Detection of loci associated with water-soluble carbohydrate accumulation and environmental adaptation in white clover (*Trifolium repens* L.)

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

Doctor of Philosophy

in

Plant Biology

at Massey University, Palmerston North, New Zealand

Sofie Margaret Pearson

2021

Abstract

White clover (*Trifolium repens* L.) is an economically important forage legume in New Zealand/Aotearoa (NZ). It provides quality forage and a source of bioavailable nitrogen fixed through symbiosis with soil *Rhizobium* bacteria. This thesis investigated the genetic basis of two traits of significant agronomic interest in white clover. These were foliar water-soluble carbohydrate (WSC) accumulation and soil moisture deficit (SMD) tolerance. Previously generated divergent WSC lines of white clover were characterised for foliar WSC and leaf size. Significant ($p < 0.05$) divergence in foliar WSC content was observed between five breeding pools. Little correlation was observed between WSC and leaf size, indicating that breeding for increased WSC content could be achieved in large and small leaf size classes of white clover in as few as 2 – 3 generations. Genotyping by sequencing (GBS) data were obtained for 1,113 white clover individuals (approximately 47 individuals from each of 24 populations). Population structure was assessed using discriminant analysis of principal components (DAPC) and individuals were assigned to 11 genetic clusters. Divergent selection created a structure that differentiated high and low WSC populations. Outlier detection methodologies using PCAdapt, BayeScan and KGD-F_{ST} applied to the GBS data identified 33 SNPs in diverse gene families that discriminated high and low WSC populations. One SNP associated with the starch biosynthesis gene, *glgC* was identified in a genome-wide association study (GWAS) of 605 white clover individuals. Transcriptome and proteome analyses also provided evidence to suggest that high WSC levels in different breeding pools were achieved through sorting of allelic variants of carbohydrate metabolism pathway genes. Transcriptome and proteome analyses suggested 14 gene models from seven carbohydrate gene families (*glgC*, *WAXY*, *glgA*, *glgB*, *BAM*, *AMY* and *ISA3*) had responded to artificial selection. Patterns of SNP variation in the *AMY*, *glgC* and *WAXY* gene families separated low and high WSC individuals. Allelic variants in these gene families represent potential targets for assisted breeding of high WSC levels. Overall, multiple lines of evidence corroborate the importance of *glgC* for increasing foliar WSC accumulation in white clover. Soil moisture deficit (SMD) tolerance was investigated in naturalised populations of white clover collected from 17 sites representing contrasting SMD across the South Island/Te Waipounamu of NZ. Weak genetic differentiation of populations was detected in analyses of GBS data, with three genetic clusters identified by ADMIXTURE. Outlier detection and environmental association analyses identified 64 SNPs significantly ($p < 0.05$) associated with environmental variation. Mapping of these SNPs to the white clover reference genome, together with gene ontology analyses, suggested some SNPs were associated with genes involved in carbohydrate

metabolism and root morphology. A common set of allelic variants in a subset of the populations from high SMD environments may also identify targets for selective breeding, but this variation needs further investigation.

Key words: ADMIXTURE, BayeScan, BayeScEnv, discriminant analysis of principal components, environmental association analysis, genome-wide association study, genomics, genotyping by sequencing, KGD-F_{ST}, leaf area, local adaptation, New Zealand, OutFLANK, outlier detection, PCAdapt, phenotyping, population genetics, proteomics, selection, signatures of selection, single nucleotide polymorphism, soil moisture deficit, transcriptomics, *Trifolium*, water-soluble carbohydrate, white clover.

Acknowledgements

This PhD research and thesis would not have been possible without many people, who I am extremely thankful to. First and foremost, my sincere gratitude to my supervisors, Peter Lockhart, Marty Faville, Andrew Griffiths and Jennifer Tate for their immense knowledge, relentless support, patience and guidance throughout the course of my doctoral study. I am immensely grateful to have the opportunity to study under the guidance of this outstanding advisory and supervisory team whose wealth of both academic and practical knowledge contributed to not only this thesis but also my professional development.

Throughout this PhD I have had help from many incredible staff at AgResearch. Thank you to Paul Maclean, Ruy Jauregui, Poppy Miller, Abdul Baten, Catherine MacKenzie (formerly AgResearch), Siva Ganesh (formerly AgResearch), Rachael Ashby, Ken Dodds and Sai Arojju for their help with bioinformatics and statistics. To Paul, I am very thankful for making yourself available to me whenever I needed at ask “just one quick question”. I am also grateful for collaborative efforts from Marni Tausen at Aarhus University, Denmark and Limei Zhang at Nanchang University, Jiangxi, China.

Thank you to Anna Larking, Craig Anderson, Prue Taylor, Won Hong, Narsaa Na, Zulfi Jahufer, Peter Moran, Steve Odering, Andrew Faram, Tore Rayner, Rachel Tan, Hong Xue, Zac Beechey-Gradwell, Emmy Bethel, Grace Ehoche, Emma Griffiths, Yulia Morozova, Jana Schmidt, Chaewon Song, Kulatunga Tennakoon, Hayley Ridgeway for their help and technical assistance. Thank you to the AgResearch Forage Genetics team, John Ford (PGG Wrightson Seeds) for sharing your knowledge, Prashant Joshi (Massey University) for help in the lab and the LoST lab members.

A special thank you to Ken Olsen and his lab for hosting me on my trip over to St Louis. I am grateful for collaborative efforts from Mehdi Mirzaei, Dana Pascovivi and Jemma Wu at Macquarie University, Australia. I would also like to thank the land owners in the South Island/Te Wai Pounamu of New Zealand/Aotearoa who took time out of their day to meet with me and provided access to their pastures. I would also like to thank Tony Conner and Marcelo Carena for their support with my project.

I would like to acknowledge the financial support from AgResearch (MBIE programme “Genomics for Production and Security in a Biological Economy”, contract number C10X1306) and Massey University (Marsden Fund “Improved modelling in evolutionary transcriptomics and proteomics will advance understanding of plant

adaptation”, contract number MAU1707). I would like to thank AgResearch for allowing me to complete my PhD at the organisation.

Finally, I wish to thank my close friends and family for the emotional support, unconditional love, patience, always providing encouragement and for telling me how proud you are of me, it really means a lot.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AA	Homozygote for reference allele
Aa	Heterozygote
aa	Homozygote for alternate allele
adj	Adjusted
AFLP	Amplified fragment length polymorphism
AI	Aridity index
AMOVA	Analysis of molecular variance
ANOVA	Analysis of variance
ANTH	Anthrone
AP	Annual precipitation
AT	Arrowtown
AV	Awatere Valley
BIC	Bayesian information criterion
BLAST	Basic Local Alignment Search Tool
BLASTP	Protein Basic Local Alignment Search Tool
bp	Base pairs
BP	Biological process
CF	Cape Foulwind
Chr	Chromosome
CL	Clarence
DAP	Differentially abundant protein
DAPC	Discriminant analysis of principal components
DET	Differentially expressed transcript
df	Degrees of freedom
DF	Discriminant function
DM	Dry matter
DNA	Deoxyribonucleic acid
EAAs	Environmental association analyses
End	End generation
F	Ford
FDR	False discovery rate
FH	FNZLL-High-End
FL	FNZLL-Low-End
FL	Fruitlands (Chapter 5)
FNZLL	Ford New Zealand large leaf
FNZSL	Ford New Zealand small leaf
F_{ST}	Fixation index
g	Gram
GATK	Genome Analysis Toolkit
Gbp	Gigabase pair
GBS	Genotyping by sequencing
GO	Gene ontology

GS	Genomic selection
GWAS	Genome-wide association study
H	High
ha	Hectare
HA	Haast
H_E	Expected heterozygosity
HMW	High molecular weight
H_O	Observed heterozygosity
HSI	Hyperspectral imaging
HWE	Hardy-Weinberg equilibrium
IDM	Identity match
IDs	Identifiers
ioSNPs	Intergenic outlier SNPs
K	Genetic clusters
Kbp	Kilobase pair
K_E	Number of latent factors
KEGG	Kyoto Encyclopedia of Genes and Genomes
kg	Kilogram
KGD	Kinship using Genotyping by sequencing with Depth adjustment
KJ	Kumara Junction
KK	Kaikoura
K_P	Putative number of ancestral populations (Chapter 5)
K_{PC}	Number of principal components
L	Low
LA	Leaf area
LD	Linkage disequilibrium
LDA	Linear discriminant analysis
LFC	Log_2 fold change
LFMM	Latent factor mixed model
LG	Landscape genomics
LL	Large leaf
Ll	Leaf length
LMW	Low molecular weight
LT	Lower Takaka
Lw	Leaf width
MAF	Minor allele frequency
MAS	Marker-assisted selection
Mbp	Megabase pair
MF	Molecular function
Mid	Middle generation
min	Minimum
MM	Middlemarch
MR	Makarora
MSA	Mean successful assignment
n	Number
N	Nitrogen

Na	Number of alleles per locus
Ne	Effective population size
NIRS	Near infra-red reflectance spectroscopy
NIWA	National Institute of Water and Atmospheric Research
NZ	New Zealand/Aotearoa
OM	Omarama
p	*p*-value
Parent	Parent generation
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PET	Annual potential evapotranspiration
PEV	Proportion of explained variance
Pl	Petiole length
Pop	Population
PWC	Pairwise comparison
Q-Q	Quantile-Quantile
QTL	Quantitative trait loci
r	Pearson correlation coefficient
r^2	Coefficient of determination
RIN	RNA Integrity Number
RMSE	Root mean square error
RNA	Ribonucleic acid
RQS	RNA quality score
RS	Rahu Saddle
RV	Rai Valley
SB	Southbridge
SD	Standard deviation
SE	Standard error
SL	Small leaf
SMD	Soil moisture deficit
SNP	Single nucleotide polymorphism
srad	Solar radiation
SSMP	Starch and sucrose metabolism pathway
SSR	Simple sequence repeat
SSS	Soluble sugars and starches
STAR	Spliced Transcripts Alignment to a Reference
SWATH MS	Sequential window acquisition of all theoretical fragment ion spectra mass spectrometry
TASSEL	Trait Analysis by aSSociation, Evolution and Linkage
TF	Transcription factor
Tr_{To}	White clover *Trifolium occidentale*-derived subgenome
Tr_{Tp}	White clover *Trifolium pallescens*-derived subgenome
US	United States of America
vapr	Water vapour
VCF	Variant call format

W	Widdup
WH	WNZLL-High-End
WK	Waikuku
WL	WNZLL-Low-End
WNZLL	Widdup New Zealand large leaf
WNZSL	Widdup New Zealand small leaf
WP	Waipara
WR	Whataroa
WSC	Water-soluble carbohydrate
WUSLL	Widdup United States of America large leaf

CHAPTER 1

Introduction

1.1 White clover morphology, origin and history

The Fabaceae is the third largest family of flowering plants, consisting of 727 genera and ca. 19,325 species (Lewis *et al.*, 2005). The clover genus, *Trifolium* L., is one of the largest genera within the family, with 250 – 300 species (Gillett & Taylor, 2001) and is so called for the characteristic from of the leaf, consisting of three green leaflets (trifoliolate). All species within the *Trifolium* genus are herbaceous perennials or annuals and are often prostrate and stoloniferous (Ellison *et al.*, 2006). White clover (*T. repens* L.) is an allotetraploid ($2n = 4x = 32$) member of the Fabaceae and is agronomically the most important legume of grazed pastures world-wide (Williams *et al.*, 2012).

The *Trifolium* genus originated in the Mediterranean region during the early Miocene, 16 – 23 million years ago (Ellison *et al.*, 2006). The highest level of diversity within the genus occurs in the Mediterranean region, with over half of the *Trifolium* species occurring there (Williams, 1987a). Sub-Saharan African species originated via three independent dispersal events from the Mediterranean region, while North and South American species had a single introduction and origin by dispersal from the Mediterranean region (Ellison *et al.*, 2006). White clover also originated in the Mediterranean region (Vavilov, 1992; Ellison *et al.*, 2006) approximately 15,000 to 28,000 years ago during the last European glaciation, through allopolyploidisation (Griffiths *et al.*, 2019). Polyploidy is the presence of three or more chromosome sets in an organism (Grant, 1981) and can be classified into two main types: allopolyploids and autopolyploids (Soltis, Soltis & Tate, 2003). Allopolyploids form between different species by the union of unreduced gametes or gametes that undergo doubling in the F1 generation, while autopolyploids form within a species by inheritance of unreduced gametes or by somatic doubling (Soltis *et al.*, 2003). Extant diploid relatives of the maternal and paternal progenitors of white clover are *T. pallescens* and *T. occidentale*, respectively (Ellison *et al.*, 2006; Williams *et al.*, 2012; Griffiths *et al.*, 2019). These alpine (*T. pallescens*) and coastal (*T. occidentale*) species are thought to have co-occurred in glacial refugia and through multiple independent hybridisation events produced white clover (Griffiths *et al.*, 2019). The spread of white clover through Europe and Western Asia has likely been facilitated by animal dispersal, via the digestive tract of migrating birds and animals (Ahmad & Uniyal, 2016) and through its domestication by humans.

Clover was documented in historical records during medieval times in Moorish Andalusia, Southern Spain (Bolens, 1981) but early domestication did not occur until

about 1000 AD in Southern Spain (Kjærgaard, 2003). By the mid-16th century, domesticated white clover was grown in Lombardy and the Netherlands, and by the end of the 17th century clover was widespread throughout England (Kjærgaard, 2003). By the 19th century white clover could be found throughout most of Europe, from where it migrated with European settlers to various continents and where it is now naturalised (Zeven, 1991). Selective breeding started in 1917 for this species (Williams, 1987b), with breeding beginning in New Zealand/Aotearoa in 1927 (Williams, Easton & Jones, 2007). White clover has a broad adaptive range from low to high latitudes and altitudes, and has naturalised in temperate regions where the annual rainfall is greater than 750mm (Jahufer, Rogers & Rogers, 2001; Williams *et al.*, 2012). A common feature of the diverse habitats in which white clover is found is high solar radiation. Few clover species tolerate shade (Ellison *et al.*, 2006).

1.1.1 Morphology, breeding system and growth of white clover

The basic structural unit of a mature white clover plant is the stolon, a horizontal creeping above-ground stem consisting of a sequence of internodes separated by nodes. Each node gives rise to a trifoliate leaf, two root primordia and an axillary bud, which is able to grow into a lateral stolon during vegetative growth (**Figure 1.1**). If a node comes into contact with damp soil, adventitious roots may form from the root primordia closest to the ground (Thomas, 1987a).

1.1.1.1 Leaf markings and morphological variability in habitat variations

White clover leaflets are generally egg-shaped or elliptical and have minute serrations around the margin. Leaves are uniformly green or have whitish “V” marks and/or red flecks of anthocyanin pigments on the upper epidermis (Thomas, 1987a). The most common leaf marking is the white “V” mark which is dominant over a recessive allele that confers absence of the “V” mark (Brewbaker, 1955; Carnahan *et al.*, 1955). A diverse range of morphologies in natural populations of white clover are associated with variation in habitat (Ellis & Young, 1967; Caradus *et al.*, 1990). A gradient of plant characters is present across the Mediterranean basin which is linked with changes in soil moisture and temperature (Caradus *et al.*, 1990). Populations of white clover that come from regions of low altitude and latitude are characteristically highly cyanogenic (Daday, 1965; Caradus *et al.*, 1990), have a higher incidence of “V” leaf markings, larger leaflet size, increased flower abundance and a high proportion of plants flowering early (Caradus *et al.*, 1990). Characters that decline with an increase in altitude above 300m include late flowering (flowers per plant), plant persistence and summer growth.

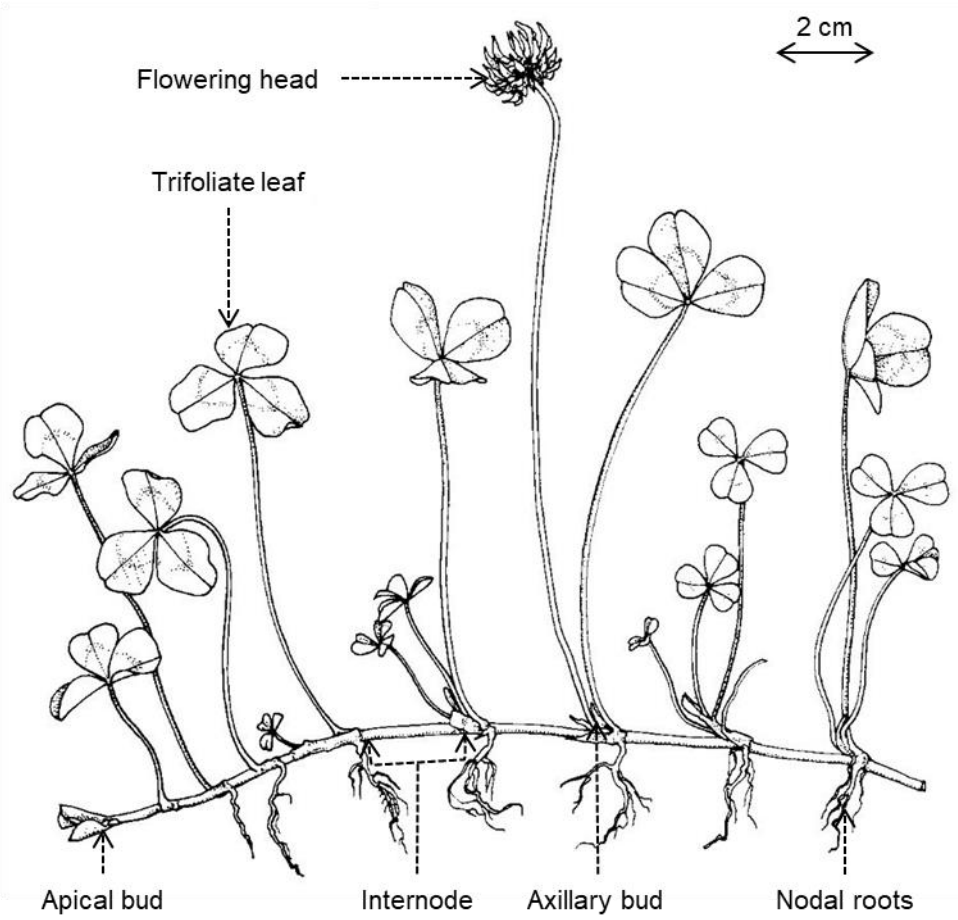


Figure 1.1 The white clover (*Trifolium repens* L.) plant. Adapted from Turkington and Burdon (1983).

1.1.1.2 Leaf size categories of white clover

Leaf size in white clover can range from very small leaflets (< 1 cm long) in prostrate, short petiole types to large leaflets (> 2 cm long) in more erect, longer petiole types (Thomas, 1987a; Frame, 2003). Cultivars (plant variety that have been produced in cultivation by selective breeding) of white clover are classified according to leaf size classes including: small (e.g., 'Tahora'), small-medium (e.g., 'Demand' and 'Prestige'), medium (e.g., 'Avoca', 'Bounty' and 'Pitau'), medium-large (e.g., 'Sustain' and 'Tribute'), and large (e.g., 'Aran', 'Kopu II', 'Kotare', 'Legacy' and 'Mainstay'). Leaf size is linked with several other important features. Large-leaved clovers are normally associated with upright plant growth, larger roots and flower heads, thicker stolons, but have fewer stolons and a lower stolon density (Charlton & Stewart, 1999). These cultivars are

predominantly used and are most productive in rotationally grazed dairy pastures, where livestock are regularly moved to fresh paddocks allowing for sufficient recovery time of the pasture plants (Brock & Hay, 1996). However due to their low stolon density, these cultivars have a reduced ability to regenerate and persist as stolons are more likely to be removed by grazing due to their large size. For this reason, large-leaved cultivars are normally mixed with medium-leaved cultivars to ensure clover persistence in closely grazed pastures (Charlton & Stewart, 1999).

Medium-leaved cultivars are intermediate in features and a typical example of this type is the original 'Grasslands Huia'. These cultivars perform well under a range of grazing managements, except under continuous grazing where they are outperformed by the small-leaved cultivars, and under lax grazing where large-leaved cultivars are better (Caradus *et al.*, 1995). Larger leaf sizes generally have more potential yield, however the higher stolon density of the medium-leaved plants offers a better tolerance to adverse conditions such as drought, pests or pugging (Finch & Percival, 2017). Small-leaved cultivars are low growing, with many small leaves and thin, multi-branched stolons. These plants are best used under continuous grazing, particularly by sheep, as their high stolon density and low-growing habit makes it difficult for animals to remove the plants (Charlton & Stewart, 1999).

1.1.1.3 White clover flowers, seeds and reproduction systems

White clover is a perennial species, capable of both sexual reproduction through seed production and dispersal; and asexual (or vegetative) reproduction through the generation of stolons. Clonal propagation is important for sward maintenance while propagation by seed is crucial for the colonization of new areas.

White clover flowers are produced from active apical buds. The inflorescences are supported by long peduncles and are globular racemes. Each inflorescence consists of 20 – 40 florets, which are white but normally tinged with pink (Turkington & Burdon, 1983; Australian Government, 2008). Flowers turn brown and hang from the seed head as the seed matures (Gibson & Hollowell, 1966). White clover is cross-pollinated by a range of insects, with bumble bees (*Bombus spp.*) and honey bees (*Apis spp.*) being the most common pollinator (Turkington & Burdon, 1983). The strong self-incompatibility system prevents self-fertilisation or, fertilisation by a close relative, which has resulted in an obligate outcrossing system (Wright, 1939; Atwood, 1940). The seeds, 3 – 6 seeds per pod, ripen three to four weeks after pollination and are smooth, heart-shaped and range in colour from bright yellow to brown (Frame, 2003).

After seed germination, white clover has two distinct morphological phases of growth (Brock *et al.*, 2000). The first is a seminal tap rooted phase that lasts for about two years until the taproot dies. The taproot phase has two stages, the first where the primary stem produces 10 – 15 leaves with very compressed internodes and the plant forms a rosette-like habit (Erith, 1924; Thomas, 1987a, 1987b). The second stage consists of axillary buds growing outwards on the primary stem to produce phytomers with elongated internodes which forms a system of stolons (Erith, 1924; Westbrooks & Tesar, 1955). The stolons are genetically identical to the parent plant (Crush *et al.*, 2005). Roots develop on the nodes of the lateral stems which eventually support the new clonal plants that form upon the death of the taproot at the end of this first phase (Westbrooks & Tesar, 1955). Following the death of the taproot, the second clonal phase is initiated where the plant splits into small fragments that are each dependent upon their own nodal root systems (Brock *et al.*, 2000). The clonal phase of white clover persists in the field and allows the plant to spread along the ground (Brock *et al.*, 2000).

1.1.2 History of white clover in New Zealand/Aotearoa and its importance

White clover was introduced to New Zealand/Aotearoa (NZ) from the United Kingdom in the mid-19th century and has since become the most important pasture legume in NZ agriculture (Williams, 1983, 1987a). It was used as a component of a mixed sward (two or more species grown in a pasture) to oversow (sow seed where species are already present) cleared forest and tussock grassland areas. However the species is often considered ‘weedy’ as it grows naturally along pastures, roadsides, lawns and where soil is disturbed (Daday, 1958). White clover came to NZ via numerous imports of seed from throughout Europe, but specific source locations are not known. The assessment of white clover strains and ecotypes in NZ began in the late 1920’s, where four types of clover were identified (Williams, 1983). One of the types was found to be preferable for farming throughout NZ and for 30 years improvements were made at the Grasslands Research Centre in Palmerston North, until a selection was made from the initial line and named ‘Grasslands Huia’, a medium-leaved cultivar found to grow well in many situations (Williams, 1983). Since the 1920’s, more than 250 synthetic cultivars and ecotypes of white clover have been released worldwide (Caradus, 1986) but the most successful has been the ‘Huia’ cultivar. Released in 1964, it has accounted for 70% of white clover seed exported from NZ (Pyke, Rolston & Woodfield, 2004). Clover seed export contributes to the NZ economy, earning NZ$14.2 million in 2011 (NZIER, 2016).

