations and DT are again reduced when the genetic effects of FPC are excluded from the analysis (Table 4.5). Weak positive genetic correlations can still be observed

between DT and P4 (0.36) and P5 (0.31) concentrations, however these correlations remain low meaning that breeders may still be able to select for low epitope varieties without significantly compromising DT in their cultivars.

Water Absorption is another valuable trait in the milling and baking industries and can also be used as an indicator for dough strength with some research groups choosing to use it as a way to classify flours as being either weak, medium or strong (Serna-Saldivar, 2012). The WA also influences flour characteristics such as loaf volume, bread yield, machinability, proofing and shelf life and higher values for WA are generally favoured over lower values (Awuchi et al., 2019; Puhr & D'Appolonia, 1992; Zghal et al., 2001). In this study, individual gluten epitope concentrations and total gluten epitope concentrations (P1-P6) again showed only weak-moderate positive phenotypic and genetic correlations with WA (Table 4.3, Table 4.4). In contrast, the mean genetic correlation between FPC and WA was relatively high (0.78). Additionally, mean genetic correlations between epitope concentrations and WA were again reduced when the genetic effects of FPC were excluded (Table 4.5). These data again suggest that WA may have some genetic associations with FPC, but no major associations appear to exist between epitope concentrations and WA. This is consistent with other research which has shown moderate positive correlations between total flour protein and WA (Park, 2006; Rakszegi et al., 2014).

Two key metrics were obtained from the bake testing trials, these being loaf volume and loaf texture. Loaf volume is often considered one of the most important metrics for determining bread quality. Studies have shown that lower volume breads are more at risk of staling, have a poorly aerated crumb and poor gas retention (Axford et al., 1968; Sahi, 2014). Loaf texture on the other hand, is a measure of the bread cell structure and dough streaking. Loaf texture is primarily a sensory trait and breads with a lower texture score will have poorer consistency in structure and are considered less desirable by consumers (Annett et al., 2007; Lassoued et al., 2008). Moderate, positive phenotypic correlations were observed between FPC and both loaf volume and loaf texture (Table 4.3). These data are consistent with results from other groups who have shown similar associations between these traits (He, 1992; Johansson et al., 2001; Różyło, 2011). In most cases, weak positive phenotypic correlations were observed between epitope concentrations and both loaf volume and loaf texture (Table 4.3). In contrast, mean genetic correlations were generally higher, particularly for total epitope concentrations (P1-P6) and concentrations of P4, P5 and P6 (Table 4.4). However, in the partial relationship analysis of the individual effects of FPC and epitope concentrations on loaf volume, mean genetic correlations between epitope concentrations and loaf volume were reduced to below 0.35 in all cases, suggesting that much of the genetic associations observed between these traits may have been due to the genetic effects of FPC. These data further indicate that breeders may be able to select towards low gluten epitope traits in their wheat varieties while also maintaining acceptable baking quality characteristics. Furthermore, by breaking down the analysis to look at individual epitope concentrations, it

appears that P3 concentrations for example have little to no associations with any of the tested baking quality metrics, including FPC. Therefore, breeders can be confident that the baking quality attributes of their cultivars will not be compromised by low concentrations of the P3 epitope. Similar results can be seen for concentrations of the P5 epitope which showed only weak positive phenotypic and genetic correlations with FPC, DT and loaf volume. This is especially encouraging considering that the P5 epitope is considered the most immunodominant of the tested epitopes (Schalk et al., 2017). The FPC has consistently shown to be moderately-strongly associated with each of the baking quality metrics tested in this study, in line with many other studies assessing similar characteristics(Field et al., 1983; Johansson et al., 2001; Peña et al., 2005). In most cases however, epitope concentrations appear to have much weaker associations with baking quality, although associations with FPC do appear to exist in some cases, particularly for P4 concentrations. Therefore, a potential strategy for optimising breeding selections towards lines with lower levels of gluten epitope whilst retaining acceptable levels of baking quality, may be to select towards wheat varieties with a high protein to epitope ratio with a particular emphasis on P4 concentrations.

4.6 References

American Association of Cereal Chemists. 2000. Approved Methods ofthe AACC, 10th ed. Method 54-21.01. TheAssociation: St. Paul, MN.

Anderson, O. D., Huo, N., & Gu, Y. Q. (2013). The gene space in wheat: the complete gamma-gliadin gene family from the wheat cultivar Chinese Spring. *Funct Integr Genomics*, *13*(2), 261-273. https://doi.org/10.1007/s10142-013-0321-8

Annett, L. E., Spaner, D., & Wismer, W. V. (2007). Sensory Profiles of Bread Made from Paired Samples of Organic and Conventionally Grown Wheat Grain. *Journal of Food Science*, *72*(4), 254-260. https://doi.org/https://doi.org/10.1111/j.1750-3841.2007.00331.x

Awuchi, C., Victory, I., & Echeta, C. (2019). The Functional Properties of Foods and Flours. *International Journal of Advanced Academic Research*, *5*(11), 139-160.

Axford, D. W. E., Colwell, K. H., Cornford, S. J., & Elton, G. A. H. (1968). Effect of loaf specific volume on the rate and extent of staling in bread. *Journal of the Science of Food and Agriculture*, *19*(2), 95-101. https://doi.org/https://doi.org/10.1002/jsfa.2740190208

Barak, S., Mudgil, D., & Khatkar, B. S. (2013). Relationship of gliadin and glutenin proteins with dough rheology, flour pasting and bread making performance of wheat varieties. *LWT - Food Science and Technology*, *51*(1), 211-217. https://doi.org/https://doi.org/10.1016/j.lwt.2012.09.011

Biesiekierski, J. R. (2017). What is gluten? *Journal of Gastroenterology and Hepatology*, *32*, 78-81. https://doi.org/10.1111/jgh.13703

Camarca, A., Anderson, R. P., Mamone, G., Fierro, O., Facchiano, A., Costantini, S., . . . Gianfrani, C. (2009). Intestinal T cell responses to gluten peptides are largely heterogeneous: implications for a peptide-based therapy in celiac disease. *J Immunol*, *182*(7), 4158-4166. https://doi.org/10.4049/jimmunol.0803181

Cato, L., & Mullan, D. (2020). Wheat quality: Wheat breeding and quality testing in Australia. In S. P. Cauvain (Ed.), *Breadmaking (Third Edition)* (pp. 221-259). Woodhead Publishing. https://doi.org/https://doi.org/10.1016/B978-0-08-102519-2.00008-6

Cauvain, S. (2012). 2 - Breadmaking: an overview. In S. P. Cauvain (Ed.), *Breadmaking (Second Edition)* (pp. 9-31). Woodhead Publishing. https://doi.org/https://doi.org/10.1533/9780857095695.9

Cheng, H., Fernando, R. L., & Garrick, D. J. (2018). JWAS: Julia implementation of whole-genome analysis software. Proceedings of the World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand.

Chung, Ohm, J.-B., Caley, Seabourn, M., & Seabourn, B. (2001). Prediction of Baking Characteristics of Hard Winter Wheat Flours Using Computer-Analyzed Mixograph Parameters. *Cereal Chemistry*, *78*. https://doi.org/10.1094/CCHEM.2001.78.4.493

Dewar, D., Pereira, S. P., & Ciclitira, P. J. (2004). The pathogenesis of coeliac disease. *The International Journal of Biochemistry & Cell Biology*, *36*(1), 17-24. https://doi.org/10.1016/s1357-2725(03)00239-5

Field, J. M., Shewry, P. R., & Miflin, B. J. (1983). Solubilisation and characterisation of wheat gluten proteins: Correlations between the amount of aggregated proteins and baking quality. *Journal of the Science of Food and Agriculture*, *34*(4), 370-377. https://doi.org/https://doi.org/10.1002/jsfa.2740340409

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*(1), 186. https://doi.org/10.1186/1471-2105-12-186

He, H., & Hoseney, R. C. (1992). Effect of quantity of wheat flour protein on bread loaf volume. *Cereal Chemistry Journal*, *69*(1), 17-19.

Ivarsson, A. (2005). The Swedish epidemic of coeliac disease explored using an epidemiological approach--some lessons to be learnt. *Best Pract Res Clin Gastroenterol*, *19*(3), 425-440. https://doi.org/10.1016/j.bpg.2005.02.005

Ivarsson, A., Hernell, O., Stenlund, H., & Persson, L. Å. (2002). Breast-feeding protects against celiac disease. *The American Journal of Clinical Nutrition*, *75*(5), 914-921. https://doi.org/10.1093/ajcn/75.5.914

Johansson, E., Prieto-Linde, M. L., & Jönsson, J. Ö. (2001). Effects of Wheat Cultivar and Nitrogen Application on Storage Protein Composition and Breadmaking Quality. *Cereal Chemistry*, *78*(1), 19-25. https://doi.org/https://doi.org/10.1094/CCHEM.2001.78.1.19

Kivelä, L., Caminero, A., Leffler, D. A., Pinto-Sanchez, M. I., Tye-Din, J. A., & Lindfors, K. (2021). Current and emerging therapies for coeliac disease. *Nature Reviews Gastroenterology & Hepatology*, *18*(3), 181-195. https://doi.org/10.1038/s41575-020-00378-1

Kumar, A., Mantovani, E. E., Simsek, S., Jain, S., Elias, E. M., & Mergoum, M. (2019). Genome wide genetic dissection of wheat quality and yield related traits and their relationship with grain shape and size traits in an elite × non-adapted bread wheat cross. *PLoS ONE*, *14*(9), e0221826. https://doi.org/10.1371/journal.pone.0221826

Kumar, P., Yadava, R. K., Gollen, B., Kumar, S., Verma, R. K., & Yadav, S. (2011). Nutritional Contents and Medicinal Properties of Wheat: A Review. *Life Sciences and Medicine Research*.

Labuschagne, M. T., Lindeque, R. C., & van Biljon, A. (2016). Dough mixing characteristics measured by Mixsmart software as possible predictors of bread making quality in three production regions of South Africa. *Journal of Cereal Science*, *70*, 192-198. https://doi.org/https://doi.org/10.1016/j.jcs.2016.06.008

Lassoued, N., Delarue, J., Launay, B., & Michon, C. (2008). Baked product texture: Correlations between instrumental and sensory characterization using Flash Profile. *Journal of Cereal Science*, *48*(1), 133-143. https://doi.org/https://doi.org/10.1016/j.jcs.2007.08.014

Lu, Y., Vandehaar, M. J., Spurlock, D. M., Weigel, K. A., Armentano, L. E., Staples, C. R., . . . Tempelman, R. J. (2015). An alternative approach to modeling genetic merit of feed efficiency in dairy cattle. *J Dairy Sci*, *98*(9), 6535-6551. https://doi.org/10.3168/jds.2015-9414

Park, S. H., Bean, S. R., Chung, O. K. & Seib, P. A. (2006). Levels of Protein and Protein Composition in Hard Winter Wheat Flours and the Relationship to Breadmaking. *Cereal Chemistry Journal*, *83*(4), 418-423.

Peña, E., Bernardo, A., Soler, C., & Jouve, N. (2005). Relationship between common wheat (Triticum aestivum L.) gluten proteins and dough rheological properties. *Euphytica*, *143*(1), 169-177. https://doi.org/10.1007/s10681-005-3157-z

Puhr, D. H., & D'Appolonia, B. L. (1992). Effect of Baking Absorption on Bread Yield, Crumb Moisture, and Crumb Water Activity. *Cereal Chemistry Journal*, *69*(5), 582-586.

Rakszegi, M., Balázs, G., Bekes, F., Harasztos, A., Kovács, A., Láng, L., . . . Tömösközi, S. (2014). Modelling Water Absorption of Wheat Flour by Taking into Consideration of the Soluble Protein and Arabinoxylan Components. *Cereal Research Communications*, *42*, 629-639. https://doi.org/10.1556/CRC.2014.0007

Różyło, R., & Laskowski. J. (2011). Predicting Bread Quality (Bread Loaf Volume and Crumb Texture). *Polish Journal of Food and Nutrition Sciences*, *61*(1), 61-67.

Sahi, S. S., Little, K., and Ananingsih, V.K. (2014). Quality Control. In W. Zhou (Ed.), *Bakery Products Science and Technology* (2nd Edition ed., pp. 490-500). John Wiley & Sons, Ltd.

Schalk, K., Lang, C., Wieser, H., Koehler, P., & Scherf, K. A. (2017). Quantitation of the immunodominant 33-mer peptide from alpha-gliadin in wheat flours by liquid chromatography tandem mass spectrometry. *Sci Rep*, *7*, 45092. https://doi.org/10.1038/srep45092

Scherf, K. A., Koehler, P., & Wieser, H. (2016). Gluten and wheat sensitivities – An overview. *Journal of Cereal Science*, *67*, 2-11. https://doi.org/10.1016/j.jcs.2015.07.008

Serna-Saldivar, S. O. (2012). *Cereal Grains: Laboratory Reference and Procedures Manual* (1st ed.). CRC Press, Taylor & Francis Group, LLC. https://doi.org/https://doi.org/10.1201/b11726

Shan, L., Molberg Ø Fau - Parrot, I., Parrot I Fau - Hausch, F., Hausch F Fau - Filiz, F., Filiz F Fau - Gray, G. M., Gray Gm Fau - Sollid, L. M., . . . Khosla, C. (2002). Structural basis for gluten intolerance in celiac sprue. *Science*, *297*(5590), 2275-2279.

Sharma, N., Bhatia, S., Chunduri, V., Kaur, S., Sharma, S., Kapoor, P., . . . Garg, M. (2020). Pathogenesis of Celiac Disease and Other Gluten Related Disorders in Wheat and Strategies for Mitigating Them. *Front Nutr*, *7*, 6. https://doi.org/10.3389/fnut.2020.00006

Shewry, P. (2019). What Is Gluten-Why Is It Special? *Front Nutr*, *6*, 101. https://doi.org/10.3389/fnut.2019.00101

Sun, J., Yang, W., Liu, D., Zhao, J., Luo, G., Li, X., . . . Zhang, A. (2015). Improvement on Mixograph test through water addition and parameter conversions. *Journal of Integrative Agriculture*, *14*(9), 1715-1722. https://doi.org/https://doi.org/10.1016/S2095-3119(15)61078-4

Tosi, P., Parker, M., Gritsch, C. S., Carzaniga, R., Martin, B., & Shewry, P. R. (2009). Trafficking of storage proteins in developing grain of wheat. *J Exp Bot*, *60*(3), 979-991. https://doi.org/10.1093/jxb/ern346

van den Broeck, H. C., van Herpen, T. W., Schuit, C., Salentijn, E. M., Dekking, L., Bosch, D., . . . van der Meer, I. M. (2009). Removing celiac disease-related gluten proteins from bread wheat while retaining technological properties: a study with Chinese Spring deletion lines. *BMC Plant Biol*, *9*, 41. https://doi.org/10.1186/1471-2229-9-41

Wieser, H. (2007). Chemistry of gluten proteins. *Food Microbiology*, *24*(2), 115-119. https://doi.org/https://doi.org/10.1016/j.fm.2006.07.004

Xue, C., Matros, A., Mock, H.-P., & Mühling, K.-H. (2019). Protein Composition and Baking Quality of Wheat Flour as Affected by Split Nitrogen Application. *Frontiers in Plant Science*, *10*. https://www.frontiersin.org/articles/10.3389/fpls.2019.00642

Zghal, M. C., Scanlon, M. G., & Sapirstein, H. D. (2001). Effects of Flour Strength, Baking Absorption, and Processing Conditions on the Structure and Mechanical Properties of Bread Crumb. *Cereal Chemistry*, *78*(1), 1-7. https://doi.org/https://doi.org/10.1094/CCHEM.2001.78.1.1

Chapter 5

Genome wide association study for gluten epitope concentrations in bread wheat (*T. aestivum*)

5.1 Abstract

Gluten epitopes are peptides found in wheat grain and are known to trigger coeliac disease in affected consumers. A Genome Wide Association Study (GWAS) was used to identify significant genetic markers for concentrations of 6 different gluten epitopes (P1-P6) to allow for the development of genomic based breeding tools for low gluten epitope wheat cultivars. Two methods were used, the first being a single marker analysis of 26,386 Single Nucleotide Polymorphism (SNP) markers and the second being a Bayesian window approach using 14,385 mapped SNPs. The GWAS was conducted using 471 wheat lines sampled over three consecutive growing seasons. Genotypic information was collected using a 90K Illumina wheat SNP chip. Between 2-4 markers were identified as being significantly associated with concentrations of each of the P2, P3, P4 and P6 epitopes. Significant differences in mean epitope concentrations were identified between populations with opposing alleles for each of those markers ($P<0.001$). Those markers therefore have the potential to be utilised for Marker Assisted Selection (MAS) for wheat lines with reduced concentrations of the relevant epitopes. The results of the Bayesian window analysis were summarised by genomic location in terms of the window posterior probability of association (WPPA) for non-overlapping windows each containing 20 contiguous SNPs. This approach is able to capture the causal mutation effects that may be distributed across each of the markers in the relevant window and can therefore identify Quantitative Trait Loci (QTLs) that may not be identified in a single marker analysis. Three genomic windows were identified as having significant associations with epitope concentrations (WPPA > 0.8). Two windows, one spanning 559.4-593.7 Mb on chromosome 2B and the other spanning 14.8-17.7 Mb on chromosome 1B were shown to be significantly associated with concentrations of both the P3 and P5 epitopes. A third window at 423.8-435.3 Mb on chromosome 1B was significantly associated with P1 concentrations. These significant windows did not co-locate with the primary gliadin loci, suggesting that their effect may be regulatory rather than structural. No significant associations were observed between any of the genomic windows and concentrations of the P2, P4 or P6 epitopes. Nor were any associations identified between the genomic windows and any quality traits including Flour Protein Content, Mixograph quality or baking quality. Individual marker effects for epitope concentrations were generally small, particularly for markers within the windows that were identified as being significantly associated with epitope concentrations. The genomic windows identified in this study are likely to represent significant loci for associations with gluten epitope concentrations and further investigation of these regions may allow for MAS of low epitope wheats in the future.

5.2 Introduction

The key factors that determine the quality of wheat flour for milling and baking are the amount and quality of gluten proteins present in the grain (Biesiekierski, 2017). Many research groups have attempted to develop effective breeding strategies for improving the composition of gluten in their cultivars (Kiszonas & Morris, 2017). Research has shown total gluten content to be negatively correlated with grain yield, making it a difficult trait to target selection for improved flour quality without compromising on yield (Kibite & Evans, 1984; Simmonds, 1995). Therefore, gluten quality is often considered to be a more appealing target than total gluten content for wheat improvement. Gluten quality generally refers to the composition of gliadins and glutenins, the two primary subunits of gluten. Measuring gliadin and glutenin contents, however, is a time consuming and costly exercise and often impractical for commercial breeding programs. To overcome this, a number of research groups and breeders have begun to shift away from traditional selection methods in favour of genomic based selection (Lemeunier et al., 2022; Michel et al., 2018; Zhang-Biehn et al., 2021).

Gluten epitopes are specific peptide sequences that exist primarily in the α-gliadin fraction of gluten and are known to trigger coeliac disease in affected consumers (Sharma et al., 2020; Taranto et al., 2020). It is believed that reduced exposure to gluten epitopes may reduce the risk of susceptible consumers developing coeliac disease while also reducing the symptoms of gluten intolerance in affected consumers (Shewry & Tatham, 2016). Low gluten epitope grain is expected to attract a premium price from consumers with concerns around dietary gluten intake. Therefore, wheat breeders and researchers are increasingly interested in tools for developing low gluten epitope cultivars. Gluten epitope concentrations have been shown to be moderately to highly heritable with significant variations in epitope concentrations known to exist between cultivars (Chapter 3). This suggests that selective breeding practices are likely to be effective in developing low epitope wheats. However, measuring epitope concentrations is a costly and time-consuming exercise and unlikely to ever be suitable as a high-throughput method for selecting low epitope wheats. The aim of this study has been to identify genomic regions associated with concentrations of 6 different gluten epitopes in order to determine the practicality of genomic based selection methods for breeding low epitope wheat cultivars as an alternative to more traditional breeding methods.

Genome Wide Association Studies (GWAS) are an effective method for identifying loci associated with quantitative traits in diverse populations (Alseekh et al., 2021). They have become more practical with the development of high-throughput, low-cost single nucleotide polymorphism (SNP) genotyping. Bayesian methods are being regularly applied to conduct GWAS as they offer advantages over traditional regression models that tend to focus only on individual markers which have been shown in some cases to overestimate quantitative trait loci (QTL) effects and produce spurious associations (Kemper et al., 2018; Korte & Farlow,

2013; Wolc & Dekkers, 2022). In contrast, Bayesian approaches can be applied to specific genomic windows and be used to determine the window posterior probability of association (WPPA) as a way to manage the level of false positives (Fernando & Garrick, 2013). These methods tend to provide more accurate estimates of effect sizes, they can better account for linkage disequilibrium and the results are not reliant on sequence resolution markers (Tsai et al., 2020). The wheat genome contains a high proportion of repetitive sequences and many regions of non-coding DNA (Garbus et al., 2015; Zhang et al., 2021). Additionally, a high proportion of recorded SNPs remain unmapped, leaving a small number of useful markers for conducting GWAS for positional studies (Wen et al., 2017). Therefore, in this study we have used a Bayesian approach to conduct GWAS analyses to identify genomic regions associated with gluten epitope concentrations.

5.3 Materials & methods

5.3.1 Field trials

A population of 471 wheat lines were grown and phenotyped over three successive growing seasons (2019-2020 = Year 1; 2020-2021 = Year 2; 2021-2022 = Year 3), as described in Chapter 3. The trials were conducted in Lincoln, Canterbury with seeds sown in mid-May and plots harvested in late January. The genotypes used in this study were acquired from the germplasm resources at the New Zealand Institute for Plant & Food Research. The tested population included advanced breeding material crossed from as early as 1986 up until 2018, as well as commercial and ex-commercial cultivars from New Zealand, Australia, the United Kingdom or Canada. Climate data for the three seasons were presented in Chapter 3.

5.3.2 Phenotyping and genotyping

A range of phenotypic data were collected across the three growing seasons. Flour Protein Content (FPC) and concentrations of 6 gluten epitopes (P1-P6) measured as ug of epitope per g of flour were measured in all three seasons. Mixograph quality parameters were measured in Year 2, those traits included Midline Peak Time (MPT; mins), Midline Peak Width (MPW; %) and Midline Peak Value (MPV; %). Farinograph testing and bake testing was conducted in Year 3, those traits include Water Absorption (WA; %), Development Time (DT; mins), loaf volume and loaf texture. Genotyping was carried out using a 90K Illumina SNP bead chip on all of the phenotyped lines. All phenotyping and genotyping methods were described in Chapters 3 and 4. The presence or absence of awns was also recorded by visual observation.

5.3.3 Statistical analysis

A single marker GWAS for associations between individual SNPs and gluten epitope concentrations was conducted using the statgenGWAS package in R Studio (van Rossum & Kruijer, 2023). The analysis included unmapped markers and excluded markers with a minor allele frequency of <0.05. The Bonferroni correction was used to determine a conservative significance threshold of 5.7 ($-\log_{10}$) based on the 26,386 markers

tested and a p value of 0.05. The threshold was used to minimise the false detection rate (FDR) of associations with epitope concentrations. The aim of this analysis was to identify significant genetic markers that may act as effective tools for Marker Assisted Selection (MAS) of low gluten epitope wheat varieties. Effects of significant markers were quantified by comparing differences in concentrations of the relevant epitope between populations that were separated based on the allele carried for the relevant marker. Significant differences in mean epitope concentrations between populations carrying opposing alleles were determined using a two-tailed t-test. Sample sizes of the populations being compared were considered equal if the variance ratio was less than 4.

Marker effects for GWAS analysis of windows of 20 contiguous SNPs for gluten epitope concentrations were estimated using the BayesC method (Habier et al., 2011) in the Julia for Whole-genome Analyses Software (JWAS, version 1.1.2) package (Cheng et al., 2018) run in a Julia computing environment (version 1.7.0). A Markov chain Monte Carlo (MCMC) approach was conducted on single traits using the model equation:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Ma} + \boldsymbol{e}$$

Where ***y*** was the vector of observed phenotypes, ***b*** was the vector of fixed effects, ***a*** was the vector of random additive SNP effects, ***e*** was the vector of residuals,***X*** was the design matrix corresponding to ***b*** and ***M*** was the matrix whose columns are additive dosage covariates representing 0, 1 or 2 copies of the Illumina B allele at a particular locus. Fixed effects were the batch in which epitope concentrations were measured and the year that samples were grown. These analyses were run with Pi = 0.99 and a chain length of 100,000 samples after a burn-in of 10,000 samples that were discarded. Every 10th samplewas retained after the burn-in period.The value for Pi used in the analysis was selected due to epitope concentrations being a complex quantitative trait. This means that epitope concentrations are believed to be influenced by many small effect genes and therefore it can be expected that at least 1% of markers will have non-zero effects (Lee et al., 2017). Convergence of the MCMC chain was assessed by visual inspection. The MCMC data were assumed to have converged if the mean in the first portion of the chain was not different to the mean at the end of the chain and no prolonged variations in the mean were obvious. In all cases, MCMC chains exhibited evidence of convergence.

A GWAS analysis was performed on concentrations of 6 gluten epitopes (P1-P6) individually as well as a trait defined by the sum of all these epitope concentrations. This analysis was done for each of three growing seasons and again in a multi-year analysis with data combined from all three seasons. The GWAS analysis was limited to mapped markers with a minor allele frequency of less than 5% and these markers were assigned to non-overlapping windows of 20 contiguous SNPs for summarising QTL effects based on the methods proposed by Fernando and Garrick (2013) and Fernando et al. (2014). That approach accounts for the linkage disequilibrium within the population and aims to minimise false positives. A total of 14,385

markers were included in the GWAS, representing 729 windows across the genome. Windows on the tail end of each chromosome contained from 3 to 19 markers. The GWAS results were used to calculate the window posterior probabilities of association (WPPA) and the proportion of genetic variance (GV) attributed to each window. Manhattan plots were produced using the StatsPlots (version 0.15.4) package in Julia. Genomic windows with WPPA > 0.8 were considered to be significantly associated with the tested phenotype. This threshold has previously been shown to be effective for declaring significant associations between genomic windows and a range of phenotypes (Wang et al., 2020). Window Breeding Values (BVs) were calculated for genomic windows that were associated with concentrations of multiple epitopes. The window BVs were determined by multiplying the corresponding allelic values of the markers at the relevant genomic window by the posterior mean effects of the markers within that window. Pearson's correlation coefficient was used to determine the strength and direction of the window effects for epitope concentrations that were associated with the same genomic window.

Population structure was assessed based on SNP marker composition by Principal Component Analysis (PCA) using singular value decomposition in the vcfR package in R Studio (Knaus&Grünwald, 2017). Markers with a minor allele frequency of less than 5% were removed prior to that analysis. Epitope concentrations and FPC are compared with the distribution of Principal Components 1 and 2 in Appendix Figure 8.13.

5.4 Results

5.4.1 Single marker analysis

Manhattan plots for the single marker analysis of associations between SNPs and epitope concentrations are presented in Figure 5.1 and Figure 5.2. Markers were identified as being significantly associated with P2, P3, P4 and P6 concentrations, however, no major loci with multiple significant markers were observed (Figure 5.1, Figure 5.2). Summary statistics for effects of significant markers on gluten epitope concentrations are presented in Table 5.1. Violin plots are presented in Figures 5.3-5.6 to illustrate differences in epitope concentrations for populations carrying opposing alleles for significant markers.