In NZ farm systems today, white clover is used primarily as a component in a binary mixed sward with a grass, typically perennial ryegrass (*Lolium perenne* L.). The grasses provide more forage during the cool seasons and clovers produce forage during warmer summer conditions when grass growth is less vigorous compared to clover (Hoglund & Brock, 1978; Charlton & Stewart, 1999; Brock & Hay, 2001). This complementary association works very well in temperate climates such as NZ. Although both species do not perform well in late summer when conditions are hot and dry (Charlton & Stewart, 1999; Brock & Hay, 2001). Other advantages of these mixed swards is the reduction in weed encroachment and erosion, greater stand longevity than grass or legume monocultures (Casler, 1988), reduced risk of bloat for grazing animals and significant reduction in nitrogen (N) fertiliser application due to white clover's N fixation capability (Ledgard & Steele, 1992). Nitrogen can be transferred directly from the roots of clovers to roots of perennial ryegrass through mycorrhizal fungi interconnecting the root system of both plants, or indirect transfer which involves the death and decomposition of clover which transfers N compounds to the soil which are taken up by neighbouring plants (Brock & Hay, 2001; Rasmussen *et al.*, 2007; Annicchiarico *et al.*, 2014). Reducing N runoff is a key aspect for NZ as improving the quality of fresh water is a strong current focus of the NZ Government (Larned *et al.*, 2004; Rivas *et al.*, 2017).

Harris *et al.* (1998) determined that the optimum clover content in a mixed perennial ryegrass/white clover diet for milk yield in NZ was 62%, coinciding with the dietary preference shown by cattle for 700g white clover kg^{-1} total dry matter (DM) in their diet (Penning *et al.*, 1995). Higher clover content increases the nutritive value of the diets, resulting in increased energy and protein intakes. However, achieving such high clover contents in a grazed pasture situation may prove difficult and unrealistic as high clover content pastures have a lower annual DM production, a higher risk of bloat (build-up of gas in the rumen), and maintenance difficulties (Harris *et al.*, 1997). Therefore, a clover content of 400 g kg^{-1} total DM may be a more realistic option as the cows would still produce 95% of a maximum possible milk yield (Harris *et al.*, 1998). This correlates with an ideal clover content in mixed swards of 20 – 30%, with 50 – 65% needed for near maximum milk production (Harris *et al.*, 1997; Brock & Hay, 2001).

One of the main issues for NZ pastures is achieving and maintaining the optimal clover content in pastures. There is a natural ecological cycle in the proportion of clover in mixed pastures, mainly driven by the way perennial ryegrass (hereafter referred to as ryegrass) and clover respond to N and how this affects their ability to capture light.

Excessive use of N fertiliser (> 200 kg N ha^{-1} $year^{-1}$) gives a competitive advantage to ryegrass as mineral uptake of N requires less energy than the combined approach white clover takes (N uptake and N fixation) (Harris *et al.*, 1996). The improved ryegrass performance suppresses clover growth as clover grows closer to the ground and does not compete well for light. However, as the ryegrass uses the soil N, and reduces it to lower concentrations, the advantage switches to clover due to its stoloniferous growth habit and its ability to fix N. This becomes a cyclic process as the accumulation of N from clover fixation reaches the point where ryegrass growth is favoured and starts outcompeting clover again. Clover will typically perform best under rotational grazing as keeping ryegrass short minimises the light competition (Brock & Hay, 1996). In most stable pasture systems, N fixation needs only to replace losses via leaching, transfers, and volatilisation (Brock & Hay, 2001).

NZ's dairy cows and livestock (principally cattle, sheep and deer) have unrestricted access to pasture at all times throughout the year. Pasture (ryegrass, clover, plantain, chicory, browntop, cocksfoot, tall fescue and other species) is essential to the NZ economy and provides around 90 – 95% of the dietary/energy requirements for livestock (NZIER, 2019). White clover contributes to 12% of all livestock nutrient requirements, making it one of NZ's preeminent economic plants (NZIER, 2016). NZ is the largest exporter of dairy products and produces 3% of all the milk in the world (DairyNZ & LIC, 2018). The value of dairy exports from NZ (milk powder, butter, cheese, and casein) have increased significantly from 2007 to 2012, with exports increasing 72% (NZ$12.5 billion) (Statistics New Zealand, 2012). In 2017 – 2018, dairy farming contributed 28% to NZ exports and earned NZ$15.1 billion (NZIER, 2019). Intensification has put greater pressure on clover performance and fitness for these farming systems (Clark, Mathew & Crush, 2001; Lambert, Clark & Litherland, 2004), and has highlighted the need to develop clover cultivars that are better adapted to intensive cattle grazing systems (Woodfield & Caradus, 1996).

1.1.3 Increasing pasture productivity and environmental sustainability through improving white clover traits

1.1.3.1 Increasing foliar water-soluble carbohydrate content of white clover to reduce the environmental impact and improve the nutritional quality of pasture systems in New Zealand/Aotearoa

Forage digestion in ruminants occurs due to the symbiotic relationship between the host animal and gut microflora (bacteria, fungi and protozoa). The majority of digestion

occurs in the rumen through fermentation and the remainder in the lower gut, mainly the caecum and colon (Moran, 2005). Forages are mainly comprised of less-digestible, structural carbohydrates, namely cellulose and hemicellulose (collectively neutral detergent fibre) and non-structural fractions including starch and soluble cell contents. Soluble cell contents consist primarily of water-soluble carbohydrates (WSC) and crude protein, but small amounts of fats, minerals and vitamins are also present (Humphreys, 1989). Carbohydrates are the main source of digestible energy for ruminants but the relative amounts of digestible fibre and soluble cell contents affects the rate of forage intake (energy utilised from the feed). Carbohydrates ingested by the ruminant are broken down by microbes in the rumen to products that can be used by both the ruminant and microbes. Microbial fermentation breaks carbohydrates down to simple sugars and the end products are volatile fatty acids (mainly acetate, propionate and butyrate) which provide a major source (70%) of energy for the ruminant (Moran, 2005). Depending on whether a ruminant is fed pasture (high in structural carbohydrates) or a high-grain diet (higher concentrations of readily digestible carbohydrate than structural carbohydrates) will determine which volatile fatty acids are produced. Soluble and storage carbohydrates are fermented faster than structural carbohydrates (Ulyatt, Lancashire & Jones, 1977). Bacteria that digest starchy feeds (e.g., cereal grains) produce mainly propionic acid and lactic acid (a biproduct of starch fermentation) which creates a more acidic environment (pH 5.5). This increase in acidity can overwhelm the ruminant's ability to buffer and absorb these acids and lead to metabolic acidosis. The acidic environment leads to tissue damage within the rumen and can lead to ulcerations of the rumen wall. Furthermore, the acidity change can cause a shift in the microbial population as it supresses the bacteria that digest neutral detergent fibre. This has negative implications as the bacteria that digest neutral detergent fibre produce a large proportion of acetic acid, which is important in the production of milk fat. As these bacteria are sensitive to acidity in the rumen, if the rumen becomes too acidic through feeding rapidly digestible carbohydrates, the growth rate of the bacteria declines and they can be completely eliminated. Reduction or elimination of these bacteria can result in a reduction of the digestibility of the feed and may reduce the ruminant's intake of feed. If ruminants are fed feeds high in soluble sugars (e.g., molasses, and sugar cane) there are generally fewer problems with increased acidity in the rumen than starchy feeds but these sugary feeds need to be introduced slowly into the ruminant's diet (Moran, 2005).

There are two types of protein available for ruminant use: protein from the feed and protein from the microbes that inhabit the rumen. During digestive contractions,

some microbes are washed from the rumen to the abomasum (equivalent to a non-ruminant's stomach), where they are digested like other proteins in acid, thus creating a source of protein for the animal (Moran, 2005). The amount of crude protein the animal digests is split into two fractions: soluble and insoluble (also known as "rumen bypass protein") (Beever, 1996). Rumen microbes break down the soluble protein into peptides, amino acids and ammonia, which are used by the microbes along with energy from carbohydrate digestion for growth and reproduction. Excess ammonia is absorbed by the rumen wall and converted to urea in the liver, where it is excreted by the body as urine or it is returned in the blood to non-protein nitrogen (N). Rumen microbes do not digest insoluble protein, instead it is digested in the abomasum and absorbed through the small intestine and used by the ruminant as a protein source. The amount of protein utilised by the ruminant depends on the extent to which dietary protein is degraded in the rumen and on the growth and outflow of the microbes from the rumen (Moran, 2005). There is a strong interaction between carbohydrate and protein metabolism as microorganisms digest most of the feed (Nocek & Russell, 1988). Increasing simple carbohydrates can bolster the supply of readily available energy to support microbial degradation of crude protein in the rumen (Kingston-Smith & Theodorou, 2000). This improves the incorporation of protein into microbial protein which is utilised by the ruminant when the microbes are washed into the abomasum and eventually the small intestine (Cosgrove *et al.*, 2009). Furthermore, less amino acid is converted to ammonia and lost as urea through urine and dung (Nocek & Russell, 1988; Edwards *et al.*, 2007). Reduction of N output onto soils can reduce nitrous oxide (N_2O) production. Soil microbial transformations of nitrification and denitrification contribute approximately 70% of the annual N_2O production worldwide (Gödde & Conrad, 2000). N_2O has a long atmospheric lifetime (over 100 years) and traps heat 300 times more effectively than carbon dioxide, and hence plays a considerable role in contributing to climate change (Intergovernmental Panel on Climate, 2014). Thus, reducing the overall loss of N to the environment can lead to more productive pastures and more sustainable farming.

Land used in the agricultural sector in NZ is primarily utilised for pasture production which provides animals year-round access to forage. Grain supplements are not widely produced and are an expensive supplement for NZ farmers. One way of increasing the productivity of pastures in NZ, and concurrently reducing environmental footprint, is through increasing the nutritional quality of forages. Forages bred for high sugar content can be used to increase readily available energy, without the requirement for expensive supplementary feeds (Cosgrove *et al.*, 2007). As forages contain sufficient fibre, increasing the sugar content should not have negative implications for the acidity

of the rumen. As previously mentioned, white clover provides a high quality, protein rich feed to grazing animals, however increasing readily available energy in the form of soluble sugars (e.g., fructose, glucose and sucrose) and starch is suggested to improve protein utilisation, which results in increased animal and environmental benefits (Kingston-Smith & Theodorou, 2000). This is because more digested N can be utilised for milk production in dairy cows and live weight gain in sheep, with less N lost in urine and dung which reduces the environmental impact of the system (Lee *et al.*, 2001; Moorby *et al.*, 2006; Easton *et al.*, 2009). Success in increasing milk yield per cow and live weight gain in sheep has been observed through increasing the WSC content of ryegrass (Lee *et al.*, 2001; Moorby *et al.*, 2006; Easton *et al.*, 2009). Furthermore protein utilisation in dairy cows fed high WSC clover was improved compared to dairy cows fed a low WSC white clover diet (Higgs *et al.*, 2010). However there has been conflicting evidence for the impact of high WSC on production, with studies where ryegrass WSC content varied by 24 – 32 g kg^{-1} DM having no significant effect on milk production (Tas *et al.*, 2005; Taweel *et al.*, 2005; Tas *et al.*, 2006; Taweel *et al.*, 2006; Cosgrove *et al.*, 2007). Furthermore, some have argued that the ratio of WSC to protein content in forage is the key parameter for increasing protein metabolism and beneficial protein partitioning (to production and growth rather than excretion) rather than increasing WSC content on its own (Cosgrove *et al.*, 2007; Edwards *et al.*, 2007). Edwards *et al.* (2007) suggest that a WSC:protein ratio in forage needs to be greater than 0.7 to improve efficiency of protein utilisation for milk production. This ratio has been achieved in white clover populations bred for increased WSC levels, however the protein content of the populations was still above optimal levels for animal diets (Widdup *et al.*, 2010). Furthermore, the high WSC white clover populations exhibited a reduction in vigour but this was attributed to inbreeding depression resulting from selective breeding (Widdup *et al.*, 2010). As pasture in NZ usually contains a mixture of grass and clover, the protein content of the overall pasture is adequate for animal consumption (Pacheco & Waghorn, 2008) but increasing the overall WSC content, hence increasing WSC in white clover as well as grass, could result in increased overall pasture productivity and improved environmental footprint.

Although there is conflicting evidence of whether increasing WSC content will provide a substantial increase in milk production and live-weight gain, there is considerable evidence suggesting that increasing WSC content in forage can mitigate the environmental impact of pasture systems. As well as mitigating greenhouse gas emissions, as described above, manipulating WSC content of pastures could also reduce the environmental impact of pasture agriculture by reducing N losses to

waterways. Runoff of N and phosphorus causes excessive algae growth which disrupts the natural river ecosystem. In the past two decades there has been significant improvement in the reduction of N in NZ's rivers, but agricultural activity still results in a higher nutrient level in river catchments (Larned *et al.*, 2016). White clover's N fixing ability significantly reduces the amount of N fertiliser applied to pastures in NZ, but further reducing N runoff from pastures can be achieved by reducing the N that is produced by the animals (Ledgard & Steele, 1992; Harris *et al.*, 1998). Previous studies have shown that increased levels of WSC in perennial ryegrass have reduced N loss through urine and dung from ruminants (Miller *et al.*, 2000; Miller *et al.*, 2001; Moorby *et al.*, 2006). Therefore, it is thought that increasing the overall WSC of the pasture system, including that of white clover, will decrease the loss of N to the land from ruminants (Edwards *et al.*, 2007; Widdup *et al.*, 2010).

Elucidating the genetic control of WSC in white clover could improve breeding efficiency for producing clover plants with increased WSC. Selection programmes have been undertaken at AgResearch to produce divergent foliar WSC white clover populations. Two discrete breeding programmes, one running between 2000 – 2004 over four generations of selection in three breeding pools (Widdup *et al.*, 2010) and the other between 1999 – 2004 over six generations in two breeding pools (Mr John Ford, pers comm), were utilised for this thesis. In all breeding pools, divergent selection was undertaken to create populations with low or high levels of WSC. Details of that process are provided in Chapter 2 (section 2.2.3). Overall, a total of five breeding pools, consisting of three "Widdup" (W) and two "Ford" (F) pools, were available for use: WNZLL, WNZSL, WUSLL, FNZLL and FNZSL. These pools consisted of nine populations in each of the Widdup pools and thirteen populations in each of the Ford pools, giving 53 populations in total. After four generations of selection, WSC levels in the high and low WSC Widdup populations diverged significantly ($p < 0.05$) from the initial unselected Parent population. By the fourth generation the high WSC populations had 35% greater WSC concentrations and the low WSC populations had 35% lower WSC concentrations than the generation average (Widdup *et al.*, 2010). The high WSC populations showed greater WSC concentrations compared to commercial varieties and there was an increase in total WSC from 230 g kg^{-1} DM to 310 g kg^{-1} DM from the Parent population to the fourth generation high WSC population. The crude protein (CP) decreased from 320 g kg^{-1} DM to 250 g kg^{-1} DM, giving a WSC:CP ratio of 1.2 (Widdup *et al.*, 2010), higher than the 'ideal' 0.7 WSC:CP ratio, and the CP levels still remained higher than optimal levels (200 g kg^{-1}) for animal diets (Pacheco & Waghorn, 2008). Using a subset of these populations as a genetic resource to identify genomic regions

associated with WSC accumulation could ultimately aid future breeding programmes by providing WSC-linked molecular markers to use in marker assisted selection. Furthermore, identifying expressed transcripts and proteins that are related to WSC accumulation can identify genes involved in WSC accumulation. In addition to identifying genomic controls of foliar WSC, from an evolutionary point of view, this study system has the potential to provide insight into the evolutionary and adaptive response of plants to selection. That is, questions can be addressed such as how many generations are required before a significant shift in phenotype is observed in response to artificial selection? By what mechanism is the response mediated – by sorting allelic variants of carbohydrate biosynthetic genes, and/or selection for trans-acting and/or post-translational regulation of carbohydrate metabolism? Will different pools of white clover exhibit the same physiological and biochemical response? Answering these questions, which are addressed in this thesis, could provide insight into ways of increasing the effectiveness and productivity of breeding programmes for white clover.

1.1.3.2 Utilise local adaptation in contrasting environments to aid discovery of genomic regions associated with soil moisture deficit tolerance

Climate change is expected to increase the rate of extreme weather events (Reisinger *et al.*, 2010). One way to minimise the impact of the possible extreme weather events on agricultural production is to further our understanding of the genetic control of plant persistence in unfavourable environmental conditions, for example drought, and leverage that understanding to breed climate-resilient plant varieties. The magnitude and frequency of droughts are expected to increase in a warmer climate as evapotranspiration increases and is not compensated by a simultaneous increase in precipitation (Reisinger *et al.*, 2010). Droughts in NZ can have substantial short-term macroeconomic impacts because of dependency of the NZ economy on the agricultural sector, which is itself dependent on rainfall. Examples of the negative economic consequences of drought include those in 1998, 2008 and 2013, which had a considerable direct impact on the agricultural industry (Kamber, McDonald & Price, 2013). Costs to the NZ economy associated with the 2013 drought are estimated at NZ$1.5 billion (Frame *et al.*, 2020), while severe moisture deficits in 2008 cost the economy $2.8 billion (Clark, Mullan & Porteous, 2011). Drought resistance is therefore an important breeding target for enhancing pasture productivity under water deficit conditions.

Drought stress limits white clover persistence in pastures. Identification of white clovers with enhanced drought tolerance could increase the legume content in pastures which in turn enhances pasture quality. Drought tolerance is a complicated trait with many white clover studies focusing on root architecture (van Den Bosch *et al.*, 1993; Annicchiarico & Piano, 2004), carbohydrate storage (Karsten & MacAdam, 2001), compatibility solutes and photosynthetic pigments (Kim *et al.*, 2004; Lee *et al.*, 2009), drought-inducible dehydrins (Vaseva *et al.*, 2011; Vaseva, Anders & Feller, 2014), protective flavonoids (Ballizany *et al.*, 2012) such as quercetin glycosides (Hofmann & Jahufer, 2011) and kaempferol glycosides (Hofmann *et al.*, 2000) and their role in drought tolerance. Selection for extensive, deep root systems in white clover has been recommended for better tolerance to less severe intermittent drought stress (Collins, 2002). However, it has also been shown that selection for root morphology is less effective than screening plants in drought-prone environments to find plants that exhibit additional physiological characteristics that enable adaptation to drought (van Den Bosch *et al.*, 1993). Because droughts impose strong selective pressure on natural plant populations, investigating drought adaptation in naturalised populations can be informative for efforts to produce drought tolerant populations (Mickelbart, Hasegawa & Bailey-Serres, 2015).

The environmental heterogeneity observed in the South Island/Te Waipounamu of NZ provides an exciting opportunity to study the molecular basis of local adaptation. NZ is located in the Pacific Ocean and has a distinct maritime character to the climate (Sturman & Wanner, 2001). Mountain chains extending the length of NZ provide a barrier from the prevailing westerly winds, and also divide the country into different climatic regions (Gibson & Cullen, 2015). The entire western portion of the South Island/Te Waipounamu is dominated by the Southern Alps/Kā Tiritiri o te Moana, a relatively straight mountain range orientated North-East to South-West, rising over 3000 m in height (Sturman & Wanner, 2001). Generally, the West Coast/Te Tai Poutini of the South Island/Te Waipounamu has wetter and cooler conditions, whereas the area to the east of the main divide has the driest and warmer conditions. Precipitation gradients are extreme across the South Island, ranging from over 12,000 mm per year on the West Coast/Te Tai Poutini to less than 55 mm per year in the rain shadow areas on the East Coast (Sturman & Wanner, 2001). Local adaptation occurs when individual plants exhibit superior fitness in their home environment compared to a transplanted individual. Local adaptation is driven by natural selection as it determines which characteristics are favourable under certain environmental conditions (e.g., changes in soil moisture deficit, temperature, soil type and solar radiation) (Kawecki & Ebert, 2004; Sork *et al.*, 2013).

There is evidence that local adaptation can occur over a short period of time (tens of years) e.g., Grant, and Grant (2002); Saccheri *et al.* (2008); Umina *et al.* (2005). Annuals and short-lived perennials may adapt more readily to a changing environment than long-lived herbaceous and woody perennials due to their short generation times (Anderson, Willis & Mitchell-Olds, 2011). Local adaptation is most likely to occur in species with large effective population sizes as they have the ability to harbour large amounts of existing genetic variation. White clover has large effective population sizes (Griffiths *et al.*, 2019) that would be expected to facilitate local adaptation.

A previous study by van Ham *et al.* (2016) identified 26 white clover populations from both West and East sides of the South Island/Te Waipounamu that were present in "long-term" pasture, which was arbitrarily defined as more than 10 years uninterrupted pasture (i.e., clover was not resown during that period). Use of a subset of these white clover populations from environmentally contrasting sites is an ideal resource for identifying genomic regions associated with adaptation to high SMD in white clover. Due to the hypothesised strong environmental selective pressure and the short generation time of white clover, it is expected that local adaptation has driven evolution of drought tolerant white clover populations. Hence, assessment of the genetic variation of these white clover populations may identify regions of the genome associated with drought tolerance.

1.2 Approaches to identifying signatures of selection

1.2.1 Signatures of selection

Genetic variation within populations contributes significantly to phenotypic variation on which selection acts and adaptation is based. Selection can act upon existing (standing) genetic variation or new mutations, with allele frequencies changing in the adapting populations (Barrett & Schluter, 2008). Evolution from standing genetic variation is likely to be faster than from new mutations as there may already be a high starting frequency of the beneficial allele(s) (Barrett & Schluter, 2008). A selective sweep is the reduction or elimination of linked genetic variation resulting from a beneficial allele reaching fixation due to strong positive selection (Ewing *et al.*, 2011). A selective sweep can occur when a rare or new allele that increases the fitness of the individual relative to other members of the population, increases rapidly in frequency due to natural or artificial selection. Because other less fit alleles are displaced from the population, the nucleotide diversity at the selected locus decreases. As the prevalence of this allele increases, polymorphic variants linked to the selected allele are also dragged to fixation, reducing

nucleotide diversity in the adjacent regions as well (Nurminsky, 2001). This process is known as genetic hitchhiking and was first proposed by Smith, and Haigh (1974). The size of the genomic segment affected by hitchhiking is proportional to the strength of selection driving the selective sweep. The larger the hitchhiked segment, the higher the probability of detecting a selective sweep in the region (Nurminsky, 2001).

The hitchhiking effect is expected to decrease with distance from the advantageous allele, which produces a “valley” of reduced genetic variation and high linkage disequilibrium (LD) around the site of selection. The extent to which natural or artificial selection acting upon a beneficial allele will affect levels of variation at neighbouring alleles is correlated with the strength of selection (Rose *et al.*, 2011). Selective sweeps are generally defined in three categories. The first is a “hard selective sweep” where a selective sweep originates from a single new mutation, and all ancestral neutral variation that is tightly linked to the selected allele will be eliminated by hitchhiking. Ancestral variation can be preserved only if there is recombination between the polymorphic locus and the selection target during the selective phase. In contrast, strong selection often leads to a “soft sweep,” as it increases the probability of multiple forms or haplotypes of a beneficial mutation being swept. The second type of selective sweep is a “soft sweep from multiple origins” (multiple copies that are derived from independent origins). These sweeps are frequent if the mutation rate at the population level is sufficiently high and selection favours their fixation. The final category is a “soft sweep from existing genetic variation” and occurs when previously neutral mutations become beneficial due to an environmental change. These mutations may exist at low frequency in numerous genomic backgrounds and become more frequent, even for a very low mutation rate, if the mutant alleles have a high relative selective advantage. The sweep pattern linked to these genetic variants depends on the strength of the deleterious selection that the allele experienced in the old environment (Hermisson & Pennings, 2005). The effect of a selective sweep is expected to decline with time. This recovery is due to the eventual accumulation of new mutations and recombinational events that eventually occur between adjacent loci and disrupt the linkage between them (Rose *et al.*, 2011).

The importance of selective sweeps in molecular evolution and artificial selection is becoming increasingly evident. Several studies in animals and plants in the early 2000s detected regions of low variation and high LD when investigating variation at the genomic level, exhibiting tell-tale signs of recent selective sweeps (Schlenke & Begun, 2004; Sabeti *et al.*, 2006; Raquin *et al.*, 2008) and sweeps in maize from domestication

(Vigouroux *et al.*, 2002a; Tenaillon *et al.*, 2004). A more recent study explored the evolution of organophosphorus insecticide resistance in the Australian sheep blowfly, *Lucilia cuprina* (Diptera: Calliphoridae), which provided a useful model to test predictions of the impact of natural selection on the patterns of variation at linked loci. Responses to insecticides are convenient to study as selection is typically strong and relatively recent, permitting the study of sweeps before population recovery (Rose *et al.*, 2011). In another study, Mathew *et al.* (2015) investigated the date palm (*Phoenix dactylifera* L.), which is one of the oldest cultivated trees. They suggested that the origin of date palm domestication was in North Africa and the Arabian Gulf. They concluded this on the basis of genomic regions with high densities of geographically-segregating single nucleotide polymorphisms (SNPs) identified using genotyping by sequencing (GBS) analyses. They observed higher levels of allele fixation on the X-chromosome than on the autosomes. One explanation for this observation is the existence of selective sweeps that preserved domesticated alleles on the X chromosome. Mathew *et al.* (2015) suggested that these alleles were adapted to the geographic regions of cultivation. Many other recent studies have investigated selective sweeps as a consequence of artificial selection in plant species such as rice (*Oryza sativa* ssp. *japonica* and *Oryza sativa* ssp. *Indica*) (Yuan *et al.*, 2017), wild cabbage (*Brassica oleracea*) (Purugganan, Boyles & Suddith, 2000), wheat (*Triticum aestivum* L.) (Raquin *et al.*, 2008), as well as the sheep (*Ovis aries*) breeds: Zel and Lori-Bakhtiari (Moradi *et al.*, 2012).