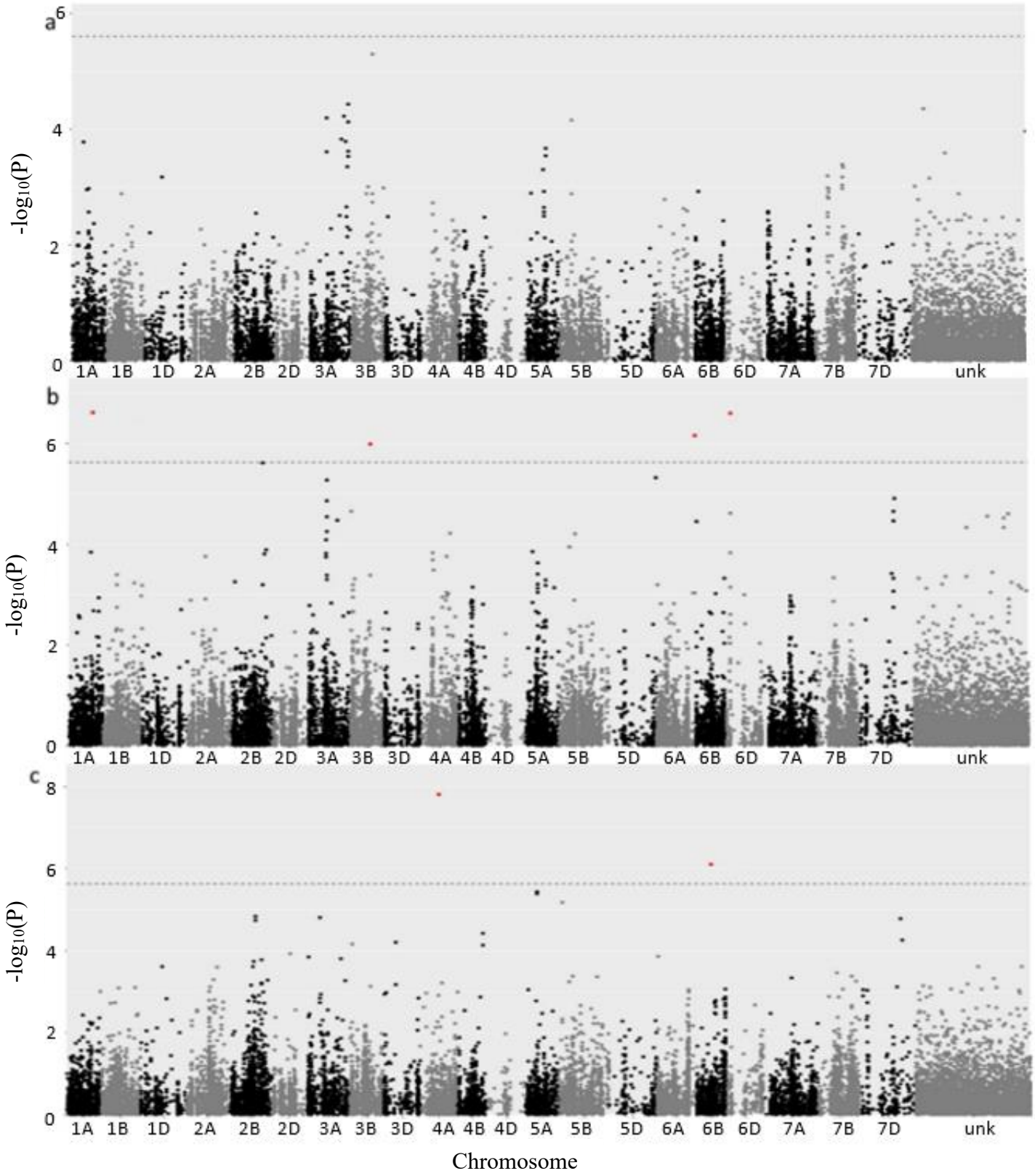

Figure 5.1: Manhattan plots for single marker analysis of SNPs associated with P1-P3 (a-c) concentrations (unk = unmapped markers)

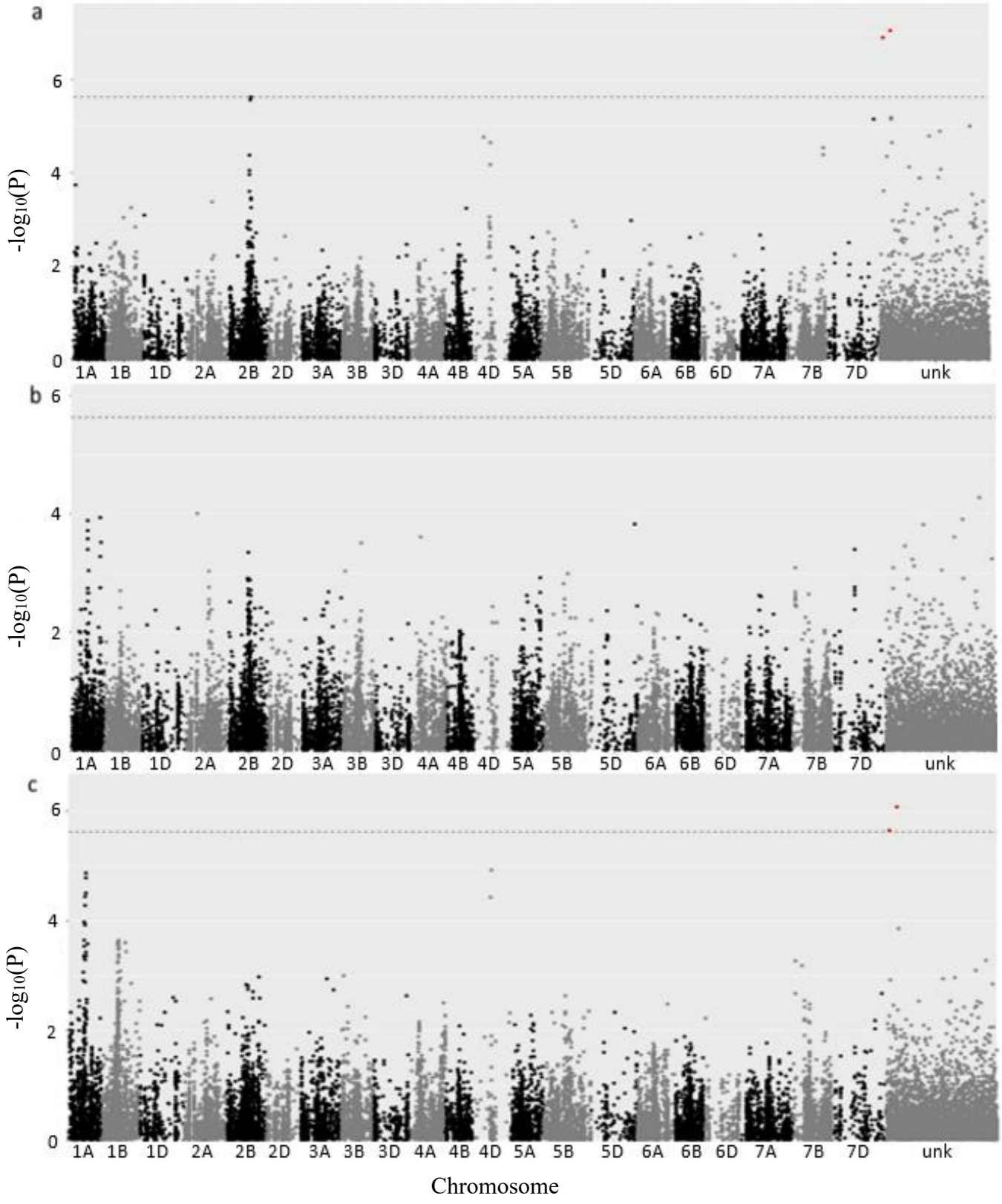

Figure 5.2: Manhattan plots for single marker analysis of SNPs associated with P4-P6 (a-c) concentrations (unk = unmapped markers)

*Table 5.1: Summary statistics for effects of significant markers on gluten epitope concentrations. (Chr = Chromosome, cM = marker position, Mean = gluten epitope concentration reported as batch adjusted value, * = p<0.05, *** = p<0.001)*

				Sample number		Mean		Standard Deviation		Difference	Variance
Trait	Marker ID	Chr	cM	Allele A	Allele B	Allele A	Allele B	Allele A	Allele B	in means	ratio
P2	BS00058861_51	3B	10.88406	268	203	2.65	-3.31	11.8	16.2	5.96***	1.90
	BobWhite_c35443_249	6B	82.01527	433	38	-0.18	5.02	14.0	13.6	5.21***	1.06
	Kukri_c13136_1363	1A	52.54881	196	275	2.21	-1.43	13.4	14.3	3.64***	1.13
	Ex_c882_534	6D	24.76628	17	454	-25.89	1.09	28.0	12.3	26.98***	5.13
P3	Kukri_c33278_298	6B	42.75277	395	76	6.16	-17.53	36.0	4.7	23.69***	59.04
	Kukri_c50720_397	4A	73.64594	161	310	10.71	-1.38	39.5	30.8	12.09***	1.65
P4	Excalibur_rep_c66565_108	unk	unk	336	135	-3.45	8.15	13.4	15.8	11.60***	1.39
	Excalibur_c6972_2300	unk	unk	345	126	-3.48	9.07	13.8	14.3	12.55***	1.07
P6	Kukri_c77282_63	unk	unk	194	277	-49.70	16.99	105.7	116.5	66.69***	1.22
	RFL_Contig5277_1283	unk	unk	126	345	8.77	-15.29	83.6	125.7	24.06*	2.26

A total of 4 single markers were identified as being significantly associated with concentrations of the P2 epitope (Figure 5.1). The differences in P2 concentrations between populations carrying opposing alleles for significant SNPs can be observed in Figure 5.3. For each of the 4 markers identified as being significantly associated with P2 concentrations, a significant difference in mean P2 concentration between populations with opposing alleles can be observed (Table 5.1, $p<0.001$). The largest difference in mean P2 concentration is observed for populations separated by alleles of the marker 'Ex_c882_534', however these results may be unreliable due to the uneven number of lines carrying each of the opposing alleles (Table 5.1). None of the identified markers appear to explain the low P2 outlier group (Figure 5.3).

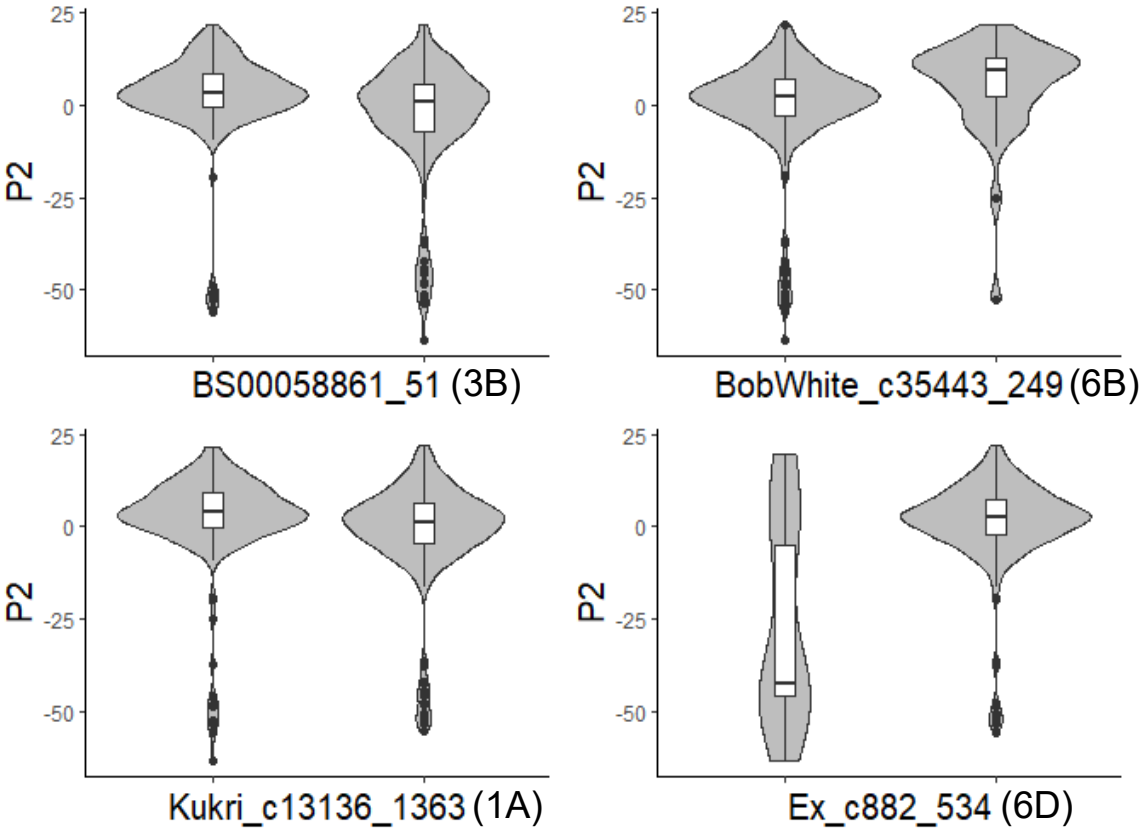

Figure 5.3 Violin plots for P2 concentrations (reported as batch adjusted values) in populations with opposing alleles for each of the markers identified as being significantly associated with P2 concentrations. The x-axes are labelled with marker IDs and the chromosome that they are associated with is in brackets.

For P3 concentrations, 2 markers were identified as being significant (Figure 5.1). In each case, a significant difference between the mean concentrations of P3 for populations with each of the opposing alleles can be observed (Table 5.1). The largest difference in means appears to be explained by the marker 'Kukri_c33278_298', however the sample sizes are again uneven and therefore may be unreliable (Table 5.1, Figure 5.4).

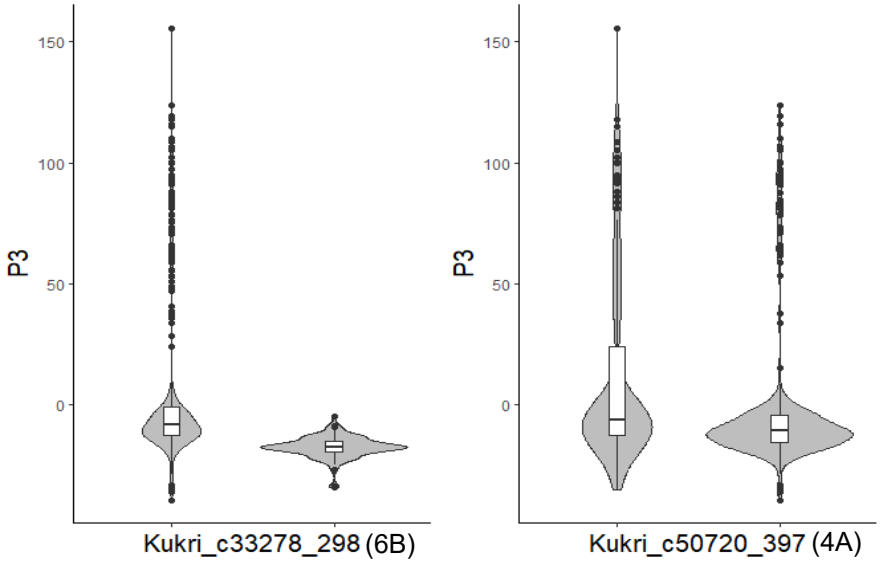

Figure 5.4 Violin plots for P3 concentrations (reported as batch adjusted values) in populations with opposing alleles for each of the markers identified as being significantly associated with P2 concentrations. The x-axes are labelled with marker IDs and the chromosome that they are associated with is in brackets.

Violin plots for differences in P4 concentrations between populations with opposing alleles for the two significant markers identified as being associated with P4 concentrations are presented in Figure 5.5. In this case, the sample sizes are considered equal (Table 5.1) giving confidence that the differences in mean P4 concentrations between populations with opposing alleles are real. For both of the significant markers, the differences in mean P4 concentrations between populations is highly significant (Table 5.1, $p<0.001$).

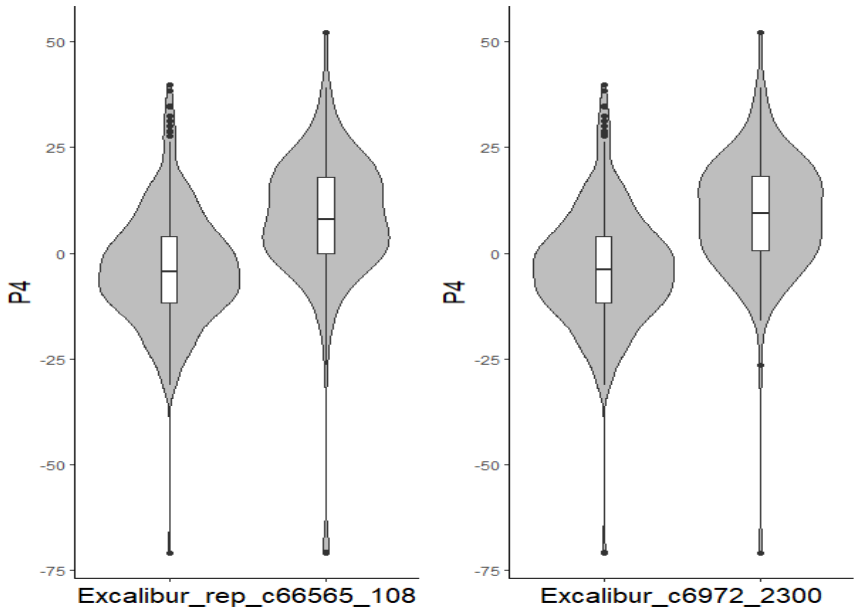

Figure 5.5 Violin plots for P4 concentrations (reported as batch adjusted values) in populations with opposing alleles for each of the markers identified as being significantly associated with P2 concentrations. The x-axes are labelled with marker IDs, both of which are unmapped.

Significant differences in P6 concentrations can be observed between populations with opposing alleles for the two markers identified as being significantly associated with P6 concentrations (Table 5.1). Variations in P6 concentrations between populations can be observed in Figure 5.6.

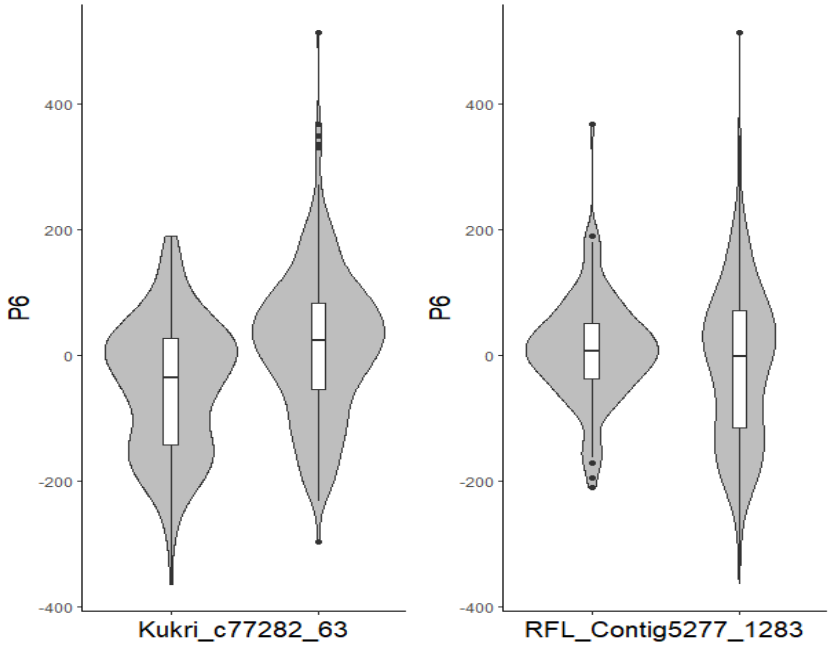

Figure 5.6 Violin plots for P5 concentrations (reported as batch adjusted values) in populations with opposing alleles for each of the markers identified as being significantly associated with P2 concentrations. The y-axes are labelled with marker IDs, both of which are unmapped.

5.4.2 Genome wide association study for windows of 20 contiguous SNPs

The WPPA values and proportion of genomic variation explained by each window of 20 contiguous SNPs are presented in Table 5.2 for those windows above or close to the WPPA threshold of 0.8.

Table 5.2: Genomic windows above or near significance for associations with concentrations of different gluten epitopes (GV = proportion of genomic variance, WPPA = Window Posterior Probability of Association)

Window number	Trait	Year	Chromosome	Window start (Mb)	Window end (Mb)	GV (%)	WPPA
50	P3	Year 2	1B	14.8	17.7	3.5	0.51
50	P3	Year 3	1B	14.8	17.7	5.1	0.85
50	P3	3 Years	1B	14.8	17.7	3.2	0.63
50	P5	Year 1	1B	14.8	17.7	2.7	0.66
50	P5	Year 2	1B	14.8	17.7	4.1	0.90
50	P5	Year 3	1B	14.8	17.7	3.1	0.69
50	P5	3 Years	1B	14.8	17.7	4.1	0.77
197	P1	3 Years	2B	423.8	435.3	2.3	0.90
203	P3	Year 2	2B	559.4	593.7	13.7	0.92
203	P3	Year 3	2B	559.4	593.7	7.1	0.94
203	P3	3 Years	2B	559.4	593.7	9.8	0.96
203	P5	Year 1	2B	559.4	593.7	3.7	0.78
203	P5	Year 2	2B	559.4	593.7	4.7	0.94
203	P5	Year 3	2B	559.4	593.7	3.7	0.75
203	P5	3 Years	2B	559.4	593.7	4.2	0.78
251	P4	Year 1	2D	649.8	650.7	1.1	0.73
497	P2	3 Years	5B	550.2	551.7	1.9	0.76
632	P3	Year 2	7A	11.1	15.6	2.5	0.77
684	P3	Year 2	7B	185.6	203.7	2.1	0.78

Across all of the single year and multi-year analyses, three genomic windows were shown to be associated with concentrations of the epitopes P1, P3 and P5 (WPPA > 0.8; Table 5.2) while no windows (WPPA < 0.8) appear to be associated with concentrations of the remaining epitopes (P2, P4 and P6) or the total concentration (sum of P1-P6). Window number 197 on chromosome 2B between 423.8-435.3 Mb showed a strong association with P1 concentrations in the multi-year analysis (Table 5.2, Figure 5.7). However, that window was not identified as significant in any of the single year analyses (WPPA: Year 1 = 0.23, Year 2 = 0.24, Year 3 = 0.14). No other windows were identified as being significantly associated with P1 concentrations. Single marker effects for SNPs potentially associated with P1 concentrations in window 197 were identified as being non-significant in all cases.

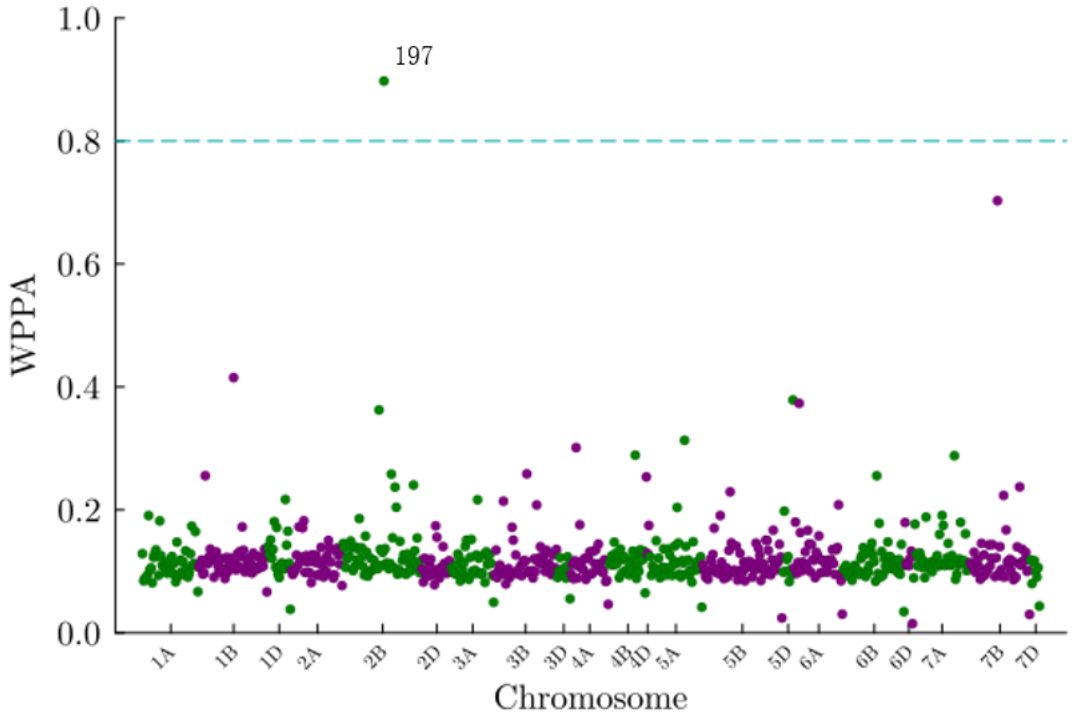

Figure 5.7: Manhattan plot showing window posterior probabilities of association (WPPA) for P1 gluten epitope concentrations in the multi-year analysis. The window number is annotated for windows with a WPPA value greater than 0.8.

Two separate windows were shown to be significantly associated with P3 concentrations. Window 50 on chromosome 1B between 14.8-17.7 Mb was shown to be significantly associated with P3 concentrations in the Year 3 analysis (Table 5.2, Figure 5.8) but not in any of the other analyses (WPPA: Year 1 = 0.18, Year 2 = 0.51, multi-year = 0.63). Window 203 on chromosome 2B between 559.4-593.7 Mb was shown to be significantly associated with P3 concentrations in Year 2, Year 3 and multi-year analyses (Table 5.2, Figure 5.8) but not in the Year 1 analysis (WPPA = 0.23). Single marker effects for SNPs potentially associated with P3 concentrations in windows 50 and 203 were identified as being non-significant in all cases.

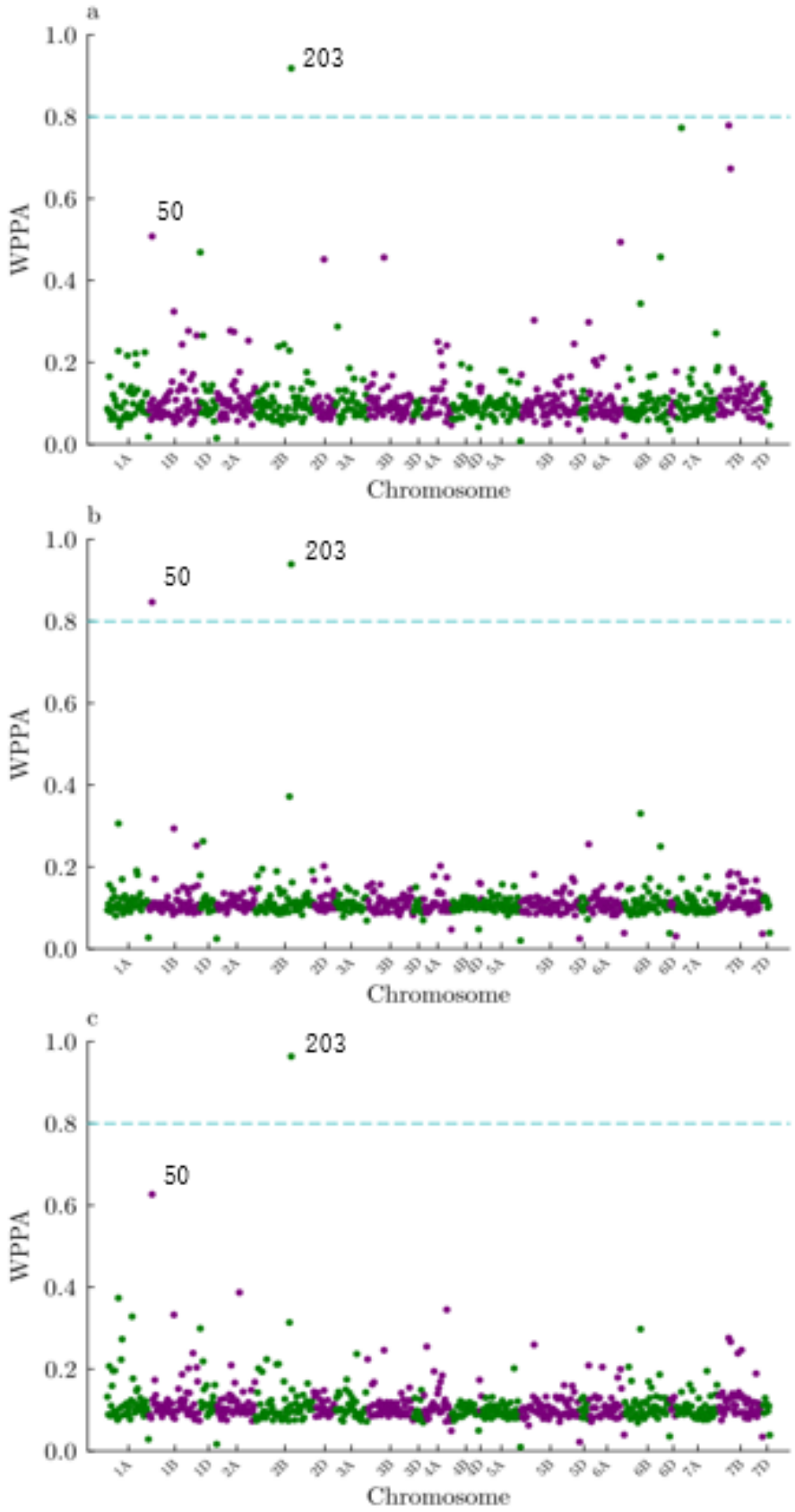

Figure 5.8: Manhattan plot showing window posterior probabilities of association (WPPA) for P3 gluten epitope concentrations in the Year 2 (a), Year 3 (b) and multi-year (c) analysis. Annotations are for windows 50 and 203.

Window numbers 50 and 203 were also shown to be significantly associated with P5 concentrations. Window 50 was associated with P5 concentrations in the Year 2 analysis (Table 5.2, Figure 5.9). No significant associations between window 50 and P5 were observed in the remaining analyses, although WPPA values were close to significance (WPPA: Year 1 = 0.66, Year 3 = 0.69 and multi-year = 0.77). Window 203 was also associated with P5 concentrations in the Year 2 analysis (Table 5.2, Figure 5.9). No significant associations were observed between window 203 and P5 concentrations in any of the other analyses but WPPA values were almost high enough to have been significant (WPPA: Year 1 = 0.78, Year 3 = 0.75, multi-year = 0.78). Single marker effects for SNPs potentially associated with P5 concentrations in windows 50 and 203 were identified as being non-significant in all cases. Strong positive correlations can be observed between the window BVs for P3 and P5 concentrations for both window 50 ($r=0.89$, $p<0.001$) and window 203 ($r=0.84$, $p<0,001$).

Other possible locations of interest include window 684 located between 185.6-203.7 Mb on chromosome 7B where a WPPA of 0.78 was observed for P3 concentrations in Year 2 (Table 5.2). That window also contributed to 2.1% of the genomic variance (Table 5.2). Additionally, the WPPA for window 632 at 11.1-15.6 Mb on chromosome 7A was 0.77 for P3 concentrations in Year 2 and accounted for 2.5% of the genomic variance (Table 5.2). Window 497 between 550.2-551.7 Mb on chromosome 5B produced WPPA results of 0.76 in the multi-year analysis of P2, explaining 2.0% of the genomic variance and window 251 between 649.8-650.9 Mb on chromosome 2D produced WPPA results of 0.73 in the Year 1 analysis of P4 concentrations, explaining 1.1% of the genomic variance (Table 5.2).

A GWAS analysis was performed on FPC for the three individual growing seasons and in a combined three-year analysis. Additionally, GWAS analysis was performed on mixograph quality parameters (MPT, MPV and MPW) in Year 2 as well as on Farinograph and bake test results in Year 3. No significant associations were identified between any genomic windows and the measured phenotypes in any of those analyses.

An additional GWAS analysis was performed on the presence/absence trait defined by awned vs awnless wheats. In the full population of 471 lines, 327 were observed to have awns while the remaining 144 lines were observed to be awnless. The observed marker effects for the SNP 'BobWhite_c8266_227' on chromosome 5A were shown to explain all of the variation observed for the awned/awnless trait. That marker has previously been shown to be in close proximity to the Tipped B1 locus in wheat (Mellers et al., 2020), one of the three dominant loci known to inhibit awn development (Li et al., 2023). These results therefore act as a reliable check on the genotypic and phenotypic data included in our analyses of other traits.

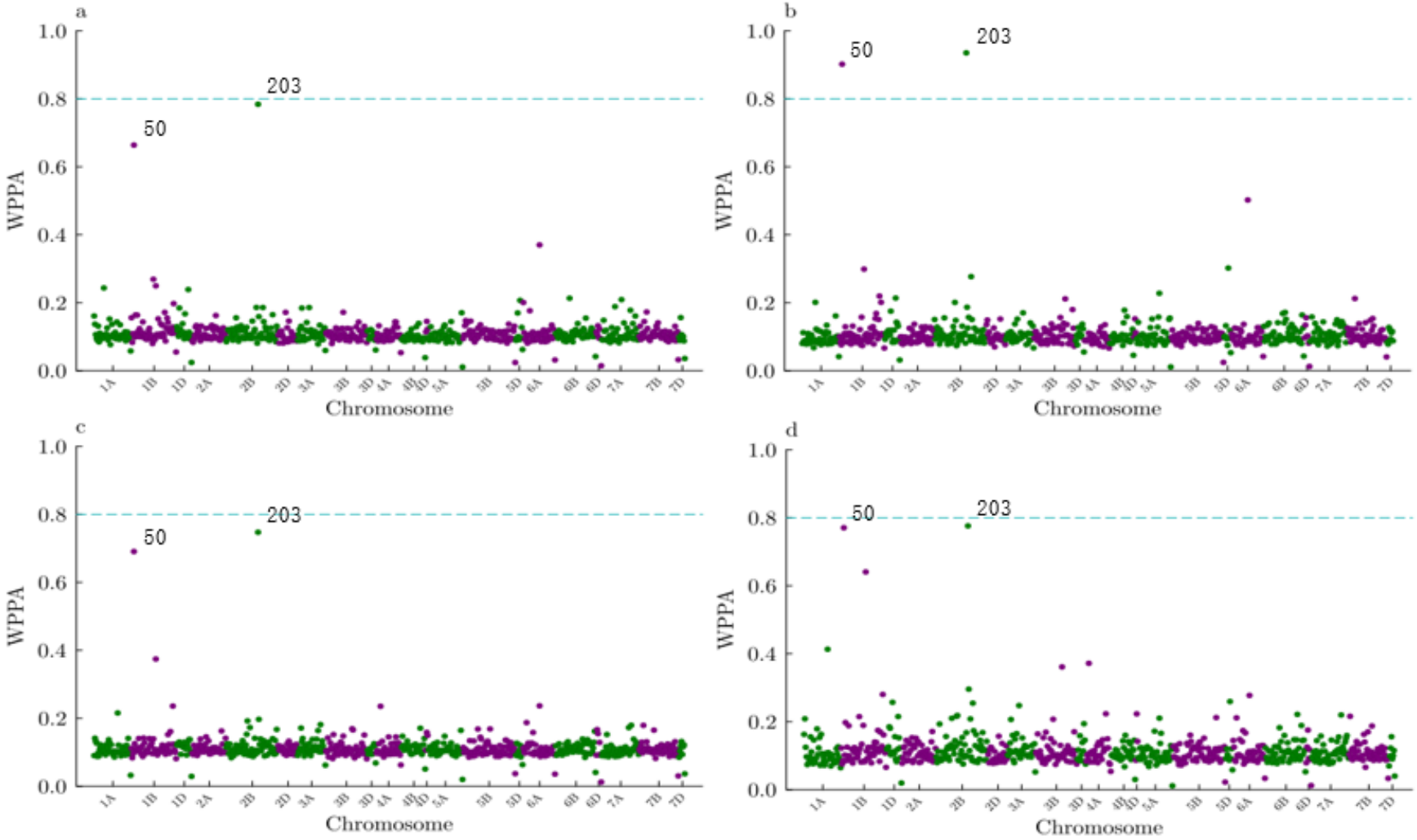

Figure 5.9: Manhattan plot showing window posterior probabilities of association (WPPA) for P5 gluten epitope concentrations in the Year 1 (a), Year 2 (b), Year 3 (c) and multi-year (d) analysis. Annotations are for window numbers 50 and 203.

5.5 Discussion

A single marker GWAS was used to identify SNPs that were significantly associated with concentrations of each of the measured gluten epitopes. No significant markers were identified as being associated with concentrations of the P1 or P5 epitopes. Between 2 and 4 markers were identified as being associated with each of the P2, P3, P4 and P6 epitopes, however no major loci with multiple significant markers were identified. Each of the identified markers were shown to have a significant effect on the mean epitope concentrations of populations separated by the alleles of each of the significant markers (Table 5.1). However, the mean effects of the markers 'Ex_c882_534' and 'Kukri_c33278_298' on P2 and P3 concentrations may be unreliable due to the unequal population sizes (Table 5.2). For each of the remaining markers, it is likely that these could be employed through MAS as a breeding tool for selecting towards wheat lines with reduced concentrations of P2, P3, P4 and P6 epitopes.

A GWAS was conducted using a Bayesian window approach with the aim of identifying significant QTLs for epitope concentrations that were not identified in the single marker analysis. The Bayesian window approach relies on the relevant genetic effects being partly captured by each of the markers in the analysis and can identify QTLs that may otherwise be missed. This is possible because none of the markers are individually in high linkage disequilibrium with the causal mutation(s) in the window. Additionally, analyses of single marker effects tend to be enriched with type one errors leading to spurious associations (Fernando et al., 2017; Fernando & Garrick, 2013). Three genomic windows were identified as having significant associations with concentrations of certain epitopes. Window 203 on chromosome 2B appears to have the greatest impact on epitope concentrations. That window was associated with both the P3 and P5 epitopes (Table 5.2). While the association between that window and P5 concentrations was only deemed significant in the Year 2 analysis, WPPA values in the Year 1, Year 3 and multi-year analysis bordered on the significance threshold (0.75-0.78). The association between that window and P3 concentrations was also significant in the Year 2, Year 3 and multi-year analyses with WPPA values between 0.92-0.96 (Table 5.2). These data suggest that alleles in this window that were not genotyped on the chip are associated with key factors in determining the concentrations of P3 and P5 in a wheat cultivar. The proportion of genomic variance explained by window 203 for concentrations of the P3 and P5 epitopes was relatively large, ranging from 7.1-13.7% for P3 and 3.7-4.7% for P5. These results provide further evidence that the genetic effects at window 203 are likely to be major factors contributing to P3 and P5 concentrations.

Further investigation of the QTL at window 203 may be able to uncover the specific genetic effects that are influencing P3 and P5 concentrations. Although the single marker effects observed here were low, window 203 is still likely to be an important QTL for P3 and P5 concentrations due to the high WPPA values observed. This is because the causal mutation effect is being partly captured by each marker so is

distributed across them. A strong positive correlation (r=0.84, p<0.001) exists between the BVs of window 203 for P3 and P5 concentrations. This suggests that the genetic effects for window 203 are acting in the same directions for both epitopes. Therefore, breeding selections based on the genotypes at window 203 are likely to be effective at simultaneously selecting for low P3 and low P5 cultivars. However, in Chapter 3, a significant negative genetic correlation (-0.39, P>0.001) between P3 and P5 concentrations was reported. This suggests that other important genetic factors may also exist that are influencing P3 and P5 concentrations in opposite directions.

Both the P3 and P5 epitopes are part of the α-gliadin fraction of gluten and therefore it was expected that the α-gliadin coding loci may be associated with epitope regulation. However, those loci exist on chromosomes 1 & 6 of wheat. Therefore, our results suggest that the genomic factors determining the concentrations of the P3 and P5 epitopes may be more regulatory than structural. This is in line with work from Taranto et al. (2020), who were unable to observe any marker trait associations between epitope concentrations and either the Gli-A1 or Gli-B1 loci in Durum wheat. The biochemistry involved in protein regulation is highly complex and influenced by a myriad of environmental and genomic variables and more research is required to fully understand the biochemical pathways involved in determining P3 and P5 concentrations.

A second genomic window (window 50) on chromosome 1B was also shown to have associations with P3 and P5 concentrations. Those associations were only significant (WPPA > 0.8) in the Year 3 analysis for P3 and the Year 2 analysis for P5. However, WPPA values were still greater than 0.5 in the Year 2 and multi-year analysis for P3 being 0.51 and 0.63 respectively and each explaining 3.5% and 3.2% of the genomic variance. WPPA values were higher still for P5 in the Year 1 (WPPA = 0.66), Year 3 (WPPA = 0.69) and multi-year analysis (0.77) explaining 2.7%, 3.1% and 4.1% of the genomic variance respectively (Table 5.2). These data suggest that this genomic window on chromosome 1B is also likely to have some association with P3 and P5 concentrations, although to a lesser extent than the aforementioned window on chromosome 2B (window 203). A strong positive correlation (r=0.89, p<0.001) exists between the window BVs of window 50 for P3 and P5 concentrations, suggesting that this window is affecting P3 and P5 concentrations in the same direction. As for window 203, this indicates that selections based of the alleles within this window may allow for simultaneous selection of low P3 and low P5 wheat lines.

The only other genomic window showing a significant association with epitope concentrations was window 197 on chromosome 2B. That window was associated with P1 concentrations in the multi-year analysis (WPPA = 0.9) where it explained 2.3% of the genomic variance. While no association was observed between this window and P1 concentrations in any of the single year analyses, the multi-year analyses considers the effect of genotype in an average year whereas individual year analyses can be influenced by genotype-

environment interactions. Therefore, the multi-year results are still likely to provide reliable data in determining the genomic associations with concentrations of P1. The consistently low proportion of genomic variance explained by each of the significantly associated windows further suggests that many small effect genes and genetic interactions are responsible for determining epitope concentrations.

Several additional genomic windows were identified as candidates for further investigation. WPPA values of 0.73-0.78 were observed between these windows and concentrations of different epitopes and while these values are not technically deemed significant, they are approaching significance and those that are truly significant will show up stronger in different populations or with more data. Two of those windows, one between 11.1-15.6 Mb on chromosome 7A (number 632) and the other between 185.6-203.7 Mb on chromosome 7B (number 684) produced WPPA results of 0.77 (GV=2.5%) and 0.78 (GV=2.1%) respectively in the Year 2 analysis of P3 concentrations (Table 5.2). The next most associated window, between 550.2-551.7 Mb on chromosome 5B (number 497) produced a WPPA result of 0.76 (GV=1.9%) in the multi-year analysis of P2 concentrations and the window between 649.8-650.7 Mb on chromosome 2D (number 251) produced a WPPA result of 0.73 (GV=1.1%) in the Year 1 analysis of P4 (Table 5.2). These two windows were the only ones to produce WPPA results above 0.7 for P2 and P4 concentrations and no windows were observed to be significantly associated with either P6 concentrations or total concentrations of P1-P6.

Our analysis utilized marker information from the 90K Illumina wheat SNP chip in a total of 471 different wheat varieties. Due to the moderate-high heritability's and genetic correlations observed in previous analyses for epitope concentrations between different growing seasons (Chapter 3), it was expected that a higher number of genomic windows would be significantly associated with concentrations of individual gluten epitopes. Unfortunately, due to the high complexity, allopolyploid and large genome size of wheat, only a relatively small proportion of SNPs (52%) are currently mapped meaning that many markers were unable to be included in our GWAS analysis, reducing the likelihood of identifying significant genomic windows for all of the tested epitopes (Guan et al., 2020). Additionally, many of the key gliadin coding genes have been identified on the D genome of wheat which contains the least number of mapped markers of the three sub-genomes in wheat due to the lower genetic diversity involved in the ancestral hybridisation event (Alipour et al., 2019). This further suggests that important loci associated with epitope concentrations may not have been captured in the analysis. Most of the significant genomic associations with epitope concentrations were observed on the B genome. However, it remains unclear as to why this is the case as none of the identified loci appear to co-locate with any of the known gliadin regulating loci on the B genome. The effectiveness of GWAS is greatly improved with an increased number of individuals (Plavsin et al., 2021). The number of individuals assessed in this study is consistent with the numbers used in many other genomic based studies of wheat (Gao et al., 2021; Khan et al., 2022; Yang et al., 2020), but a greater number of individuals would produce more reliable characterisations of association. It also appears that

epitope concentrations in wheat are highly complex traits, the likely interactions between many small effect genes makes it difficult to understand the genetic factors involved. Therefore, large amounts of high quality phenotypic and genotypic data are essential to understanding these genetic factors.

A number of lines tested in this study were identified as being outliers for either low concentrations of P2 (22 lines) or high concentrations of P3 (29 lines, Chapter 3). The large differences in P2 and P3 concentrations in these outlier groups compared to the population mean, suggests that there may be major gene effects involved. It was therefore expected that these could be identified in a GWAS to identify the genetic factors involved. However, no such factors were able to be identified here. One possible explanation is that the small number of outlier lines in this study meant that the effects were not picked up in the analysis. It is also possible that these outlier groups are determined by combinations of rare alleles that were discarded from analysis due to the low allele frequency. Another explanation for not being able to identify any major gene effects for P2 and P3 concentrations could be due to the absence of the relevant genotypic information. These factors are supported by the PCA in Appendix Figure 8.13 where the low P2 and high P3 outlier lines appear to be evenly spread across the training population, suggesting that the genetic factors that may be linked to low P2 or high P3 concentrations are not explained by either PC1 or PC2.

A GWAS was performed on a range of baking quality traits including FPC, MPT, MPV and MPW. In each of those analyses, no genomic windows were significantly associated with any of the measured traits. While this suggests that utilising SNP chip data for genomic selection of baking quality traits in wheat may be difficult, it also indicates that no major associations appear to exist between epitope concentrations and baking quality characteristics. This is in line with the generally low genetic correlations observed between concentrations of some epitopes and baking quality characteristics seen in Chapter 4. These results suggest that breeders may be able to select for reduced gluten epitope levels in wheat cultivars without any significant reductions in baking quality characteristics.

5.6 References

Alipour, H., Bai, G., Zhang, G., Bihamta, M. R., Mohammadi, V., & Peyghambari, S. A. (2019). Imputation accuracy of wheat genotyping-by-sequencing (GBS) data using barley and wheat genome references. *PLoS ONE*, *14*(1), e0208614. https://doi.org/10.1371/journal.pone.0208614

Alseekh, S., Kostova, D., Bulut, M., & Fernie, A. R. (2021). Genome-wide association studies: assessing trait characteristics in model and crop plants. *Cellular and Molecular Life Sciences*, *78*(15), 5743-5754. https://doi.org/10.1007/s00018-021-03868-w

Biesiekierski, J. R. (2017). What is gluten? *Journal of Gastroenterology and Hepatology*, *32*, 78-81. https://doi.org/10.1111/jgh.13703

Cheng, H., Fernando, R. L., & Garrick, D. J. (2018). JWAS: Julia implementation of whole-genome analysis software. Proceedings of the World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand.

Fernando, R., Toosi, A., Wolc, A., Garrick, D., & Dekkers, J. (2017). Application of Whole-Genome Prediction Methods for Genome-Wide Association Studies: A Bayesian Approach. *Journal of Agricultural, Biological and Environmental Statistics*, *22*(2), 172-193. https://doi.org/10.1007/s13253-017-0277-6

Fernando, R. L., & Garrick, D. (2013). Bayesian Methods Applied to GWAS. In C. Gondro, J. van der Werf, & B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction* (pp. 237-274). Humana Press. https://doi.org/10.1007/978-1-62703-447-0_10

Fernando, R. L., Toosi, A., Garrick, D. J., & Dekkers, J. C. M. (2014). Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. Proceedings of the 10th world congress on genetics applied to livestock production, Vancouver.

Gao, L., Meng, C., Yi, T., Xu, K., Cao, H., Zhang, S., . . . Zhao, Y. (2021). Genome-wide association study reveals the genetic basis of yield- and quality-related traits in wheat. *BMC Plant Biology*, *21*(1), 144. https://doi.org/10.1186/s12870-021-02925-7

Garbus, I., Romero, J. R., Valarik, M., Vanžurová, H., Karafiátová, M., Cáccamo, M., . . . Echenique, V. (2015). Characterization of repetitive DNA landscape in wheat homeologous group 4 chromosomes. *BMC Genomics*, *16*(1), 375. https://doi.org/10.1186/s12864-015-1579-0

Guan, J., Garcia, D. F., Zhou, Y., Appels, R., Li, A., & Mao, L. (2020). The Battle to Sequence the Bread Wheat Genome: A Tale of the Three Kingdoms. *Genomics, Proteomics & Bioinformatics*, *18*(3), 221-229. https://doi.org/https://doi.org/10.1016/j.gpb.2019.09.005

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*(1), 186. https://doi.org/10.1186/1471-2105-12-186

Kemper, K. E., Bowman, P. J., Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2018). A multi-trait Bayesian method for mapping QTL and genomic prediction. *Genetics Selection Evolution*, *50*(1), 10. https://doi.org/10.1186/s12711-018-0377-y

Khan, H., Krishnappa, G., Kumar, S., Mishra, C. N., Krishna, H., Devate, N. B., . . . Singh, G. P. (2022). Genome-wide association study for grain yield and component traits in bread wheat (Triticum aestivum L.). *Front Genet*, *13*, 982589. https://doi.org/10.3389/fgene.2022.982589

Kibite, S., & Evans, L. E. (1984). Causes of negative correlations between grain yield and grain protein concentration in common wheat. *Euphytica*, *33*(3), 801-810. https://doi.org/10.1007/BF00021906

Kiszonas, A. M., & Morris, C. F. (2017). Wheat Breeding for Quality: A Historical Review. *Cereal Chemistry Journal*. https://doi.org/10.1094/cchem-05-17-0103-fi

Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*(1), 44-53. https://doi.org/https://doi.org/10.1111/1755-0998.12549

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, *9*(29). https://doi.org/10.1186/1746-4811-9-29

Lee, J., Cheng, H., Garrick, D., Golden, B., Dekkers, J., Park, K., . . . Fernando, R. (2017). Comparison of alternative approaches to single-trait genomic prediction using genotyped and non-genotyped Hanwoo beef cattle. *Genetics Selection Evolution*, *49*(1), 2. https://doi.org/10.1186/s12711-016-0279-9

Lemeunier, P., Paux, E., Babi, S., Auzanneau, J., Goudemand-Dugué, E., Ravel, C., & Rincent, R. (2022). Training population optimization for genomic selection improves the predictive ability of a costly measure in bread wheat, the gliadin to glutenin ratio. *Euphytica*, *218*(8), 111. https://doi.org/10.1007/s10681-022-03062-4

Li, J., Xin, X., Sun, F., Zhu, Z., Xu, X., Yang, J., . . . Ma, J. (2023). Copy number variation of B1 controls awn length in wheat. *The Crop Journal*, *11*(3), 817-824. https://doi.org/https://doi.org/10.1016/j.cj.2022.10.007

Mellers, G., González Aguilera, J., Bird, N., Bonato, A. L., Bonow, S., Caierão, E., . . . Cockram, J. (2020). Genetic Characterization of a Wheat Association Mapping Panel Relevant to Brazilian Breeding Using a High-Density Single Nucleotide Polymorphism Array. *G3*, *10*. https://doi.org/10.1534/g3.120.401234

Michel, S., Kummer, C., Gallee, M., Hellinger, J., Ametz, C., Akgöl, B., . . . Buerstmayr, H. (2018). Improving the baking quality of bread wheat by genomic selection in early generations. *Theoretical and Applied Genetics*, *131*(2), 477-493. https://doi.org/10.1007/s00122-017-2998-x

Plavsin, I., Gunjaca, J., Satovic, Z., Sarcevic, H., Ivic, M., Dvojkovic, K., & Novoselovic, D. (2021). An Overview of Key Factors Affecting Genomic Selection for Wheat Quality Traits. *Plants (Basel)*, *10*(4). https://doi.org/10.3390/plants10040745

Sharma, N., Bhatia, S., Chunduri, V., Kaur, S., Sharma, S., Kapoor, P., . . . Garg, M. (2020). Pathogenesis of Celiac Disease and Other Gluten Related Disorders in Wheat and Strategies for Mitigating Them. *Front Nutr*, *7*, 6. https://doi.org/10.3389/fnut.2020.00006

Shewry, P. R., & Tatham, A. S. (2016). Improving wheat to remove coeliac epitopes but retain functionality. *Journal of Cereal Science*, *67*, 12-21. https://doi.org/https://doi.org/10.1016/j.jcs.2015.06.005

Simmonds, N. W. (1995). The relation between yield and protein in cereal grain. *Journal of the Science of Food and Agriculture*, *67*(3), 309-315.

Taranto, F., D'Agostino, N., Catellani, M., Laviano, L., Ronga, D., Milc, J., . . . Francia, E. (2020). Characterization of Celiac Disease-Related Epitopes and Gluten Fractions, and Identification of Associated Loci in Durum Wheat. *Agronomy*, *10*(9). https://doi.org/10.3390/agronomy10091231

Tsai, H. Y., Janss, L. L., Andersen, J. R., Orabi, J., Jensen, J. D., Jahoor, A., & Jensen, J. (2020). Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. *Sci Rep*, *10*(1), 3347. https://doi.org/10.1038/s41598-020-60203-2

van Rossum, B., & Kruijer, W. (2023). *statgenGWAS: Genome Wide Association Studies.*https://doi.org/https://biometris.github.io/statgenGWAS/index.html

Wang, Z., Chapman, D., Morota, G., & Cheng, H. (2020). A Multiple-Trait Bayesian Variable Selection Regression Method for Integrating Phenotypic Causal Networks in Genome-Wide Association Studies. *G3 Genes|Genomes|Genetics*, *10*(12), 4439-4448. https://doi.org/10.1534/g3.120.401618

Wen, W., He, Z., Gao, F., Liu, J., Jin, H., Zhai, S., . . . Xia, X. (2017). A High-Density Consensus Map of Common Wheat Integrating Four Mapping Populations Scanned by the 90K SNP Array. *Frontiers in Plant Science*, *8*. https://www.frontiersin.org/articles/10.3389/fpls.2017.01389

Wolc, A., & Dekkers, J. C. M. (2022). Application of Bayesian genomic prediction methods to genome-wide association analyses. *Genetics Selection Evolution*, *54*(1), 31. https://doi.org/10.1186/s12711-022-00724-8

Yang, Y., Chai, Y., Zhang, X., Lu, S., Zhao, Z., Wei, D., . . . Hu, Y.-G. (2020). Multi-Locus GWAS of Quality Traits in Bread Wheat: Mining More Candidate Genes and Possible Regulatory Network. *Frontiers in Plant Science*, *11*. https://www.frontiersin.org/articles/10.3389/fpls.2020.01091

Zhang-Biehn, S., Fritz, A. K., Zhang, G., Evers, B., Regan, R., & Poland, J. (2021). Accelerating wheat breeding for end-use quality through association mapping and multivariate genomic prediction. *Plant Genome*, *14*(3). https://doi.org/10.1002/tpg2.20164

Zhang, Y., Fan, C., Chen, Y., Wang, R. R. C., Zhang, X., Han, F., & Hu, Z. (2021). Genome evolution during bread wheat formation unveiled by the distribution dynamics of SSR sequences on chromosomes using FISH. *BMC Genomics*, *22*(1), 55. https://doi.org/10.1186/s12864-020-07364-6

Chapter 6

Bayesian approach for genomic prediction of gluten epitope concentrations in bread wheat (*T. aestivum*)

6.1 Abstract

Concentrations of 6 gluten epitopes (P1-P6) were measured in the primary training population (471 lines) in each of three growing seasons (2020-2021, 2021-2022, 2022-2023) in Lincoln, New Zealand. Epitope concentrations were also measured in an external validation population (199 lines) in just the third growing season (2022-23). The primary training population represents a diverse set of wheat germplasm while the external validation population represents advanced breeding lines from the New Zealand Institute of Plant & Food Research (PFR) milling wheat breeding program. Genotype data were obtained for all lines using a 90K wheat Illumina SNP chip. Three scenarios for genomic prediction of epitope concentrations were tested using variations of the training and validation populations. Scenario 1 employed data from the primary training population to predict epitope concentrations in the external validation population. This scenario aimed to determine the likely prediction accuracies achievable in the PFR breeding program using the primary training data. The same analysis was repeated in Scenario 2 but with a reduced training population, where lines that were non-representative of the external validation population were excluded from analysis. That scenario aimed to improve the prediction accuracies observed in scenario 1. In scenario 3, randomly assigned internal validation populations were tested and this was repeated for 100 randomly formed populations. These internal validation populations consisted of about one-fifth of the data, namely 94 lines from the primary training population, but training data were only used from the remaining four-fifths of the data comprising 377 lines. This scenario aimed to determine the potential prediction accuracies for epitope concentrations in a target population that was genetically representative of the training population. The structure of each population in scenarios 1-3 were assessed by Principal Component Analysis (PCA) based on SNP genotypes. A BayesC approach was applied to obtain Estimated Breeding Values (EBV's) and Prediction Error Variances (PEV's) for epitope concentrations in each scenario. Pearson's correlation coefficient was used to determine the empirical accuracy (r) of EBV predictions in terms of measured values for epitope concentrations in each validation population. Empirical prediction accuracies in scenarios 1 and 2 were similar for concentrations of most epitopes. Mean prediction accuracies in scenario 3 were always higher than those observed in either scenarios 1 or 2. This was particularly the case for P6 concentrations where empirical prediction accuracies were over three times higher than when prediction was applied to a population of advanced breeding lines somewhat disparate from the training population. Empirical prediction accuracies were consistently highest for P5 concentrations ranging from 0.43-0.67 across scenarios 1-3. Realised accuracies of EBVs rather than phenotypes were calculated by adjusting for the heritability of each trait to determine how well the EBV's represented the true Breeding Value (BV). In scenario 3, the realised accuracies for epitope concentrations ranged from 0.43 for P2 concentrations to 0.82 for P6 concentrations. These results indicate that genomic prediction will be an

effective tool for identifying and selecting low epitope wheat cultivars, provided that the training data has sufficient size and relevance for the lines comprising the target population.

6.2 Introduction

Wheat is one of the most widely grown and consumed products worldwide (Shiferaw et al., 2013). It is a highly versatile and relatively low input crop allowing it to be grown in a wide range of environments (Shiferaw et al., 2013). Wheat is a crucial ingredient in many food products, yet some people are unable to safely consume it. The gluten proteins in wheat are known to trigger coeliac disease (CD) in an increasing number of consumers and are also believed to be linked to other symptoms of wheat intolerance (King et al., 2020). Gluten epitopes are specific peptides that exist primarily in the α-gliadin fraction of gluten proteins (Gujral et al., 2012). These epitopes survive proteolytic digestion and are recognised as antigens in the small intestine and are presented to gluten sensitive T-cells in affected consumers (Gujral et al., 2012). That process triggers an inflammatory immune response that eventually leads to the breakdown of intestinal villi and the symptoms of CD (Gujral et al., 2012). Damaged villi are poor at absorbing essential nutrients meaning that coeliac patients often suffer from associated symptoms such as malnutrition (Bernardo & Peña, 2012; Green & Cellier, 2007). This can have significant and lifelong effects for children and young adults suffering for CD during critical stages of physical development. There is no cure for CD and the only treatment is a lifelong gluten free diet (Lindfors et al., 2019). Even the tiniest amount of dietary gluten can trigger vomiting, diarrhoea, nausea and other symptoms of CD. In most countries, products must be shown to have less than 20ppm of gluten to be considered safe for coeliac consumers (Abdi et al., 2023). The prevalence of CD varies worldwide, however in most populations, around 1-2% of people are believed to suffer from the disease (Singh et al., 2018). Susceptibility to CD is hereditary in the sense that it can only occur when the patient carries at least one copy of either the DQ2 or DQ8 haplotype for Human Leukocyte Antigens (HLA); (Romanos et al., 2009). In the United States, it is believed that around 35% of people are carriers of these haplotypes and are therefore at risk of developing CD (Brown et al., 2019). Additional risk factors have been suggested to explain why a minority of the DQ2 or DQ8 carriers ever develop CD. Some evidence suggests that a period of sickness, infection, pregnancy or other condition that may negatively impact the immune system can be a trigger for an at-risk consumer to become coeliac (Gujral et al., 2012; Lebwohl & Rubio-Tapia, 2021). Additionally, studies have shown that people who consume greater quantities of gluten in childhood are at higher risk of developing CD (Aronsson et al., 2019; Lund-Blix et al., 2019; Mårild et al., 2019). It has been proposed that if the quantity of gluten consumed, and more specifically particular gluten epitopes, can be reduced, it may be possible to reverse or at least slow the current trend of an increasing incidence of CD (Molberg et al., 2005). Reduced dietary exposure to gluten epitopes may also reduce the chances of at-risk consumers developing CD early in life when it can have the most significant impacts on their development.