As described earlier, the theory of genetic hitchhiking predicts that the level of genetic variation is greatly reduced at the site of strong directional selection and increases as the recombinational distance from the site of selection increases (Kim & Stephan, 2002). This characteristic pattern can be used to detect recent selection from DNA polymorphism data. Numerous methods have been developed for detecting signatures of natural selection (Oleksyk, Smith & O'Brien, 2010), and in the last decade there has been a shift from the analysis of a limited number of anonymous markers or candidate genes to genome-wide studies that encompass thousands of SNPs or entire genomes (Andrew *et al.*, 2013). These genomic analyses have been assisted by advances in next-generation sequencing technology, which allows for the extension of genomic analyses to organisms with no reference genome (Catchen *et al.*, 2011; Lu *et al.*, 2013). One of the major difficulties associated with this technology is that population genetic approaches typically require sampling from many individuals from multiple populations. A cost-effective way to estimate genomic diversity at the population level is to pool multiple individuals per population. It should also be suitable for the identification of candidate genes through approaches such as genome-wide environmental

associations (Schoville *et al.*, 2012). It has been shown that next-generation sequencing analysis of pooled individuals is effective for SNP discovery and provides accurate estimates of allele frequencies (Futschik & Schlötterer, 2010; Gautier *et al.*, 2013; Rellstab *et al.*, 2013).

A selective sweep alters the allele frequencies of SNPs in the vicinity of the selected allele, causing a distorted pattern of genetic variation that can be useful for detecting selection. The predicted reduction in sequence diversity at a neutral locus that is closely linked to a beneficial allele is often analysed by testing a neutral model against (strong) selection with the assumption that the beneficial allele is co-dominant (Smith & Haigh, 1974). Several statistical tests have been proposed for inferring a selective sweep event based on predicted effects relative to the standard neutral model. These include (1) an excess of rare alleles compared to the standard neutral model (expanding from Tajima's Tajima (1989) *D* statistic) (Braverman *et al.*, 1995; Fu, 1997), (2) a depression of expected heterozygosity relative to divergence at the target of selection (Hudson, Kreitman & Aguadé, 1987), (3) an excess of derived variants at high frequency (derived refers to the non-ancestral state as determined from an outgroup) (Fay & Wu, 2000), and (4) an excess of LD (Przeworski, 2002; Kim & Nielsen, 2004). Since these signatures are localized to regions adjacent to the targets of selection, it seems appropriate to attempt to identify loci subject to recent directional selection by analyzing genomic patterns of presumably neutral polymorphism (Harr, Kauer & Schlötterer, 2002; Kim & Stephan, 2002; Vigouroux *et al.*, 2002b). With reduced representation technologies, only a portion of the genome is sampled and marker density is often insufficient to identify nucleotide polymorphisms indicative of a selective sweep. Hence, alternative methodologies have been implemented to identify signatures of selection.

Highly differentiated SNPs can also be identified through genetic differentiation analyses to identify SNPs and genomic regions with elevated genetic differentiation between populations, an approach termed "outlier analyses". These approaches typically include F_{ST}-based tests and principal component analyses to identify loci that are distinct from those under neutral expectations. Outlier analyses do not require phenotypic information and therefore numerous markers can be screened to identify candidate genes. Hierarchical genetic structure can result in false positive loci (type 1 errors) in outlier analyses (Excoffier, Hofer & Foll, 2009) because a hierarchical structure of populations leads to a narrow null distribution of F_{ST} values, which leads to an excess of false significant loci (Excoffier *et al.*, 2009). One way to mitigate false discovery is to consider SNPs detected by at least two outlier detection tests. It is generally thought

that those identified in both tests as outliers are more likely to be true positives (Rellstab *et al.*, 2015; Berthouly-Salazar *et al.*, 2016).

Identifying patterns of genetic variation that might suggest selection can also involve using the F statistics proposed by Wright (1951) which are most commonly used to estimate and interpret the genetic structure of populations. The fixation index (F_{ST}) compares expected heterozygosity (H_E) within subpopulations to H_E among all populations, collectively treated as one population. Values of 0 represent a group of populations with no genetic differentiation (perfectly mixed); while values of 1 indicate high levels of genetic variation and the populations are differentiated. F_{ST} is calculated by **Equation 1.1**. Actual F_{ST} values are seldom 0 or 1 and so require interpretation to be biologically comprehensible. For biallelic marker systems (including SNPs), Wright (1978) suggests that values from 0 – 0.05 indicate little differentiation, 0.05 – 0.15 moderate differentiation, 0.15 – 0.25 great differentiation, and values above 0.25 indicate very great differentiation (Balloux & Lugon-Moulin, 2002; Hartl & Clark, 2007). Calculating pairwise F_{ST} values for all SNPs can identify SNPs with elevated differentiation between populations.

$$F_{ST} = \frac{H_{ET} + H_{ES}}{H_{ET}} = 1 - \frac{H_{ES}}{H_{ET}} \qquad \textbf{Equation 1.1}$$

Where: H_{ES} is the mean heterozygosity averaged over the expected heterozygosity of a sub-population (**Equation 1.2**), and H_{ET} is the expected total heterozygosity for the overall total population (**Equation 1.3**).

$$H_{ES} = p_1(1 - p_1) + p_2(1 - p_2) \qquad \textbf{Equation 1.2}$$

Where: p_1 is the allele frequency for one allele in sub-population 1, and p_2 is the allele frequency for the same allele in sub-population 2, assuming equal population sizes and a diploid locus.

$$H_{ET} = H_{ES} + \frac{\delta^2}{2} \qquad \textbf{Equation 1.3}$$

Where: H_{ES} is the mean heterozygosity averaged over the expected heterozygosity of each sub-population (**Equation 1.2**) and $\delta = p_1 - p_2$.

1.2.2 Identifying genes under artificial selection

The phenotype of an organism is an observable characteristic or trait, including its morphological, biochemical or physiological properties, metabolism, and behaviour and has been the main driver of conventional breeding. An observed phenotype depends on environmental factors and the interactions between genotype and environment. Genetic diversity is positively associated with phenotypic variation and may be derived from naturally occurring or artificially-selected breeding populations (Moose & Mumm, 2008).

Discovering the genes and genetic mechanisms that contribute to phenotypic changes associated with selective breeding and natural selection is of considerable importance. This is because their identification has the potential to make breeding more efficient and enable novel breeding strategies that include techniques of biotechnology and genetic engineering, as well as marker-assisted selection. Furthermore, such studies highlight the importance of incorporating natural genetic variation found in wild species and early landrace varieties (earliest form of cultivars and the first step in the process of domestication). These genetic resources can help expand the genetic base of crop plants and provide greater genetic flexibility for the future (McCouch, 2004). There are two main approaches to achieve this goal, starting at opposing ends of the phenotype-genotype ladder (**Figure 1.2**). To elucidate this variation, the majority of research to date has adopted a top-down approach, which begins with a trait of interest and uses genetic analyses such as association mapping (also known as genome-wide association study, GWAS) and quantitative trait loci (QTL) analysis to identify causative genomic regions. Alternatively, a bottom-up approach might be used, which starts with genetic signatures of adaptation observed in genomic, transcriptomic or proteomic data and then aims to infer a phenotype (**Figure 1.2**). In this latter approach, population genetic and phylogenetic methods can be used to identify signatures of selection across the genome, transcriptome and proteome. Such patterns of variation can help identify candidate genes for target breeding. Ontology analyses can also be used to infer the phenotypes and cryptic physiologies of species (Ross-Ibarra, Morrell & Gaut, 2007; Voelckel, Gruenheit & Lockhart, 2017).

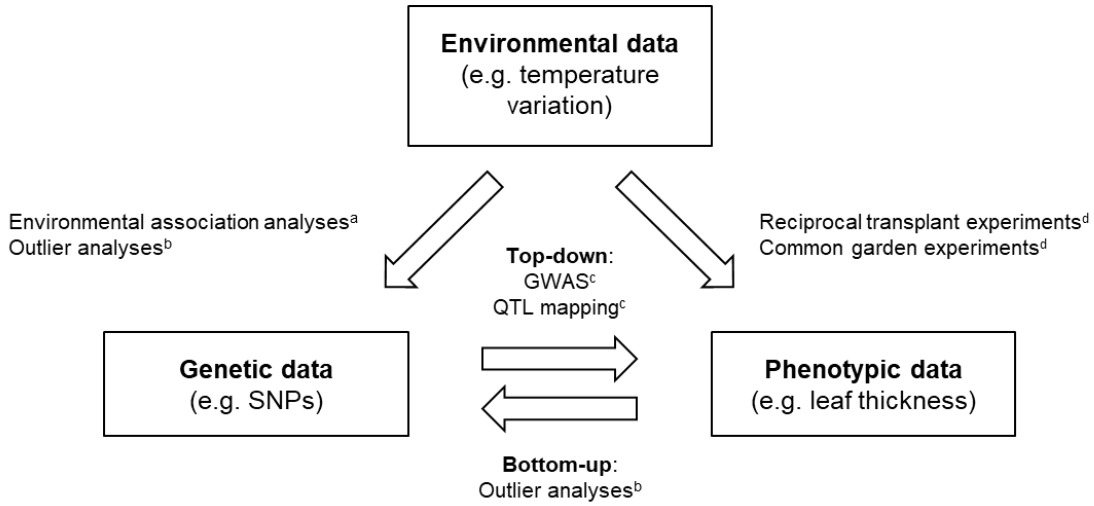


Figure 1.2 Overview of common approaches to detect genes and environmental factors involved in local adaptation using environmental, genotypic and phenotypic data. The simplified diagram is modified from Sork *et al.* (2013) and Rellstab *et al.* (2015) and includes top-down and bottom-up approaches of the phenotype-genotype hierarchy (Ross-Ibarra *et al.*, 2007).
[a] Environmental association analyses aim to correlate environmental data and genotypic data and is the main analytical method for landscape genomics.
[b] Analyses identify loci under selection through a bottom-up approach but rely only on genetic data.
[c] Genome-wide association studies (GWAS) and quantitative trait locus (QTL) mapping identify loci underlying phenotypic traits through a top-down approach.
[d] Reciprocal transplant and common garden experiments explore the phenotypic differences of individuals living in different environments.

Identifying genes underlying adaptive changes as a consequence of artificial selection has been most successful via top-down approaches, beginning with the phenotype and using genetic analyses to discover genomic regions and ultimately candidate genes responsible for the phenotype of interest. One of the most successful methods for identifying these genes has been QTL mapping, but association (or LD) methods are becoming more popular in the plant genomics community. QTL mapping was the first method available for localising the genetic basis of a trait of interest (Sax, 1923). Biparental QTL mapping was a widely used method and led to the successful identification and cloning of genes underlying domestication traits (Paterson *et al.*, 1988). Some well-known examples of QTL location and subsequent mapping include: *Arabidopsis* flowering time (El-Din El-Assal *et al.*, 2001; Werner *et al.*, 2005) and root morphology (Mouchel, Briggs & Hardtke, 2004); maize plant architecture (Doebley, Stec & Gustus, 1995; Doebley, Stec & Hubbard, 1997); rice heading date (Yano *et al.*, 2000; Takahashi *et al.*, 2001; Kojima *et al.*, 2002; Doi *et al.*, 2004); tomato fruit sugar content (Fridman, Pleban & Zamir, 2000; Fridman *et al.*, 2004), fruit shape (Liu *et al.*, 2002) and

fruit weight (Frary *et al.*, 2000; Cong, Liu & Tanksley, 2002); and white clover seed production (Barrett, Baird & Woodfield, 2005) and salt stress tolerance (Wang *et al.*, 2010). However due to the rise of reduced representation sequencing technologies, association mapping has become more prevalent in recent studies, particularly for those species with moderate or no genomic resources.

Association mapping is an alternative method to identify genomic regions that contribute to phenotypes, and can be separated into two types that focus on different levels of genetic analysis. The first type of association analysis aims to identify genome-wide variation that affiliates with phenotypic trait variation (e.g. GWAS). These studies measure the genetic variability of markers that represent the majority of the genome and test the phenotype-genotype association of each marker. The other type attempts to determine the causative genetic mutation(s) that effect phenotype, which typically focus on variation in a few candidate genes rather than the whole genome. The main advantage of association mapping is that there is no need for crosses and producing large numbers of progeny as it can rely on unstructured population samples. This is beneficial for long-lived perennial species and allows studies to proceed rapidly. However, distinguishing true associations from statistical noise due to population and geographic structure can be a challenge. This is because false associations between genotype and population structure/geographic region, instead of genotype and phenotype, can occur. Therefore identifying true associations from statistical noise requires large sample sizes for increased statistical power and to correct multiple tests (Long & Langley, 1999; Macdonald & Long, 2004).

GWAS is widely used in diploid species to study complex traits in breeding populations, but is considered to have limited application for highly heterozygous outcrossing and polyploid species, such as white clover, due to the necessity for developing a very high number of SNP markers because of their large genome sizes. The number of markers required is dependent on the rate of LD decay, which in turn depends on the material used. In natural populations of outcrossing species, LD decays very quickly so there are a very high number of small LD blocks. For this reason, it has been suggested that several million markers will be required for GWAS analyses of white clover (Hayes *et al.*, 2013). Currently such data are unavailable, but with the developments in next generation sequencing technology and numerous promising reports in polyploid species such as cocksfoot, potato, sugar cane, sunflower and wheat (Rosyara *et al.*, 2016; Berdugo-Cely *et al.*, 2017; Zhao *et al.*, 2017; Bock *et al.*, 2018; Phan *et al.*, 2018), GWAS analyses of polyploid species such as white clover have the

potential to be realised in the near future (Brazauskas *et al.*, 2011). Furthermore, numerous GWAS studies have reported significant associations in outcrossing species obtained with just thousands of markers (Arojju *et al.*, 2016; Biazzi *et al.*, 2017; Sakiroglu & Brummer, 2017). Breeding populations derived from a relatively small number of founders can aid QTL and LD mapping in a single population whilst minimising spurious associations due to population structure (Flint-Garcia *et al.*, 2005; Breseghello & Sorrells, 2006; Yu & Buckler, 2006).

Bottom-up approaches start by identifying signatures of adaptation, either genomic or transcriptomic variation, and then use genetic tools (including outlier loci tests and PCA based tests) to identify differentiating SNP loci. This approach is relatively new, with many methodologies still being developed. The ability to detect the signal of adaptation depends critically on the history and strength of selection, the demographic history of the population, and the analysis method. The largest disadvantage of bottom-up approaches is that genetic variation of candidate genes may not be easy to associate with a phenotypic trait in non-model species. This is in contrast to the situation for many model species for which many genetic tools and databases for gene expression data; genetic maps; and partial or complete genome sequences to connect candidate genes to a phenotype are available (Ross-Ibarra *et al.*, 2007). The link from gene to phenotype for non-model species poses some challenges but with the cost of *de novo* genome sequencing decreasing with time, there is increasing availability of genome resources for non-model species, thus making bottom-up approaches more attainable (Ellegren, 2014; Unamba, Nag & Sharma, 2015). Other limitations for bottom-up approaches include levels of genetic diversity, polyploidy and population structure. Polyploidy makes population genetic analysis difficult, requiring careful separation of homoeologues and their independent histories, however the white clover parental subgenomes are well-defined, thus avoiding this problem (Griffiths *et al.*, 2013).

Each of these different methods to identify candidate genes or regions requires verification by additional functional characterization (Weigel & Nordborg, 2005). In many cases, this last step, to connect a candidate gene to a phenotype via functional studies, is often not much easier for top-down approaches than for candidate genes identifying signatures of adaptation. There is an important and pressing need for a broad-based initiative to implement bottom-up approaches in important crop and agricultural species as new sequencing technologies make such an initiative relatively inexpensive.

Like bottom-up approaches, top-down approaches have limitations. Generation time (as for QTL studies) is often a disadvantage but the severe limitation of top-down approaches is identifying a phenotype *a priori* (based on theoretical deduction rather than empirical observation) when studying adaptation. Bottom-up approaches are free from this constraint and also have several advantages for finding genes that contribute to adaptive traits and that will be useful in agronomic context. These advantages include: (1) far fewer plants are needed: < 100 individual samples (Teshima, Coop & Przeworski, 2006) c.f. hundreds or thousands for LD mapping (Long & Langley, 1999); (2) bottom–up approaches can be used for species that lack genetic tools; (3) segregating variation is not required to identify genes of interest, and (4) they shed light on demographic history and provide historical insights into the selection process (Ross-Ibarra *et al.*, 2007).

1.2.3 Identifying adaptive genes responding to environmental heterogeneity

Landscape genetics has been around since the 1950s and typically uses less than one hundred molecular markers to test the effects of landscape variables on genetic population structure and gene flow (Holderegger *et al.*, 2010; Storfer, Patton & Fraik, 2018). In the past decade an increase in the number of molecular markers available for use has led to the development of the field of landscape genomics (LG). This is a new field with fewer than 40 articles published in this timeframe (Li *et al.*, 2017b). Thousands of loci (often SNPs) and even full genomes (e.g., complete transcriptomes or genomes) can be utilised to detect candidate genes under selection that are indicative of local adaptation, which is the aim of LG studies (Storfer *et al.*, 2018). Traditional local adaptation studies typically involve translocation experiments or greenhouse trials under controlled environmental conditions (i.e., common garden experiments). These studies aim to dissect the environment effects from phenotypes but require significant time, labour and financial resources, and the number of individuals studied are often limited (**Figure 1.2**). A bottom-up approach can be used to identify signatures of adaptive genetic variation and relate them to environmental variation. Bottom-up approaches are slightly different in an LG context as environmental data, instead of phenotypic data, are used (**Figure 1.2**). Examples of a bottom-up approach include correlating environmental variables and genotypic data (e.g., environmental association analysis) and outlier analyses to identify regions of the genome under selection (**Figure 1.2**). These bottom-up approaches are used in LG studies to identify significant differences in allele frequencies between populations *in situ* (without removing individuals or seed from their natural habitat), thereby indicating that individuals in the population are experiencing

selection pressure. Outlier detection methods are popular and powerful and can identify loci or regions with extreme differentiation, but they do not include the actual driver of selection (the environment) in the analysis. Hence, for landscape genomics studies, environmental association analyses are used to correlate abiotic/biotic data with genomic data to identify environmental factors and loci that are putatively involved in this process (Rellstab *et al.*, 2015).

1.3 Research question, aims and objectives

The question of which genomic regions are most responsible for foliar WSC accumulation and soil moisture deficit (SMD) tolerance in white clover motivated the work undertaken in this thesis. This question was addressed with two specific aims. The first aim was to identify regions of the white clover genome associated with foliar WSC accumulation, while the second aim was to identify regions linked to SMD tolerance. To achieve these aims the following four objectives were identified:

1. Phenotype 25 white clover populations for WSC content using near infra-red reflectance spectroscopy that have been selectively bred for divergent WSC levels. In addition, leaf area will be measured as it is a potential confounding factor.
2. Use genotyping by sequencing (GBS) technology and appropriate genetic analyses to identify genomic regions associated with WSC accumulation.
3. Examine expressed transcripts and proteins in a subset of the white clover population material to identify genes important for foliar WSC accumulation.
4. Investigate naturalised white clover populations from contrasting environmental habitats, using genotyping by sequencing data, to assess genetic variation and identify genomic regions associated with local adaptation.

The structure of the thesis is shown in **Figure 1.3** and these objectives are addressed in the following four chapters:

Chapter 2 describes the phenotypic assessment of 25 white clover populations for WSC and leaf area using near-infrared spectroscopy (NIRS) and image analysis, respectively. This chapter addresses the first objective. As a significant portion of this thesis is focused on the genetic analysis of white clover populations bred for differing levels of foliar-expressed WSC, it was critical to confirm that divergently selected populations within each of the five pools examined did indeed have significantly different foliar WSC levels.

Chapter 3 identifies genomic regions associated with WSC accumulation using SNPs generated from a double restriction enzyme GBS protocol. The same white clover populations as described in Chapter 2 are utilised, which also allows for a small genome-wise association study (GWAS) to be implemented. Outlier detection approaches are used for identification of loci putatively associated with WSC accumulation.

Chapter 4 identifies expressed transcripts and proteins related to WSC accumulation in a subset of four of the white clover populations used in Chapters 2 and 3. This chapter is complimentary to Chapter 3 and uses standard pairwise comparisons of differentially expressed transcripts and proteins between a subset of high and low WSC populations to identify genes important for foliar WSC accumulation. Concurrence between the three analyses (genomics, transcriptomics and proteomics) is assessed to provide evidence for genomic regions and genes involved in WSC accumulation. In this chapter the third objective was addressed and it provided insight into whether similar selection pressure on the two pools resulted in a convergent phenotypic outcome (i.e., high or low WSC) through selection for allelic variants.

Chapter 5 investigates local adaptation of white clover populations in the South Island/Te Waipounamu of New Zealand/Aotearoa from contrasting SMD locations. Naturalised populations from 17 sites were used to identify genetic signatures associated with drought tolerance through persistence of populations in high SMD areas.

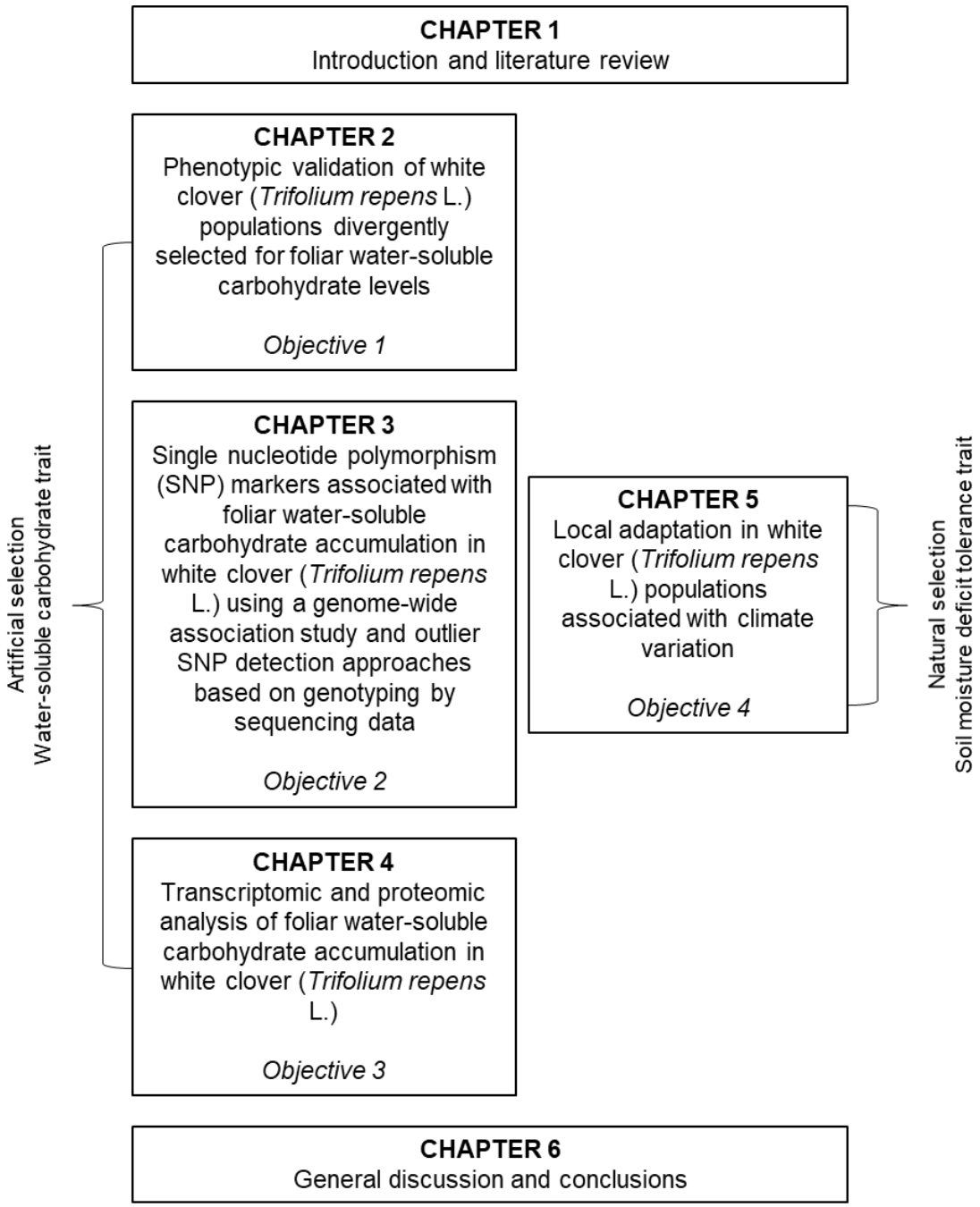


Figure 1.3 Thesis structure with selection type and trait investigated highlighted.