A number of approaches have been suggested for reducing the concentrations of gluten epitopes in wheat-based products. One approach is based on new strategies for processing wheat products in ways that break down digestive resistant epitopes so that they are unable to trigger an immune response in the gut. Examples of such methods include sourdough fermentation for breads using specific bacterial strains of *Lactobacillus*(Greco et al., 2011) or the use of malted grains where the initiation of germination has also begun to breakdown gluten peptides (Loponen et al., 2007). Other approaches involve adding specific peptidases to break down epitopes (Walter et al., 2014) or separating out the gliadin fraction and retaining the remaining gluten fractions (van den Broeck et al., 2011). Each of these approaches show promise in developing low gluten epitope wheat products but none are likely to be scaled up to the extent of existing commercial baking methods. A further approach is to select new wheat cultivars that contain lower concentrations of gluten epitopes but retain acceptable levels of other gluten peptides as well as acceptable baking quality characteristics. A number of researchers have attempted this approach using transgenic techniques such as CRISPR-Cas9 or mutagenesis with mixed results (Gil-Humanes et al., 2014; Sanchez-Leon et al., 2018; Wieser et al., 2007). In most cases, these techniques have required eliminating significant portions of gliadins and therefore the challenge has been in maintaining acceptable baking quality characteristics. Significant natural variation in gluten epitope concentrations has been shown to exist in wheat (Chapter 3, Salentijn et al., 2012, 2013). Therefore, it is proposed that a genomic selection (GS) approach may be an effective way for developing new lines of wheat with reduced levels of gluten epitope while also retaining acceptable baking quality. This approach would allow breeders to select for a balance between low epitope concentrations and acceptable baking quality, while also considering other important agronomic traits. In this study, we investigate the accuracy of a GS model for identifying low epitope wheat cultivars.

6.3 Materials & methods

6.3.1 Training and validation populations

A primary training population of 471 wheat lines was grown each year over three consecutive seasons (Year 1 = 2019-2020, Year 2 = 2020-2021 and Year 3 = 2021-2022) at the New Zealand Institute of Plant & Food Research (PFR) in Lincoln, New Zealand. An external validation population of 199 wheat lines was also grown in Year 3 with a single, two row plot (1m x 40cm) sown for each entry. The validation population were sown alongside the training population in a randomized fashion and treated with identical inputs of water, fertilizer and fungicide. The primary training population was a diverse selection of germplasm including PFR breeding lines developed between 1986-2018 along with wheat lines from Australia, the United Kingdom and Canada. The external validation population is made up of advanced breeding lines from New Zealand. In this study, prediction accuracies for genomic selection of gluten epitope

concentrations are compared using three different strategies (scenarios 1-3). Scenario 1 attempts to predict epitope concentrations in the external validation population using the genotypic and phenotypic data collected on the primary training population (Table 6.1). In scenario 2, data from 234 lines in the primary training population were removed from analysis. That population will henceforth be referred to as the reduced training population. The lines that were removed are those that were developed outside of New Zealand and those that represented a specific double haploid population of 81 lines. Scenario 2 attempts to predict epitope concentrations in the external validation population using the genotypic and phenotypic data collected on the reduced training population (Table 6.1). In scenario 3, randomly assigned internal validation populations were generated and tested to understand the effects of population structure on the prediction accuracies achieved for epitope concentrations. A total of 94 lines representing about 20% of the primary training population were generated for each iteration of internal validation populations. In each case, data from the remaining 377 lines representing about 80% of the primary training population were used to determine EBV's for the internal validation population (Table 6.1). For each replicate, the lines comprising the internal validation population were selected using the random numbers function in the Julia programming language standard library (v1.9.2, Bezanson et al., 2017). Unless otherwise stated, prediction accuracies reported for scenario 3 are the mean values from the results of all replicates of the internal validation population.

Table 6.1: Population ID's and number of lines in the training and validation populations tested in scenarios 1-3

	Population ID	**Number of lines**	**Population ID**	**Number of lines**
Scenario 1	Primary Training	471	External Validation	199
Scenario 2	Reduced Training	237 (of 471)	External Validation	199
Scenario 3	Internal Training	377 (of 471)	Internal Validation	94 (of 471)

6.3.2 Phenotyping & genotyping

The concentrations of 6 gluten epitopes (P1-P6) were measured in each of the three seasons for all members of the training population (n=471) but only in year 3 for the external validation population (n=199). Flour Protein Content (FPC) was measured in all three seasons but only in the primary training population. The methods for measuring epitope concentrations and FPC were described in Chapter 3. Genotyping of all lines in the training and validation populations was conducted using a 90K Illumina SNP bead chip. The genotyping methods used were described in Chapter 3.

6.3.3 Statistical analysis

Population structure was assessed based on SNP marker composition by Principal Component Analysis (PCA) using singular value decomposition in the vcfR package in R Studio (Knaus & Grünwald, 2017). Markers with a minor allele frequency of less than 5% were removed prior to that analysis.

Variance components were estimated for individual traits using the BayesC method (Habier et al., 2011) in the Julia Whole-genome Analyses Software (JWAS) package (Cheng et al., 2018) run in a Julia computing environment. This involved fitting the model equation:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Ma} + \boldsymbol{e}$$

Where **y** was the vector of observed phenotypes, **b** was the vector of fixed effects, **a** was the vector of random additive SNP effects, ***e*** was the vector of residuals, **X** was the design matrix corresponding to **b** and **M** was the matrix of additive dosage covariates representing 0, 1 or 2 copies of Illumina B alleles. In the case of gluten epitope concentrations, fixed effects were the batch in which phenotypes were measured and estimates of the batch effect were used to adjust gluten epitope concentration for reporting the data. This adjustment accounts for technical variability between Liquid Chromatography Mass spectrometry (LC-MS) batches, within years. Gluten epitope concentrations were measured as ug of epitope per g of flour but are reported as deviations that represent batch-adjusted values. A Markov chain Monte Carlo (MCMC) approach was used to sample plausible values of variance components,fixed effects and breeding values (BV). The posterior means were used to produce point value estimates of variance components (reported for each trait in Chapter 3) and estimated breeding values (EBV). Prediction error variances (PEV) of EBVs were obtained from the variance of the samples of plausible values for BVs. All analyses were run with a chain length of 100,000 samples after a burn-in of 10,000 samples that were discarded. Every 10th post burn-in sample was retained for further inference. Convergence of the MCMC chain was assessed by visual inspection. The MCMC data were assumed to have converged if the mean in the first portion of the chain was not different to the mean at the end of the chain and no prolonged variations in the mean were obvious. In all cases, MCMC chains exhibited evidence of convergence.

Reliabilities of EBV predictions (R^2) were calculated for each training and validation line from analysis of the training data using the formula:

$$\boldsymbol{R^2 = 1 - \left(\frac{PEV}{genetic\ variance}\right)}$$

The PEV for a line reflects the variation in the samples of the BV for that line that result from variation in the samples of the marker effects.

In addition to the EBV reliability, an empirical accuracy (r) for predicting validation phenotype was determined using Pearson's correlation coefficient calculated between EBV's of traits in the relevant validation population and their measured values. All plots were produced using the ggplot2 package in R Studio (Wickham, 2016).

Since the correlation between EBV and adjusted phenotype cannot theoretically exceed the square root heritability, empirical accuracy of the BVs were scaled by dividing by the root heritability as in (Montesinos-López et al., 2021), to aid in interpretation.

$$\boldsymbol{Realised\ Accuracy} = \left(\frac{\boldsymbol{r}}{\sqrt{\boldsymbol{h^2}}}\right)$$

6.4 Results

6.4.1 Epitope concentrations in the validation population

The distribution of epitope concentrations as deviations from the batch means (P1-P6) for lines in the validation population are presented as violin plots in Figure 6.1. Epitope concentrations in the validation population are within the range of epitope concentrations seen in the training population (Figure 3.2, Figure 6.1). There also appears to be fewer extreme outliers for epitope concentrations in the validation population than in the training population (Figure 3.2, Figure 6.1).

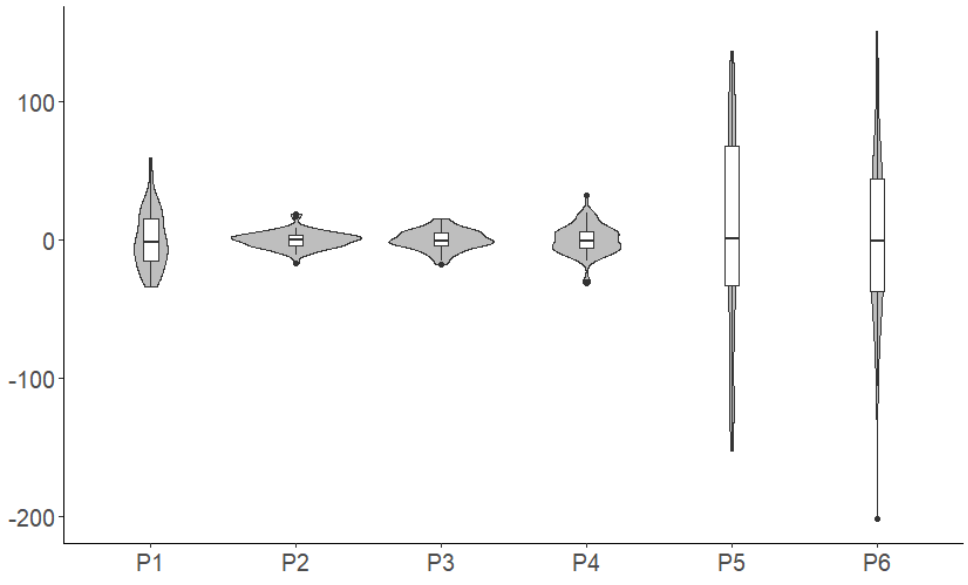

Figure 6.1: Violin plots illustrating the distribution in batch adjusted epitope concentrations (P1-P6) for the validation population

6.4.2 Principal component analysis

The percentages of variation explained by the 10 largest principal components identified in the PCA of SNP marker data are presented in Table 6.2. Almost all of the variation in SNP data was explained by the first principal component (PC1).

Table 6.2: Percent of variation explained by the 10 most important principal components (PC1-10) in the PC analysis of SNP data in the training and validation populations.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
Percent variation explained	77.08	3.92	3.11	2.34	2.06	1.85	1.79	1.63	1.38	1.21

The PCA in Figure 6.2 illustrates the genetic variation within the external validation population and primary training population (scenario 1) between the first two principal components (PC1 and PC2). Figure 6.3 illustrates the genetic variation within the external validation population and reduced training population (scenario 2) between PC1 and PC2.

In scenario 1, the large and diverse primary training population was used to predict epitope concentrations in the external validation population. The external validation population was comprised of advanced PFR breeding lines meaning that this scenario aims to determine the effectiveness of employing the primary training data in the actual modern PFR breeding program. The PCA indicates that the genetic diversity in the primary training population was much greater than that in the external validation population (Figure 6.2).

In scenario 2, lines in the primary training population that were developed outside of New Zealand or from a specific double haploid population were removed from analysis. The reduced training population appears to be more genetically representative of the external validation population than the primary training population (Figure 6.3). Scenario 2 aims to determine the differences in prediction accuracies for epitope concentrations in the external validation population using the more representative reduced training population compared to the primary training population. This strategy was employed to determine the most effective method for implementing genomic selection in the PFR breeding population using the existing training data.

Figure 6.2: Principal Component Analysis from 90K Illumina SNP chip grouped by external validation population and primary training population (scenario 1)

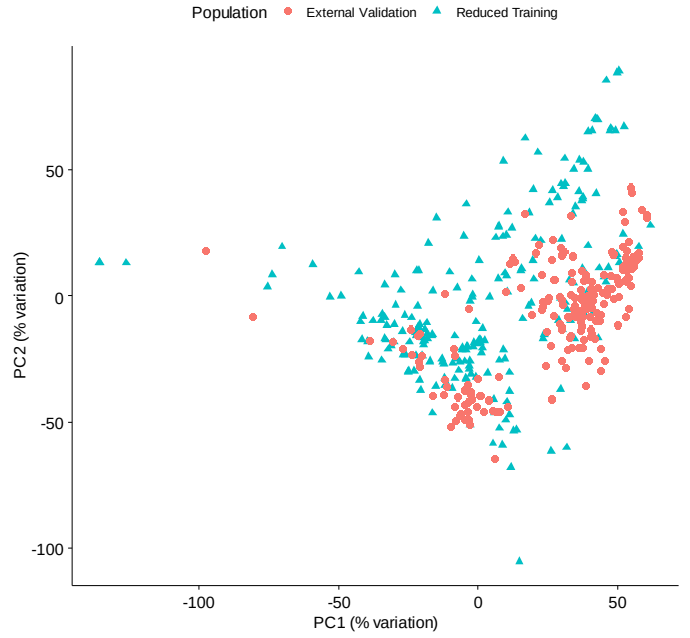

Figure 6.3: Principal Component Analysis from 90K Illumina SNP chip grouped by external validation population and reduced training population (scenario 2)

Scenario 3 was designed to determine the prediction accuracies that could be achieved using training population data that was representative of the target breeding population. This approach was employed to better understand the effect of population structure on prediction accuracies of epitope concentrations. In this scenario, approximately one fifth of the lines in the primary training population were assigned to the internal validation population. A total of 100 replicates of the internal validation population were tested for

each epitope. For each replicate, the lines comprising the internal validation population were randomly assigned and data from the remaining lines (internal training population) were used to determine their EBV's for epitope concentrations.

6.4.3 Prediction accuracies for gluten epitope concentrations

Scatter plots illustrating the empirical prediction accuracies for epitope concentrations in scenario 1 are presented in Figures 6.4-6.5. Scatter plots for empirical prediction accuracies in scenario 2 are presented in the Appendix (Figures 8.14-8.15). Empirical prediction accuracies for epitope concentrations in scenario 1 ranged from 0.16 for P6 to 0.53 for P5 concentrations (Figure 6.5). Reliabilities of EBV prediction (R^2) are depicted using different colours. Mean R^2 values for predictions of epitope concentrations in scenario 1 ranged from 0.4 for P1 to 0.67 for P5 (Figure 6.5).

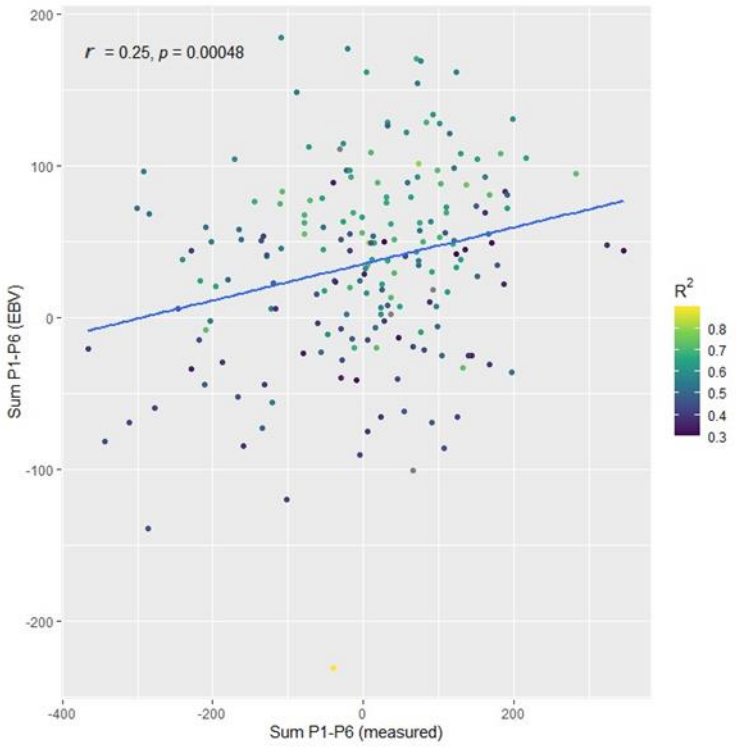

Figure 6.4: Scatter plot for empirical prediction accuracies (r) of total gluten epitope concentrations (Sum P1-P6) in the external validation population using training data from the primary training population (scenario 1). (R2 = Reliability of predictions; EBV = Estimated Breeding Value)

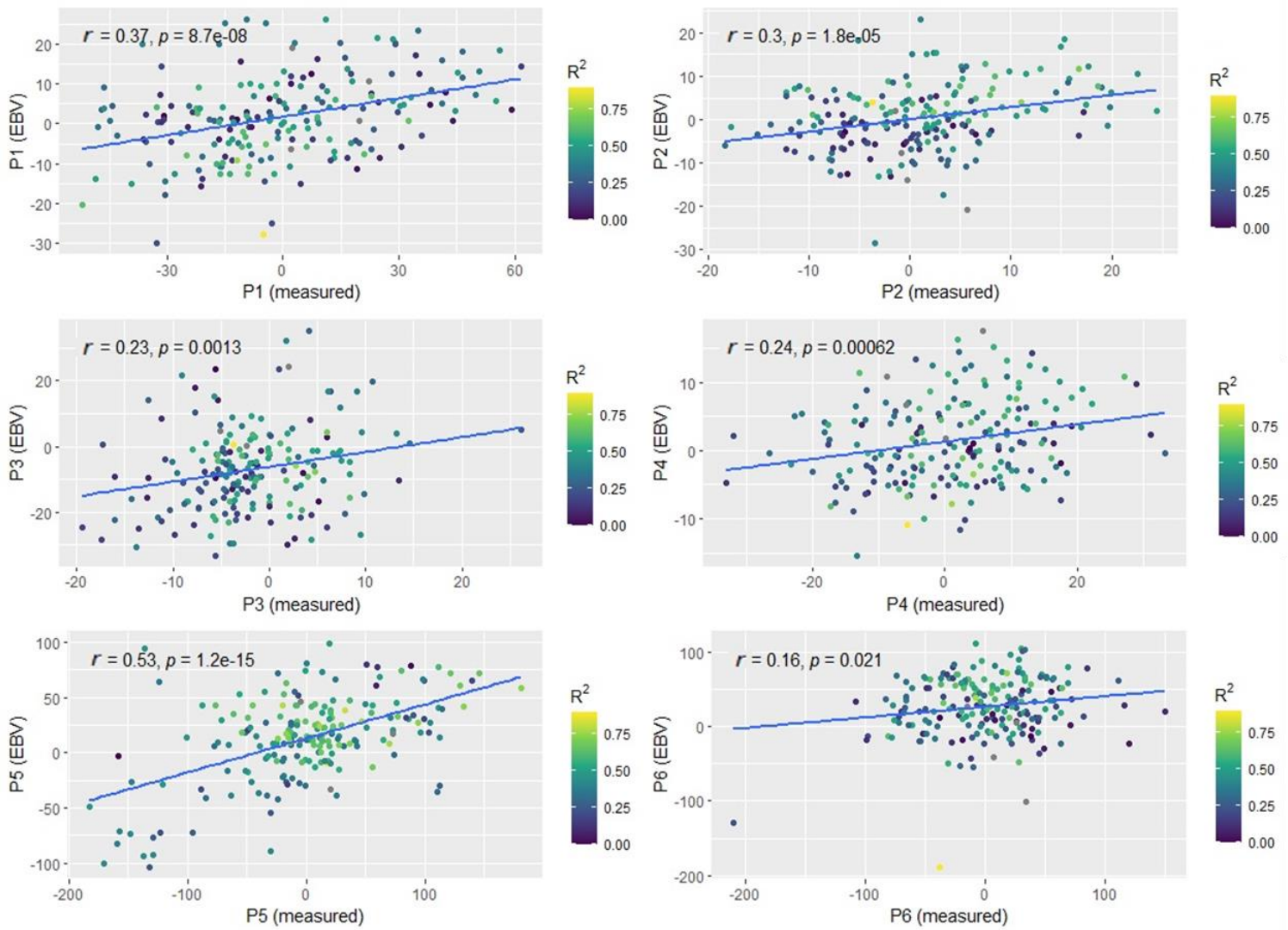

Figure 6.5: Scatter plot for empirical prediction accuracies (r) of gluten epitope concentrations (P1-P6) in the external validation population using training data from the primary training population (scenario 1). (R2 = Reliability of predictions; EBV = Estimated Breeding Value)

Empirical prediction accuracies for epitope concentrations scenarios 1-3 are presented in Figure 6.6. Empirical prediction accuracies for epitope concentrations in scenarios 1 and 2 appear similar in all cases, except when predicting P5 concentrations where empirical accuracies were lower in scenario 2 (Figure 6.6). Mean empirical prediction accuracies for epitope concentrations in scenario 3 were always higher than those observed in either scenarios 1 or 2 (Figure 6.6). Empirical prediction accuracies were higher in scenario 3 than those seen in scenarios 1 and 2 by between 0.04 for P1 up to 0.42 for P6 (Figure 6.6). Empirical prediction accuracies for P6 concentrations were over three times higher in scenario 3 than those observed in scenario 1 or 2 (Figure 6.6). The highest prediction accuracies were consistently observed for P5 concentrations, ranging from 0.43 in scenario 2 up to 0.67 in scenario 3 (Figure 6.6).

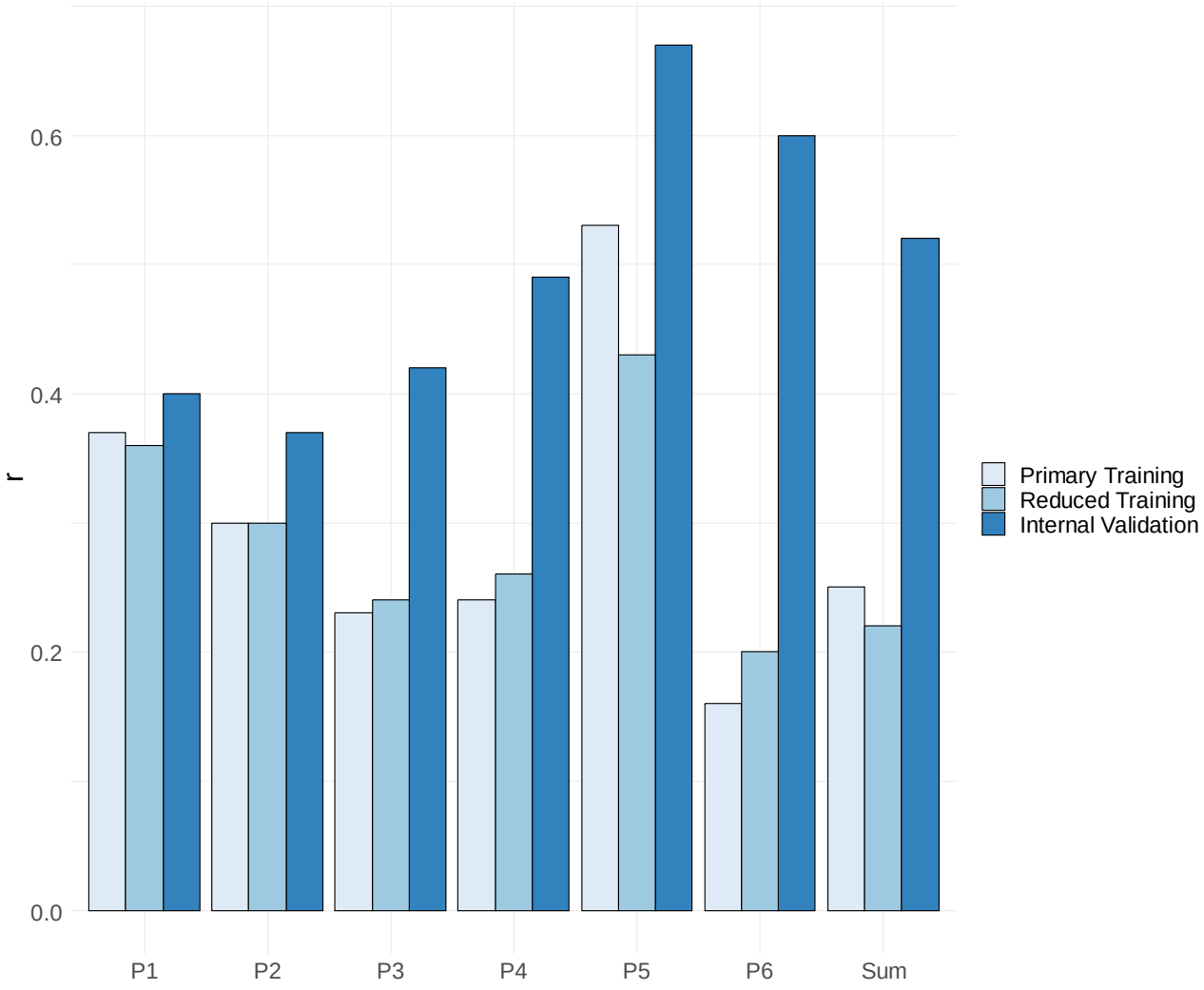

Figure 6.6: Empirical prediction accuracies (r) for individual gluten epitope concentrations (P1-P6) and total P1-P6 concentrations (Sum) in scenarios 1-3 (Scenario 1 = Primary Training, Scenario 2 = Reduced Training, Scenario 3 = Internal Validation). Empirical prediction accuracies in Scenario 3 are the mean accuracies from each iteration of the internal validation population.

Mean empirical prediction accuracies and 95% confidence intervals for iterations of the internal validation population are presented in Table 6.3. Empirical prediction accuracies were highest for FPC (r=0.75). Mean prediction accuracies were also high for P5 (r=0.67) and P6 (r=0.60) concentrations while accuracies for concentrations of the P1-P4 epitopes ranged from 0.37-0.49 (Table 6.3).

Table 6.3: Mean empirical accuracies (r) and 95% confidence intervals (CI upper, CI lower) for individual epitope concentrations (P1-P6), total P1-P6 concentrations (Sum) and Flour Protein Content (FPC) in the internal validation populations (scenario 3) from 100 independent iterations.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**Sum**	**FPC**
Mean	0.40	0.37	0.42	0.49	0.67	0.60	0.52	0.75
CI upper (95%)	0.48	0.49	0.56	0.55	0.74	0.74	0.71	0.79
CI lower (95%)	0.32	0.25	0.27	0.43	0.58	0.47	0.37	0.71

The distribution of empirical prediction accuracies for each iteration of the validation population is presented for epitope concentrations and FPC in Figure 6.7. Prediction accuracies appear to be most consistent in the internal validation populations for FPC, and concentrations of P4 and P5 (Figure 6.7). In most cases, empirical prediction accuracies for epitope concentrations in the internal validation populations appear to be normally distributed (Figure 6.7).

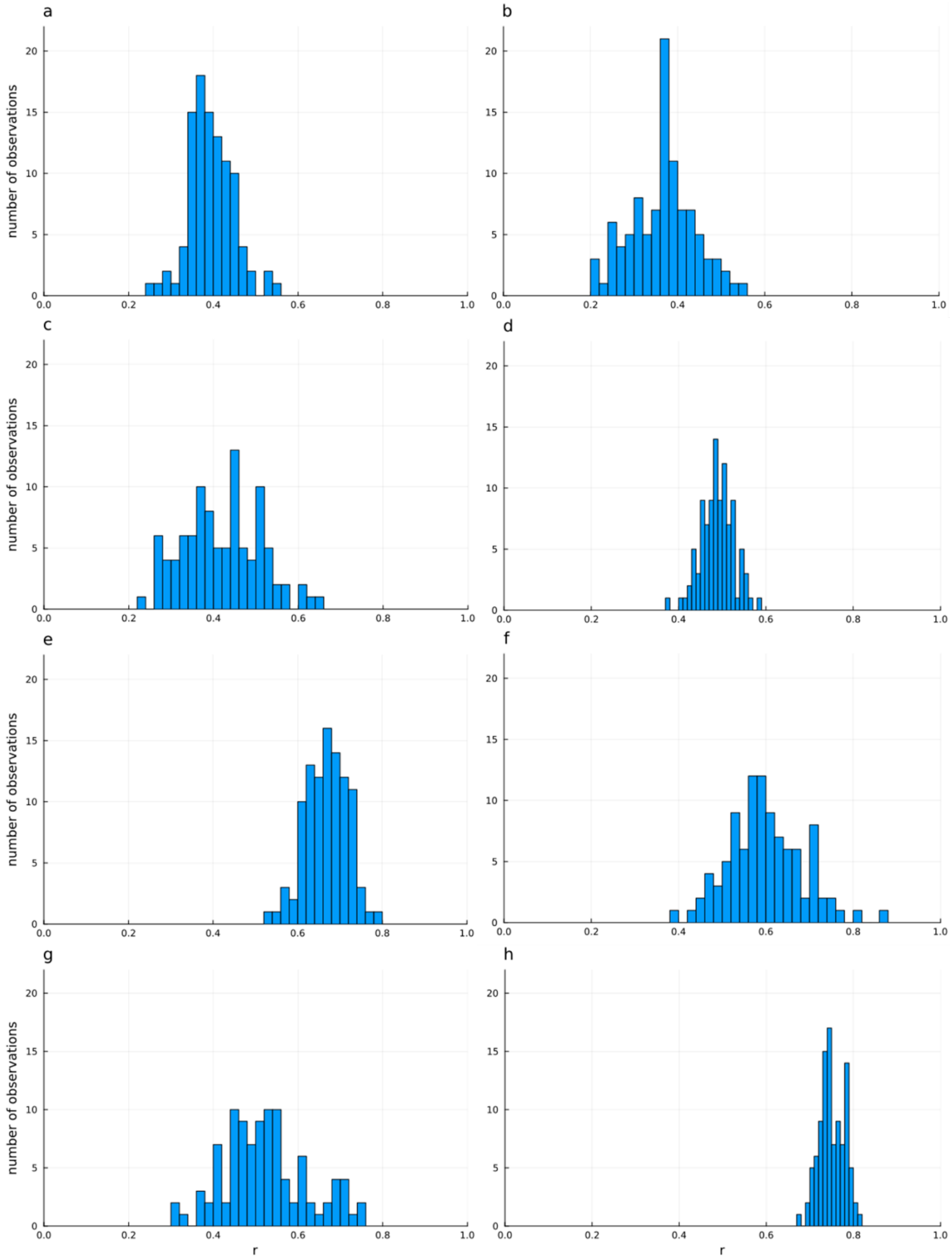

Figure 6.7: Distribution of empirical prediction accuracies (r) forindividual gluten epitope concentrations (P1-P6) total P1-P6 concentrations (Sum) and Flour Protein Content (FPC), (a-f = P1-P6, g = Sum, h = FPC), for each iteration of the internal validation population (scenario 3).

Realised prediction accuracies for epitope concentrations in Scenarios 1-3 are presented in Table 6.4. Realised accuracies ranged from as low as 0.22 for P6 concentrations in Scenario 1 up to 0.82 for P6 concentrations in scenario 3 (Table 6.4). Realised accuracies were consistent with the empirical accuracies presented in Figure 6.6.

Table 6.4: Realised accuracies for individual gluten epitope concentrations (P1-P6) and total P1-P6 concentrations (Sum) in Scenarios 1-3 (Scenario 1 = Primary Training, Scenario 2 = Reduced Training, Scenario 3 = Internal Validation).

	P1	P2	P3	P4	P5	P6	Sum
Primary Training	0.56	0.35	0.28	0.33	0.61	0.22	0.35
Reduced Training	0.54	0.35	0.29	0.35	0.50	0.27	0.31
Internal Validation	0.60	0.43	0.51	0.67	0.77	0.82	0.73

Empirical prediction accuracies for epitope concentrations were determined in the external validation population in scenario 1 using training data from individual seasons. These results are compared to the empirical prediction accuracies observed in the multi-year analysis (Figure 6.8). Empirical prediction accuracies were always highest using the combined years data (Figure 6.8).

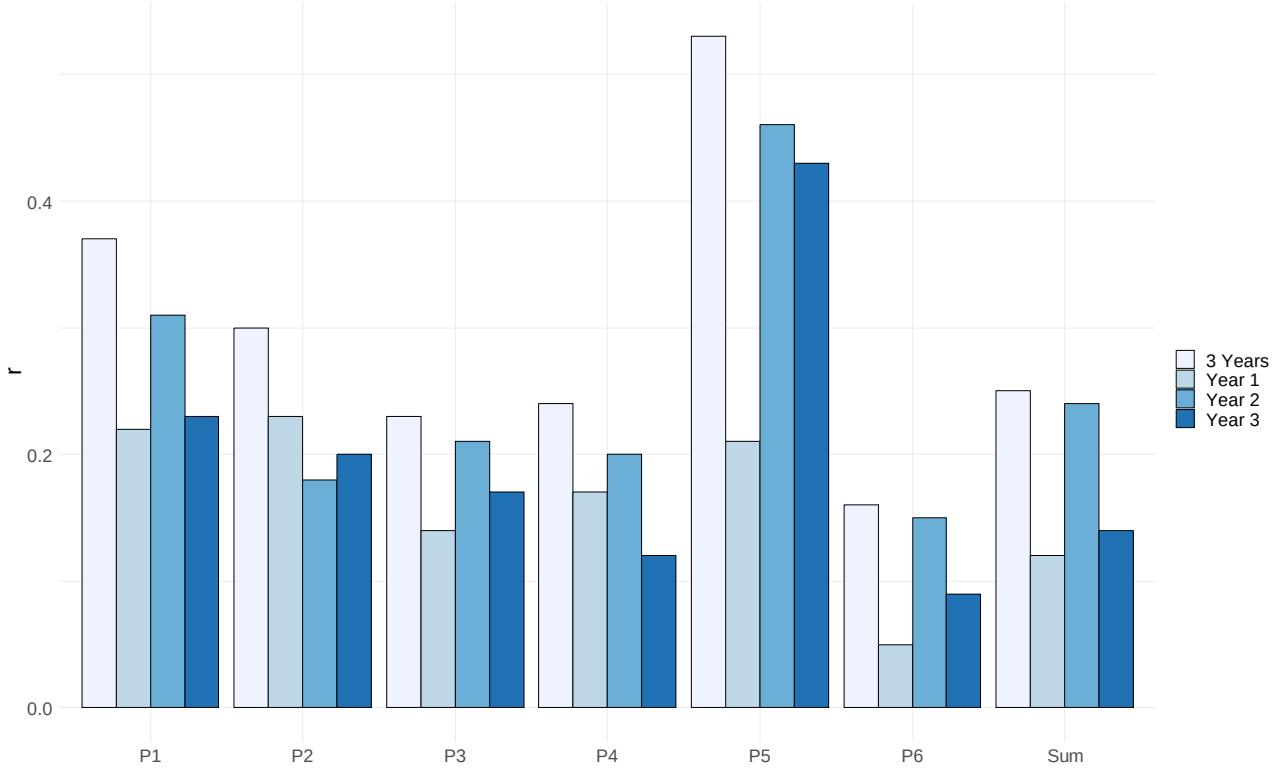

Figure 6.8: Empirical prediction accuracies (r) for individual gluten epitope concentrations (P1-P6) and total P1-P6 concentrations (Sum) in single year analyses (Year 1, Year 2, Year 3) and multi-year analyses (3 Years) for the external validation population in scenario 1.

6.5 Discussion

The results of this study indicate that there is potential for employing genomic breeding techniques to identify wheat lines with lower concentrations of gluten epitopes. However, there are a number of factors to consider that appear to affect the accuracy and therefore utility of genomic based selections. First, using multiple seasons data to predict epitope concentrations appears to provide more accurate predictions than any of the single year analyses (Figure 6.8). These results are as expected seeing that a multi-year dataset tends to average out the effects of genotype-by-year interactions and allows the model to identify patterns that may not be recognisable in a smaller dataset (Norman et al., 2018). It was expected that the year 3 data may produce higher prediction accuracies in scenario 1 than any of the other single year data. This is due to the external validation population being grown in year 3, directly alongside the training population. Therefore, the validation population would have experienced near identical environmental conditions to the training population in year 3. However, it does not appear that the year 3 prediction accuracies were any greater in scenario 1 than either the year 1 or year 2 predictions (Figure 6.8). In order to obtain accurate predictions for a trait, it is also important that the genotypes of the target population are representative of the training population. This means that the predictive model can more accurately consider the effects of most alleles present in the target populations as in most cases, those same alleles will be present in the training population (Crossa et al., 2017). In this study, the primary training population appeared to have a much greater range of genetic diversity than the external validation population tested in scenario 1 (Figure 6.2). The empirical prediction accuracies observed for epitope concentrations in this scenario were generally low, ranging from 0.16 for P6 to 0.53 for P5 (Figure 6.5). Scenario 2 was designed to determine if the empirical prediction accuracies observed in scenario 1 could be increased by using only training data that was representative of the external validation population. These scenarios give an indication of the expected prediction accuracies that could be achieved if the data obtained for the primary training population were used to predict epitope concentrations in the PFR breeding program. The empirical prediction accuracies observed in scenario 2 differed by no more than 0.05 to those in scenario 1, except for P5 concentrations where the empirical prediction accuracy was higher by 0.1 in scenario 1 than scenario 2 (Figure 6.6). The absence of any major improvements in prediction accuracies in scenario 2 may be due to the smaller size of the reduced training population which consisted of only 237 lines compared to the 471 lines in the primary training population. Scenario 3 was employed to further understand the effects of population structure on prediction accuracies for epitope concentrations. By employing a randomly assigned internal validation population, it is assumed that in most cases, this population will be generally representative of the internal training population. Additionally, this strategy is able to retain a high number of training lines for determining EBV's in the internal validation population. This structure is expected to be the most effective method for producing high prediction accuracies that can be employed for genomic

selection (GS) in a breeding program. In scenario 3, mean empirical prediction accuracies were always higher than those observed in scenarios 1 and 2 (Figure 6.6). This is particularly the case for the P6 epitope, where prediction accuracies were over 3 times higher than those observed in scenarios 1 or 2 (Figure 6.6). Empirical prediction accuracies were only 0.04 higher for P1 concentrations and 0.07 for P2 concentrations in scenario 3 than in scenarios 1 or 2 (Figure 6.6). In all other cases, empirical prediction accuracies for epitope concentrations were at least 0.18 higher in scenario 3 than either scenarios 1 or 2 (Figure 6.6). These results suggest that the population structure of the training and target populations is highly important in determining the accuracy of genomic prediction. This is an important consideration for breeders as it means that ongoing genotyping and phenotyping is required to maintain a suitable training population for calculating EBV's in breeding lines. Additionally, breeders may need to continue phenotyping breeding lines to determine the ongoing reliability of their predictions when employing GS. Therefore, in order to most effectively employ GS in the PFR wheat breeding program, more training data that is representative of the breeding population is required. These are significant factors for breeders to consider when aiming to implement methods for genomic selection. Large amounts of data, preferably collected across multiple seasons, are required to achieve useful prediction accuracies and the training population needs to be representative of the breeding population (Crossa et al., 2017). The training population also needs to be sufficiently diverse to sample as much genetic diversity for the trait of interest as is possible. Genomic prediction can be a useful tool for breeders to reduce expensive phenotyping costs and improve the accuracy of their selections for multiple quantitative traits (Ibba et al., 2020). However, the establishment costs to sufficiently phenotype a large enough training population is a significant barrier to entry (Budhlakoti et al., 2022). Additionally, as the breeding population progresses over subsequent breeding cycles, it is likely that the genetic diversity of the population will continue to shrink in response to selection criteria based on the information from a static training population (Schulthess et al., 2022). Therefore, breeders will need to continually add new phenotyped and genotyped material to their training population and continually refine their genomic model in order to maintain sufficient genetic diversity in the breeding program and ensure continued improvement in their selections.

The application of genomic prediction in wheat breeding programs is slowly gaining traction around the world thanks to increasingly cost-effective, high-throughput SNP genotyping and improved mapping of the wheat genome (Sandhu et al., 2022; Sun et al., 2020). However, several unique challenges exist for wheat breeders. First, wheat is an inbred crop, meaning that many years of repeated self-pollination is required to achieve a homozygous cultivar (Alahmad et al., 2022). Therefore, early generation breeding lines are at least partially heterozygous for many important alleles and segregation is common between years. This means that selections made on early generation breeding material can be unreliable as the genetic makeup of the population can change between generations. However, this becomes less of a problem as each subsequent

generation becomes increasingly fixed. Therefore, to apply methods of genomic prediction in a wheat breeding program, breeders need to strike a balance between making selections on early generation material versus later generation material (Bassi et al., 2016). Selections at an early generation may allow breeders to cut down on material early and increase the efficiency of their program but these selections may eventually prove unreliable due to segregation in following generations. In contrast, by waiting to genotype breeding lines at a later generation, breeders are required to carry through larger numbers of lines to a later stage and therefore may miss many of the potential advantages provided by genomic prediction. Furthermore, the wheat genome is large and relatively complex and much of it remains poorly understood. These characteristics make identifying causal genes particularly challenging, thus limiting the ability of breeders to make informed decisions based only on genotypic data (Schulthess et al., 2022; Walkowiak et al., 2020). A common approach to overcoming some of these challenges is to employ double haploid (DH) populations in the breeding program (Humphreys & Knox, 2015). This process allows breeders to produce a homozygous line in a single generation. This means that genomic selection can be employed in the F_2 generation and used to immediately reduce the size of the breeding population before any major field trials are sown (Song et al., 2017). However, the main challenges of a DH program include the high cost of producing a sufficient breeding population and the occasional occurrence of recalcitrant genotypes with low responsiveness to DH crossing (Humphreys & Knox, 2015).

Despite these challenges, genomic prediction continues to show promise for a number of complex, quantitative traits in wheat (Sandhu et al., 2021). The nutritional composition of wheat is becoming an important focus for breeders globally but many such traits, for example, gluten epitope concentrations, are difficult, costly and time consuming to measure (Govindan et al., 2022; Shewry & Tatham, 2016). These characteristics mean that genomic prediction may offer significant advantages over traditional selection methods. In this study, given that particular requirements for the amount of training data and the composition of the training and validation populations are met (scenario 3), relatively high empirical prediction accuracies can be achieved for epitope concentrations (Figure 6.6). These results correspond with realised accuracies of between 0.43 for P2 concentrations to 0.82 for P6 concentrations (scenario 3, Table 6.4). Realised accuracies consider both the strength of the relationship between the EBV and BV, and the strength of the relationship between BV and phenotype. This allows more accurate comparisons of prediction accuracy between traits with different heritability's considering that empirical prediction accuracies cannot theoretically exceed the square root heritability of a trait. The high realised accuracies seen for P5 and P6 concentrations in particular in scenario 3 provide confidence that the EBV's for these traits are relatively close to the true BV. This is particularly encouraging for breeding towards low P5 wheat lines as this epitope has shown to be the most immunodominant of all the identified gluten epitopes. Therefore selections towards new wheat cultivars with reduced P5 concentrations are likely to offer the

greatest improvement towards wheat that is more suited to those with gluten sensitivity(Shewry & Tatham, 2016).

Flour Protein Content is an important trait in wheat which is often used as an indicator of baking quality (Laidig et al., 2022; Xue et al., 2019). Although FPC is a relatively easy trait to measure, it is also known to be moderately-highly heritable (Kramer, 1979; Suprayogi et al., 2009). Evidence of this can be seen in the high mean empirical prediction accuracy observed for FPC in scenario 3 (0.75, Table 6.3). This suggests that using a genomic prediction model to select for FPC in wheat is likely to be an effective breeding tool. Relatively strong associations between FPC and concentrations of some epitopes, as well as between FPC and baking quality, were reported in Chapter 4. Therefore, it is suggested that the most effective selections towards new wheat cultivars with low concentrations of gluten epitopes but high levels of baking quality, may be based on targeting lines with a low epitope to high protein ratio. Due to the relatively reliable predictions for FPC and concentrations of the P5 and P6 epitopes in particular, such selections based on genomic data have the potential to be highly effective for these traits. As the interactions between epitope concentrations and baking quality become better understood, along with additional research into the genetic controls and the immunogenicity of different gluten epitopes, selection indices can begin to be developed to allow breeders to employ GS in wheat for multiple traits. These tools are currently well established in livestock breeding and are likely to become defining characteristics of modern wheat breeding programs.

6.6 References

Abdi, F., Zuberi, S., Blom, J.-J., Armstrong, D., & Pinto-Sanchez, M. I. (2023). Nutritional Considerations in Celiac Disease and Non-Celiac Gluten/Wheat Sensitivity. *Nutrients*, *15*(6).

Alahmad, S., Rambla, C., Voss-Fels, K. P., & Hickey, L. T. (2022). Accelerating Breeding Cycles. In M. P. Reynolds & H.-J. Braun (Eds.), *Wheat Improvement: Food Security in a Changing Climate* (pp. 557-571). Springer International Publishing. https://doi.org/10.1007/978-3-030-90673-3_30

Andrén Aronsson, C., Lee, H.-S., Hård af Segerstad, E. M., Uusitalo, U., Yang, J., Koletzko, S., . . . for the, T. S. G. (2019). Association of Gluten Intake During the First 5 Years of Life With Incidence of Celiac Disease Autoimmunity and Celiac Disease Among Children at Increased Risk. *JAMA*, *322*(6), 514-523. https://doi.org/10.1001/jama.2019.10329

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., & Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). *Plant Science*, *242*, 23-36. https://doi.org/https://doi.org/10.1016/j.plantsci.2015.08.021

Bernardo, D., & Peña, A. S. (2012). Developing strategies to improve the quality of life of patients with gluten intolerance in patients with and without coeliac disease. *European journal of internal medicine*, *23*(1), 6-8. https://doi.org/10.1016/j.ejim.2011.09.016

Brown, N. K., Guandalini, S., Semrad, C., & Kupfer, S. S. (2019). A Clinician's Guide to Celiac Disease HLA Genetics. *Official journal of the American College of Gastroenterology | ACG*, *114*(10). https://journals.lww.com/ajg/Fulltext/2019/10000/A_Clinician_s_Guide_to_Celiac_Disease_HLA_Genetics.9.aspx

Budhlakoti, N., Kushwaha, A. K., Rai, A., Chaturvedi, K. K., Kumar, A., Pradhan, A. K., . . . Kumar, S. (2022). Genomic Selection: A Tool for Accelerating the Efficiency of Molecular Breeding for Development of Climate-Resilient Crops [Review]. *Frontiers in Genetics*, *13*. https://www.frontiersin.org/articles/10.3389/fgene.2022.832153

Cheng, H., Fernando, R. L., & Garrick, D. J. (2018). JWAS: Julia implementation of whole-genome analysis software. Proceedings of the World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., . . . Varshney, R. K. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, *22*(11), 961-975. https://doi.org/https://doi.org/10.1016/j.tplants.2017.08.011

Gil-Humanes, J., Pistón, F., Altamirano-Fortoul, R., Real, A., Comino, I., Sousa, C., . . . Barro, F. (2014). Reduced-Gliadin Wheat Bread: An Alternative to the Gluten-Free Diet for Consumers Suffering Gluten-Related Pathologies. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0090898

Govindan, V., Michaux, K. D., & Pfeiffer, W. H. (2022). Nutritionally Enhanced Wheat for Food and Nutrition Security. In M. P. Reynolds & H.-J. Braun (Eds.), *Wheat Improvement: Food Security in a Changing Climate* (pp. 195-214). Springer International Publishing. https://doi.org/10.1007/978-3-030-90673-3_12

Greco, L., Gobbetti, M., Auricchio, R., Di Mase, R., Landolfo, F., Paparo, F., . . . Auricchio, S. (2011). Safety for Patients With Celiac Disease of Baked Goods Made of Wheat Flour Hydrolyzed During Food Processing. *Clinical Gastroenterology and Hepatology*, *9*(1), 24-29. https://doi.org/https://doi.org/10.1016/j.cgh.2010.09.025

Green, P. H. R., & Cellier, C. (2007). Celiac Disease. *New England Journal of Medicine*, *357*(17), 1731-1743. https://doi.org/10.1056/NEJMra071600

Gujral, N., Freeman H, J., & Thomson, A. B. (2012). Celiac disease: prevalence, diagnosis, pathogenesis and treatment. *World Journal of Gastroenterology*, *18*(42), 6036-6059. https://doi.org/10.3748/wjg.v18.i42.6036

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*(1), 186. https://doi.org/10.1186/1471-2105-12-186

Humphreys, D. G., & Knox, R. E. (2015). Doubled Haploid Breeding in Cereals. In J. M. Al-Khayri, S. M. Jain, & D. V. Johnson (Eds.), *Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools* (pp. 241-290). Springer International Publishing. https://doi.org/10.1007/978-3-319-22521-0_9

Ibba, M. I., Crossa, J., Montesinos-López, O. A., Montesinos-López, A., Juliana, P., Guzman, C., . . . Poland, J. (2020). Genome-based prediction of multiple wheat quality traits in multiple years. *The Plant Genome*, *13*(3), e20034. https://doi.org/https://doi.org/10.1002/tpg2.20034

King, J. A., Jeong, J., Underwood, F. E., Quan, J., Panaccione, N., Windsor, J. W., . . . Kaplan, G. G. (2020). Incidence of Celiac Disease Is Increasing Over Time: A Systematic Review and Meta-analysis. *Official journal of the American College of Gastroenterology | ACG*, *115*(4). https://journals.lww.com/ajg/Fulltext/2020/04000/Incidence_of_Celiac_Disease_Is_Increasing_Over.9.aspx

Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*(1), 44-53. https://doi.org/https://doi.org/10.1111/1755-0998.12549

Kramer, T. (1979). Environmental and genetic variation for protein content in winter wheat (Triticum aestivum L.). *Euphytica*, *28*(2), 209-218. https://doi.org/10.1007/BF00056577

Laidig, F., Hüsken, A., Rentel, D., & Piepho, H.-P. (2022). Protein use efficiency and stability of baking quality in winter wheat based on the relation of loaf volume and grain protein content. *Theoretical and Applied Genetics*, *135*(4), 1331-1343. https://doi.org/10.1007/s00122-022-04034-x

Lebwohl, B., & Rubio-Tapia, A. (2021). Epidemiology, Presentation, and Diagnosis of Celiac Disease. *Gastroenterology*, *160*(1), 63-75. https://doi.org/https://doi.org/10.1053/j.gastro.2020.06.098

Lindfors, K., Ciacci, C., Kurppa, K., Lundin, K. E. A., Makharia, G. K., Mearin, M. L., . . . Kaukinen, K. (2019). Coeliac disease. *Nature Reviews Disease Primers*, *5*(1), 3. https://doi.org/10.1038/s41572-018-0054-z

Loponen, J., Sontag-Strohm, T., Venäläinen, J., & Salovaara, H. (2007). Prolamin Hydrolysis in Wheat Sourdoughs with Differing Proteolytic Activities. *Journal of Agricultural and Food Chemistry*, *55*(3), 978-984. https://doi.org/10.1021/jf062755g

Lund-Blix, N. A., Mårild, K., Tapia, G., Norris, J. M., Stene, L. C., & Størdal, K. (2019). Gluten Intake in Early Childhood and Risk of Celiac Disease in Childhood: A Nationwide Cohort Study. *Official journal of the American College of Gastroenterology | ACG*, *114*(8). https://journals.lww.com/ajg/Fulltext/2019/08000/Gluten_Intake_in_Early_Childhood_and_Risk_of.21.aspx

Mårild, K., Dong, F., Lund-Blix, N. A., Seifert, J., Barón, A. E., Waugh, K. C., . . . Norris, J. M. (2019). Gluten Intake and Risk of Celiac Disease: Long-Term Follow-up of an At-Risk Birth Cohort. *Official journal of the American College of Gastroenterology | ACG*, *114*(8). https://journals.lww.com/ajg/Fulltext/2019/08000/Gluten_Intake_and_Risk_of_Celiac_Disease_.22.aspx

Molberg, Ø., Uhlen, A. K., Jensen, T., Flæte, N. S., Fleckenstein, B., Arentz–Hansen, H., . . . Sollid, L. M. (2005). Mapping of gluten T-cell epitopes in the bread wheat ancestors: Implications for celiac disease. *Gastroenterology*, *128*(2), 393-401. https://doi.org/https://doi.org/10.1053/j.gastro.2004.11.003

Montesinos-López, A., Runcie, D. E., Ibba, M. I., Pérez-Rodríguez, P., Montesinos-López, O. A., Crespo, L. A., . . . Crossa, J. (2021). Multi-trait genomic-enabled prediction enhances accuracy in multi-year wheat breeding trials. *G3 (Bethesda)*, *11*(10). https://doi.org/https://doi.org/10.1093/g3journal/jkab270

Norman, A., Taylor, J., Edwards, J., & Kuchel, H. (2018). Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3 Genes|Genomes|Genetics*, *8*(9), 2889-2899. https://doi.org/10.1534/g3.118.200311

Romanos, J., van Diemen, C. C., Nolte, I. M., Trynka, G., Zhernakova, A., Fu, J., . . . Wijmenga, C. (2009). Analysis of HLA and Non-HLA Alleles Can Identify Individuals at High Risk for Celiac Disease. *Gastroenterology*, *137*(3), 834-840.e833. https://doi.org/https://doi.org/10.1053/j.gastro.2009.05.040

Salentijn, E. M. J., Esselink, D. G., Goryunova, S. V., van der Meer, I. M., Gilissen, L. J. W. J., & Smulders, M. J. M. (2013). Quantitative and qualitative differences in celiac disease epitopes among durum wheat varieties identified through deep RNA-amplicon sequencing. *BMC Genomics*, *14*(1), 905. https://doi.org/10.1186/1471-2164-14-905

Salentijn, E. M. J., Mitea, D. C., Goryunova, S. V., van der Meer, I. M., Padioleau, I., Gilissen, L. J. W. J., . . . Smulders, M. J. M. (2012). Celiac disease T-cell epitopes from gamma-gliadins: immunoreactivity depends on the genome of origin, transcript frequency, and flanking protein variation. *BMC Genomics*, *13*(1), 277. https://doi.org/10.1186/1471-2164-13-277

Sanchez-Leon, S., Gil-Humanes, J., Ozuna, C. V., Gimenez, M. J., Sousa, C., Voytas, D. F., & Barro, F. (2018). Low-gluten, nontransgenic wheat engineered with CRISPR/Cas9. *Plant Biotechnol J*, *16*(4), 902-910. https://doi.org/10.1111/pbi.12837

Sandhu, K. S., Merrick, L. F., Sankaran, S., Zhang, Z., & Carter, A. H. (2022). Prospectus of Genomic Selection and Phenomics in Cereal, Legume and Oilseed Breeding Programs [Review]. *Frontiers in Genetics*, *12*. https://www.frontiersin.org/articles/10.3389/fgene.2021.829131

Sandhu, K. S., Mihalyov, P. D., Lewien, M. J., Pumphrey, M. O., & Carter, A. H. (2021). Combining Genomic and Phenomic Information for Predicting Grain Protein Content and Grain Yield in Spring Wheat [Original Research]. *Frontiers in Plant Science*, *12*. https://www.frontiersin.org/articles/10.3389/fpls.2021.613300

Schulthess, A. W., Kale, S. M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., . . . Reif, J. C. (2022). Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. *Nature Genetics*, *54*(10), 1544-1552. https://doi.org/10.1038/s41588-022-01189-7

Shewry, P. R., & Tatham, A. S. (2016). Improving wheat to remove coeliac epitopes but retain functionality. *Journal of Cereal Science*, *67*, 12-21. https://doi.org/https://doi.org/10.1016/j.jcs.2015.06.005

Shiferaw, B., Smale, M., Braun, H.-J., Duveiller, E., Reynolds, M., & Muricho, G. (2013). Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. *Food Security*, *5*(3), 291-317. https://doi.org/10.1007/s12571-013-0263-y

Singh, P., Arora, A., Strand, T. A., Leffler, D. A., Catassi, C., Green, P. H., . . . Makharia, G. K. (2018). Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis. *Clinical Gastroenterology and Hepatology*, *16*(6), 823-836.e822. https://doi.org/https://doi.org/10.1016/j.cgh.2017.06.037

Song, J., Carver, B. F., Powers, C., Yan, L., Klápště, J., El-Kassaby, Y. A., & Chen, C. (2017). Practical application of genomic selection in a doubled-haploid winter wheat breeding program. *Molecular Breeding*, *37*(10), 117. https://doi.org/10.1007/s11032-017-0715-8

Sun, J., Khan, M., Amir, R., & Gul, A. (2020). Genomic selection in wheat breeding. In M. Ozturk & A. Gul (Eds.), *Climate Change and Food Security with Emphasis on Wheat* (pp. 321-330). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-819527-7.00023-6

Suprayogi, Y., Pozniak, C. J., Clarke, F. R., Clarke, J. M., Knox, R. E., & Singh, A. K. (2009). Identification and validation of quantitative trait loci for grain protein concentration in adapted Canadian durum wheat populations. *Theoretical and Applied Genetics*, *119*(3), 437-448. https://doi.org/10.1007/s00122-009-1050-1

van den Broeck, H. C., Gilissen, L. J., Smulders, M. J., van der Meer, I. M., & Hamer, R. J. (2011). Dough quality of bread wheat lacking α-gliadins with celiac disease epitopes and addition of celiac-safe avenins to improve dough quality. *Journal of Cereal Science*, *53*(2), 206-216.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., . . . Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature*, *588*(7837), 277-283. https://doi.org/10.1038/s41586-020-2961-x

Walter, T., Wieser, H., & Koehler, P. (2014). Degradation of gluten in wheat bran and bread drink by means of a proline-specific peptidase. *J Nutr Food Sci*, *4*(5), 1-6.