CHAPTER 2

Phenotypic validation of white clover (*Trifolium repens* L.) populations divergently selected for foliar water-soluble carbohydrate levels

2.1 Abstract

White clover is an economically important forage legume in New Zealand/Aotearoa (NZ) due to its nitrogen-fixing ability. Improving water-soluble carbohydrate (WSC) content in white clover is important for nutritional quality and reducing environmental impacts of dairy farms. Previous NZ breeding efforts have resulted in five breeding pools, each selectively bred for divergent (low or high) WSC content. These pools were utilised to identify genomic regions and genes underpinning the trait but before any genetic work could be performed, phenotyping was required to ensure that truly divergent phenotypes were used in the analysis. Therefore, WSC and leaf size, a potential factor influencing WSC, were evaluated in a subset of the divergent material from each pool. Significantly different ($p < 0.05$) WSC levels were observed when comparing low and high WSC populations in all five pools (mean difference of 78.3 g kg^{-1} dry matter). This confirms that divergent selection was successful in altering WSC levels in each of the five breeding pools. Although leaf size varied amongst the divergent populations in three of the pools, there was no statistically significant relationship observed between the two variables. The results from this chapter underpin the genetic, transcriptomic and proteomic analyses in Chapters 3 and 4.

2.2 Introduction

2.2.1 Study system

White clover (*Trifolium repens*, L.) is a member of the Fabaceae and is an obligately outcrossing, herbaceous perennial plant species. It is a recent allotetraploid (2*n* = 4*x* = 32) that formed from two diploid progenitors, the extant relatives of the paternal and maternal progenitors being *T. occidentale* and *T. pallescens,* respectively (Griffiths *et al.*, 2019). Pollination is facilitated via bees (*Apis mellifera*). White clover reproduces by seed and vegetatively by stolons that can form large clonal mats (Gibson & Hollowell, 1966). White clover is the third most economically important plant species in New Zealand/Aotearoa (NZ) and the second most economically important forage species in NZ agriculture, contributing NZ$2.3 billion to NZ's annual Gross Domestic Product (GDP) (NZIER, 2016). White clover is sown in temperate pastures due to its nitrogen-fixing ability which enables it to improve soil nitrogen content and improve the quality of forage eaten by ruminants (Ulyatt, 1997).

2.2.2 Water-soluble carbohydrate trait

Carbohydrates play an important role in plants, including supporting growth and development, as a metabolic "fuel" source, in stress responses and as signal molecules (Calenge *et al.*, 2006). They also play roles as signals to regulate metabolic pathways, such as anthocyanin biosynthesis, nitrogen metabolism and photosynthesis (Gibson, 2005). Carbohydrates in plants can be split into structural and non-structural functions. Structural carbohydrates make up the plant cell wall and consist of polysaccharides such as cellulose, hemicelluloses and pectin, and are less digestible than the non-structural carbohydrates (Ruckle *et al.*, 2018). Non-structural carbohydrates include starch and water-soluble carbohydrates (WSC), which consist of mono and disaccharides but also larger polymers (oligosaccharides and polysaccharides). Monosaccharides are the basic units of carbohydrates and cannot be split by hydrolysis into simpler sugars. The size of the sugar determines whether it is categorised as high or low molecular weight. Low molecular weight (LMW) WSC in plants include monosaccharides (for example, glucose, fructose) and disaccharides (for example, pinitol and sucrose); while high molecular weight (HMW) WSC are polysaccharides, e.g., fructan (oligo- and polysaccharides based on fructose molecules), which typically act as storage carbohydrates (Rosa *et al.*, 2009). The level of soluble sugars varies depending on a range of parameters, including species, growth conditions, and organ, but generally reflects the balance between photosynthesis and growth (Calenge *et al.*, 2006). The

most abundant soluble sugar in white clover foliage is sucrose but glucose and fructose are also found (Ruckle *et al.*, 2018). Breeding and selection for higher sugar levels has been applied in a range of different plant species such as peach (Cirilli, Bassi & Ciacciulli, 2016), corn (Lertrat & Pulam, 2007), perennial ryegrass (Easton *et al.*, 2009) and sugarcane (Huang *et al.*, 2010).

White clover foliage has a high concentration of crude protein but a relatively low concentration of WSC (Cosgrove *et al.*, 2006). Foliar WSC is important because it provides readily available energy to the rumen microbiome which improves the efficiency of crude protein utilisation (Woodfield *et al.*, 2001). Higher levels of WSC available for consumption by ruminant microbes enables a shift in the partitioning of digested nitrogen, with less excreted as urea and more going to animal growth and production (Easton *et al.*, 2009). Breeding for increased foliar WSC can therefore have beneficial effects for the environment as less nitrogen is lost via urine and dung (Luo *et al.*, 2015; Selbie, Buckthought & Shepherd, 2015). In addition to the potential for increased WSC levels to improve nutritional quality and positively affect environmental outcomes, WSC has been found to be important in conferring cold tolerance and drought resistance in plants due to its role in osmotic adjustment (Kerepesi & Galiba, 2000; Dalmannsdóttir, Helgadóttir & Gudleifsson, 2001; Livingston, Hincha & Heyer, 2009). Quantification of WSC content in white clover plants has been extensively researched (Michell, 1973; Ruckle *et al.*, 2018; Kagan *et al.*, 2020). However, little is known about the genetic control of foliar WSC accumulation in white clover. Understanding the genetic control is important so breeding tools such as marker-assisted selection and genomic selection can be developed for the trait. Previous research has focused more on WSC in stolons and their role in cold tolerance (Dalmannsdóttir *et al.*, 2001; Inostroza *et al.*, 2018), and only one study has addressed genetic control of WSC accumulation (Inostroza *et al.*, 2018). Recent breeding programmes have generated white clover populations with improved foliar WSC concentrations compared to commercial varieties (Widdup *et al.*, 2010). Selective breeding to create divergent levels of WSC in white clover foliage has been undertaken in five discrete breeding pools, three from Widdup *et al.* (2010) and two from an unpublished study (Mr John Ford, pers comm). These were utilised for this study and are described below.

2.2.3 Breeding programmes utilised for the current study

Widdup *et al.* (2010) describe the use of the conventional plant breeding practice of recurrent selection to develop divergent white clover lines with modified (low or high)

WSC concentrations. Details of that process are provided here. High and low WSC lines were developed within three white clover pools: United States of America large leaf (USLL), New Zealand large leaf (NZLL) and New Zealand small leaf (NZSL). In each pool, the divergent lines were created starting from a Parent population. The Parent population was initiated from a polycross (cross-pollination) of plants taken from different cultivars with either large or small leaves (LL or SL), New Zealand/Aotearoa or United States of America material (NZ or US). One hundred to one hundred and forty plants were randomly sampled from each Parent population and assessed for leaf WSC in late spring using near infra-red reflectance spectroscopy (NIRS). Based on those data, 20 – 25 plants with high WSC and 20 – 25 plants with low WSC were identified from each Parent population. The three high WSC and three low WSC groups were placed into separate isolations in crossing cages with bumblebees (*Bombus* sp.) to carry out cross pollination, ultimately resulting in six populations of seed that were harvested and stored separately for each population. These six populations were the first generation of selection. Three more generations for both high and low WSC lines were produced in the same manner, resulting in four generations of selection in each of the three pools (**Figure 2.1**). This same methodology was applied to a further two white clover breeding pools (NZLL and NZSL), however in these there were six divergent generations created (Mr John Ford, pers comm). Overall, a total of five breeding pools, consisting of three "Widdup" (W) and two "Ford" (F), were available for use: WNZLL, WNZSL, WUSLL, FNZLL and FNZSL. These pools consisted of nine populations in each of the Widdup pools and thirteen populations in each of the Ford pools, giving 53 populations in total (**Figure 2.1**). Seed from all these populations is stored in the Margot Forde Forage Germplasm Centre (MFFGC; AgResearch Grasslands Research Centre, Palmerston North, New Zealand). A subset of these 53 populations, specifically the Parent population, middle generation (Mid) and final generation (End) populations from each low WSC and high WSC divergent lines in each pool, was used for phenotype evaluation and subsequent genotyping (**Figure 2.1**).

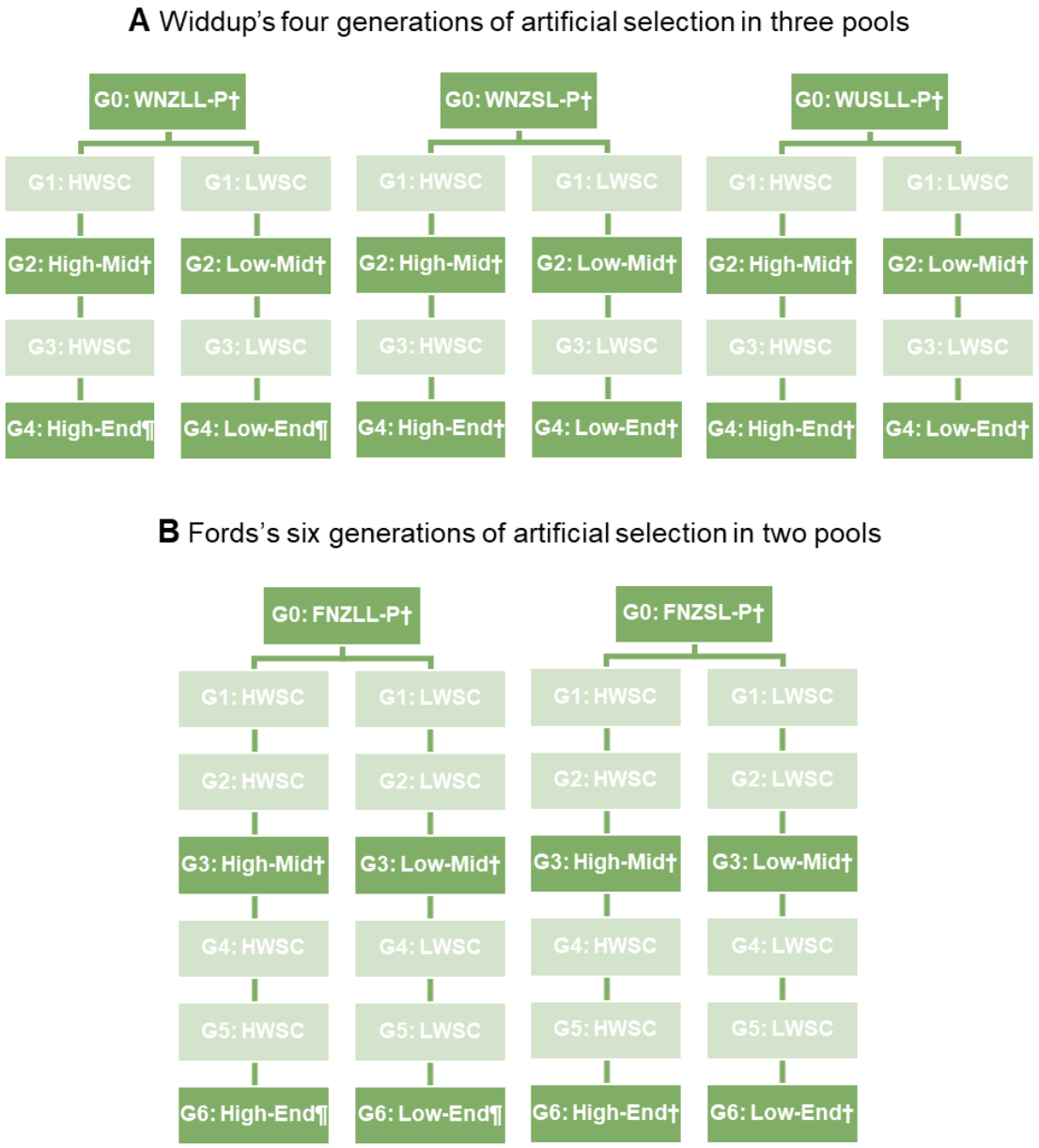


Figure 2.1 Schematic representation of white clover populations for Widdup (**A**) and Ford (**B**).
Note: G = generation, W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL= large leaf, SL= small leaf, P = Parent generation, HWSC and High = high water-soluble carbohydrate (WSC); LWSC and Low = low WSC, Mid = Middle generation, and End = End generation.
† denotes populations used in phenotyping and genotyping studies.
¶ denotes populations used in phenotyping, genotyping, transcriptomic and proteomic studies.

The principal objective of the work described in this chapter was two-fold: firstly, the WSC phenotypes of the populations used in the study needed to be confirmed, to ensure that the subsequent genetic studies (Chapters 3 and 4) were performed on truly divergent phenotypes. Similarly, because the overall goal of the research was to identify genomic regions and genes associated with WSC accumulation *per se*, it was important to establish whether or not the selective breeding programmes had altered WSC independently of changing leaf area (LA). Hence, the relationship between WSC and LA was investigated using correlation and regression analyses. These principal objectives were supported by two additional experiments designed to investigate specific issues that had a bearing on the main experiment. Because NIRS was the method chosen to measure WSC levels in the main experiment, it was important to assess the accuracy of the NIRS calibrations used against an acceptable wet chemistry methodology. Two NIRS calibrations were used for WSC determination. The first was developed using data from a range of forage species and sample types (including white clover) to determine total soluble sugars and starch (Corson *et al.*, 1999). This calibration was used by both Widdup *et al.* (2010) and John Ford (Mr John Ford, pers comm) for WSC determination in white clover foliage. The other available calibration was developed using perennial ryegrass (*Lolium perenne* L.) samples only, to determine total, HMW- and LMW-WSC fractions (Cosgrove *et al.*, 2009). The LMW-WSC fraction targets sucrose, fructose and glucose while the HMW-WSC fraction targets fructans, which do not occur in white clover. In support of the LA study, the number of leaves measured per plant in order to obtain an accurate estimate of mean plant LA needed to be optimised. This is because there is a trade-off between the number of leaves collected and the time it takes to make the LA measurements. Furthermore, the age of the leaf was considered to ensure accurate comparisons within and between individual plants to ensure they were measured at the same developmental stage.

2.3 Materials and methods

The following terminology is used to describe the white clover populations and how they group together. Firstly, a “pool” refers to a multi-parent breeding population within which divergent selection has occurred. There are sub-populations within each pool, hereafter referred to as “populations”, that are the result of divergent selection. For example, the FNZLL pool consists of thirteen populations, five of which were used in the present study (as indicated by the † in **Figure 2.2**). Each population refers to a group of individuals under the same selection (low or high water-soluble carbohydrate, WSC) at a certain point in time (Parent, Middle or End). For example, FNZLL-High-Mid is the population

resulting from selection for high WSC, mid-way through the selection programme. Secondly, a “generation” refers to a single or collection of populations from certain point in time (Parent, Mid or End), irrespective of the direction of selection. For example, FNZLL-High-End and FNZLL-Low-End are collectively the FNZLL-End generation. Lastly, a “line” refers to a collection of populations from the same type of selection, irrespective of the generation. For example, the FNZSL-Low-Mid and FNZSL-Low-End are the FNZSL-Low line, as they have both been bred for low WSC (**Figure 2.2**).

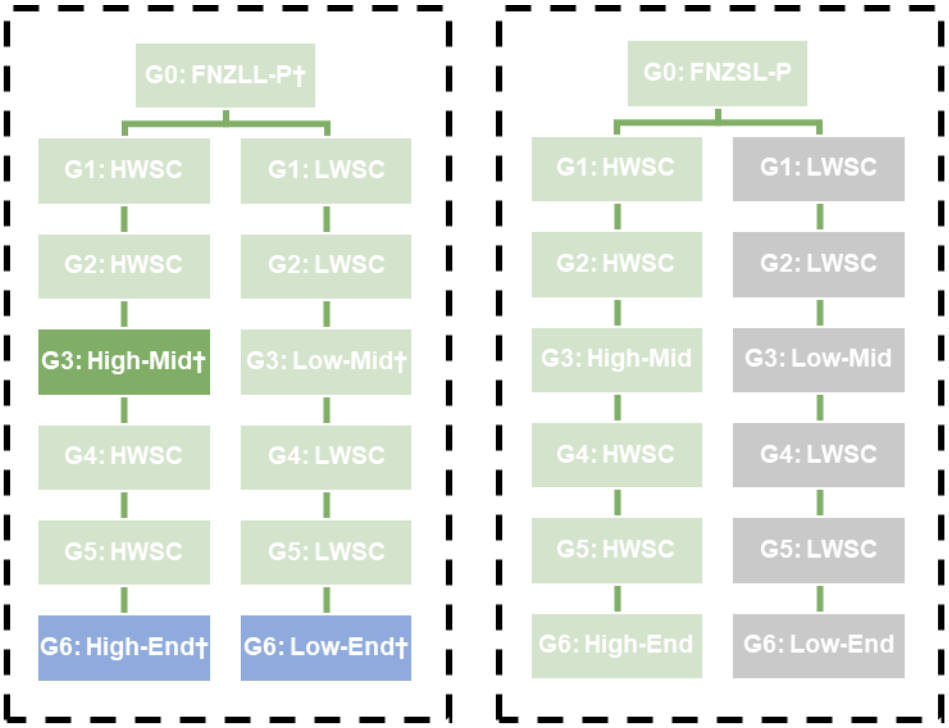


Figure 2.2 Diagram demonstrating the use of terminology used in subsequent chapters. Each dashed box represents a pool e.g., the FNZLL and FNZSL pools. The dark green box represents one population, e.g., FNZLL-High-Mid. The two blue boxes represent a generation, e.g., FNZLL-End. The grey boxes represent a line, e.g., FNZSL-Low.
Note: G = Generation, F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, SL = small leaf, High and HWSC = high water-soluble carbohydrate, Low and LWSC = low water-soluble carbohydrate, P = Parent generation, Mid = Middle generation, and End = End generation.
† denotes populations from the FNZLL pool that were used in the present study.

2.3.1 Establishment of white clover populations

Selective breeding for improved levels of WSC in white clover foliage has been undertaken in two discrete breeding programmes, one running between 2000-2004 over four generations of selection in three breeding pools (Widdup *et al.*, 2010) and the other between 1999-2004 over six generations in two pools (Mr John Ford, pers comm). In all

breeding pools, divergent selection has been undertaken to create populations with low or high levels of WSC. Seed from all these populations is stored in the MFFGC (AgResearch Grasslands Research Centre, Palmerston North, New Zealand) and was utilised for this experiment (**Figure 2.1**).

White clover seed coats are sometimes impermeable and require scarification and stratification before germination. Approximately 100 seeds from each population (53 populations in total) were lightly scarified for 2 – 3 seconds using fine sandpaper (P120) and germinated on petri-dishes containing pre-dampened filter paper. To synchronise germination, seeds were kept at 4˚C for 48 hours and then incubated at 20 – 25˚C for 24 hours. From 5,845 seeds, 5,321 germinated seeds (90% average germination success rate) were planted into propagation trays containing a mix of peat and sand with a 3-month slow release Osmocote fertiliser and maintained under standard glasshouse conditions in March 2017.

2.3.2 Experimental design

All plants were maintained in a glasshouse for approximately two months to allow for seedling establishment. After this, a total of 900 plants from 25 populations to be used in the phenotyping study (plus 5 spare plants for each population) were potted into 2L pots containing a mix of peat and 8 – 9-month slow-release fertiliser. Plants for the phenotyping experiment consisted of 60 plants per Parent population (generation 0), 30 plants from the Middle populations (generation 2 or 3 depending on the breeding pool) and 30 plants from the End populations (generation 4 or 6) (**Figure 2.1**). Potted plants were kept in the glasshouse to acclimatize for two weeks before placing outside in a randomized Latin square design (**Table S2.1**, Appendix 1) constructed in Genstat v 18 (VSN International, 2015) with three replicate blocks in late May 2017 (**Figure S2.1**, Appendix 1). The block was set up on a concrete pad 9 x 9 m and plants were spaced evenly apart at 30 cm spacing (centre-to-centre).

2.3.3 Water-soluble carbohydrate phenotyping using near infra-red reflectance spectroscopy

Plants were maintained outside until mid-November 2017 when they were large enough to harvest for WSC determination. Leaves from the 900 plants were sampled over three consecutive days. One replicate block was harvested per day during a narrow interval, to minimise diurnal variation in WSC levels. To minimise age differences among leaves, only fully expanded healthy leaves were sampled from each plant. At sampling,

approximately 30 leaf lamina were removed from each plant between 8:00 and 10:00 am (when WSC levels are lowest in the leaves therefore controlling for diurnal variation (Widdup *et al.*, 2010; Ruckle *et al.*, 2018)). From a practical standpoint, this is also the time on NZ dairy farms that cows generally return to their paddocks after milking. Each sample was sealed in a plastic bag and snap frozen in liquid nitrogen. Samples were stored in a -20˚C freezer before being freeze-dried, finely ground to pass through a 1 mm sieve and re-dried at 60°C for three hours prior to submitting samples to the commercial laboratory, as per their instructions. Samples were analysed using an MPA Brucker near infra-red spectrophotometer (Nutrition Laboratory, School of Food and Advanced Technology, Massey University, Palmerston North) for a range of nutritional quality attributes, including WSC concentration (g kg^{-1} dry matter, DM). Two NIRS calibrations were used to determine WSC, one determined total soluble sugars and starch (SSS) (Corson *et al.*, 1999) and was originally used by Widdup *et al.* (2010) and John Ford (Mr John Ford, pers comm). The other determined the total, high molecular weight (HMW)- and low molecular weight (LMW)-WSC fractions but was based on samples from ryegrass only (Cosgrove *et al.*, 2009). Two of the replicate blocks (2 and 3) were initially analysed, giving a total of 600 samples. The analysis and timing (season and time of day) of the NIRS experiment coincides with the methodology originally used by the plant breeders when developing these divergent WSC selections (Widdup *et al.*, 2010).

2.3.4 Near infra-red reflectance spectroscopy data validation using wet chemistry

2.3.4.1 Plant material and abbreviations description

A sub-sample of 160 plants was selected from the 600 plants (ca. 27%) used for near infra-red reflectance spectroscopy (NIRS) WSC determination (see section 2.3.3), for anthrone (ANTH) WSC determination. The sub-sample was selected to include individuals from each of the five pools. With NIRS-SSS data ordered from high to low WSC in each pool, every fifth sample was chosen. In addition, a further four of the highest and four of the lowest WSC samples were selected, giving a total of 32 samples per pool used in the ANTH-WSC determination (Jermyn, 1956).

Definitions of subsequent WSC determination methodology and WSC fraction expressions are as follows. "*ANTH-Total*" is the total WSC determined by anthrone, "*ANTH-HMW*" is the high molecular weight WSC determined by anthrone, "*ANTH-LMW*" is the low molecular weight WSC determined by anthrone, "*NIRS-SSS*" is the soluble sugars and starches determined by near infra-red reflectance spectroscopy using the

calibration developed by Corson *et al.* (1999), "*NIRS-Total*" is the total WSC determined by near infra-red reflectance spectroscopy using the calibration developed by Cosgrove *et al.* (2009), "*NIRS-HMW*" is the high molecular weight WSC determined by near infra-red reflectance spectroscopy using the calibration developed by Cosgrove *et al.* (2009), and "*NIRS-LMW*" is the low molecular weight WSC determined by near infra-red reflectance spectroscopy using the calibration developed by Cosgrove *et al.* (2009).

2.3.4.2 Water-soluble carbohydrate extraction

Low molecular weight (LMW) and high molecular weight (HMW) WSC was extracted via the following protocol**:** 25 – 30 mg of freeze-dried tissue was weighed into 2 mL screw cap tubes (Sarstedt, Nümbrecht, Germany). To each tube, 1 mL 80% ethanol was added and then heated at 65˚C for 30 min, shaking at 1,400 rpm (Thermomixer comfort, Eppendorf, Hamburg, Germany). Tubes were centrifuged at 17,900 rcf for 10 min. Supernatant was collected into 2 mL snap lock tubes (Eppendorf, Hamburg, Germany). The solution was re-extracted with 80% ethanol and the two supernatants combined (total 2 mL). This fraction contained LMW-WSC. One mL of distilled H_2O was then added to the residual solution in the screw cap tubes, heated at 65˚C for 30 minutes with shaking at 1,400 rpm. Tubes were centrifuged at 17,900 rcf for 10 min. Supernatant was collected and re-extracted once more with H_2O using the same procedure. The two aqueous supernatants were combined to give the HMW fraction (2 mL total). The extracted carbohydrate solutions were stored at 4˚C until quantified.

2.3.4.3 Water-soluble carbohydrate quantification using colorimetric anthrone assay

Standards were prepared for the LMW and HMW fractions, using sucrose and inulin, respectively (100 µg mL^{-1}, 75 µg mL^{-1}, 50 µg mL^{-1}, 40 µg mL^{-1}, 30 µg mL^{-1}, 20 µg mL^{-1}, 10 µg mL^{-1} and 0 µg mL^{-1}). The anthrone reagent was made by cooling 30 mL absolute ethanol on ice, then slowly adding 50 mL concentrated H_2SO_4 (final ratio of ethanol:H_2SO_4 of 3:5) while stirring. The mixture was warmed to room temperature and 100 mg of anthrone was added and mixed until dissolved. The required quantity of reagent was prepared daily. For the LMW samples, 12 µL extract was transferred to a Nunc™ 96-well microwell plate (300 µL capacity) and diluted by adding 188 µL H_2O (dilution factor = 16.67). For the HMW samples, the extract was diluted by adding 160 µL H_2O to 40 µL extract (dilution factor = 5). Samples were mixed by pipetting up and down. Forty µL of diluted extract and 40 µL of each standard were transferred in triplicate to another Nunc™ 96-well microwell plate, and 200 µL anthrone reagent was added to each sample in a fume hood. The plate was incubated at 65˚C for 25 min. Finally, the

absorbance of the samples was measured at 620 nm using a VersaMax tunable microplate reader and viewed using SoftMax Pro v4.8.

2.3.4.4 Data analysis

Calibration graphs were made by plotting concentrations of the standard solutions on the *x*-axis against the averaged absorbance on the *y*-axis, in Minitab v 18.1 (Minitab LLC, 2017). The absorbance of the samples (mean of three replicates per sample) were converted into concentration (µg mL^{-1}) using linear regression and then converted into mg g^{-1} dry matter (DM) using **Equation 2.1**. Total WSC was calculated by simply adding the LMW and HMW fractions together for a sample. As mg g^{-1} DM is equivalent to g kg^{-1} DM, the concentration of ANTH-Total, ANTH-LMW and ANTH-HMW could be directly compared to the NIRS-SSS, NIRS-Total, NIRS-LMW and NIRS-HMW (all g kg^{-1} DM). Prior to correlation analysis, data were evaluated for normality using the Shapiro-Wilk Normality Test implemented in R using the function *shapiro.test()* from the basic "*stats*" v 3.6.1 package (Royston, 1995; R Core Team, 2019). Data that did not follow a normal distribution as determined by the Shapiro-Wilk test were transformed based on the Box-Cox Transformation analysis in the R package "*MASS*" v 7.3-51.4 (Venables & Ripley, 2002) using the function *boxcox()*. The optimal data transformation for each phenotypic dataset was determined from the lambda (λ) value corresponding to the maximum log-likelihood (Box & Cox, 1964; Osborne, 2010). Shapiro-Wilk Normality Tests were re-run on transformed data and visual assessment of data spread before and after transformation was evaluated with histograms and Q-Q plots. As a result of normality checking, no transformations were required for the NIRS data but square root transformations were applied to all three ANTH data. ANTH-Total, ANTH-LMW and ANTH-HMW data were then compared with the NIRS WSC data for each sample by Pearson correlation coefficient (r) in DeltaGen v 0.02 (Jahufer & Luo, 2018) and graphed in R using the package "*corrplot*" v 0.84 (Wei & Simko, 2017). All analyses using R statistical software were carried out in version 3.6.1 (R Core Team, 2019).

$$\mathrm{C} = \frac{(Y \times 2 \times 40)}{(W \times E)} \qquad \textbf{Equation 2.1}$$

Where: *C* is the concentration of sample (mg g^{-1}); *Y* is the concentration from the standard linear regression equation (µg mL^{-1}); *2* is the volume (mL) of supernatant extracted for each fraction; *40* is the volume (µL) of diluted extract added to the read plate; *W* is the weight of freeze-dried tissue of the sample (mg); and *E* is the volume

(µL) of extract used for analysis (3 µL for the LMW fraction or 10 µL for the HMW fraction).

2.3.5 Leaf collection strategy to determine number of leaves required for accurate leaf size measurements in white clover populations

2.3.5.1 Plant material

A representative sample of plants (total *n* = 33) from the 900 plants in the main experiment were used in a pilot study to analyse variation of leaf size within a plant and amongst plants from the same population. Plants were selected for the pilot experiment based on inclusion of both small- and large-leaved plants, high and low WSC lines, and generation (Parent, Middle and End generation). Eleven populations and three individuals per population were chosen based on these variables (WNZLL-Parent, WUSLL-Parent, WNZSL-Parent, FNZLL-Parent, FNZSL-Parent, WUSLL-Low-Mid, WUSLL-High-Mid, WNZSL-Low-End, FNZSL-High-End, FNZLL-Low-End and FNZLL-High-End).

2.3.5.2 Leaf area determination

The assessment of leaf size was conducted by estimating leaf area (LA). Montgomery (1911) suggested a formula for LA measurement (**Equation 2.2**). This method is non-destructive, quick and easy. However, the formula is not uniform for all plants as the coefficient *b* differs among species and leaf shapes. There is extensive variety in leaf shape among white clover plants (**Figure S2.2**, Appendix 1) and furthermore, LA varies with age as developing leaves tend to be smaller than older ones (personal observation). To reduce this size variation, for each plant the first leaf from the stolon growing tip with fully opened laminae was measured. Despite these limitations, for the purpose of determining how many leaves per plant are required to give an accurate representation of average leaf size for that plant, this LA formula was considered to be sufficient. For the full LA phenotyping experiment (section 2.3.6) a more accurate analysis of LA, using high resolution digital imagery techniques, was conducted.

$$LA = b \times Ll \times Lw \qquad \textbf{Equation 2.2}$$

Where: *LA* is the leaf area (cm^2), *b* is the leaf shape coefficient, *Ll* is the length of the leaf (cm), and *Lw* is the width of the leaf (cm).

2.3.5.3 Data analysis

The length and width at the widest point of the middle leaflet, and the length of the petiole were measured from five representative leaves (first fully opened leaf on a stolon) per plant (**Figure S2.3**, Appendix 1), giving 495 measurements from 165 leaves (33 plants) and 165 petiole measurements. A maximum of five leaves were sampled from each plant due to time constraints. For Montgomery's (1911) formula (**Equation 2.2**), a co-efficient (*b*) of 0.7 was used for all the leaves to calculate LA. Prior to sample size determination, data were assessed for normality by Shapiro-Wilk test, as described in section 2.3.4.4. As a result of normality checking, data for all four measurements were $\log_{10}$ transformed to meet the assumption of normality. Variance of population and individual means were calculated for each dataset (leaf length, leaf width, leaf area and petiole length) in GenStat v 18 (VSN International, 2015). Data were grouped into two categories (large leaf and small leaf) for each dataset, and the difference in the means was used to infer a numeric value for the size of response that is detectable. Using the mean calculated variance and the inferred values for the response to be detected, the sample size calculation module in GenStat was used to determine the number of leaves and petioles per plant and the number of plants per population, to give an accurate representation of LA and growth form. Simple analysis of variance (ANOVA) tests were performed in RStudio using the function *anova()* from the "*stats*" package to determine if the large leaf (LL) and small leaf (SL) means were significantly different at $\alpha = 0.05$ for each trait. Large leaf and SL means were then back transformed to obtain values on the original cm scale. The difference between the back transformed means was calculated by subtracting the back transformed SL mean from the back transformed LL mean.

2.3.6 Leaf area phenotyping

Having previously identified the optimal sampling methodology for LA assessment (section 2.4.3), a larger phenotyping study was completed to identify LA for each of the five pools and investigate whether there is a relationship between LA and WSC content. Leaf area sampling occurred one week prior to WSC phenotyping (section 2.3.3), which was as close to the WSC leaf harvest as possible. A total of 450 out of 900 plants were sampled across all three blocks (150 plants per block). Four leaves per plant were collected to give an accurate representation of leaf size for each plant. In order to reduce the influence of age and environmental variation within and among plants, the first leaf from the stolon tip with a fully opened laminae was collected. Leaves were plucked at the base of the laminae to reduce petiole inclusion in the area analysis, sealed in plastic

bags, and kept on ice to prevent turgor loss in cells. The undersides of the leaves were blotted dry with tissue paper, glued flat to 1mm graph paper and scanned immediately to a JPEG file. The scanned images were converted to binary in ImageJ (Schindelin *et al.*, 2012), and a global calibration curve was set to convert from pixel distance to actual distance (cm). LA was then estimated for 1,800 leaves using the calibration. The mean from four leaf measurements per plant was used to give one value per plant, and used as a proxy for average plant LA.

2.3.7 Phenotype statistical analyses and correlation investigation

Estimated Marginal means for the NIRS-SSS data (hereafter referred to as WSC) and LA were calculated for each population using a linear mixed model in the R package "*emmeans*" v 1.4.2 that accounted for treatment and spatial variation (**Equation 2.3**).

$$y \sim TrtC + (1 \mid Block) + (1 \mid Block:Column) + (1 \mid Block:Row) \qquad \textbf{Equation 2.3}$$

Where: *y* is the phenotypic trait, and *TrtC* is the interaction between Pool, Generation (e.g., Parent, Mid and End) and *Trt* (e.g., None, high WSC and low WSC).

Leaf area data were transformed by square root so residuals conformed to constant variance and normality (**Figure S2.4** right, Appendix 1). WSC required no transformation as residual assumptions were met (**Figure S2.4** left, Appendix 1). For both traits, spatial variation was adequately explained by block, column and row effects (**Figure S2.5**, Appendix 1) and assumptions of residual independence and constant variance was met (**Figure S2.6**, Appendix 1). Leaf area data were back-transformed to get population-fitted values on the original cm^2 scale. Fitted values for each population were visualised in R using the package "*ggplot2*" v 3.2.1 with the standard error of the mean determined in "*emmeans*". The Parent population was used as the baseline for comparisons between populations within each pool. Significant differences in these comparisons were investigated using "*emmeans*" at $\alpha = 0.05$. The Low-End population was then used as the baseline for the WSC dataset so differences between the Low-End and High-End populations for each pool could be determined.

Two hundred and ninety-seven plants were used for correlation analysis between WSC and LA (mean of four leaves per plant). This dataset was also split into five datasets corresponding to the five pools (WNZLL, WNZSL, WUSLL, FNZLL and FNZSL) within which correlations between the two phenotypic variables were

investigated. Prior to correlation statistical analysis, LA and WSC data were evaluated for normality by Shapiro-Wilk test, as described in section 2.3.4.4. As a result of normality checking, log transformations were applied to the LA data for the WNZSL and FNZSL pools, and a square root transformation was applied to the LA combined pool dataset. No transformations were required for the WSC datasets. Leaf area and WSC were compared by Pearson correlation (r) and scatter plots were graphed using the *ggscatter()* function from the package "*ggpubr*" v 0.2.3 (Kassambara, 2019). Correlations were completed using both transformed and non-transformed data to verify that there was no qualitative change in the results. Models were created for each pool using R statistical software and the *lm()* function from the *"stats*" package, and adjusted coefficient of determination (r^2) values of the lines of best fit were used to initially compare the predictive abilities of LA for WSC. Regression analysis amongst populations was then analysed using linear mixed models that were constructed separately for each pool. Models were constructed with WSC as the dependent variable and the interaction between LA and population was used as the independent variable. The Parent population for each pool was used as the baseline for comparison. Adjusted r^2 was reported for both types of linear models (WSC ~ LA and WSC ~ 0 + Population * LA) to avoid false inflation of r^2 values, which can happen simply by adding extra predictors. Adjusted r^2 is a modified version of r^2 that is adjusted for the number of predictors in the model (**Equation 2.4**). Residuals were checked for normality and assumptions were met; therefore, no data transformation was required.

$$r^2_{adj} = 1 - \left[\frac{(1 - r^2)(n - 1)}{n - k - 1}\right] \qquad \textbf{Equation 2.4}$$

Where: *n* is the number of data points, *k* is the number of variables in the given model, and r^2 is the coefficient of determination.

2.4 Results

2.4.1 Near infra-red reflectance spectroscopy data validation using wet chemistry

Low molecular weight (LMW) and high molecular weight (HMW) water-soluble carbohydrate (WSC) was successfully extracted for all 160 samples. WSC determination using wet chemistry was conducted over two days, with both LMW- and HMW-WSC measured on each of the two days, so four standard curves were calculated and used to determine WSC concentrations (g kg^{-1} dry matter, DM) (**Table S2.2**,

Appendix 1). Two plates for each of LMW- and HMW-WSC were analysed on the first day and ten plates were analysed on the following day (five plates for each LMW- and HMW-WSC). From the 160 samples, LMW-WSC was the most abundant WSC fraction, with a mean accumulation of 52.2 g kg^{-1} DM and a range of 8.8 g kg^{-1} DM to 101.7 g kg^{-1} DM. HMW-WSC on the other hand was negligible, with a mean accumulation of 2.7 g kg^{-1} DM and a range of 0.0 g kg^{-1} DM to 6.7 g kg^{-1} DM (**Table S2.3**, Appendix 1).

Data were checked for normality and constant variance prior to correlation analysis. All near infra-red reflectance spectroscopy (NIRS) data conformed but all anthrone (ANTH) data required square root transformation. Results from the correlation analysis showed that Pearson correlation coefficients (r) for transformed and non-transformed data were identical (data not presented), therefore the non-transformed data were used subsequently. When the ANTH-LMW and ANTH-HMW fractions are combined, a perfect correlation ($r = 1.00$, $p < 0.01$) was observed between ANTH-Total and ANTH-LMW, while a moderate correlation was observed between ANTH-Total and ANTH-HMW ($r = 0.51$, $p < 0.01$) (**Table 2.1**). When the ANTH-Total was compared to the NIRS WSC measures, a moderate correlation was observed with both NIRS-SSS and NIRS-Total ($r = 0.56$, $p < 0.01$ and $r = 0.48$, $p < 0.01$, respectively) (**Table 2.1**). There was no significant correlation ($\alpha = 0.05$) between the ANTH-HMW and any of the NIRS variables, but there was a moderate correlation between ANTH-LMW and NIRS-LMW ($r = 0.62$, $p < 0.01$). A summary of the Pearson correlation coefficients and *p*-values of all variables is depicted in **Figure 2.3**.

Table 2.1 Correlation statistics amongst total water-soluble carbohydrates (total WSC), low molecular weight (LMW) and high molecular weight (HMW) WSC measured by anthrone (ANTH) determination; and soluble sugars and starches (SSS), total WSC, HMW and LMW WSC determined by near infra-red reflectance spectroscopy (NIRS) in white clover leaves. A subset of samples was taken from across the five pools (32 samples per pool, 160 samples in total) for comparisons between the two WSC determination methods.

Response variable	Predictor variable	Pearson correlation coefficient (r)	Coefficient of determination (r^2)	*p*-value
ANTH-Total	ANTH-LMW	1.00	1.00	<0.00001*
NIRS-SSS	NIRS-Total	0.96	0.92	<0.00001*
NIRS-SSS	ANTH-Total	0.56	0.31	<0.00001*
NIRS-Total	ANTH-Total	0.48	0.23	<0.00001*
NIRS-LMW	ANTH-LMW	0.62	0.38	<0.00001*
NIRS-HMW	ANTH-HMW	0.09	0.008	0.24
NIRS-LMW	ANTH-HMW	0.33	0.11	<0.00001*
NIRS-LMW	ANTH-HMW	0.43	0.18	<0.00001*

Note: Total = total WSC, NIRS- = determined by near infra-red reflectance spectroscopy, ANTH- = determined by the anthrone method, LMW = low molecular weight WSC, and HMW = high molecular weight WSC.

* $p < 0.05$ significance threshold

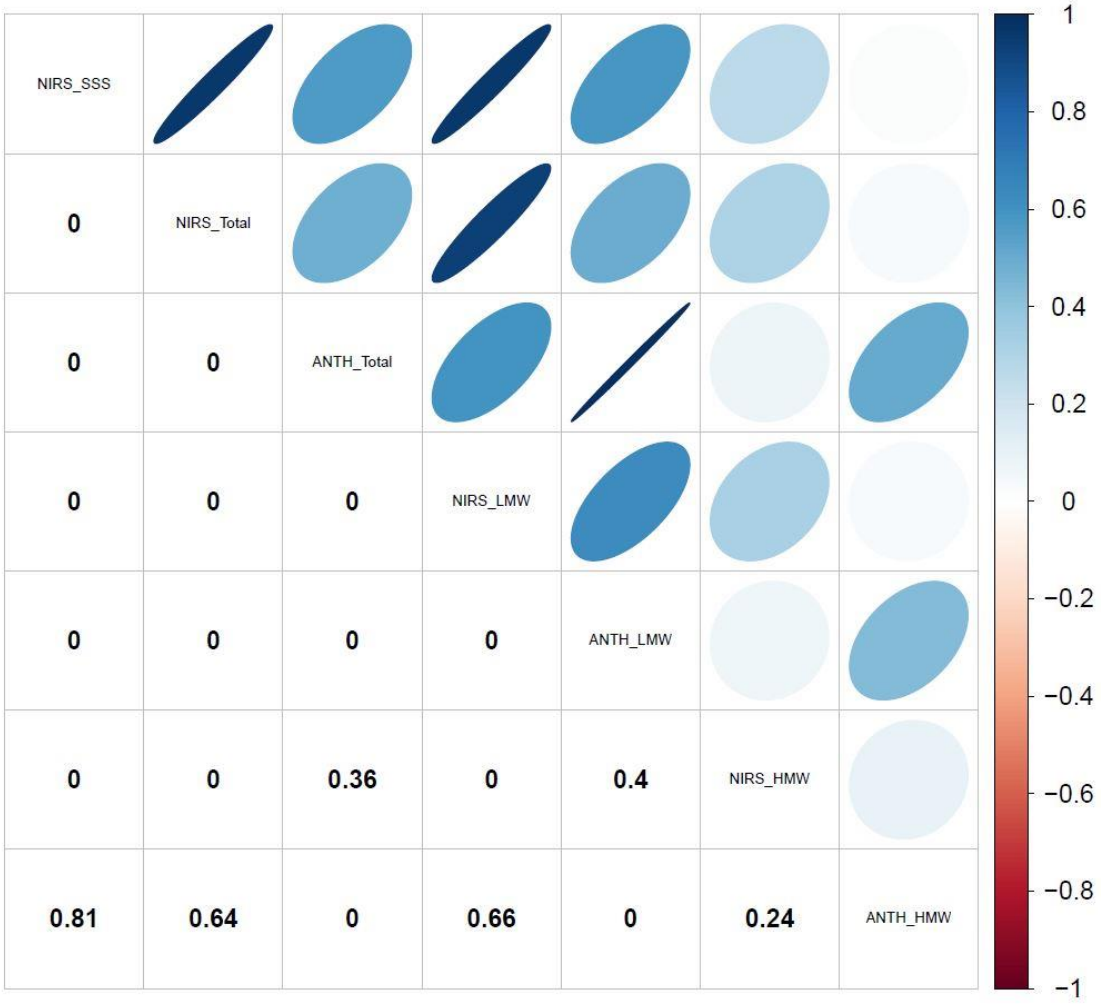


Figure 2.3 Pearson correlation matrix for NIRS and ANTH water-soluble carbohydrate (WSC) determination based on data from 160 white clover leaf samples. Ellipses show the strength of the correlation of two variables by the width (thinner ellipses indicate very strong correlations and wider, more circular shapes indicate very weak correlations) and colour (dark blue represents positive correlation of 1 and dark red represents a negative correlation of -1). Numerical values in the bottom left triangle are *p*-values associated with the correlation. Variable names are located along the diagonal. *Note*: SSS = soluble sugars and starches, NIRS = near infra-red reflectance spectroscopy, ANTH = WSC determined by the anthrone method, Total = total WSC, LMW = low molecular weight WSC, and HMW = high molecular weight WSC.
p-value of 0 = <0.00001.

2.4.2 Water-soluble carbohydrate phenotyping using near infra-red reflectance spectroscopy

Water-soluble carbohydrate (WSC) content, in g kg^{-1} dry matter (g kg^{-1} DM), was measured by NIRS for 600 plants from two replicate blocks, giving 120 measurements per pool. The NIRS validation results showed the best NIRS predictor of ANTH-Total was NIRS-SSS (see section 2.4.1), therefore the NIRS-SSS dataset was used for WSC determination in this experiment and is subsequently referred to as WSC, in this section and section 2.4.5. Importantly, the NIRS-SSS calibration was also used by Widdup *et al.* (2010), therefore direct comparison could be made between the results presented here and the results from Widdup *et al.* (2010). Population fitted values (adjustment for treatment, block, row and column effects) are presented with standard error (**Figure 2.4**, top) and with 95% confidence intervals for WSC (**Table S2.4**, Appendix 1). Within each pool, population fitted values were compared to the Parent population to determine if WSC increased or decreased with selection from the baseline (**Table 2.2**). There was an overall trend for WSC content to increase with selection for higher WSC levels and, conversely, decrease with selection for lower WSC levels. Across all pools there were 12 significant ($\alpha = 0.05$) changes in WSC compared to the Parent population. The WNZLL and FNZSL pools showed the same pattern of WSC change. There was an increase in WSC content in the High-Mid and High-End populations (+36.4 g kg^{-1} DM, $p < 0.01$ and +57.5 g kg^{-1} DM, $p < 0.01$ for WNZLL; and +57.8 g kg^{-1} DM, $p < 0.01$ and +71.6 g kg^{-1} DM, $p < 0.01$ for FNZSL), and a decrease in WSC content in the Low-End population (-42.2 g kg^{-1} DM, $p < 0.01$ for WNZLL and -27.1 g kg^{-1} DM, $p = 0.012$ for FNZSL), but no significant change from the Parent population to the Low-Mid population. The WNZSL and FNZLL pools also showed a pattern of a significant increase in WSC from Parent to High-Mid (+48.7 g kg^{-1} DM, $p < 0.01$ for WNZSL and +25.2 g kg^{-1} DM, $p = 0.029$ for FNZLL) and from Parent to High-End (+69.6 g kg^{-1} DM, $p < 0.01$ for WNZSL and +41.4 g kg^{-1} DM, $p < 0.01$ for FNZLL), but non-significant changes in WSC from Parent to Low-Mid (+0.443 g kg^{-1} DM, $p = 1.0$ for WNZSL and -22.8 g kg^{-1} DM $p = 0.07$ for FNZLL) and from Parent to Low-End (-0.098 g kg^{-1} DM, $p = 1.0$ for WNZSL and -13.4 g kg^{-1} DM, $p = 0.82$ for FNZLL). WUSLL on the other hand had a significant decrease in WSC in the low WSC populations (-34.2 g kg^{-1} DM, $p < 0.01$ for Low-Mid and -45.1 g kg^{-1} DM, $p < 0.01$ for Low-End) but there was a non-significant increase in the high WSC populations (+13.4 g kg^{-1} DM, $p = 0.82$ for High-Mid and +23.6 g kg^{-1} DM, $p = 0.054$ for High-End). When averaged across all pools, there were significant differences ($p < 0.01$) for all comparisons, with the low WSC populations lower in WSC than the Parent, and the high WSC populations with higher WSC content than the Parent (**Table 2.2**).

Although high and low WSC populations did not always differ significantly from the Parent population in most of the pools, there were consistently distinct and significant ($p < 0.05$) differences when comparing just the Low-End and High-End populations (mean difference of 78.3 g kg^{-1} DM) for all five pools when the Low-End populations were used as the baseline (**Table 2.2**). These results confirm that breeding for divergent WSC in these pools was successful and resulted in a mean 76.9% increase in WSC from the Low-End to High-End populations (**Table S2.4**, Appendix 1).