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

Wieser, H., Koehler, P., Folck, A., & Becker, D. (2007). Characterization of wheat with strongly reduced α-gliadin content. *Gluten proteins*, 13-16.

Xue, C., Matros, A., Mock, H.-P., & Mühling, K.-H. (2019). Protein Composition and Baking Quality of Wheat Flour as Affected by Split Nitrogen Application. *Frontiers in Plant Science*, *10*. https://www.frontiersin.org/articles/10.3389/fpls.2019.00642

Chapter 7

General discussion

7.1 Implementation of genomic selection in wheat

Breeding via genomic selection (GS) has proven to be a highly effective method for the rapid improvement of a range of complex, quantitative traits in livestock and a number of plant species (Atanda et al., 2021; Rutkoski et al., 2016; Stock & Reents, 2013). Uptake for GS has been relatively slow for wheat breeding, primarily due to a number of unique challenges presented by the wheat genome (Bassi et al., 2016). These include its large genome size (c.17 Gb), its combination of three distinct sub-genomes, and the presence of a high number of transposable elements and repetitive sequences (Uauy, 2017). These characteristics have meant that mapping of the wheat genome has been relatively slow and arduous. However, a number of high quality, high density and high throughput genotyping methods are now available to wheat breeders (Allen et al., 2017; Rasheed & Xia, 2019). The understanding of the genome is now at a point where GS is becoming increasingly practical and effective in commercial wheat breeding programs (Kumar et al., 2021). In fact, relatively high prediction accuracies have been shown for several valuable agronomic traits including grain yield and protein content and a number of breeders are now regularly implementing GS in the programs (Lozada et al., 2019; Tessema et al., 2020). In this study, we have shown that moderate-high prediction accuracies can be achieved for estimating the concentrations of specific gluten epitopes, provided that the training data has sufficient quantity and relevance for the target population (Chapter 6). Therefore, it is proposed that genomic prediction may indeed be an effective tool for developing low epitope wheat cultivars. Quantifying the concentrations of gluten epitopes in wheat is a highly time consuming and expensive exercise. This means that GS offers significant advantages over more traditional phenotype-based selections as it greatly reduces the amount of phenotyping required. This is an important consideration as GS tools require significant upfront investment in order to phenotype and genotype an adequate training population. Furthermore, in the long term, effective GS in a breeding program requires continually adding phenotypic and genotypic information to maintain diversity in the gene pool and maximise prediction accuracies (Larkin et al., 2019).

Developing a reliable prediction model from a large enough training population is one of the most significant barriers to entry for GS methods in breeding (Budhlakoti et al., 2022). This is simply due to the time and expense involved with collecting sufficient genotypic and phenotypic information. However, this has now been achieved in this study meaning that GS could be introduced in the New Zealand Institute for Plant & Food Research (PFR) wheat breeding program. Using this data, prediction accuracies are likely to range between 0.16-0.53 for different epitopes as seen in scenarios 1 and 2 that were tested in Chapter 6. However, if additional genotyping and phenotyping is completed on more PFR breeding lines, prediction accuracies can expect to increase to between 0.37-0.67 as seen in scenario 3, also presented in Chapter 6. Additional genotyping and phenotyping will likely be an ongoing requirement to ensure genetic diversity is captured and prediction accuracies remain reliable in the future. However, as genotyped resources

accumulate, predictions are likely to become increasingly reliable (Larkin et al., 2019). Wheat is an inbred species, meaning that it takes several years of inbreeding to establish a fixed line from an F1 cross (Rutkoski et al., 2022). Genetic segregation occurs between each successive generation, meaning the allele frequencies of a line change from one generation to the next until fixation is achieved. Ideally, GS would be applied throughout the breeding cycle so that breeders can ensure they are stacking as many beneficial alleles of interest as possible. However, in some programs, genotyping all lines in a breeding program every year is unlikely to be a practical approach. Therefore, a balance must be struck to maximise the effectiveness of genotyping that is undertaken. Genomic Selection may be most reliable when employed at a later stage due to the fact that genotypic information collected on early generation lines will be incomplete and no longer accurate in the following generation. However, an effective breeding program needs to be as efficient as possible and if GS is not being employed at early generations, then laborious phenotyping is required for making selections at these stages. The stages at which different phenotyping measurements are taken in the PFR wheat breeding program are described in Table 7.1. This table illustrates how most of the expensive and time-consuming measurements are taken at later generations. One of the key benefits of GS is that it allows for reduced phenotyping of lines and therefore increased breeding efficiency. These benefits are significantly reduced if GS is not employed until the final stages of breeding. Therefore, the most effective stage for employing GS in a traditional breeding program is likely to be around the F4-F5 generations where lines are around 94-97% fixed and segregation is greatly reduced when compared to that in the F1-F3 generations (Table 7.2). This would allow for a balance between the competing factors mentioned above, assuming that genotyping all generations is not a viable option. Alternatively, double haploid (DH) crossing could be implemented in order to produce homozygous F2 lines. This is a common strategy for large breeding programs; however, the resource requirements are often too great for smaller programs (Alahmad et al., 2022; Humphreys & Knox, 2015). For small scale wheat breeding programs, such as those that exist in New Zealand, targeted DH crossing may be useful for a small number of crosses. In the case of epitope concentrations, DH crossing could be employed to produce lines with exceptionally low concentrations of epitopes and these lines could quicky be identified through GS. These lines could then be incorporated as parents in traditional crossing blocks to reduce epitope concentrations in the breeding population whilst maintaining agronomic and baking quality characteristics.

Table 7.1: Breeding selection criteria at different generations (F2-F4, F5 and F6+) in the Plant & Food Research wheat breeding program

F2-F4	F5	F6+
● Plant type (height, straw strength) ● Disease resistance	● Disease resistance ● Grain quality (size, density) ● Dough strength (Mixograph) ● Sprouting risk ● Grain protein content	● Disease resistance ● Grain quality (size, density) ● Sprouting risk ● Grain protein content ● Milling quality (extraction rate, grain hardness) ● Baking quality (loaf volume, loaf texture, water absorption, development time)

Table 7.2: Approximate levels of homozygosity in wheat lines at different generations in a breeding program (Rutkoski et al., 2022)

Generation	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**
% Homozygous (Fixed alleles)	50%	75%	88%	94%	97%	98%	99%	100%

The genomic models employed in this study are used to provide estimated breeding values (EBV's) for individuals of interest. These values represent the genetic merit of individuals for each trait of interest. Numerous approaches have been employed for utilising EBV's to make effective selections in a breeding program. The development and implementation of selection indices has become a popular approach, particularly in the livestock industry, to allow breeders to select individuals with multiple desirable characteristics. Selection indices require breeders to assign relative values, often a dollar figure, to particular traits based on their importance and other trait characteristics such as heritability, genetic correlations and their relevance to breeding objectives. These weighted values are added together to determine the overall value of an individual and can therefore be used in selecting superior lines based on multiple traits. The development of selection indices requires large amounts of data and can quickly become highly complex due to the range of interacting phenotypic and genotypic factors. These tools are likely to be useful for selecting low gluten epitope wheats as they will allow breeders to account for the differing characteristics of individual epitopes. For example, reductions in concentrations of epitopes such as P5 which has been shown to be highly immunodominant are likely to be of more value than the same volume of reduction of less immunogenic epitopes. Therefore, through the development of selection indices, breeders will be able to account for these variations when making selections and ensure that they are maximising the value they can achieve through their breeding selections. Furthermore, selection indices are able to consider a wide range of traits, including the agronomic and quality characteristics of wheat. By

considering the genetic associations observed in this study between epitope concentrations and specific baking quality characteristics, breeders may be able to use selection indices for optimising selections so that high levels of baking quality are retained in the breeding population while epitope concentrations are reduced. The establishment of a selection index for the PFR wheat breeding program will require ongoing development and breeder observations. However, these tools are likely to eventually allow significant improvements in the accuracy and effectiveness of breeding selections overtime.

7.2 Challenges for producing low gluten epitope products

In this study, the concentrations of 6 distinct gluten epitopes were measured for genomic analyses. However, there are at least 27 epitopes present in wheat that are known to trigger coeliac disease (Chlubnová et al.). Therefore, whilst this work has been effective in understanding the genetic regulation of the 6 tested epitopes, there remains great potential to analyse additional proteins in order to further reduce total epitope concentrations in wheat. The results of this study have indicated that selective breeding, either through traditional phenotype-based methods or GS, will be effective in reducing the concentrations of certain gluten epitopes. It is expected that reduced dietary exposure to gluten epitopes is likely to have far reaching benefits for those at risk of coeliac disease (CD) or who are suffering from gluten intolerance (Shewry & Tatham, 2016). However, the extent to which epitope concentrations must be reduced in foods to achieve such benefits remains relatively unknown. Numerous studies have shown that the interaction of gluten epitopes with the immune system of coeliac patients, along with the severity of symptoms, can vary widely between individuals (Ciccocioppo et al., 2005; De Re et al., 2013; Koning, 2021; Mowat, 2003). Furthermore, the triggers and pathways involved in non-coeliac gluten sensitivity (NCGS) remain poorly understood, along with the role of gluten epitopes in this condition. This makes it difficult for breeders to establish useful breeding targets for low epitope phenotypes and also creates challenges around establishing and making health claims on low epitope food products. To overcome these challenges, it is likely that large scale and comprehensive clinical trials would be required in order to truly quantify the benefits of reduced epitope consumption. This poses significant challenges with regards to ethical considerations and resourcing such trials. It seems likely that this question will remain unanswered for some time yet.

7.3 Consumer education

The gluten free diet (GFD) has become increasingly popular across the western world over recent decades (Newberry et al., 2017). Whilst this is a requirement for those suffering from CD, it is believed that many consumers choose to avoid gluten simply due to it being perceived as unhealthy. For consumers who do not suffer from any form of gluten intolerance, a GFD may actually be detrimental to their health considering that wheat-based products provide a significant proportion of the average person's daily carbohydrates,

protein, vitamins and minerals (Newberry et al., 2017). Additionally, to compensate for the lack of functionality in gluten-free specialty products, these foods are often high in salt and other preservatives (Roman et al., 2019). This illustrates the importance of accurate and effective public communication around gluten containing foods. Public education will likely be essential in order to maximise the potential benefits of any low gluten epitope wheat cultivars or food products. Gluten is present in such a vast range of foods, from baked goods, to pastas, to cereals and many other processed products. This makes gluten avoidance highly challenging and for most consumers will never be a practical option unless absolutely necessary. Specialty low gluten epitope food products are likely to be available soon and it will be important that consumers understand more about these products. Those that are likely to benefit most from such products are the 35-40% of non-coeliac consumers who are carriers of at least one copy of either the DQ2 or DQ8 haplotypes that makes them susceptible to developing CD (Brown et al., 2019; Jihane et al., 2014). By specifically reducing their dietary exposure to gluten epitopes, it is believed that such consumers may be able to reduce their risk of developing CD (Molberg et al., 2005). This will be most beneficial for younger consumers as it is believed that high exposure to gluten at a young age is a significant risk factor for developing CD (Gujral et al., 2012; Lebwohl & Rubio-Tapia, 2021). Therefore, low gluten epitope food products may allow susceptible consumers to reduce their risk or developing CD without needing to resort to total gluten avoidance. Benefits are also expected for consumers suffering from NCGS. The triggers for NCGS are believed to be highly individual as it is a poorly understood condition that is often self-diagnosed (Molina-Infante et al., 2015). It is expected that low gluten epitope food products may only be effective in reducing symptoms in some consumers suffering from NCGS and not others. These complicated factors mean that effective public communication will be vital alongside the eventual release of low gluten epitope food products.

7.4 Optimal strategies for developing low gluten epitope products

The results of this study have demonstrated the potential for employing GS in order to create new wheat varieties with reduced concentrations of particular gluten epitopes. The goal of selective breeding for quantitative traits is to make incremental progress each generation towards target values for the traits of interest. This requires a balanced approach as breeders must simultaneously target multiple traits, some of which may be negatively correlated, such as grain yield and total protein content. This would also be the case for breeders selecting low gluten epitope phenotypes. Progress would likely be gradual and would need to ensure other important traits aren’t compromised. For example, the results presented in Chapter 4 suggest that moderate positive associations exist between the concentrations of some gluten epitopes and total flour protein content. Therefore, selecting too intensely on reduced epitope concentrations may compromise the protein content and thus baking quality.

While GS does show promise in allowing breeders to develop low gluten epitope cultivars, the gains are likely to be gradual. A number of other strategies have also been suggested to reduce epitope concentrations in wheat and wheat-based products, such as genetic engineering or mutagenesis of breeding lines or sourdough fermentation of baked products using specific *lactobacillus* bacteria (Becker et al., 2012; Gil-Humanes et al., 2014; Jouanin et al., 2018; Sanchez-Leon et al., 2018). It seems likely that the most effective approach to producing low gluten epitope products would be a combination of the tactics described above. Each approach has its own benefits and challenges as discussed in Chapter 2 but some of these may be mitigated by combining multiple approaches. It is proposed that the methods of GS described in this study would be the best first step in gradually adjusting the gene pool of breeding populations, so that epitope concentrations are considered in all breeding selections. This would provide a foundation of cultivars that remain functional both agronomically and in the bakery but with gradually reducing immunogenicity. Such cultivars could then be further improved through the application of additional strategies such as those mentioned above.

7.5 Further research

The work in this study focused on understanding the regulation of gluten epitope concentrations and associations with baking quality primarily at the population level. Whilst this focus was effective in determining general associations between traits, there remains potential to further investigate specific loci and better understand direct gene effects. Genomic selection will be useful in moving population averages towards a desired target, but there is also the potential to investigate unique characteristics that present themselves at the individual level. For example, in this study, the commercially available cultivar 'Griffin', was consistently in the bottom 10% of tested lines for P5 concentrations and is contracted by New Zealand mills as a 'Premium' grade wheat due to its strong, elastic dough. These results suggest that although some associations appear to exist between P5 concentrations and quality characteristics, as presented in Chapter 4, by using a targeted approach, breeders can expect to be able to develop low gluten epitope cultivars that retain high levels of baking quality.

The purpose of breeding for low gluten epitope traits in wheat is to reduce the overall immunogenicity of the gluten proteins. This is a factor that also needs to be considered by breeders as different epitopes have shown to have differing levels of immunogenicity (Camarca et al., 2009; Shan et al., 2002). Therefore, a small reduction in concentrations of highly immunogenic epitopes may be more beneficial than a larger reduction in concentrations of less immunogenic epitopes. This means that selections should consider individual epitope concentrations. The challenge, however, is that the differences in immunogenicity are difficult to quantify and more research is certainly required to better understand these differences.

7.6 Final summary and conclusions

The primary aim of this study was to develop genomic prediction models for selecting wheat varieties with reduced gluten epitope concentrations. To achieve this, a number of factors first required investigation. The heritability of gluten epitope concentrations has not previously been reported, meaning that the potential effectiveness of selective breeding for these traits was relatively unknown. In Chapter 3, mean heritability estimates for each of the tested epitopes are reported as ranging from 0.35-0.91. These results suggest that selective breeding of any kind would be highly effective for concentrations of some epitopes. For those epitopes with moderate heritability's, GS is likely to be the most effective form of selective breeding as phenotypic selections of lower heritability traits can often be strongly influenced by environmental factors (Larkin et al., 2019). To further support these conclusions, estimates for genetic correlations of epitope concentrations between seasons were reported. These estimates ranged from 0.62-0.97, providing evidence that in most cases, genetic factors are the primary determinant of epitope concentrations as opposed to environmental variables. This also suggests that breeding selections for epitope concentrations will generally be reliable even when phenotypes were only collected in a single environment. This provides benefits for breeding efficiency as it reduces the need for multi-site and multi-season trialling.

Gluten is the primary factor that determines baking quality traits in wheat (Biesiekierski, 2017), therefore, it was expected that gluten epitope concentrations may also present some associations with baking quality. This was investigated in Chapter 4 to determine the need for breeders to consider how selecting low gluten epitope lines may influence the baking quality of their breeding populations. In that chapter, the genetic and phenotypic correlations between epitope concentrations and a range of baking quality characteristics were reported. The strongest associations were generally observed between epitope concentrations and Flour Protein Content (FPC), however there was a wide range in the correlations between FPC and the different epitopes. For example, genetic correlations between P3 concentrations and FPC were near zero while correlations between P4 concentrations and FPC were as high as 0.8. Moderate associations were shown to exist between concentrations of most gluten epitopes and each of the different baking quality metrics, however these were always reduced when the genetic effects of FPC were excluded from analysis. Encouragingly, generally weak associations were observed between concentrations of the most immunodominant epitope (P5), (Schalk et al., 2017) and most baking quality metrics. This indicates that selections of low P5 lines are unlikely to have major effects on the baking quality of the breeding population.

With the findings in both Chapters 3 and 4 indicating that selective breeding will likely be an effective and practical method for developing low gluten epitope wheat cultivars, Chapters 5 and 6 investigated the genomic factors regulating epitope concentrations along with the quantifying the prediction accuracy of a GS model. Three genomic windows were identified as significant loci for regulating the concentrations of

the P1, P3 and P5 epitopes in Chapter 5 through a Genome Wide Association Study (GWAS). Each of these regions appeared to be distinct from the primary gliadin coding loci, suggesting that their effects may primarily be regulatory rather than structural. The individual marker effects for markers in these windows were generally low suggesting that further investigation is required to identify specific genes or alleles that may be useful for marker assisted selection of low gluten epitope wheats. Additionally, the proportion of genomic variance explained by each window for concentrations of the relevant epitopes was also generally low. This suggests that epitope concentrations are likely to be determined by many small effect genes rather than a small number of large effect genes.

A genomic prediction model was tested using three different scenarios in Chapter 6. By testing 100 iterations of an internal validation population, empirical prediction accuracies of between 0.37-0.67 for concentrations of the different tested epitopes were observed. These values correspond with realised accuracies of between 0.43-0.82. With additional phenotyping and genotyping of lines that are representative of the PFR breeding program, it is expected that these prediction accuracies could be achieved for breeding selections. However, using the data collected in this study for GS in the PFR breeding program, the expected prediction accuracies would be lower, ranging from 0.16-0.53 as seen in scenarios 1 and 2 (Chapter 6). These empirical prediction accuracies correspond with realised accuracies of between 0.28-0.61. High empirical prediction accuracies were also observed for FPC (0.75) in scenario 3, meaning that by using targeted selections, breeders may be able to apply GS for simultaneously selecting for reduced epitope concentrations and high FPC. This is likely to ensure high levels of baking quality are maintained in the breeding population. The results of this study therefore indicate that GS for gluten epitope concentrations is an effective breeding method. Reduced epitope concentrations can therefore be considered a practical breeding target for wheat breeding groups to incorporate into their own programs. The potential long-term benefits to the health of the global population are significant and the nutritional composition of wheat should not be ignored by breeders. This renewed focus on wheat nutrition has the opportunity to diversify the industry with a much needed ‘value add’ component and allow both growers and food producers to diversify their produce.

7.7 References

Alahmad, S., Rambla, C., Voss-Fels, K. P., & Hickey, L. T. (2022). Accelerating Breeding Cycles. In M. P. Reynolds & H.-J. Braun (Eds.), *Wheat Improvement: Food Security in a Changing Climate* (pp. 557-571). Springer International Publishing. https://doi.org/10.1007/978-3-030-90673-3_30

Allen, A. M., Winfield, M. O., Burridge, A. J., Downie, R. C., Benbow, H. R., Barker, G. L. A., . . . Edwards, K. J. (2017). Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (Triticum aestivum). *Plant Biotechnology Journal*, *15*(3), 390-401. https://doi.org/https://doi.org/10.1111/pbi.12635

Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., . . . Robbins, K. R. (2021). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theoretical and Applied Genetics*, *134*(1), 279-294. https://doi.org/10.1007/s00122-020-03696-9

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., & Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). *Plant Science*, *242*, 23-36. https://doi.org/https://doi.org/10.1016/j.plantsci.2015.08.021

Becker, D., Wieser, H., Koehler, P., Folck, A., Muhling, K. H., & Zorb, C. (2012). Protein composition and techno-functional properties of transgenic wheat with reduced a-gliadin content obtained by RNA interference. *Journal of Applied Botany and Food Quality*, *85*, 23-33.

Biesiekierski, J. R. (2017). What is gluten? *Journal of Gastroenterology and Hepatology*, *32*, 78-81. https://doi.org/10.1111/jgh.13703

Brown, N. K., Guandalini, S., Semrad, C., & Kupfer, S. S. (2019). A Clinician's Guide to Celiac Disease HLA Genetics. *Official journal of the American College of Gastroenterology | ACG*, *114*(10). https://journals.lww.com/ajg/Fulltext/2019/10000/A_Clinician_s_Guide_to_Celiac_Disease_HLA_Genetics.9.aspx

Budhlakoti, N., Kushwaha, A. K., Rai, A., Chaturvedi, K. K., Kumar, A., Pradhan, A. K., . . . Kumar, S. (2022). Genomic Selection: A Tool for Accelerating the Efficiency of Molecular Breeding for Development of Climate-Resilient Crops [Review]. *Frontiers in Genetics*, *13*. https://www.frontiersin.org/articles/10.3389/fgene.2022.832153

Camarca, A., Anderson, R., Mamone, G., Fierro, O., Facchiano, A., Costantini, S., . . . Troncone, R. (2009). Intestinal T Cell Responses to Gluten Peptides Are Largely Heterogeneous: Implications for a Peptide-Based Therapy in Celiac Disease. *Journal of immunology*, *182*, 4158-4166. https://doi.org/10.4049/jimmunol.0803181

Chlubnová, M., Christophersen, A. O., Sandve, G. K. F., Lundin, K. E. A., Jahnsen, J., Dahal-Koirala, S., & Sollid, L. M. Identification of gluten T cell epitopes driving celiac disease. *Science Advances*, *9*(4), eade5800. https://doi.org/10.1126/sciadv.ade5800

Ciccocioppo, R., Di Sabatino, A., & Corazza, G. R. (2005). The immune recognition of gluten in coeliac disease. *Clinical & Experimental Immunology*, *140*(3), 408-416.

De Re, V., Caggiari, L., Tabuso, M., & Cannizzaro, R. (2013). The versatile role of gliadin peptides in celiac disease. *Clinical Biochemistry*, *46*, 552-560.

Gil-Humanes, J., Pistón, F., Barro, F., & Rosell, C. M. (2014). The Shutdown of Celiac Disease-Related Gliadin Epitopes in Bread Wheat by RNAi Provides Flours with Increased Stability and Better Tolerance to Over-Mixing. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0091931

Gujral, N., Freeman H, J., & Thomson, A. B. (2012). Celiac disease: prevalence, diagnosis, pathogenesis and treatment. *World Journal of Gastroenterology*, *18*(42), 6036-6059. https://doi.org/10.3748/wjg.v18.i42.6036

Humphreys, D. G., & Knox, R. E. (2015). Doubled Haploid Breeding in Cereals. In J. M. Al-Khayri, S. M. Jain, & D. V. Johnson (Eds.), *Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools* (pp. 241-290). Springer International Publishing. https://doi.org/10.1007/978-3-319-22521-0_9

Jihane, R., Anna, R., Vinod, K., Gosia, T., Lude, F., Agata, S., . . . Prevent, C. D. G. (2014). Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants. *Gut*, *63*(3), 415. https://doi.org/10.1136/gutjnl-2012-304110

Jouanin, A., Gilissen, L. J. W. J., Boyd, L. A., Cockram, J., Leigh, F. J., Wallington, E. J., . . . Smulders, M. J. M. (2018). Food processing and breeding strategies for coeliac-safe and healthy wheat products. *Food Research International*, *110*, 11-21. https://doi.org/https://doi.org/10.1016/j.foodres.2017.04.025

Koning, F. (2021). Celiac disease: quantity matters. *Seminars in Immunopathology*, *34*(4), 541-549.

Kumar, N., Rana, M., Kumar, B., Chand, S., Shiv, A., Wani, S. H., & Kumar, S. (2021). Genomic Selection for Wheat Improvement. In S. H. Wani, A. Mohan, & G. P. Singh (Eds.), *Physiological, Molecular, and Genetic Perspectives of Wheat Improvement* (pp. 175-207). Springer International Publishing. https://doi.org/10.1007/978-3-030-59577-7_9

Larkin, D. L., Lozada, D. N., & Mason, R. E. (2019). Genomic Selection—Considerations for Successful Implementation in Wheat Breeding Programs. *Agronomy*, *9*(9).

Lebwohl, B., & Rubio-Tapia, A. (2021). Epidemiology, Presentation, and Diagnosis of Celiac Disease. *Gastroenterology*, *160*(1), 63-75. https://doi.org/https://doi.org/10.1053/j.gastro.2020.06.098

Lozada, D. N., Mason, R. E., Sarinelli, J. M., & Brown-Guedira, G. (2019). Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. *BMC Genetics*, *20*(1), 82. https://doi.org/10.1186/s12863-019-0785-1

Molberg, Ø., Uhlen, A. K., Jensen, T., Flæte, N. S., Fleckenstein, B., Arentz–Hansen, H., . . . Sollid, L. M. (2005). Mapping of gluten T-cell epitopes in the bread wheat ancestors: Implications for celiac disease. *Gastroenterology*, *128*(2), 393-401. https://doi.org/https://doi.org/10.1053/j.gastro.2004.11.003

Molina-Infante, J., Santolaria S Fau - Sanders, D. S., Sanders Ds Fau - Fernández-Bañares, F., & Fernández-Bañares, F. (2015). Systematic review: noncoeliac gluten sensitivity. (1365-2036).

Mowat, A. M. (2003). Coeliac disease – a meeting point for genetics, immunology, and protein chemistry. *Lancet*, *361*(9365), 1290-1292.