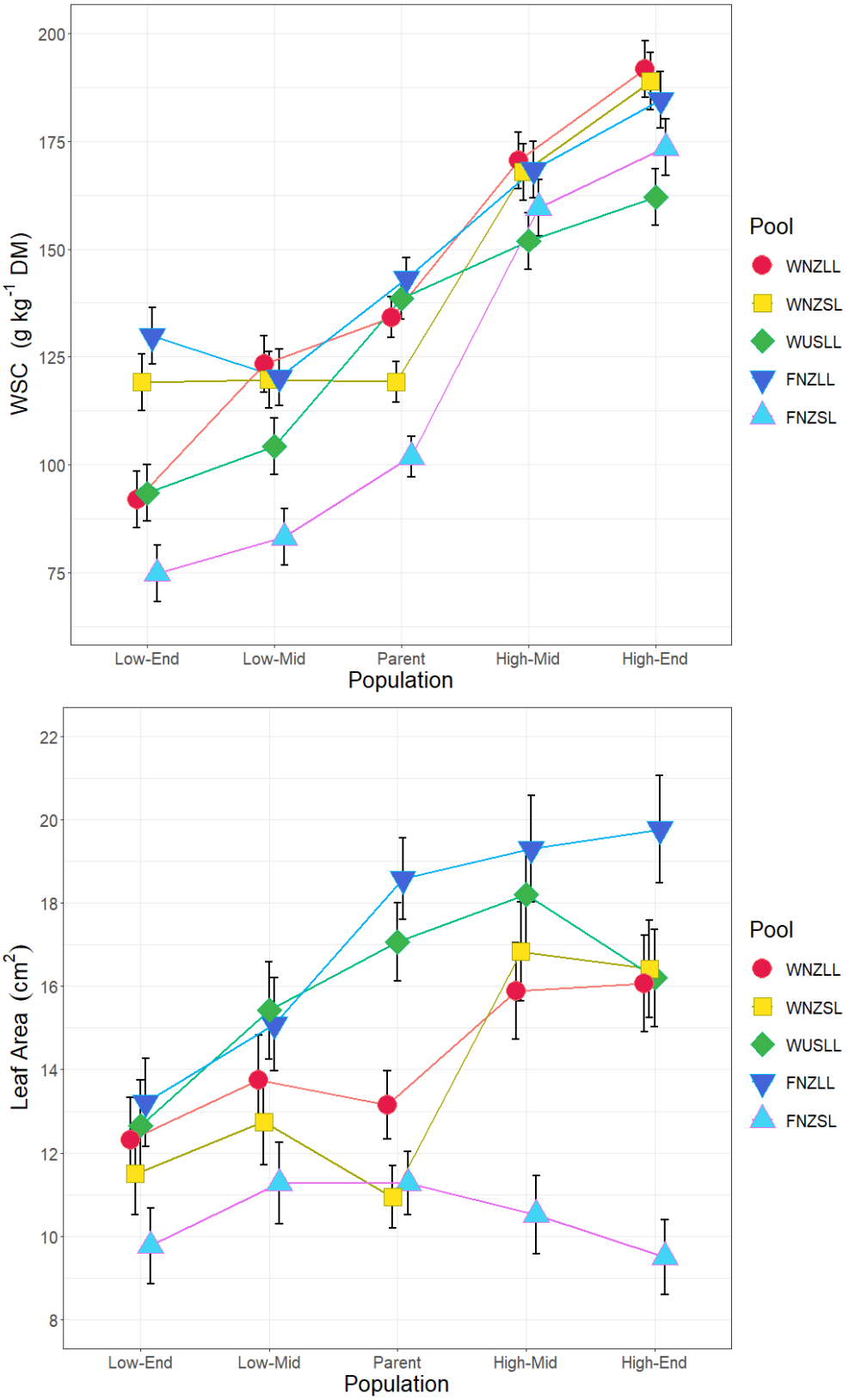


Figure 2.4 Population fitted values and standard error of water-soluble carbohydrate (WSC; top) and leaf area (bottom). Populations are grouped by pool as indicated by colour and symbol combinations. Data are ordered on the *x*-axis by generation where Low = low WSC, High = high WSC, End = End generation, Mid = Middle generation, Parent = Parent generation. Population means for WSC are based on n = 20 – 40 and population means for leaf area are based on n = 13 – 30.

Table 2.2 Estimated phenotype means for each divergently-selected population compared to the parental mean after adjusting for treatment, block, row and column effects. Leaf area (cm^2) and water-soluble carbohydrate (WSC) (grams per kilogram dry matter, g kg^{-1} DM) population fitted values compared to the Parent population are presented in the "D*ifference*" column. For WSC, the High-End population fitted values are compared to the Low-End population fitted values for each pool and are also presented in the "*Difference*" column. Standard error and *p*-values are shown for each comparison. *p*-values are adjusted for multiple comparisons calculated using the R package "*emmeans*".

	WSC (g kg^{-1} DM)			Leaf area (cm^2)		
Comparison	**Difference**	**SE**	***p*-value**	**Difference**	**SE**	***p*-value**
WNZLL-Low-End - WNZLL-Parent	-42.2	7.77	<0.01**	-0.83	0.16	1.00
WNZLL-Low-Mid - WNZLL-Parent	-10.9	7.78	0.97	0.6	0.16	1.00
WNZLL-High-Mid - WNZLL-Parent	36.4	7.77	<0.01**	2.74	0.16	0.37
WNZLL-High-End - WNZLL-Parent	57.5	7.77	<0.01**	2.92	0.16	0.27
WNZSL-Low-End - WNZSL-Parent	-0.098	7.79	1.00	0.56	0.16	1.00
WNZSL-Low-Mid - WNZSL-Parent	0.44	7.78	1.00	1.81	0.16	0.85
WNZSL-High-Mid - WNZSL-Parent	48.7	7.76	<0.01**	5.89	0.16	<0.01**
WNZSL-High-End - WNZSL-Parent	69.6	7.78	<0.01**	5.48	0.16	<0.01**
WUSLL-Low-End - WUSLL-Parent	-45.1	7.8	<0.01**	-4.41	0.17	0.02*
WUSLL-Low-Mid - WUSLL-Parent	-34.2	7.77	<0.01**	-1.65	0.16	0.99
WUSLL-High-Mid - WUSLL-Parent	13.4	7.76	0.82	1.14	0.16	1.00
WUSLL-High-End - WUSLL-Parent	23.6	7.76	0.054	-0.87	0.16	1.00
FNZLL-Low-End - FNZLL-Parent	-13.4	7.77	0.82	-5.37	0.16	<0.01**
FNZLL-Low-Mid - FNZLL-Parent	-22.8	7.79	0.07	-3.49	0.16	0.15
FNZLL-High-Mid - FNZLL-Parent	25.2	7.78	0.029*	0.72	0.16	1.00
FNZLL-High-End - FNZLL-Parent	41.4	7.79	<0.01**	1.19	0.16	1.00
FNZSL-Low-End - FNZSL-Parent	-27.1	7.76	0.012*	-1.5	0.16	0.94
FNZSL-Low-Mid - FNZSL-Parent	-18.7	7.77	0.29	0.00	0.16	1.00
FNZSL-High-Mid - FNZSL-Parent	57.8	7.78	<0.01**	-0.75	0.16	1.00
FNZSL-High-End - FNZSL-Parent	71.6	7.77	<0.01**	-1.77	0.16	0.80
Low-End - Parent	-25.6	3.47	<0.01**	-2.18	0.07	<0.01**
Low-Mid - Parent	-17.2	3.49	<0.01**	-0.43	0.07	1.00
High-Mid - Parent	36.7	3.46	<0.01**	1.95	0.07	0.01**
High-End - Parent	52.7	3.48	<0.01**	1.35	0.07	0.24
WNZLL-High-End – WNZLL-Low-End	99.6	8.99	<0.01**			
WNZSL-High-End – WNZSL-Low-End	69.7	8.98	<0.01**			
WUSLL-High-End – WUSLL-Low-End	68.7	8.98	<0.01**			
FNZLL-High-End – FNZLL-Low-End	54.8	8.98	<0.01**			
FNZSL-High-End – FNZSL-Low-End	98.7	8.98	<0.01**			

Note: Low = low WSC, High = high WSC, Parent = Parent generation, Mid = Middle generation, End = End generation, W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL = large leaf, SL = small leaf and SE = standard error.
Significance codes: ** ≤ 0.01, * = 0.01 – 0.05, no symbol ≥ 0.05 at α = 0.05.

2.4.3 Leaf collection strategy to determine number of leaves to sample

A total of 660 measurements were collected from 165 plants (five clonal reps of 33 plants) for four traits: leaf width (Lw), leaf length (Ll), leaf area (LA) and petiole length (Pl). Each trait dataset was separated into leaf size class categories (large leaf, LL and small leaf, SL) based on the pool the plants originated from. Means and differences between the LL and SL means were calculated for each trait. For this, each trait dataset

was log_{10} transformed to meet the assumption of normality but after means and differences between means were determined, these were back-transformed to obtain mean values on the original cm scale. For all traits assessed, analysis of variance (ANOVA) tests determined that mean values were significantly ($\alpha = 0.05$) different and LL plants had higher means than SL plants (**Table S2.5**, Appendix 1). The difference between LL and SL means was 0.4 cm for both for Lw and Ll, while LA and Pl had mean differences of 1.1 cm^2 and 1.4 cm (**Table S2.5**, Appendix 1), respectively, between LL and SL plants. Ideally, the statistical power for any study is considered to be 0.8 or higher (Hintze, 2008). This means that the study has a high chance of detecting a difference between accessions, if it exists. Using the mean differences and the average variance (log_{10} transformed) for both population and individual means (**Table S2.6**, Appendix 1 and **Table S2.7**, Appendix 1), the sample size required to achieve a power (or probability of detection) of 0.9 and 0.8, respectively, was calculated. The minimum number of individuals per population required to sample for LA to meet a minimum power of 0.8 was 9 – 15 individuals (**Figure S2.7**, Appendix 1) and four leaves per plant (**Figure S2.8**, Appendix 1). The variance for Pl was high (0.022 on the log scale) and to meet a minimum power of 0.8 the minimum number of individuals and leaves per plant were 21 and 7, respectively.

Based on these outcomes, 15 plants per population and four leaves per plant were recorded for LA determination in the main assessment of LA and WSC described in section 2.4.4. Due to potential spatial variability across the trial these plants were chosen from every second column in the randomised Latin square design (**Table S2.1**, Appendix 1). Pl measurements were not recorded due to time constraints as the power calculation indicated that almost all of the plants in each population and almost double the number of leaves per plant would need to be sampled.

2.4.4 Leaf area phenotyping

Leaf area (cm^2) was measured on 447 plants, with four leaves per plant sampled, giving 1,788 leaf area measurements in total and 360 leaf area measurements per pool (except WUSLL in which three plants were diseased and excluded from the analysis, resulting in 357 leaf measurements in that pool). Population fitted values are presented with standard error (**Figure 2.4**, bottom) and with 95% confidence intervals for LA (**Table S2.8**, Appendix 1). There were apparent trends for LA increasing or decreasing following selection for WSC, however there were only four changes in LA, in comparison to the starting point (Parent population), that were statistically significant ($\alpha = 0.05$) (**Table 2.2**).

There were no significant changes in LA for the WNZLL and FNZSL pools. The FNZLL and WUSLL pools showed a significant decrease in LA from the Parent to Low-End population (-5.4 cm^2, $p < 0.01$ and -4.4 cm^2, $p < 0.01$ for FNZLL-Low-End and WUSLL-Low-End, respectively). Whereas, the WNZSL pool had a significant increase in LA for both the High-Mid and High-End populations (+5.9 cm^2, $p < 0.01$ and +5.5 cm^2, $p < 0.01$, respectively).

2.4.5 Correlation and regression analysis between water-soluble carbohydrate and leaf area

A correlation analysis between WSC and LA was performed to investigate the relationship between the two traits. No transformations were required for the WSC datasets but log transformations were applied to the LA data for the WNZSL, FNZSL pools and a square root transformation was applied to the LA combined pool dataset (**Table S2.9**, Appendix 1). Correlation analysis was used to measure the strength of relationship between WSC and LA for each pool and for the combined pool dataset (**Figure 2.5**). Results from the correlation analysis showed that there was very little difference in the r^2 (coefficient of determination) values for transformed and non-transformed data, therefore the non-transformed data were used for subsequent analyses. Weak to moderate positive linear relationships between WSC and LA were observed in all pools, with Pearson's correlation coefficients (r) of: 0.38, 0.58, 0.52, 0.33, 0.046 and 0.38 for the WNZLL, WNZSL, WUSLL, FNZLL, FNZSL and combined datasets, respectively. All relationships were significant at an alpha of 0.05, except for the FNZSL pool ($p = 0.73$). Correlation analysis r^2 values (**Figure 2.5**) showed that between 1.5 – 33% of the observed WSC phenotypic variation was accounted for by LA in each of the pools (**Figure 2.5**). These low r^2 values demonstrate that the basic linear model (where WSC was the dependent variable and LA was the independent variable) provided a poor to average fit to the data. Low r^2 values such as these mean that WSC cannot be accurately predicted from LA with the given datasets using this model. The residuals were checked for each pool and the scatterplots of predicted values and residuals indicated that the data for each pool met the assumptions of homogeneity of variance and linearity, and the residuals were approximately normally distributed (see **Figure S2.9** in Appendix 1 for an example of residual plots using the WNZLL pool). Low r^2 values can occur when the linear model is missing an important predictor or interaction term. Therefore, the data were split into populations for each pool and regression analysis was used to test if LA was significantly predictive of WSC for each pool at the population level.

Linear models were constructed with WSC as the dependent variable and the interaction between LA and population was used as the independent variable. Including populations in the linear model increased the adjusted r^2 values up to 94 – 97% (**Table S2.10**, Appendix 1). This indicated that splitting the data into populations for each pool and analysing separately provided a better fit to the data than combining all data points for each pool. The majority of slope coefficients for each pool (**Figure 2.6** and **Table S2.10**, Appendix 1) were very gradual and not significant ($\alpha = 0.05$), with the exception of WUSLL-Parent (slope = 0.37, $p = 0.004$), FNZLL-Low-End (slope = 0.44, $p = 0.037$) and FNZLL-High-End (slope = 0.21, $p = 0.04$). Because all populations, except the WUSLL-Parent population, had insignificant *p*-values ($\alpha = 0.05$) for both the intercept and slope, we were unable to reject the null hypothesis, allowing the conclusion that there was no relationship between WSC and LA for all but one population (WUSLL-Parent). Flat slopes are to be expected if WSC content is similar for all individuals within a population, even if leaf area varies slightly. Each population was colour- and shape-coded so data points (individuals) could be visualised on the plot. The high WSC populations sat higher on the y-axis (WSC) than the low WSC populations, but were not located in the top-right quadrant of the graph which would have indicated a strong LA effect on WSC content. Likewise, the low WSC populations were not located in the bottom-left quadrant of the graph, (i.e., the data did not follow a forty-five-degree line).

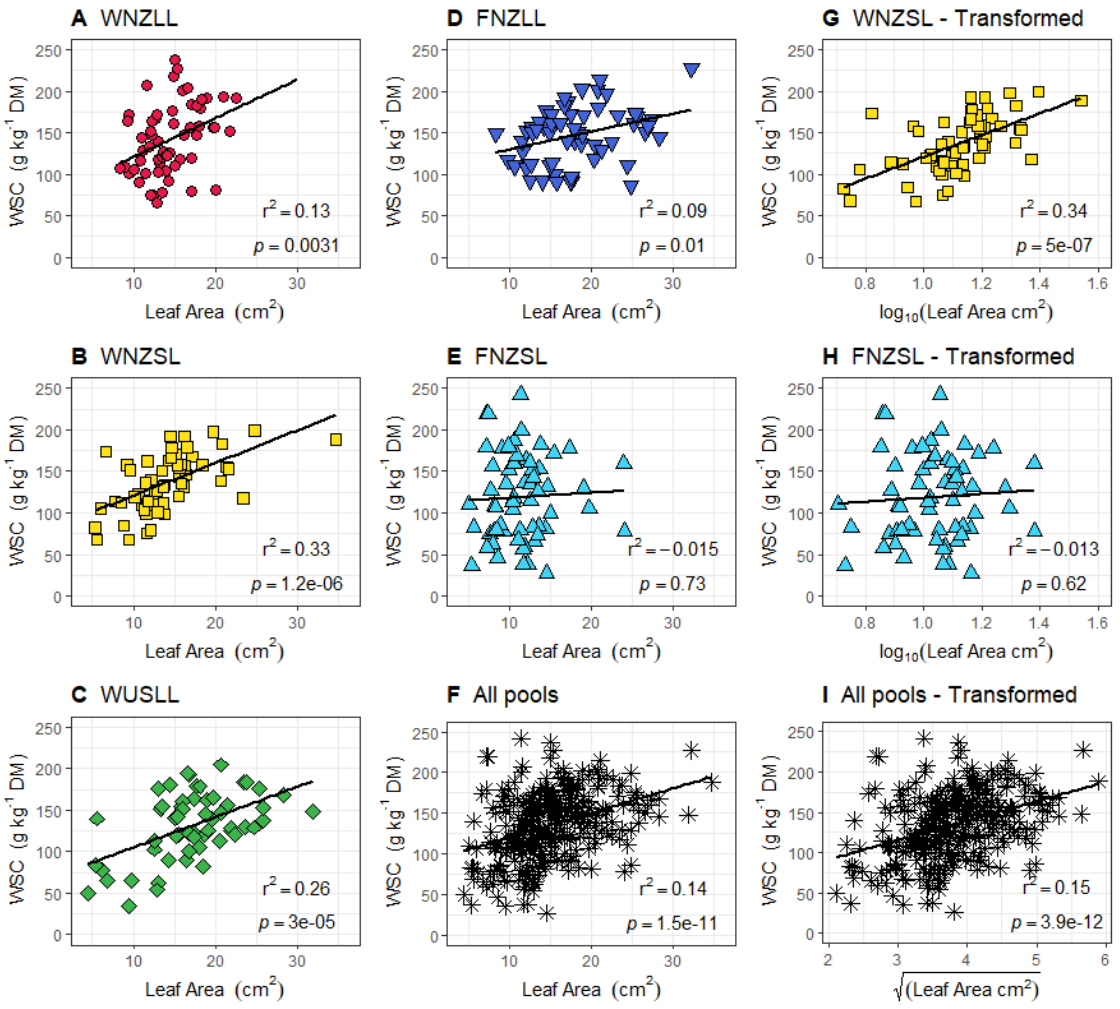


Figure 2.5 Correlation scatterplots between water-soluble carbohydrate (WSC) and leaf area for each pool and combined datasets. Raw data was used in plots A to F and log and square root transformed leaf area is presented in plots G to I. **A**) WSC and leaf area correlation for WNZLL pool, *n* = 60. **B**) WSC and leaf area correlation for WNZSL pool, *n* = 60. **C**) WSC and leaf area correlation for WUSLL pool, *n* = 57. **D**) WSC and leaf area correlation for FNZLL pool, *n* = 60. **E**) WSC and leaf area correlation for FNZSL pool, *n* = 60. **F**) WSC and leaf area for all data, *n* = 297. **G**) WSC and $\log_{10}$ leaf area correlation for WNZSL pool, *n* = 60. **H**) WSC and $\log_{10}$ leaf area correlation for FNZSL pool, *n* = 60. **I**) WSC and square root leaf area for all data, *n* = 297. Adjusted r^2 values and *p*-values for lines of best fit are shown.

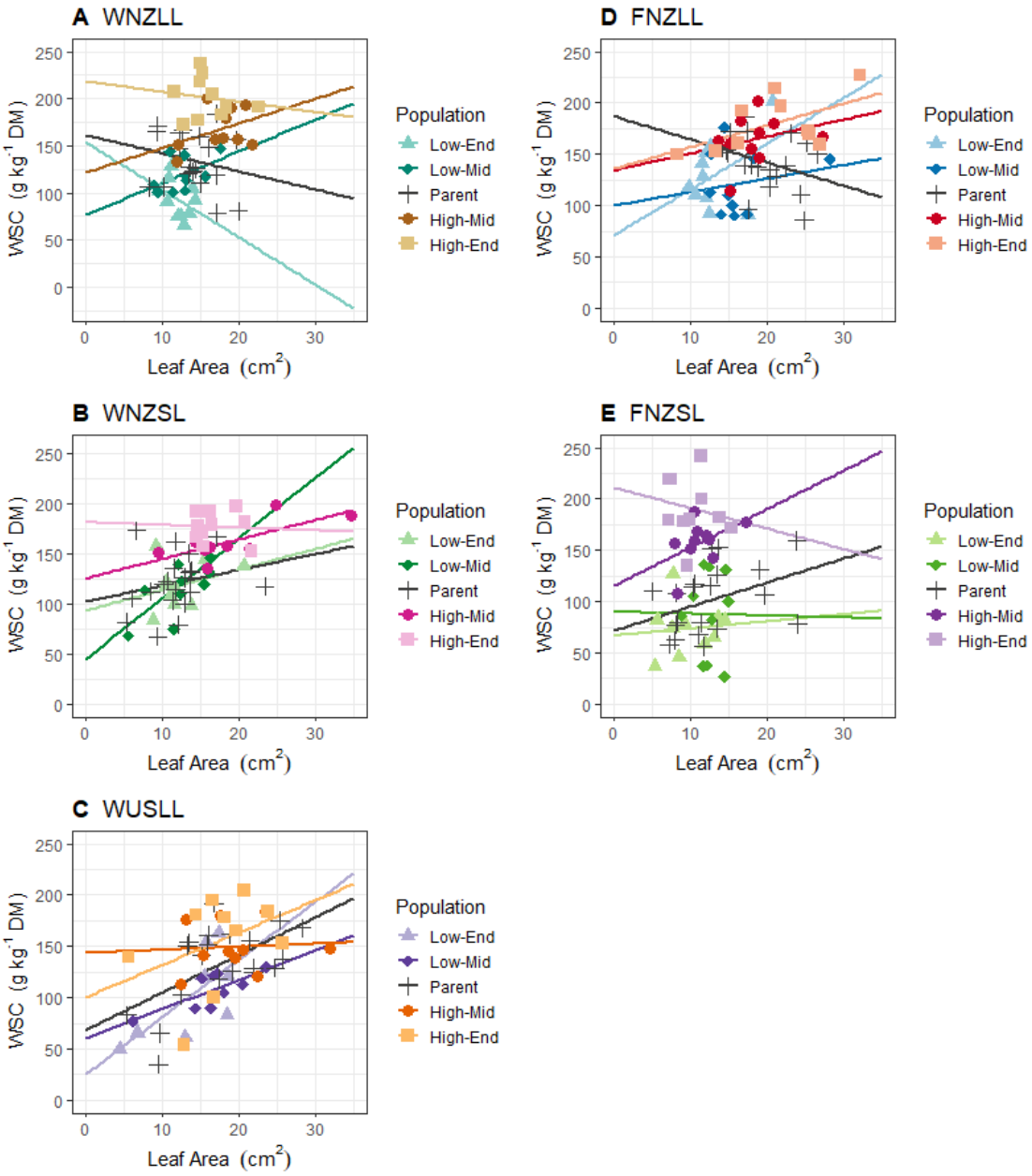


Figure 2.6 Regression scatterplots for water-soluble carbohydrate (WSC) and leaf area (LA) broken down to populations for each pool. **A**) WNZLL pool, *n* = 60, **B**) WNZSL pool, *n* = 60, **C**) WUSLL pool, *n* = 57, **D**) FNZLL pool, *n* = 60, **E**) FNZSL pool, *n* = 60. Line of best fit equations for each population can be found in **Table S2.10**, Appendix 1. Overall adjusted r^2 values and *p*-values for lines of best fit for each pool are shown in **Figure 2.5**.

2.5 Discussion

A significant portion of this thesis is focused on genetic analysis of white clover populations differing in levels of foliar-expressed water-soluble carbohydrate (WSC), with the goal of identifying genomic regions and genes associated with WSC

accumulation. The resource chosen for the analysis was a set of five breeding pools in which divergent selection for WSC levels had been historically undertaken by breeders. It was therefore critical to confirm that the divergently selected populations (Middle and End generations) within the pools had significantly different WSC levels in leaves. Furthermore, the efficacy of near infra-red reflectance spectroscopy (NIRS) for estimating WSC, as the chosen method for measuring WSC in the main phenotyping trial, needed to be evaluated. Finally, the influence of leaf size as a potential confounding factor in changing WSC levels in leaves was investigated. Leaf area (LA) was assessed to determine if any observed changes in WSC levels were linked to changing leaf size or were independent of that factor.

For all five pools of populations, comparison of the two divergent end point generations showed that breeding for divergent foliar WSC was successful. Furthermore, once population was accounted for, there was no strong relationship observed between WSC and LA. These results suggest that divergent WSC can be produced in under six generations. Below, the implications of confounding factors in experimental design, the relationship between NIRS and anthrone (ANTH) data, and the relationships between LA and WSC in white clover are discussed in detail.