Newberry, C., McKnight, L., Sarav, M., & Pickett-Blakely, O. (2017). Going Gluten Free: the History and Nutritional Implications of Today's Most Popular Diet. *Current Gastroenterology Reports*, *19*(11), 54. https://doi.org/10.1007/s11894-017-0597-2

Rasheed, A., & Xia, X. (2019). From markers to genome-based breeding in wheat. *Theoretical and Applied Genetics*, *132*(3), 767-784. https://doi.org/10.1007/s00122-019-03286-4

Roman, L., Belorio, M., & Gomez, M. (2019). Gluten-Free Breads: The Gap Between Research and Commercial Reality. *Comprehensive Reviews in Food Science and Food Safety*, *18*(3), 690-702. https://doi.org/https://doi.org/10.1111/1541-4337.12437

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., . . . Singh, R. (2016). Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. *G3 Genes|Genomes|Genetics*, *6*(9), 2799-2808. https://doi.org/10.1534/g3.116.032888

Rutkoski, J. E., Krause, M. R., & Sorrells, M. E. (2022). Breeding Methods: Line Development. In M. P. Reynolds & H.-J. Braun (Eds.), *Wheat Improvement: Food Security in a Changing Climate* (pp. 69-82). Springer International Publishing. https://doi.org/10.1007/978-3-030-90673-3_5

Sanchez-Leon, S., Gil-Humanes, J., Ozuna, C. V., Gimenez, M. J., Sousa, C., Voytas, D. F., & Barro, F. (2018). Low-gluten, nontransgenic wheat engineered with CRISPR/Cas9. *Plant Biotechnol J*, *16*(4), 902-910. https://doi.org/10.1111/pbi.12837

Schalk, K., Lang, C., Wieser, H., Koehler, P., & Scherf, K. A. (2017). Quantitation of the immunodominant 33-mer peptide from alpha-gliadin in wheat flours by liquid chromatography tandem mass spectrometry. *Sci Rep*, *7*, 45092. https://doi.org/10.1038/srep45092

Shan, L., Molberg, Ø., Parrot, I., Hausch, F., Filiz, F., Gray, G. M., . . . Khosla, C. (2002). Structural Basis for Gluten Intolerance in Celiac Sprue. *Science*, *297*(5590), 2275-2279. https://doi.org/10.1126/science.1074129

Shewry, P. R., & Tatham, A. S. (2016). Improving wheat to remove coeliac epitopes but retain functionality. *Journal of Cereal Science*, *67*, 12-21. https://doi.org/https://doi.org/10.1016/j.jcs.2015.06.005

Stock, K. F., & Reents, R. (2013). Genomic Selection: Status in Different Species and Challenges for Breeding. *Reproduction in Domestic Animals*, *48*(s1), 2-10. https://doi.org/https://doi.org/10.1111/rda.12201

Tessema, B. B., Liu, H., Sørensen, A. C., Andersen, J. R., & Jensen, J. (2020). Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat [Original Research]. *Frontiers in Genetics*, *11*. https://www.frontiersin.org/articles/10.3389/fgene.2020.578123

Uauy, C. (2017). Wheat genomics comes of age. *Current Opinion in Plant Biology*, *36*, 142-148. https://doi.org/https://doi.org/10.1016/j.pbi.2017.01.007

Chapter 8

Appendix

Table 8.1: Full list of named wheat varieties included in the training population

Alberic	Domino	Kellalac	Preston	Suneca
Amarok	Doubletake	Kennedy	Prophet	Sunguard
Aquilla	Drysdale	King Rock	Raffles	Sunlin
Aspiring	Duchess	Kingdom	Rata	Sunvale
Bakker Gold	EGA Gregory	Kohika	Reflection	Superb
Bastion	Einstein	Kotuku	Regency	SY Epson
Batavia	Empress	Lang	Reliance	Tanker
Batten Spring	Endeavour	Lillian	Resolution	Thatcher
Batten Winter	Equinox	Lincoln	Richmond	Tiritea
Beaufort	Espada	Livingston	Robigus	Tobak
Bill	Excede	Mace	Rongotea	Torch
Bolac	Forrest	Mackellar	Rosella	Torlesse
Bonnie Rock	Frame	Macro	Rubric	Tribute
Bowie	Gascoigne	Magenta	Ruapuna	Tyrian
Buster	Gator	Majestic	Sage	Vanquish
Calingiri	Genghis	Marshall	Samoa	Viceroy
Camm	Gladius	McCubbin	Sapphire	Wakanui
Carnamah	Glenn	Merinda	Saracen	Wakelin
Cascade	Graham	Monad	Savannah	Wedgetail
Centaur	Gregory	Morph	Scout	Weston
Chara	Gruner	Oakley	Sdelweisser	Westonia
Claire	Hanson	Option	Snowbird	Wyalkatchem
Commando	Hilton	Orator	Solstice	Wylah
Conqueror	Ignite	Oroua	Spitfire	Xi 19
Conquest	Inferno	Otane	Starfire	Yitpi
Consort	Janz	Pasteur	Stigg	
Copper	Katana	Pentium	Strzelecki	
Discovery	Katch 42	Phoenix	Sunco	

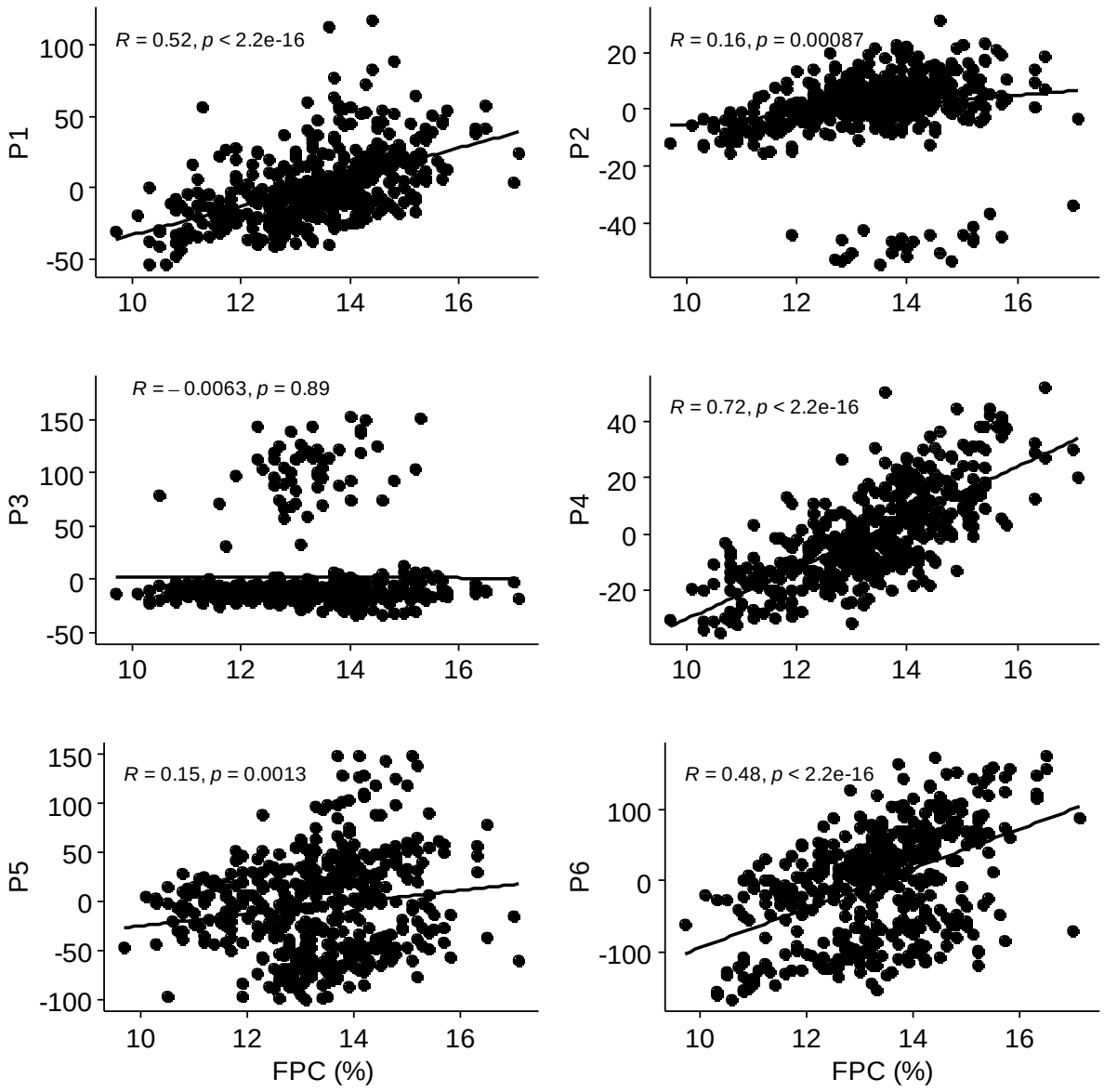

Figure 8.1: Flour Protein Content (FPC) against batch adjusted gluten epitope concentrations (P1-P6)

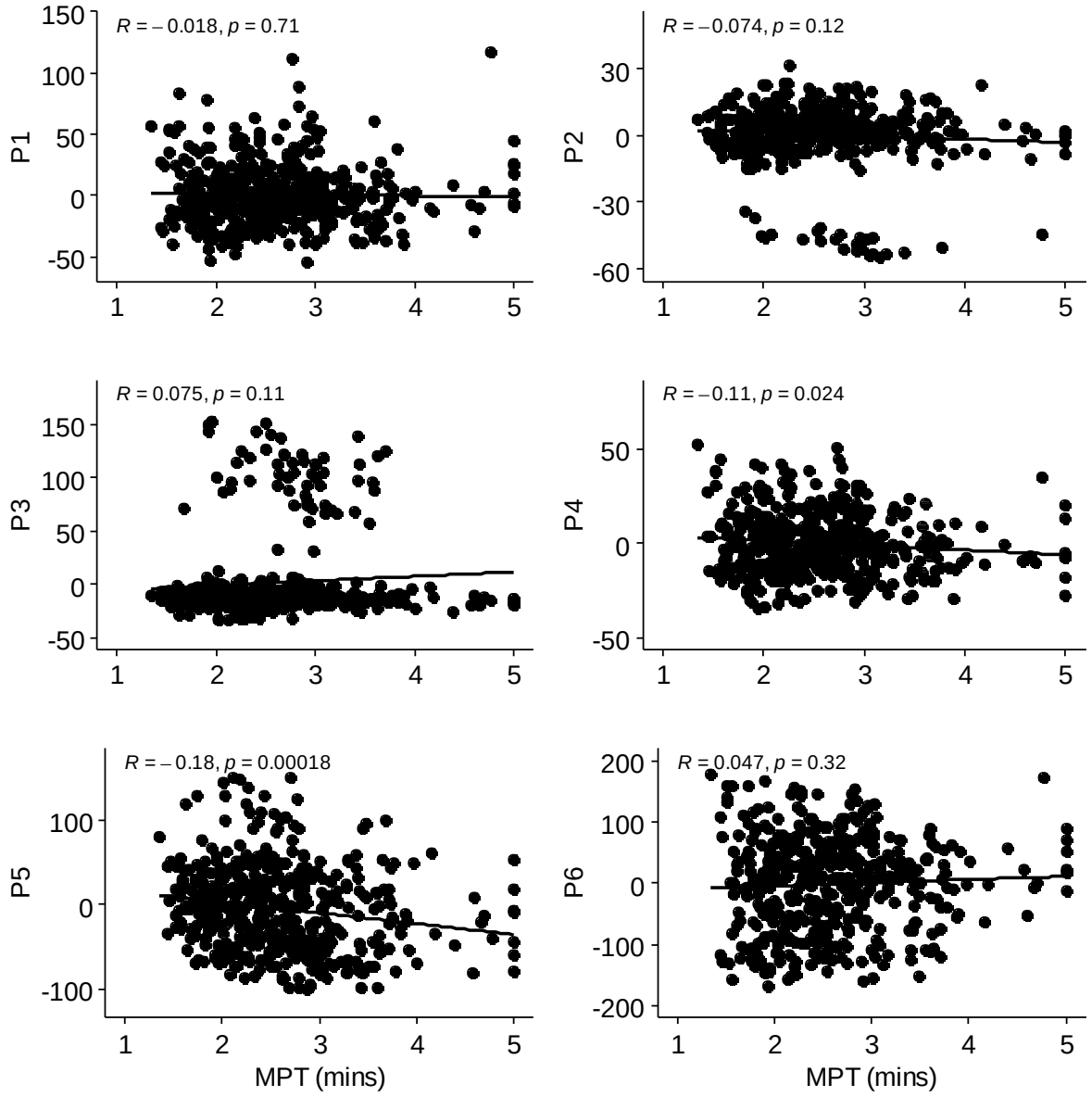

Figure 8.2: Midline Peak Time (MPT) against batch adjusted gluten epitope concentrations (P1-P6)

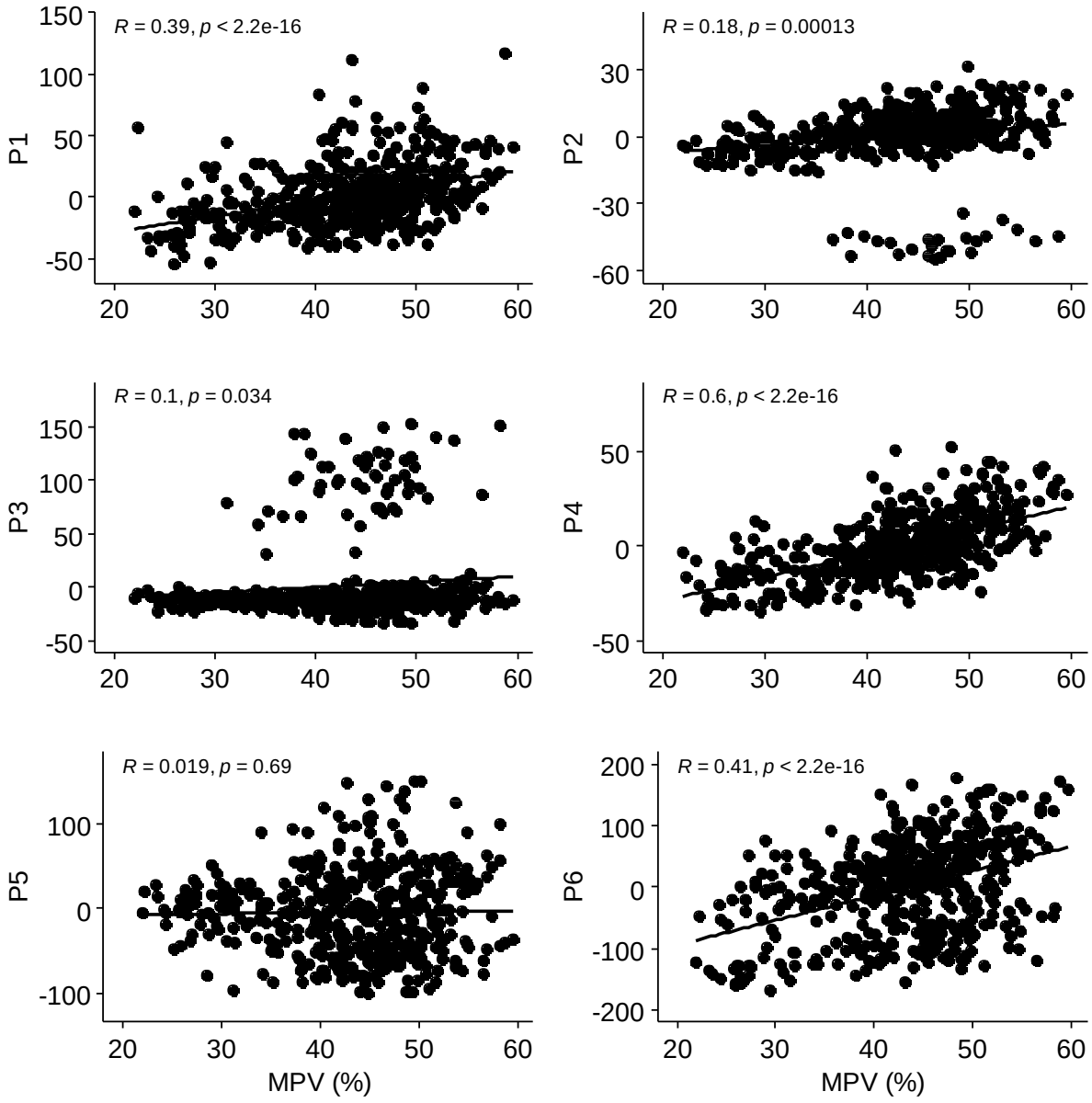

Figure 8.3: Midline Peak Value (MPV) against batch adjusted gluten epitope concentrations (P1-P6)

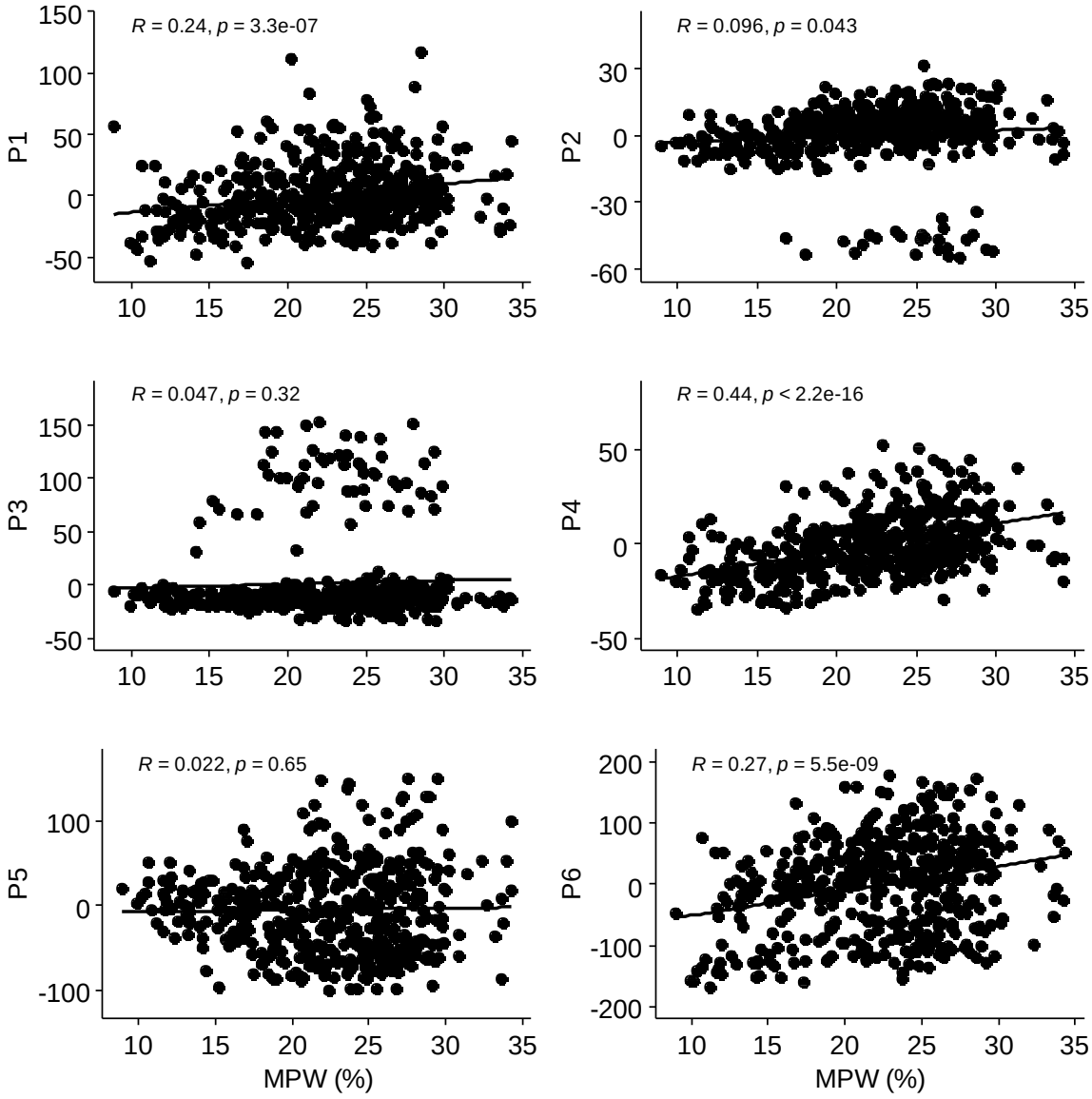

Figure 8.4: Midline Peak Width (MPW) against batch adjusted gluten epitope concentrations (P1-P6)

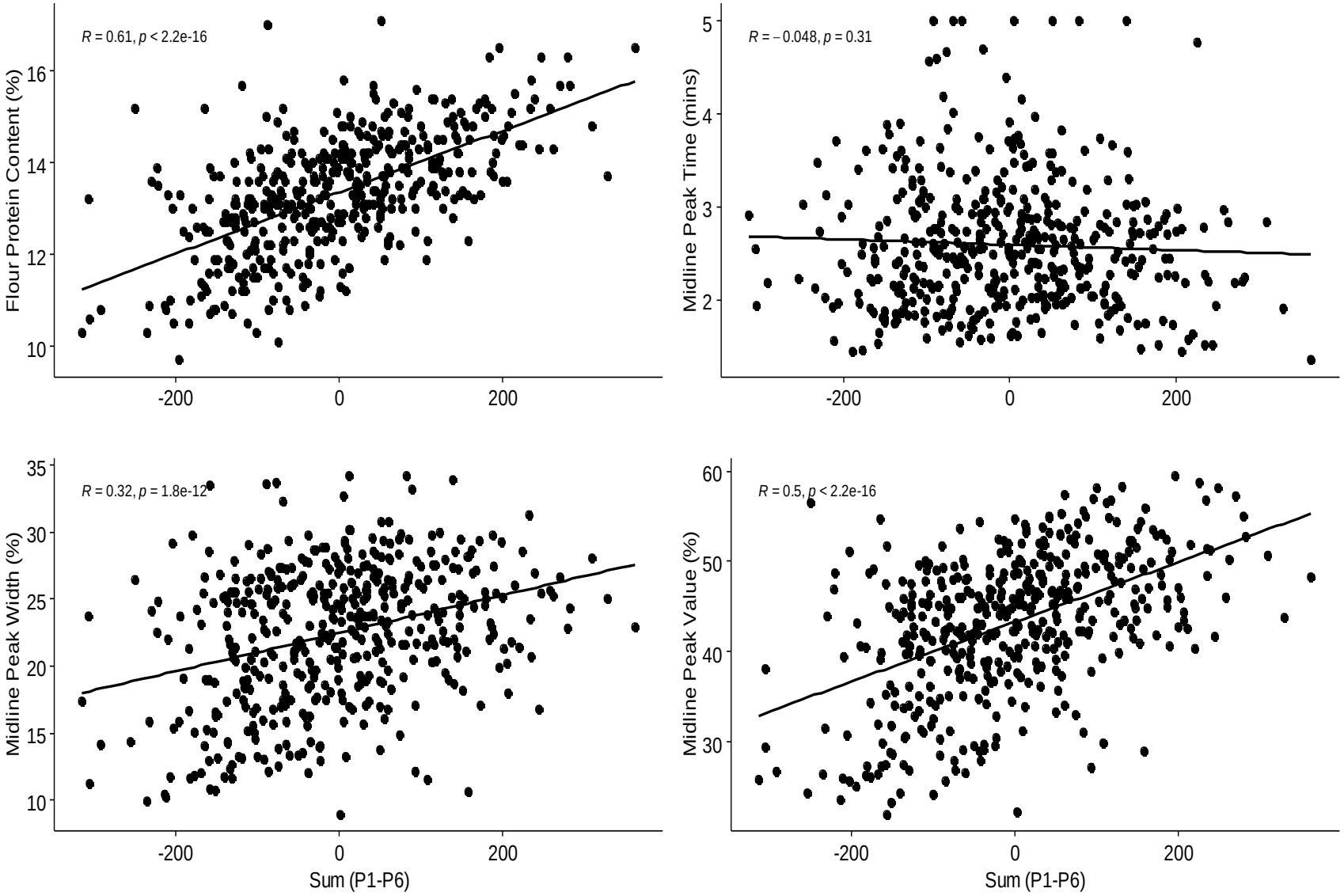

Figure 8.5: Scatter plots between total epitope concentration (P1-P6) and each of FPC, MPT, MPV and MPW

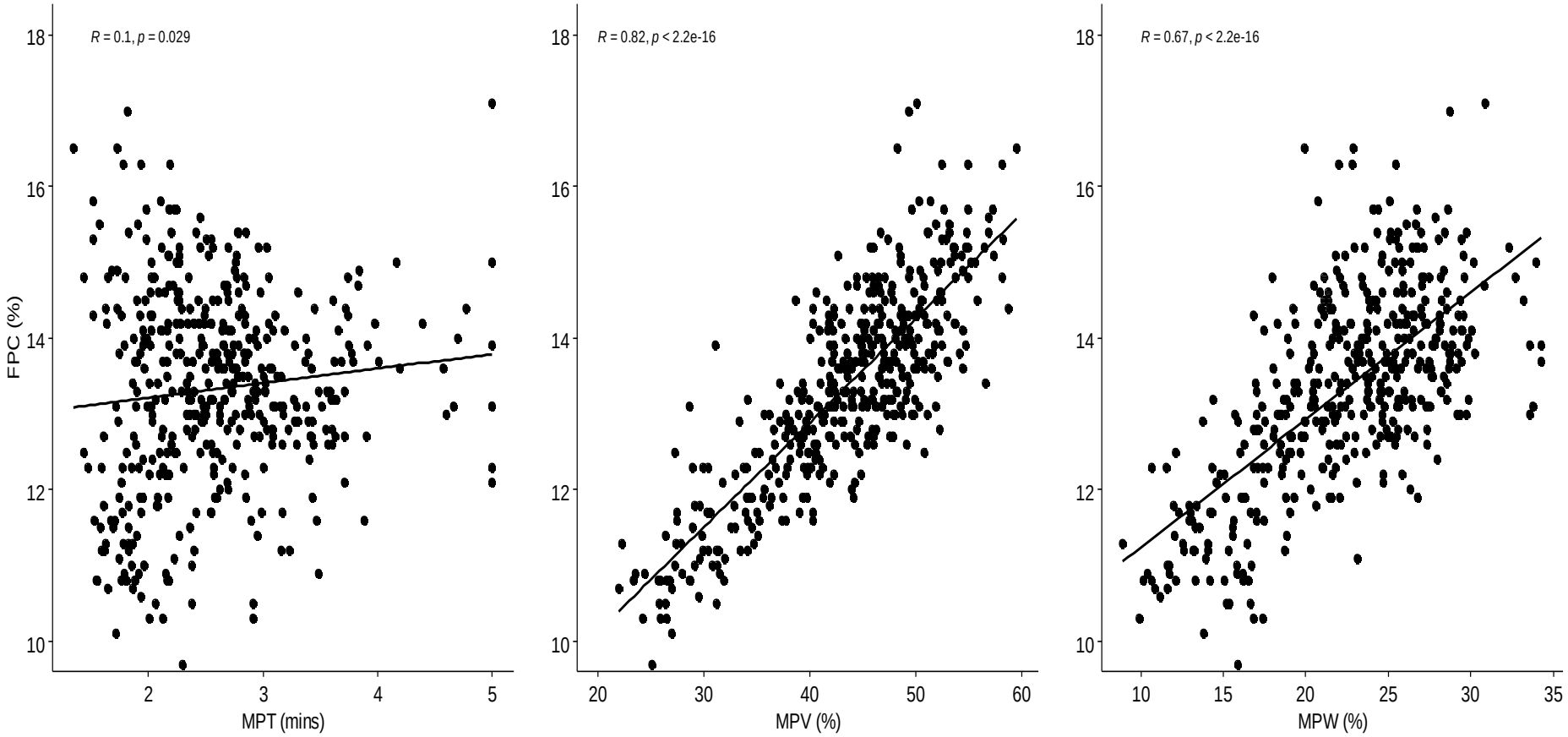

Figure 8.6: Flour Protein Content (FPC%) against mixograph quality parameters (MPT, MPV and MPW)

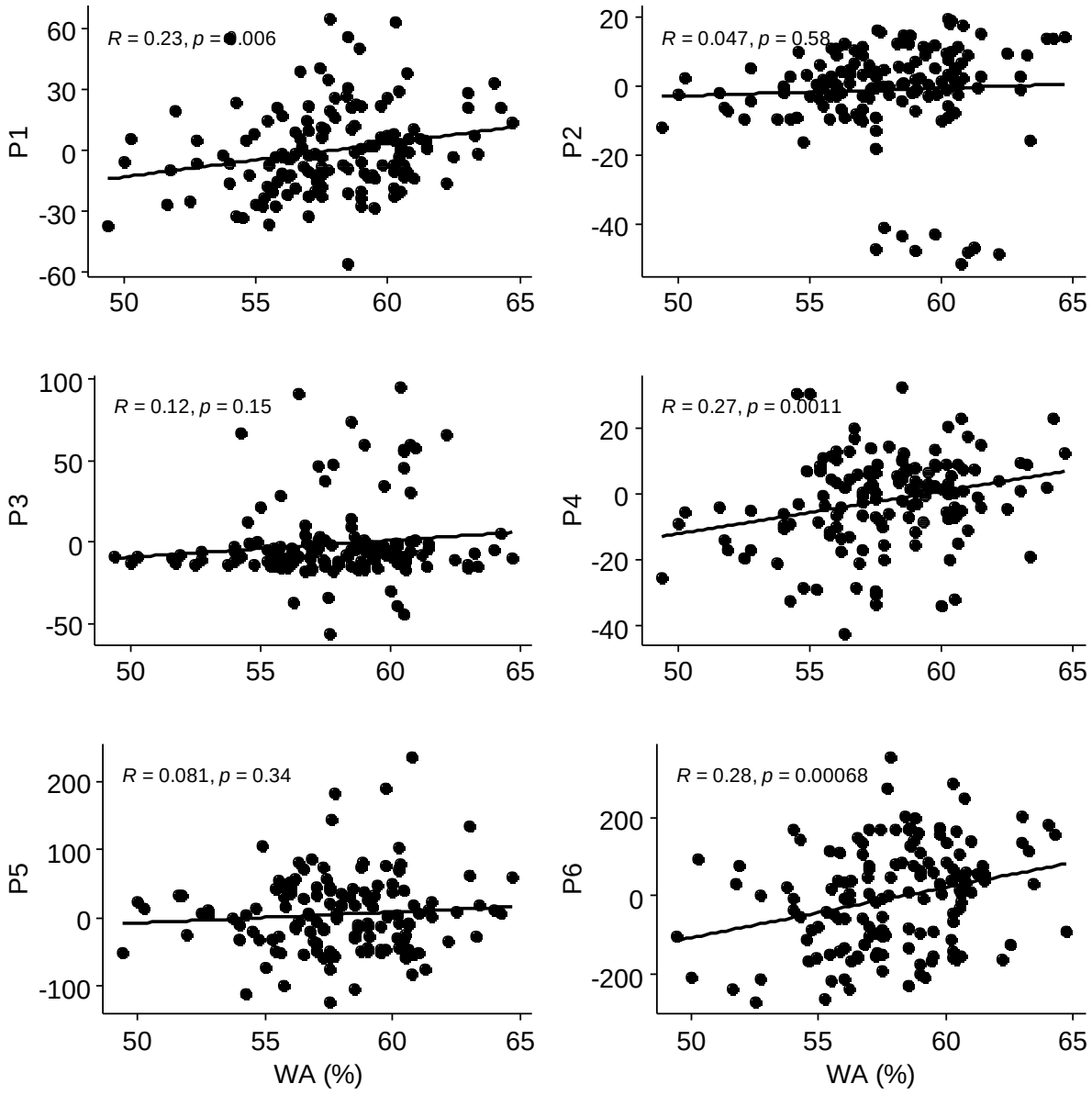

Figure 8.7: Water Absorption (WA%) against batch adjusted gluten epitope concentrations (P1-P6)