2.5.1 Controlling confounding factors for phenotype validation

Time of day

Because WSC levels are tightly linked to starch content (Ruckle *et al.*, 2018) it is important to know how both are affected over a photoperiod in order to determine the best time of day to harvest to ensure consistency between harvests on multiple days. Because starch accumulation is dependent on photosynthesis, accumulation is highly dependent on environmental factors such as temperature, day length, light intensity and harvest time (Boller & Nösberger, 1983; Ruckle *et al.*, 2017). Starch accumulates in leaves steadily throughout the day to its maximum at the end of the day and it degrades in a near-linear manner at night so that it is just about exhausted at the end of the night (EN) (Stitt & Zeeman, 2012; Ruckle *et al.*, 2018). WSC pools replenish more rapidly (peaking mid-day and linearly decreasing to the lowest level at EN) compared to starch (Gibon *et al.*, 2006). The harvesting of leaf tissue in the current study was done early- to mid-morning (8am – 10am). At this time photosynthesis is initiating, and a rapid rise in WSC levels has not commenced. Starch would also not have accumulated to a high level. The harvesting of leaves for WSC evaluation was undertaken at this time to minimise the influence of diurnal variation described above, and also to emulate on-farm

practice, whereby cows would be returning to pastures after morning milking and consuming forage. Furthermore, this was also the time of day used by the breeders to sample for the purposes of selection, therefore sampling at the same timepoint would most likely capture any differences in WSC content due to the breeding process. Because the WSC measures were made early in the day they do not reflect the maximum levels of foliar WSC, which would peak in the afternoon of sunny days (Boller & Nösberger, 1983; Ruckle *et al.*, 2017).

Age of leaves

Carbohydrates, mainly sucrose, are the major constituent in phloem sap. Translocation of phloem sap occurs from sources (areas of supply) to sinks (areas of metabolism and storage). Sources are typically mature leaves that are capable of producing excess products of photosynthesis, whereas sinks are organs that are not photosynthetic or cannot produce enough photosynthate to support their own growth or storage needs, for example developing leaves. Sucrose accumulates in leaves where there is a sink demand, therefore the age of the leaves sampled can affect the WSC levels observed. There are no existing data on white clover WSC variation due to leaf maturity but there are such data for starch (Ruckle *et al.*, 2018). Glucose and sucrose production are tightly linked to starch levels, therefore WSC levels in the mature leaves may co-vary with starch. Ruckle *et al.* (2018) assessed areas of starch deposition in white clover by staining whole plants in iodine. As expected, starch content was found to be higher in adult tissue than in young juvenile tissue. Although the starch deposition pattern at the leaf level did not affect the total starch content of the plant (Ruckle *et al.*, 2018), it highlighted the importance, when taking multiple leaf measurements across a number of plants, of sampling leaves of the same age to obtain consistent information. Therefore, to account for the differences in source/sink status, leaves of approximately the same age (fully opened and close to the stolon tip), and therefore source status, were used for the analysis in this study.

2.5.2 Total water-soluble carbohydrate determined from a perennial ryegrass near infra-red reflectance spectroscopy calibration is inaccurate in white clover

All available NIRS WSC calibrations were used to ensure the most accurate analysis was applied in the WSC validation experiment. Validation of a subset of soluble sugars and starch (SSS) and WSC data obtained via NIRS from white clover leaves was achieved using the anthrone (ANTH) wet chemistry method for quantification of WSC. Moderate correlations were found between the ANTH and NIRS data for total WSC

(ANTH-Total and NIRS-Total r of 0.56, $p < 0.05$), and SSS (ANTH-Total and NIRS-SSS r value of 0.48, $p < 0.05$) (**Table 2.1**). These results support a conclusion that the NIRS data obtained using SSS and total WSC calibrations are only moderately accurate measures of WSC in white clover leaf material. In previous studies, NIRS has been found to predict WSC levels in forages (e.g., grasses, clover, timothy grass and alfalfa) and other species such as wheat, with high accuracy (Alomar *et al.*, 2009; Nie *et al.*, 2009; Shetty *et al.*, 2012; Piaskowski, Brown & Campbell, 2016; Yang *et al.*, 2017). For example, Corson *et al.* (1999) found an r^2 value of 0.87 when comparing SSS and wet chemistry assessments of 358 samples encompassing a range of feed types. However very few studies have assessed NIRS- and ANTH-WSC in pure white clover samples. Inostroza *et al.* (2017) compared NIRS and ANTH-determined WSC for white clover stolon samples and reported a high r^2 of 0.85, which is much larger than the r^2 values estimated here (r^2 = 0.23 – 0.31, **Table 2.1**). Importantly, the NIRS calibration used by Inostroza *et al.* (2017) was specifically developed for white clover stolons. The dissimilarities may therefore be attributed to properties of the samples used to generate the NIRS calibrations, which is discussed in the following paragraphs.

Accuracy of NIRS estimation of forage composition is dependent on the data used to calibrate the instrument, with the best predictions achieved when the calibration is created for a single species or closely related species with similar chemical compositions. For example, a single NIRS calibration was created using four species in the Fabaceae (alfalfa [*Medicago sativa* L.], birdsfoot trefoil [*Lotus corniculatus* L.], red clover [*Trifolium pratense* L.] and white clover) and showed high accuracy (r^2 values greater than 0.97) for nutritive traits in all four species when compared to wet chemistry (Marten *et al.*, 1984). However, evaluating several species using the same calibration requires the inclusion of more wavelengths to develop the calibration equations. This is because there is a greater diversity of chemical structure in the sampling pool, which requires more complex statistics to generate the calibration equations for each of the traits (Marten *et al.*, 1984). This most likely explains why the NIRS-determined low molecular weight (LMW) and high molecular weight (HMW) fractions were inaccurate predictors of the equivalent ANTH-determined LMW and HMW fractions. The NIRS calibration for total, LMW- and HMW-WSC used in the present study was created using perennial ryegrass samples (Cosgrove *et al.*, 2009). Ryegrass contains a large concentration of HMW-WSC (principally fructan), but a relatively low amount of LMW-WSC (Turner *et al.*, 2006; Easton *et al.*, 2009), with HMW-WSC typically comprising 30 – 40% of total WSC in leaves (Turner *et al.*, 2006; Easton *et al.*, 2009; Rasmussen *et al.*, 2009). By contrast, white clover contains a large amount of LMW-WSC and a low

amount of HMW-WSC (**Table S2.3**, Appendix 1), with total WSC consisting of mainly LMW-WSC and low reports of HMW-WSC (Ruckle *et al.*, 2018). Furthermore, addition of the ANTH-HMW fraction to ANTH-Total reduced the correlation for ANTH-Total to NIRS-Total (r = 0.48), as the ANTH-LMW on its own showed a slightly better correlation with NIRS-Total (r = 0.50). Therefore, the differences in the chemical structure of WSC between ryegrass and white clover is likely to have caused the inaccuracy of the NIRS-HMW fraction. The ryegrass WSC calibration was used in part, to identify the most accurate measure of WSC content in white clover, but also as a useful comparator for SSS.

The relationship between ANTH-Total and NIRS-SSS had a slightly higher r^2 value of 0.31. This is in spite of the fact that the ANTH method used here only determines WSC content and does not include starch, which is also captured by SSS. Even though starch levels were expected to be low at the diurnal timepoint at which sampling was conducted (see section 2.5.1), the presence of starch may have negatively influenced the correlation between ANTH-Total and NIRS-SSS. Alternatively, the differences may be attributed to the fact that the NIRS-SSS calibration was built using mixed pasture samples, not pure white clover. This calibration was constructed using pasture, but specific species were not specified (Corson *et al.*, 1999). In a typical NZ pasture system, a higher grass to clover content is common, as clover content is normally less than 20% in the sward (Chapman, Parsons & Schwinning, 1996). Corson *et al.* (1999) suggest that incongruencies can arise because of a different feed type or if the sample composition has a composition different from the material on which the database was developed. Therefore the moderate correlation between SSS-NIRS and ANTH-Total may be in part attributed to the different sample composition in the calibration construction (grass vs clover WSC content, if these species were used in the initial calibration) as ryegrass typically has a higher overall WSC compared to white clover (up to 300 g kg^{-1} dry matter, DM) (Cosgrove *et al.*, 2009; Easton *et al.*, 2009; Rasmussen *et al.*, 2009) and compared to other legumes (Dewhurst *et al.*, 2003; Marshall *et al.*, 2004; Küchenmeister *et al.*, 2013). Although there was only a moderate predictive ability of NIRS for WSC content determination presented here, the relationship was positive and significant, indicating a proximate estimate of 'true' WSC content can be obtained by this approach. Furthermore the SSS-NIRS calibration used here is the same as that used in previous evaluations of WSC in white clover (Cosgrove *et al.*, 2006; Widdup *et al.*, 2010). Most notably, SSS was the phenotypic measure used by the breeders to support selection of the divergent populations, therefore it was

important to align the current experimental approach with theirs to enable comparisons of WSC content (section 2.5.3).

As there were distinct differences between the High-End and Low-End populations for all pools (mean difference of 78.3 g kg^{-1} dry matter), there is confidence that divergent phenotypes have been developed. However, the moderate correlation observed between the NIRS and ANTH data may have downstream ramifications on identifying genes underpinning the trait, which is discussed further in Chapter 6. It also raises the question: would greater progress be made by breeders if an alternative phenotyping methodology was used? Wet chemistry is an accurate although time- and labour-intensive way of predicting WSC content. NIRS does not require the use of chemicals (post calibration construction) and significantly reduces the test time and costs, which makes it superior over traditional chemical methods especially when a large number of samples need to be measured. Furthermore, multiple parameter estimates (e.g., crude protein, WSC, lipid and metabolizable energy) can be obtained simultaneously (Lister & Dhanoa, 1998). Finally, analyses using NIRS produce no chemical waste which reduces the costs of reagents, waste disposal and increases environmental friendliness (Yang *et al.*, 2017). However, as observed in the current study, an accurate NIRS calibration is required for the species and traits studied, which can be an issue for species that are not commonly analysed. For example, perennial ryegrass is frequently analysed for nutritional quality assessment and so many calibrations have been designed at different institutions around the world (Jafari *et al.*, 2003; Cogan *et al.*, 2005; Cosgrove *et al.*, 2009; Shetty *et al.*, 2012). White clover on the other hand has few studies assessing nutritional quality for the species using NIRS on its own (Berardo, 1997; Inostroza *et al.*, 2017). Another disadvantage of NIRS is that, although it is more cost-efficient than wet chemistry approaches, it is still a destructive assay and so requires considerable time spent in sample processing (i.e., harvest, drying, grinding then analysing). Alternative methods such as hyperspectral imaging (HSI) are showing promise for high-throughput and accurate predictions of pasture quality (Pittman *et al.*, 2016; Shorten *et al.*, 2019; Zhou *et al.*, 2019) and WSC concentration in a range of species including wheat (Dreccer, Barnes & Meder, 2014), maize (Hetta *et al.*, 2017) and perennial ryegrass (Shorten *et al.*, 2019; Smith *et al.*, 2020). Furthermore this technique is non-destructive, meaning that once the calibration is constructed, scanning and measuring plants *in situ* is rapid and accurate, with the only limitations of cost and large file sizes, which can slow image capturing and processing, presently known (Bock *et al.*, 2010). Furthermore, aerial HSI has been successful at determining nutritional traits, such as crude protein, spatially at the farm scale

(Pullanagari *et al.*, 2015; Pullanagari, Kereszturi & Yule, 2018). Determination of pasture quality using HSI could be used in the future to upscale the number of plants screened when making selections, which means higher selection intensities are possible, which would positively impact genetic gain. It also means that high enough numbers of plants can be used to mitigate inbreeding depression whilst still achieving an increase in genetic gain.

In summary, NIRS is a useful, inexpensive and high-throughput technique. However, the commercially-available NIRS calibrations, including those used by the breeders in developing these populations, are only moderately accurate predictors of true WSC levels as measured by the anthrone method. The relationship between ANTH and NIRS data was significant and still provided an estimate of WSC content that was able to confirm divergent phenotypes. For future phenotyping studies involving WSC content in white clover, an alternative methodology is suggested to improve phenotyping precision. Either generating a more accurate NIRS calibration for white clover on its own, or other methodologies such as HSI are recommended.

2.5.3 Water-soluble carbohydrate values were similar but lower to those determined concentrations in a previous experiment

The initial mean WSC concentrations in the Parent populations in the original study (referred to as the original study in this section) of Widdup *et al.* (2010) was 230 g kg^{-1} DM and rose by 35% to 310 g kg^{-1} DM in the High-End populations, with the magnitude consistent in each of the three pools: WNZLL, WNZSL, and WUSLL (Widdup *et al.*, 2010). Here, we observed a mean increase by 50.2 g kg^{-1} DM (38%) for all three Widdup pools (WNZLL, WNZSL, WUSLL) comparing the Parent populations (mean of 130.6 g kg^{-1} DM) to the High-End populations (mean of 180.9 g kg^{-1} DM). Similarly, a mean increase by 52.7 g kg^{-1} DM (41%) was observed across all five pools comparing the Parent populations (mean of 127.4 g kg^{-1} DM) to the High-End populations (mean of 180.1 g kg^{-1} DM) (**Table 2.2**). Although the absolute WSC values presented here were lower than originally reported when selections were being made, the magnitude of the increase in WSC content was consistent with the original study and the WSC content is still higher than commercial white clover varieties such as ‘Kopu II’ and ‘Tribute’ (Cosgrove *et al.*, 2006; Widdup *et al.*, 2010). Sample drying and storage procedures, environmental conditions when harvesting took place, and the age of the plants were assessed to determine why there was a large difference between the original High-End populations and the High-End populations in the current study.

Sample drying and storage procedures are unlikely to explain the differences in the observed WSC content between the original study and the current study. In the original study, samples were frozen in liquid nitrogen (N), freeze-dried, and ground prior to WSC determination via NIRS (Widdup *et al.*, 2010). In the current study, samples were frozen in liquid N, freeze-dried, ground, stored for three weeks, then oven dried at 60°C for three hours prior to WSC determination via NIRS. Sample grinding was completed at the end of December 2017 but NIRS was not performed until mid-January 2018. The length of time the samples were stored before NIRS analysis would not have contributed to the differences in WSC content observed. The evidence for this is outlined below. Changes in WSC content can occur because drying samples at lower temperatures, for example freeze drying, does not denature proteins; therefore enzymatic conversions and respiratory losses can occur when moisture is present (Smith, 1973). Couchman (1959) and Greenhill, Couchman, and De Freitas (1961) demonstrated that dry matter WSC losses in white clover, alfalfa and ryegrass herbage increase with increased storage time, increased storage temperature and increased tissue moisture, however only a small 8% DM loss occurred after 9 months stored at 36°C and 17% moisture. As only three weeks separated sample drying and analysing; and samples were stored in capped vials in air tight polystyrene containers containing silica gel, WSC concentrations should not have decreased due to the lack of moisture and short duration of storage.

Although samples were stored in a sealed polystyrene box with silica gel, samples may have absorbed atmospheric water within the sealed vial as most freeze-dried materials are hygroscopic (Baker, 1997). As small variations in moisture content in samples has been shown to produce large prediction errors of NIRS calibrations (Baker, Givens & Deaville, 1994), the additional oven drying procedure was implemented to minimise water content in the NIRS vials. It has been shown that WSC degradation post freeze drying is minimal (Pelletier *et al.*, 2010), and sample drying at 60°C has not shown a marked decrease in WSC content post pre-treatment in forage species (Deinum & Maassen, 1994; Pelletier *et al.*, 2010). Furthermore, Corson *et al.* (1999) suggest that samples should be dried at 60°C to minimise alteration in chemical composition. Therefore, sample drying procedure is also unlikely to contribute to the large differences observed between the original and current studies.

WSC is influenced heavily by the environment, therefore selection for the trait can be elusive as the temperature and sunlight need to remain consistent between sampling days and preferably over years as well. Boller and Nösberger (1983)

demonstrated that lower temperatures cause an increase in WSC content in white clover foliage. Decreases of daytime temperature by 8°C and night-time by 6°C (i.e.,18°C/13°C to 10°C/7°C) in a 16-hour photoperiod caused a substantial increase in WSC content (from 16.2 to 24.5 g m^{-1}). Environmental variables including temperature and hours of sunshine in a day were examined for the months harvesting occurred (October and November) for 2001 to 2004, when the Widdup *et al.* (2010) data were acquired and for 2017, the year of the harvest in the current experiment (2017) to determine if there were any major changes in environmental conditions between years of harvests (**Table S2.11**, Appendix 1). Climate data were taken from the Palmerston North Electronic Weather Station located at latitude -40.382, longitude 175.6092 (less than 1 km from experiment site) sourced from NIWA (2020). Simple analysis of variance (ANOVA) tests were performed in RStudio using the function *anova()* to determine if two or more means were significantly different at $\alpha = 0.05$ for each environmental variable: hours of sunlight in a day (Sun), maximum temperature (Tmax), minimum temperature (Tmin) and temperature at 9 am each day (Tdry) (See **Table S2.11**, Appendix 1 for variable descriptions). Residuals were checked for each test and no patterns were observed. There were no significant ($p = 0.197$) differences in hours of sun in a 24-hour period between the years, however there were significant differences between two or more years for Tmax, Tmin and Tdry ($p < 0.001$ for all three variables). Tukey's honest significance difference post hoc test was performed for variables with significant ANOVA *p*-values using the function *TukeyHSD()* in RStudio. There were no significant differences in temperature variables between 2004 and 2017, when the High-End population was measured for WSC in the original experiment (Widdup *et al.*, 2010) and in the current experiment (**Table S2.12**, Appendix 1). Thus, temperature and sunlight are unlikely to be the causes of mean differences in WSC observed.

The age of the plants is also unlikely to explain the differences observed. When selections were originally made, seed were germinated, plants were grown in a pot during winter in a glasshouse, placed outside in August 2004, and then sampled for WSC in late October or early November (Widdup *et al.*, 2010). Here, seed were germinated in March 2017, grown in pots, placed outside in late May, and then sampled for WSC content in mid-November. The additional two months that plants in the current study were outside and the slight age difference may have cause incongruency, however, ascertainment bias is a more likely explanation. Following the original study, Widdup *et al.* (2010) performed a field trial using the High-End populations from the WNZLL and WUSLL pools. For this field trial, seed from the fourth generation was sown in July 2005 in a glasshouse, transferred to the field in October, and then sampled for

WSC content in December and at multiple other time points. The highest observed WSC content for the field trial was measured at 200 g kg^{-1} DM in the summer (December 2005), while spring measurements in the following year showed a decrease to approximately 185 g kg^{-1} DM in October and approximately 175 g kg^{-1} DM in November, which are more like the values obtained in the current study. The original study, the current study and the field trial all used plants randomly selected from the seed pool; but both the current study and the field trial showed lower WSC content than originally reported. Therefore, it is possible that differences occurred due to ascertainment bias.

2.5.4 Leaf size was not indirectly selected for by breeding for divergent leaf water-soluble carbohydrate content

For the purpose of characterising the genetic control of foliar WSC, it is crucial that foliar WSC in the current experimental materials is not confounded by leaf size. This is because the main objective is to identify genes or genomic regions underpinning control of WSC, not leaf area. White clover cultivars are classified according to leaf size classes including: small (e.g., 'Tahora' and 'Prop'), small-medium (e.g., 'Demand' and 'Prestige'), medium (e.g., 'Avoca', 'Bounty' and 'Pitau'), medium-large (e.g., 'Sustain' and 'Tribute'), and large (e.g., 'Aran', 'Kopu II', 'Kotare', 'Legacy' and 'Mainstay'). Leaf size is linked with several other important features, the most important of which is stolon density (Charlton & Stewart, 1999). There is an inverse relationship between leaf size and stolon density that impacts the plant's ability to persist in a variety of different pasture systems (Jahufer *et al.*, 1999; Woodfield & Clark, 2009) (see Chapter 1, section 1.1.1.2*,* for more information). Larger-leaved cultivars have greater yield potential and are typically bred for use in rotationally grazed dairy pastures, whereas smaller-leaved cultivars have a better ability to persist under continuous grazing, particularly by sheep, due to their high stolon density (Charlton & Stewart, 1999). The relationship between leaf size and pasture system is very important, therefore in breeding systems it is imperative to maintain leaf size classes and plant architecture while also selecting for any other trait of interest (Caradus, Woodfield & Stewart, 1996; Woodfield *et al.*, 2001). Woodfield *et al.* (2001) tested a range of white clover cultivars in NZ and found that large-leaved cultivars tended to have higher levels of WSC than medium- and small-leaved cultivars. These data suggest that gains in foliar WSC could be achieved by breeding for increased leaf size. However, due to the strong, negative correlation between leaf size and stolon number, increasing leaf size to increase WSC content would not be feasible if breeding for sheep pasture systems as these plants would potentially have reduced stolon number and would not persist long term. Here it is shown

that leaf size remained consistent in two pools (WNZLL and FNZSL) but that there was observable and significant fluctuation in leaf size in the remaining pools. Leaf area increased in the High-Mid and High-End populations compared to the Parent population in WNZSL, whereas in the FNZLL and WUSLL pools there was a decrease in LA in the Low-End populations relative to the Parent population. The WUSLL pool in particular was affected by powdery mildew during the course of the experiment, which may have contributed to the observed decline in LA in that pool. These data suggest indirect selection for LA may have occurred while breeding for divergent foliar WSC in three pools.

Although there were weak, positive relationships between WSC and LA from the correlation analysis of each pool, the regression analysis showed that once the population was taken into account there was little relationship between the two variables. Furthermore, when splitting the data into populations, the high populations sat higher on the *y*-axis (WSC content) than the low populations, but the high populations did not sit in the top-right quadrant (high WSC and large LA) of the graph and the low populations did not sit in the bottom-left quadrant (low WSC and small LA), which would have indicated a strong LA effect on WSC content. The relationship between LA and WSC has not been extensively studied in legumes, but one study in lentil found a weak negative relationship between WSC and LA (r = -0.28) (Tahir *et al.*, 2019). In white clover, previous experiments have observed an association between WSC and LA, with larger-leaved plants containing more foliar WSC (Malinowski, Belesky & Fedders, 1998; Woodfield *et al.*, 2001). However, other studies have observed little to no correlation between WSC and foliar biomass in clovers. Little to no correlation was observed in 128 red clover genotypes between biomass and starch concentration (Pearson r = 0.1223, $p = 0.169$) (Ruckle *et al.*, 2017). Ruckle *et al.* (2018) also found no significant relationship between WSC content and aerial tissue (leaves and petioles) biomass at the end of the day, in white clover (Spearman's rank correlation coefficient (ρ) = −0.0200, $p = 0.790$). Leaf area and leaf biomass are positively correlated in many broadleaved plant families (Niklas *et al.*, 2007; Huang *et al.*, 2019a; Huang *et al.*, 2019b). Therefore, the lack of relationship observed between WSC and biomass by Ruckle et al. (2018) is comparable to the relationship observed in the present study. This lack of relationship between WSC and LA suggests that white clover plants with small leaves can accumulate WSC to comparable levels as large-leaved plants, which was demonstrated in **Figure 2.3**, and indicates that increased WSC content can be achieved without changing leaf size classes.

From the data presented here, we can conclude that the breeding programmes successfully produced populations with divergent WSC phenotypes and although there was variation in leaf size in some of the breeding pools, changes in leaf WSC did not occur as a consequence of indirect selection for LA. This highlights the importance of identifying genes underpinning WSC accumulation as once identified, more efficient breeding programmes can be developed to increase WSC content without changing leaf size class. After phenotype validation, investigation into the genetics underpinning the trait could begin, which is discussed in Chapter 3.

2.6 Conclusions

Two of the available near infra-red reflectance spectroscopy (NIRS) calibrations for water-soluble carbohydrate (WSC) were sufficiently accurate, when compared to the anthrone method, to use for phenotyping WSC in white clover. For future studies investigating WSC in white clover, an alternative phenotyping methodology is recommended or a more accurate NIRS calibration should be developed. For all five pools, breeding for divergent foliar WSC was successful, based on significant difference in WSC observed between the two populations at the divergent end time points. WSC content in the End populations sampled here contained similar, although slightly lower WSC content to values observed in Widdup *et al.* (2010). Estimation of mean leaf size in white clover can be reliably achieved by sampling four leaves per plant at the same phenological stage. Once population structure was accounted for, there was no strong relationship between WSC and LA. Clear divergent WSC phenotypes were confirmed for all five pools, thus investigation into genetics underpinning the WSC trait could be reliably explored in Chapters 3 and 4, without a confounding contribution from leaf area.