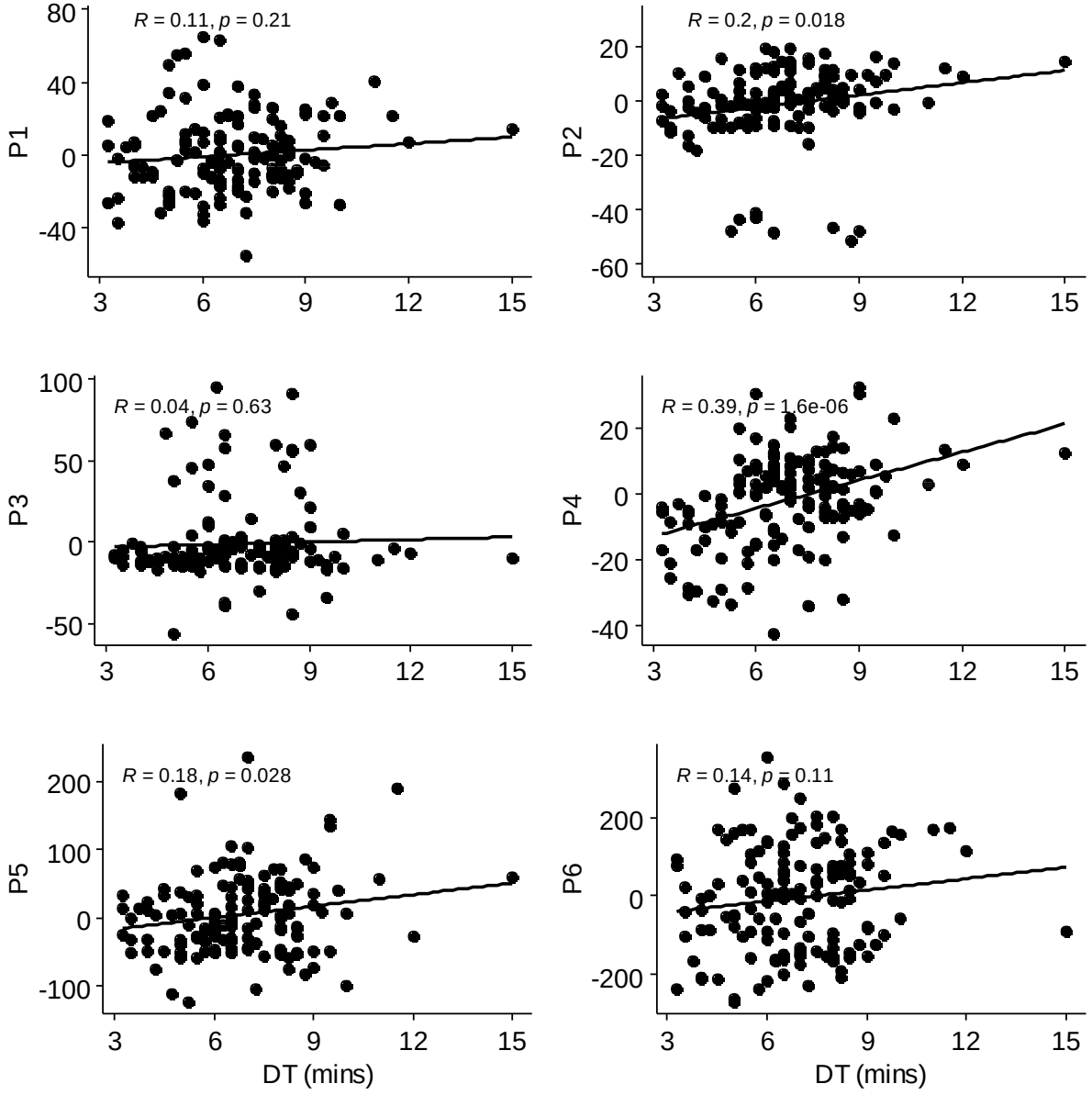

Figure 8.8: Development Time (mins) against batch adjusted gluten epitope concentrations (P1-P6)

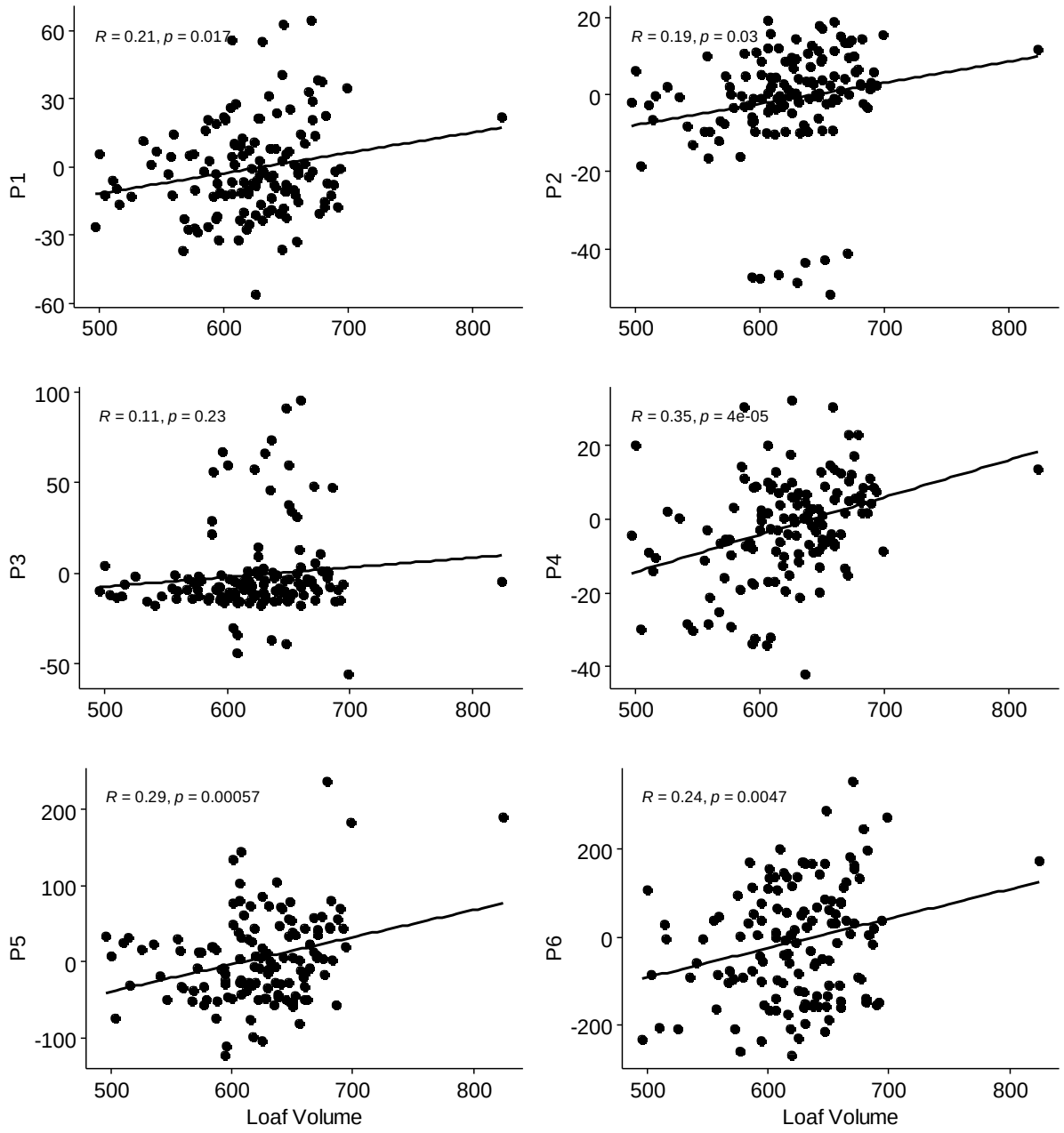

Figure 8.9: Loaf Volume against batch adjusted epitope concentrations (P1-P6)

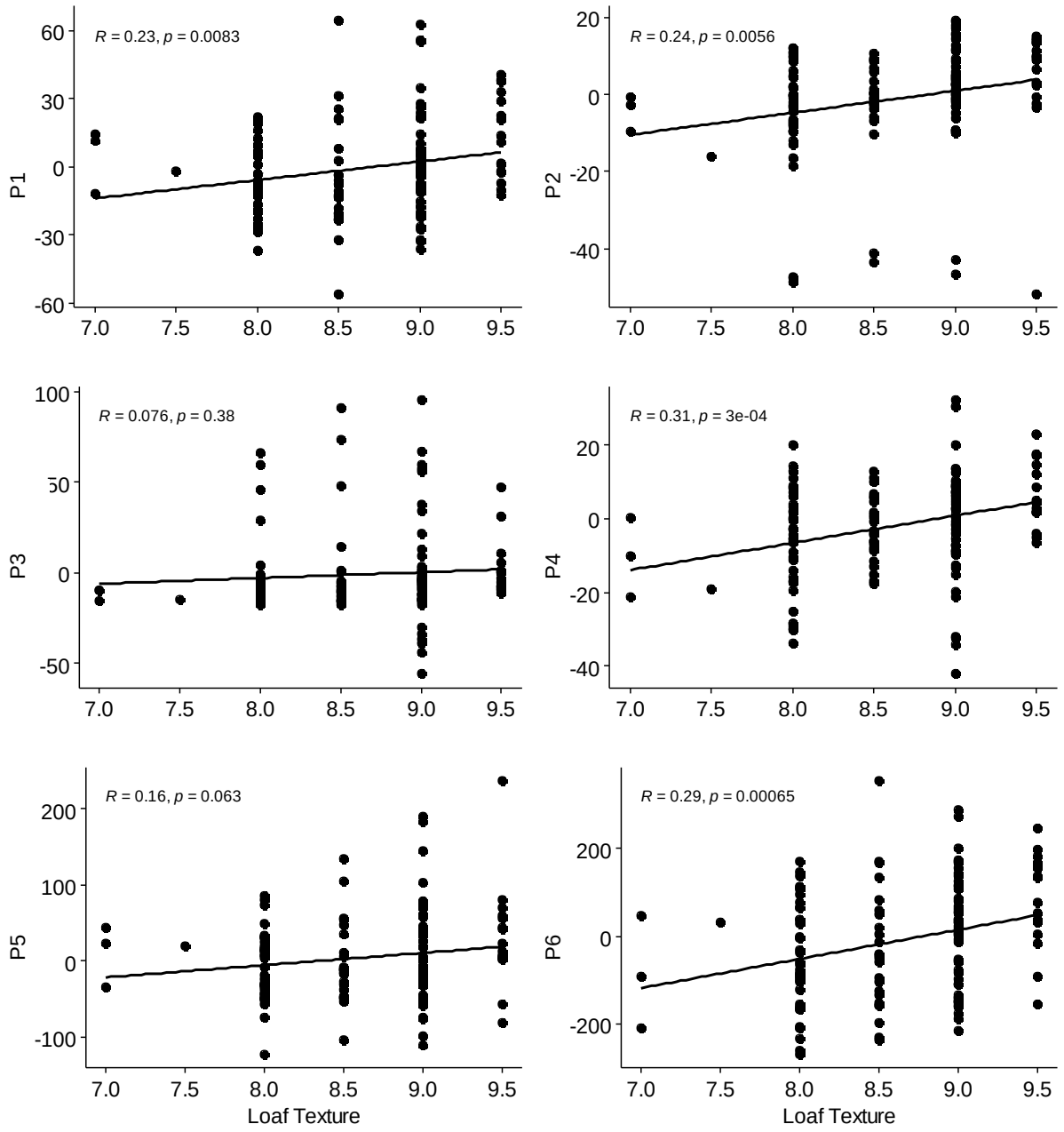

Figure 8.10: Loaf Texture against batch adjusted epitope concentrations (P1-P6)

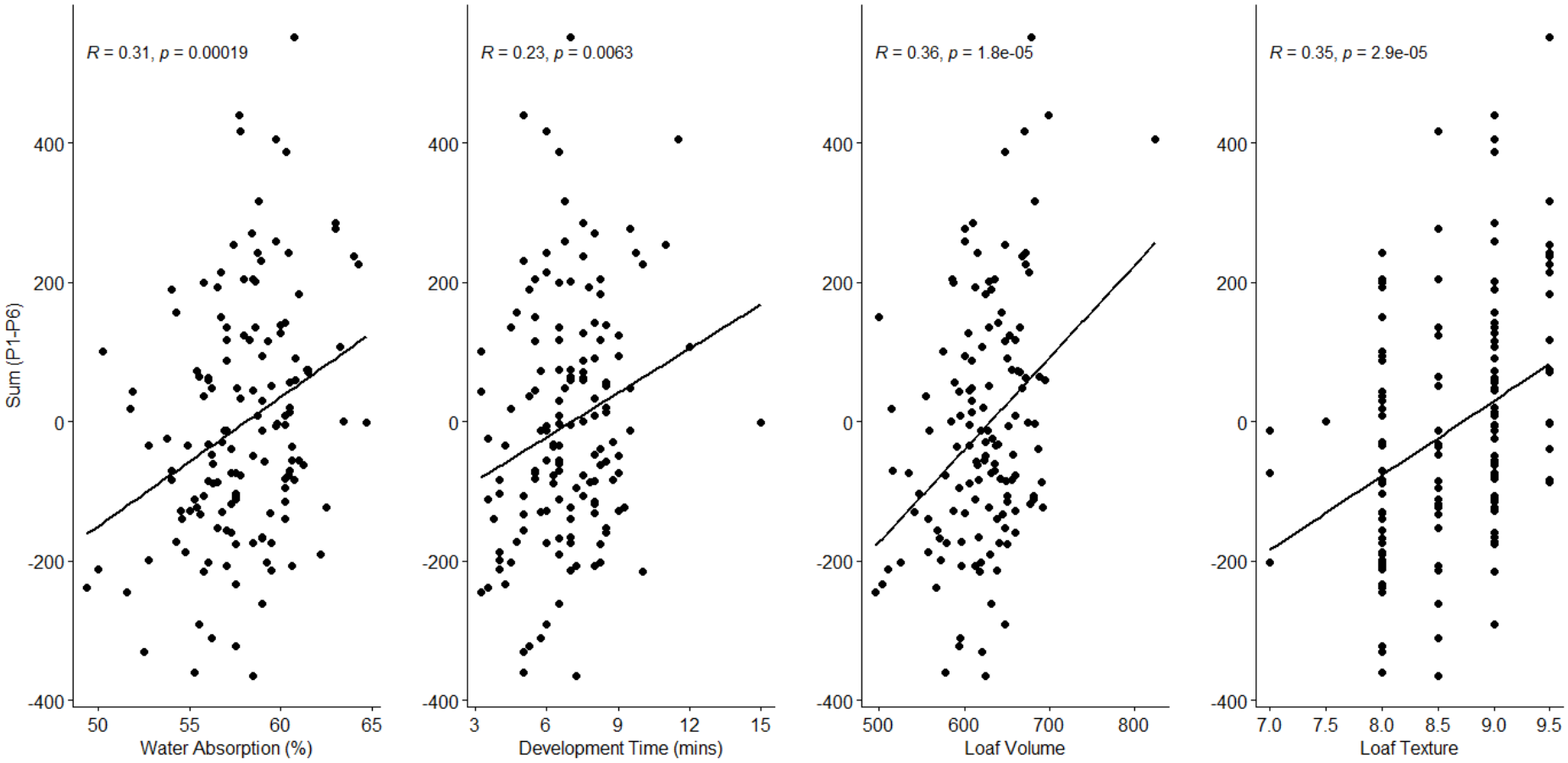

Figure 8.11: Baking quality results (Water Absorption, Development Time, Loaf Weight, Loaf Volume and Loaf Texture) against total concentrations of P1-P6 (Sum)

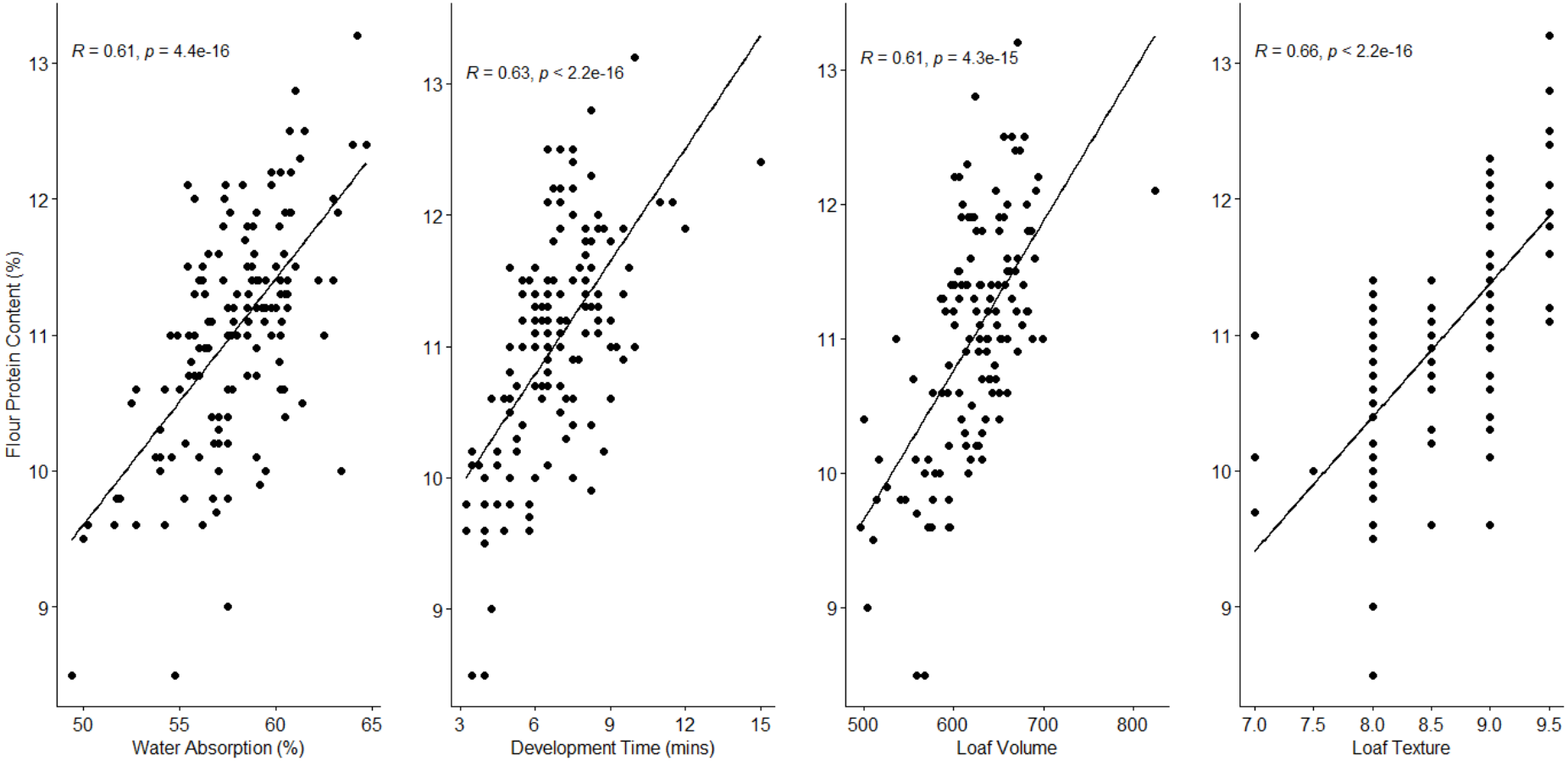

Figure 8.12: Baking quality results (Water Absorption, Development Time, Loaf Weight, Loaf Volume and Loaf Texture) against Flour Protein Content

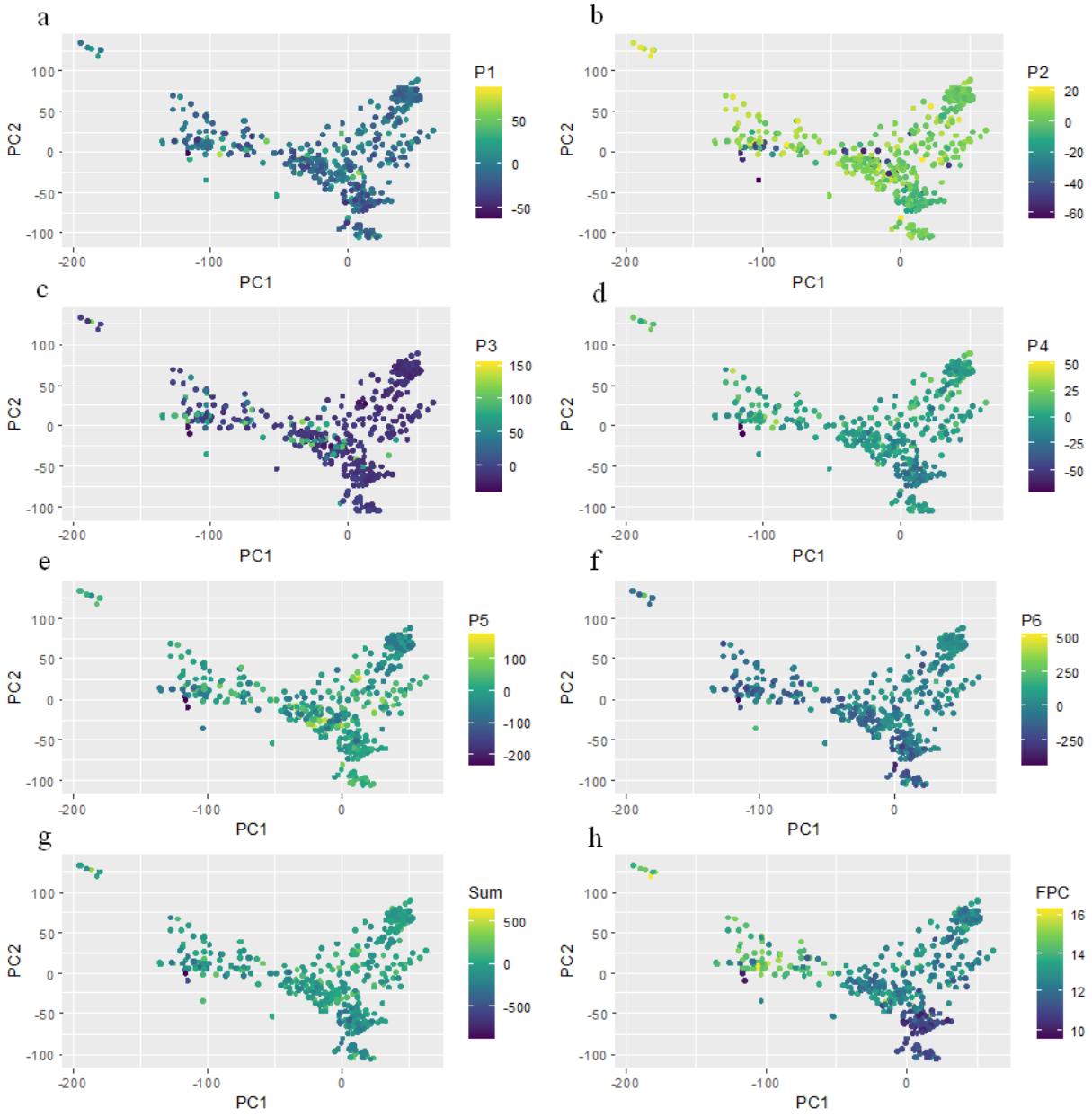

Figure 8.13: Principal Component Analysis (PCA) for the primary training population and external validation population with colour gradients representing batch adjusted epitope concentrations (P1-P6) and percent Flour Protein Content (FPC). (a-f = P1-P6, g = total P1-P6 concentrations (sum), h = FPC)

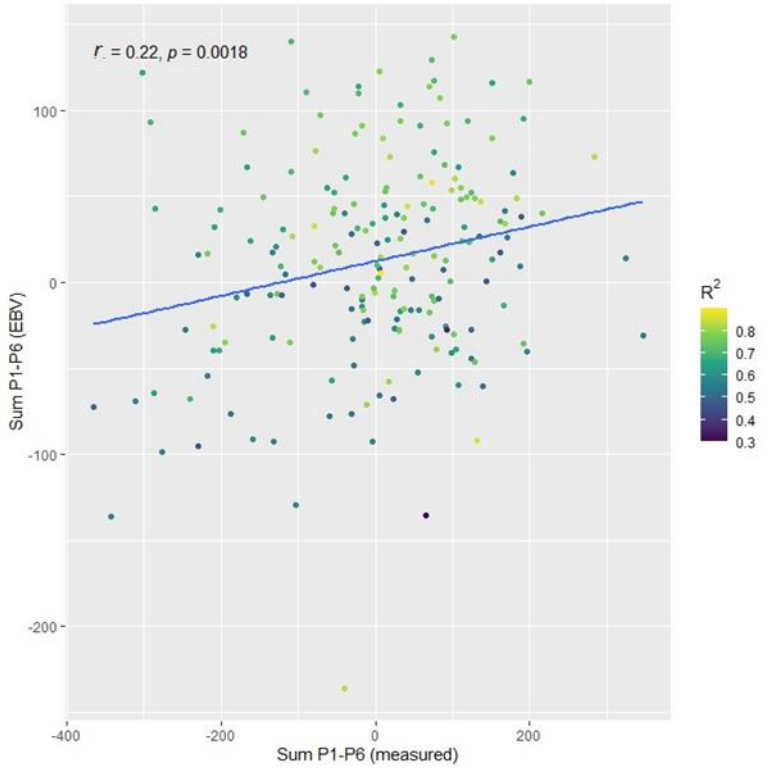

Figure 8.14: Scatter plot for empirical predictions (r) of total P1-P6 concentrations (Sum P1-P6) in the external validation population (scenario 2). (R^2 = Reliability of predictions; EBV = Estimated Breeding Value)

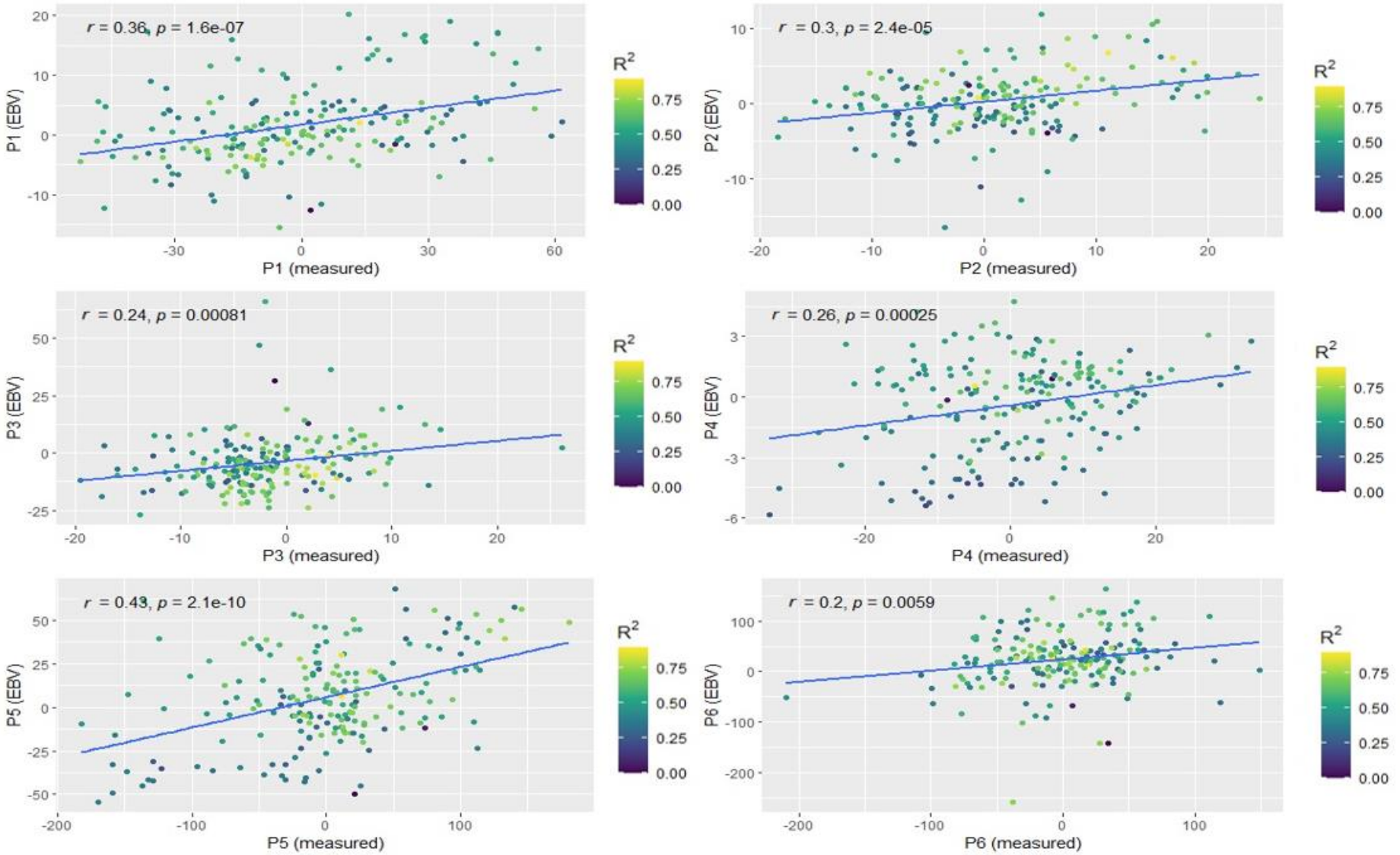

Figure 8.15: Scatter plot for predictions of gluten epitope concentrations (P1-P6) in the external validation population (scenario 2). (R^2 = Reliability of predictions; EBV = Estimated Breeding Value