2.7 Acknowledgements

I would like to thank Paul Maclean (AgResearch), Catherine McKenzie (Plant and Food Research) and Poppy Miller (AgResearch) for experimental design and statistical support. Thank you to AgResearch colleagues: Andrew Faram, Anna Larking, Steven Odering and Prue Taylor for plant care advice and nursery assistance; Anna Larking and Prue Taylor for sample processing, John Ford (PGG Wrightson Seeds) for providing a wealth of information on the white clover populations; Zac Beechey-Gradwell and Hong Xue for wet chemistry assistance and advice. Also, a huge thank you to everyone who helped with the harvest for determining WSC content: Craig Anderson, Emmy Bethel, Grace Ehoche, Emma Griffiths, Won Hong, Zulfi Jahufer, Anna Larking, Yulia Morozova, Narsaa Na, Jana Schmidt, Chaewon Song and Prue Taylor.

CHAPTER 3

Single nucleotide polymorphism (SNP) markers associated with foliar water-soluble carbohydrate accumulation in white clover (*Trifolium repens* L.) using a genome-wide association study and outlier SNP detection approaches based on genotyping by sequencing data

3.1 Abstract

Breeding for divergent water-soluble carbohydrate (WSC) levels has been successful in five white clover breeding pools, but the genetic loci responsible for the observed phenotypic variation in this trait is unknown. High-throughput sequencing using genotyping by sequencing (GBS) is a cost-effective method to generate sufficient data for aligning genotype and phenotype data to uncover genomic regions under selection in non-model species. GBS single nucleotide polymorphism (SNP) data were obtained from 1,113 white clover individuals from the 25 white clover populations utilised in Chapter 2. Both outlier detection and genome-wide association study (GWAS) approaches require a preliminary assessment of population structure as it can be a confounding factor in the analysis. Population structure was assessed using discriminant analysis of principal components (DAPC) and assigned individuals into 11 genetic clusters. Little genetic differentiation was observed between generations within the low WSC and high WSC divergent selections, respectively, and Parent individuals were assigned to one cluster. A genotype-phenotype association study using a subset of 605 white clover individuals, consisting of 25 individuals from 24 populations, was conducted and identified a single SNP associated with a starch biosynthesis gene, *glgC*. Outlier detection approaches using PCAdapt, BayeScan and KGD-F_{ST} were implemented to identify SNPs that were responsible for discriminating between high and low WSC populations. These analyses identified 33 SNPs that were under selection. One of these SNPs was located in the intron of *ERD6-like 4*, a gene coding for a sugar transporter located on the vacuole membrane. The remainder of the outlier SNPs were associated with a wide range of functions, some of which may be associated with WSC accumulation. While this study was unable to identify a universal set of SNPs that could be applied as a selection tool in all white clover populations; it was able to provide a starting point in identifying the genes underpinning foliar WSC accumulation in white clover.

3.2 Introduction

3.2.1 Study system and trait importance

White clover (*Trifolium repens* L.) is a key component of mixed grass and clover swards in pastures, particularly in New Zealand/Aotearoa (NZ), as it provides quality forage and source of bioavailable nitrogen fixed through symbiosis with soil *Rhizobium* bacteria. While the grass component provide more forage during the cool seasons, the clover produces forage during warmer summer conditions when grass growth is less vigorous compared to clover (Hoglund & Brock, 1978; Charlton & Stewart, 1999; Brock & Hay, 2001). This complementary association works very well in temperate climates such as NZ. Additional advantages of these mixed swards are the reduction of weed encroachment and erosion, greater stand longevity than grass or clover monocultures (Casler, 1988) and significant reduction in nitrogen fertiliser application due to white clover's nitrogen fixing ability (Ledgard & Steele, 1992). Higher clover content also improves forage quality of these mixed sward pastures, specifically high protein, water-soluble carbohydrate (WSC) and low fibre which increases digestibility, providing ruminant diets with increased nutritive value, resulting from increased energy and protein intakes (Harris *et al.*, 1998).

To meet the projected 100% increase in global demand for feed, fibre and food by 2050, productivity of pasture systems will require efficient increase in the speed of genetic improvement to cope with the 2% yearly increase in the world population (Tilman *et al.*, 2011). High sugar ryegrass cultivars have been produced as a means to increase animal performance and nitrogen-use-efficiency in pasture-based animal production systems (Edwards *et al.*, 2007). Additionally, elevation of WSC levels in white clover leaves may help to increase the productivity of pasture systems by increasing milk yield per cow and liveweight gain in sheep (Harris *et al.*, 1998; Edwards *et al.*, 2007; Higgs *et al.*, 2010). Despite the agronomic and environmental importance of high WSC, almost nothing is known about the genetic basis of foliar WSC accumulation in white clover. One study described increasing WSC in white clover foliage through four cycles of parallel recurrent selection resulting in a 35% increase in WSC content compared to the original (Parent) population (Widdup *et al.*, 2010). Unfortunately, when assessed in the field, the high WSC plants exhibited poor agronomic performance in the first year growth, possibly due to inbreeding depression from the four cycles of selection in a closed breeding pool (Widdup *et al.*, 2010). Improved understanding of the genetic basis of the trait could help to increase breeding efficiency to produce white clover cultivars with

increased levels of WSC, while maintaining genetic diversity, through the application of molecular breeding techniques such as marker-assisted selection (MAS) or genomic selection (GS).

3.2.2 Genomic approaches to identify loci linked to water-soluble carbohydrate

Molecular marker technologies have led to breeding strategies (e.g., MAS and GS) that can increase the precision of selection and reduce the generation interval, leading to increased genetic gain in forages (Faville *et al.*, 2012; Resende, Casler & de Resende, 2014; Faville *et al.*, 2020a). In general, MAS uses individual molecular markers in linkage disequilibrium (LD) with genes from the trait of interest. GS is a form of MAS, but integrates the effects of many thousands of markers, without necessarily knowing what they are linked to (Meuwissen, Hayes & Goddard, 2001). Both approaches make breeding a more reliable and cost-effective process as the selection accuracy can be increased with limited phenotype data (Lorenz, 2013). Reduced representation sequencing is a popular method of obtaining genotype data as single nucleotide polymorphisms (SNPs), particularly for non-model organisms lacking genomic resources. Reduced representation sequencing methods are based on restriction enzyme digestion of DNA to reduce genome complexity, allowing for a small percentage of the genome of a sample to be sequenced. There are several reduced representation sequencing methodologies, including 2b-RAD (Wang *et al.*, 2012), ddRAD (Peterson *et al.*, 2012), ezRAD (Toonen *et al.*, 2013), genotyping by sequencing (GBS) (Elshire *et al.*, 2011; Poland *et al.*, 2012b), RADseq (Baird *et al.*, 2008) and SLAF-seq (Sun *et al.*, 2013). GBS has been implemented in numerous forage species (Pembleton *et al.*, 2016; Biazzi *et al.*, 2017; Sakiroglu & Brummer, 2017; Faville *et al.*, 2018; Guo *et al.*, 2018), including white clover (Wright *et al.*, 2017; Griffiths *et al.*, 2019), and simultaneously discovers and calls genotypes for genome-wide SNPs in a cost-effective manner (Elshire *et al.*, 2011; Poland *et al.*, 2012b). Most next-generation sequencing platforms require optimisation of library preparation and SNP calling (Syed, Grunenwald & Caruccio, 2009). For example, the ideal ratio of adapters to genomic DNA for each species needs to be optimised prior to library construction (Elshire *et al.*, 2011). As AgResearch has developed GBS workflows for white clover using multiple restriction enzyme combinations, based on modified Elshire *et al.* (2011) and Poland *et al.* (2012b) protocols, pre-library optimisation was not required for the current research. A reference genome is not necessary as tools have been developed and optimised for *de novo* analysis of the sequencing (Catchen *et al.*, 2011; Glaubitz *et al.*, 2014), however a reference genome or genome of a closely related species will improve the quality of

SNPs generated and allow for gene identification (Shafer *et al.*, 2017). The white clover reference genome is currently resolved at the pseudomolecule level with genome annotations available so SNP positions can be linked to inferred genetic functions (Griffiths *et al.*, 2019).

Identification of markers associated with a trait can be achieved by several methods. There are three major genomic approaches that are used as a first step to identify loci linked to trait phenotypic variation. The first is a candidate gene approach that utilises information from theoretical and experimentally reported links between pre-specified genes of interest that underpin a given phenotype. For non-model organisms, such as perennial ryegrass (Dracatos *et al.*, 2009; Yu *et al.*, 2013; Yu *et al.*, 2015) or white clover, this method often involves the use of gene information from related model organisms such as *Arabidopsis thaliana* L.*,* alfalfa/lucerne (*Medicago sativa* L.), barrel clover (*M. truncatula* Gaertn.) and rice (*Oryza* spp.). The second approach is a genome-wide association study (GWAS), which utilises markers across the entire genome and investigates the association of phenotypic variation with genotypic variation in complex traits. This method can be used in both model and non-model organisms and has successfully pinpointed genes underlying numerous traits in forage species (Arojju *et al.*, 2016; Biazzi *et al.*, 2017; Sakiroglu & Brummer, 2017). The third approach is an outlier analysis that identifies SNPs that differentiate populations based on phenotype. These approaches typically include F_{ST}-based tests and principal component analyses (PCA) to identify loci that are distinct from those under neutral expectations.

A candidate gene approach is efficient when candidate genes have a well described function linked to the phenotype of interest. In addition, there is a lower cost as effort is focused on a set of promising genes, which is an advantage for reducing the cost as fewer markers are required (Amos, Driscoll & Hoffman, 2011). However, these studies are incapable of discovering new genes or combinations of genes that influence a given trait. The major advantage of GWAS is the ability to pinpoint genes irrespective of whether their function was previously determined (Amos *et al.*, 2011). However, GWAS experiments are often costly as both genotype and phenotype information are required for hundreds to thousands of individuals (Frayling, 2014). Outlier analyses on the other hand, do not require phenotypic information and therefore numerous markers can be screened to identify candidate genes. As no investigation into the genetic control of foliar WSC in white clover has been undertaken, an outlier analysis approach is most suitable. However, as phenotypic information was obtained (Chapter 2) for a subset of the individuals, a GWAS could also be performed. Both outlier and GWAS approaches

require a preliminary assessment of population structure as it can be a confounding factor in the analysis. When individuals are genetically related, standard association studies can identify incorrect causal variants and instead false positive associations are detected. Population structure causes a form of relatedness in the sample which can cause the statistical methodology applied to association studies to assign strong signals to SNPs that are not actually associated with the trait of interest (Sul, Martin & Eskin, 2018). Hence, failure to account for population structure can reduce power and produce false positive associations. The rate of false-positive discovery is influenced by the number of genes under selection, genetic differentiation of populations and the strength of selection (Pérez-Figueroa *et al.*, 2010). Hierarchical genetic structure can result in false positive loci (type 1 errors) in outlier analyses (Excoffier *et al.*, 2009) because a hierarchical structure of populations leads to a narrow null distribution of F_{ST} values, which in turn, leads to an excess of false significant loci (Excoffier *et al.*, 2009). One way to mitigate false discovery is to only consider SNPs detected by at least two independent outlier detection tests. Those SNPs identified as in common among the two tests should be true positives.

3.2.3 Population structure determination to reduce false positive associations

Methods for clustering individuals into populations can be divided into two categories: distance-based approaches and model-based approaches. Distance-based approaches are model-free and aim to identify clusters by analysing matrices describing genetic distances between individuals or populations derived from genotypic data. Distance-based clusters are often then identified by visualisation using a graphing methodology such as multidimensional scaling, for example, PCA. By contrast, model-based approaches are more common and assess the likelihood of multiple models from the data but have prior assumptions of the model, for example that the populations are in Hardy-Weinberg equilibrium (HWE) (Pritchard, Stephens & Donnelly, 2000). HWE is a specific assumption as it requires the population to have: a static allele frequency through infinite size; random mating; and no mutation, gene flow, nor selection occurring. Bayesian clustering and maximum-likelihood techniques are used in programmes such as ADMIXTURE (Alexander, Novembre & Lange, 2009), BAPS (Corander, Waldmann & Sillanpää, 2003) and STRUCTURE (Pritchard *et al.*, 2000). When underpinning assumptions such as that populations are in HWE, and that marker loci are unlinked and at linkage equilibrium with one another in populations, are not met, then distance-based measures are more suitable for statistical inference about population structure. For example, a discriminant analysis of principal components

(DAPC) partitions variance within and among groups without the above-mentioned assumptions on LD or HWE (Jombart, Devillard & Balloux, 2010).

Quantification of genetic variation can be achieved at both the subpopulation-population level and at the individual level. Common methods for quantifying genetic variation at the population level investigate how assumed putative subpopulations are related through the use of analysis of molecular variance (AMOVA) (Excoffier, Smouse & Quattro, 1992), F-statistics (Weir & Cockerham, 1984), or phylogenetic methods (Cavalli-Sforza & Edwards, 1967; Saitou & Nei, 1987). AMOVA is commonly used to produce estimates of variance components and F-statistic analogues to reflect the correlation of haplotype diversity at multiple levels of hierarchical subdivision, such as within populations and among populations (Excoffier *et al.*, 1992). F_{ST} is an important statistic used in population genetic studies and represents a measure of population differentiation ('genetic difference') based on differences in allele frequencies. F_{ST} is a population diversity index and cannot be used to assign individuals into clusters. As both F_{ST} and AMOVA rely on predetermined hierarchical structure, either through *a priori* information or from the results of an above-mentioned clustering method, F_{ST} and AMOVA analyses are typically performed after assessment of population structure. Weir and Cockerham's (1984) F_{ST} calculation is defined as the ratio of the expected heterozygosity of individuals within a sub-population to the total expected heterozygosity of individuals across all populations and is calculated using **Equation 1.1** (Chapter 1).

3.2.4 Identification of markers under selection

Identification of markers under selection can be obtained via numerous methods including a classical GWAS approach using software and algorithms such as rrBLUP (Endelman, 2011), GAPIT (Lipka *et al.*, 2012), TASSEL (Bradbury *et al.*, 2007), PLINK (Purcell *et al.*, 2007), EMMAX (Kang *et al.*, 2010) and P3D (Zhang *et al.*, 2010b). Alternatively, F_{ST} and PCA-based methods have shown great ability to detect regions of the genome and markers under selection in many animal species but recent studies in barley (Abebe, Naz & Léon, 2015; Reinert *et al.*, 2019), wild pearl millet (Berthouly-Salazar *et al.*, 2016) and oat (Bekele *et al.*, 2018) show increasing uptake in plants.

Once markers under selection are identified, linking these markers to genes underpinning the trait of interest requires determination of linkage disequilibrium (LD) in the individuals of interest. LD can be used to determine an estimated distance for associating SNPs with candidate genes (Flint-Garcia, Thornsberry & Buckler, 2003). LD itself determines the degree to which alleles at two different loci are non-randomly

associated and is relative to the allele frequencies at the two loci. Understanding the extent of global LD allows determination of whether a given pair of loci are likely to be in LD or linkage equilibrium (Flint-Garcia *et al.*, 2003). One of the most common measures of LD is r^2, which is simply the correlation between the pair of loci (Hill & Robertson, 1968). Values range from 0 to 1, with values of 0 indicating complete linkage equilibrium (random association), and values of 1 indicating loci are in complete LD (correlation between two loci). R^2 only equals 1 when the two SNPs have not been separated by recombination and have the same allele frequencies. As white clover is an outbreeding species, the extent of LD decay in white clover is expected to be low. However there has only been one study that has estimated LD decay in white clover, and it was reported to decay to $r^2 = 0.2$ within 134 Kbp (Inostroza *et al.*, 2018).

In this chapter, recent breeding programmes that generated white clover populations with improved foliar WSC concentrations, compared to commercial varieties, were utilised to investigate regions of the genome and loci under selection for divergent WSC accumulation. Parallel recurrent selection was implemented in these breeding programmes to create divergent lines with high or low WSC concentrations (Widdup *et al.*, 2010). Analyses were based on genome-wide GBS-derived SNP data for individuals from five breeding pools which included populations from three time points in the breeding process towards higher/lower WSC. A GWAS using rrBLUP to link phenotype and genotype in 605 individuals and outlier detection approaches using BayeScan, PCAdapt and KGD-F_{ST} in 1,113 individuals were implemented to identify SNPs separating high and low WSC populations. An LD analysis was undertaken in the populations to support insight into candidate genes potentially in LD with the detected outlier SNPs. The overall aim of this chapter was to identify a set of SNPs associated with foliar WSC accumulation, which may subsequently be developed and used for MAS in white clover populations to aid future breeding programmes for enhanced WSC.

3.3 Materials and methods

3.3.1 Plant material

Material from two discrete white clover breeding programmes, one running between 2000 – 2004 over four generations of selection in three breeding pools (Widdup *et al.*, 2010) and the other between 1999 – 2004 over six generations in two pools (Mr John Ford, pers comm), were utilised for the current experiment, as described in Chapter 2 section 2.2.3. From the five breeding pools there were 53 populations available to analyse, including nine populations in each of the Widdup pools and thirteen populations

in each of the Ford pools (**Figure 2.1**, Chapter 2), but only 25 populations were used in the current study. These were: WNZLL-Low-End, WNZLL-Low-Mid, WNZLL-Parent, WNZLL-High-Mid, WNZLL-High-End, WNZSL-Low-End, WNZSL-Low-Mid, WNZSL-Parent, WNZSL-High-Mid, WNZSL-High-End, WUSLL-Low-End, WUSLL-Low-Mid, WUSLL-Parent, WUSLL-High-Mid, WUSLL-High-End, FNZLL-Low-End, FNZLL-Low-Mid, FNZLL-Parent, FNZLL-High-Mid, FNZLL-High-End, FNZSL-Low-End, FNZSL-Low-Mid, FNZSL-Parent, FNZSL-High-Mid, and FNZSL-High-End. Where: W = Widdup, F = Ford, NZ = New Zealand, US = United States of America, LL = large leaf, SL = small leaf, P = Parent, Low = low water-soluble carbohydrate (WSC), High = high WSC, Mid = middle generation and End = end generation.

3.3.2 DNA isolation and quality control

Genomic DNA was extracted from 1,536 white clover individuals (approximately 60 individuals from each of the 25 populations) in a 96-well plate format (16 plates in total) following the freeze-dried tissue protocol described in Anderson *et al.* (2018). All samples were resolved by electrophoresis in a 0.8% lithium borate agarose (w/v) gel containing 25 µg ethidium bromide for 40 minutes at 3.3 v cm^{-1} and visualised under UV (Gel Doc™, Bio-Rad, CA, USA) to assess DNA quality. The quantity of DNA was measured using a modification of the bisbenzimidazole Hoechst 33258 DNA quantification method (Rago, Mitchen & Wilding, 1990). Briefly, DNA was diluted 1:3 in TE buffer (i.e., 15 µL DNA added to 30 µL TE in 96-well MultiMax (vwr.com) plates. The solutions for Hoechst quantification are as follows. TNE (2 M NaCl) was prepared according to Rago *et al.* (1990). Sufficient Hoechst dye reaction mixture was made for quantification of 96 DNA samples plus 8 standards in triplicate by adding 17.8 mL TNE and 200 µL Hoechst dye stock (Thermo Fisher Scientific Inc.) into a 50 mL Falcon tube. The Phage λ DNA (Boehringer Mannheim, Indianapolis) standards were diluted to a range of concentrations between 5 and 50 ng μL^{-1} to create a linear set of standards according to the following scheme:

Final concentration (ng μL^{-1}):	0	5	10	15	20	25	30	35
100 ng μL^{-1} λ DNA:	0	50	100	150	200	250	300	350
$_sH_2O$:	1000	950	900	850	8000	750	700	650

A 45 µL aliquot of Hoechst dye reaction mixture was added to each well of a black 384-well Black Assay Plate (4titude, Surrey UK). Each DNA sample was assayed

in triplicate, with 5 µL of diluted DNA added to the Hoechst dye reaction mixture and 5 µL of undiluted standard added to the Hoechst dye reaction mixture and then the 384-well plate was quantified by measuring fluorescence using a Synergy HTX Multi-Mode Microplate Reader (BioTek, VT, USA). The fluorescence values converted to ng μL^{-1} based on the Phage λ DNA standard curve. DNA samples were stored at -20°C until needed for GBS library construction.

3.3.3 Genotyping by sequencing library preparation and sequencing

A subset of 1,175 individuals, comprising 47 from each of the 25 populations used in the phenotyping study (Chapter 2) was chosen for genotyping. Each genotyping by sequencing (GBS) library can contain a maximum of 96 individuals, but one negative and one positive control were included for each library, leaving 94 positions for samples. The negative control (a water control) was used to confirm no contamination in the GBS libraries and the positive control (inbred white clover S_9 individual (Cousins & Woodfield, 2006; Griffiths *et al.*, 2019)) was used to confirm single nucleotide polymorphism (SNP) detection consistency across all libraries. As there were 47 individuals from 25 populations (1,175 samples in total), twelve and a half GBS libraries were required (1,175 samples divided by 94 wells equals 12.5 libraries). The spare 47 wells (to equate to 13 full libraries) were filled with duplicated DNA samples chosen at random from all populations. The duplicated samples included duplication of samples within a GBS library and duplication of samples between GBS libraries. The duplicated DNA samples within a GBS library and the positive S9 control samples were used to confirm consistency both within and between all GBS libraries, as explained in section 3.3.4.3.

Thirteen GBS libraries were constructed following Poland *et al.* (2012b) with some modifications. For each individual, genomic DNA concentration was normalised to 100 ng μL^{-1} for library construction and was added to a dried down 96-well adapter plate using a Nanodrop II (BioNex Solutions, CA, USA). Each well of the plate had a unique barcoded adapter to enable subsequent sample identification. The plate was sealed with an Air-O-Seal sheet (4Titude, Surrey, EU), spun down briefly and then dried down using the Savant SPD111V SpeedVac Concentrator (ThermoFisher, MA, USA). A digestion for one reaction was performed in a 20 µL volume containing 16 µL nuclease-free water, 2 µL CutSmart Buffer (New England Biolabs Inc. (NEB), Ipswich, MA, USA, No. B7204S), 1 µL 20 U μL^{-1} *PstI* (NEB, Ipswich, MA, USA, No. R0140S), and 1 µL 20 U μL^{-1} *MspI* (NEB, Ipswich, MA, USA, No. R0106S) which was added into the adapter + DNA plate. Enzymatic digestion was carried out using a Kyratech Thermo

cycler (ThermoFisher, MA, USA) and included an incubation period of 2 hours at 37°C and 30 minutes at 65°C. Ligation of adapters onto fragmented DNA was performed for each digested sample by adding a 30 µL reaction solution containing 5 µL 10X Ligase buffer (NEB, Ipswich, MA, USA, No. B0202S), 2 µL T4 DNA Ligase (NEB, Ipswich, MA, USA, No. M0202L) and 23 µL nuclease free water. Ligation was carried out using a Kyratech Thermo cycler (ThermoFisher, MA, USA) and included ligation at 22°C for one hour and incubation at 65°C for 20 minutes. A 5 µL aliquot of each digested/ligated sample was then pooled together into a single well of a PCR strip tube using a Nanodrop II (BioNex Solutions, CA, USA). The pooled samples were then transferred from the PCR strip tube into a single 5 mL Eppendorf tube (Eppendorf) with 2.5 mL CP buffer (Omega Bio-Tek, GA, USA). Libraries were purified with an E.Z.N.A. Cycle Pure Kit (Omega Bio-Tek, GA USA, No. D6492-02), per kit instructions and then eluted in 50 µL elution buffer (Omega Bio-Tek, GA, USA). To increase the amount of GBS library DNA for size selection prior to sequencing, six parallel polymerase chain reaction (PCR) amplifications were completed for each library. Restriction fragments from each library were amplified in six separate reactions, each in a 50 µL volume containing 4 µL pooled DNA fragments, 25 µL 2X Taq Master Mix (NEB, Ipswich, MA, USA, No. M0270L), 2 µL PCR primer Mix (12.5 pmol μL^{-1} each primer as described in Poland et al. (2012)), and 19 µL nuclease free water. Amplification by PCR was carried out using a Kyratech Thermo cycler (ThermoFisher, MA, USA) and included an initial denaturation at 72°C for 5 minutes and 98°C for 30 seconds; then 18 cycles at 98°C for 10 seconds, 65°C for 30 seconds and 72°C for 30 seconds; followed by a final extension at 72°C for 5 minutes. For each PCR product, all six parallel samples were pooled and purified with an E.Z.N.A. Cycle Pure Kit (Omega Bio-Tek, GA USA, No. D6492-02) per kit instructions and eluted in 35 µL elution buffer. Libraries were assessed for quantity and quality on a nanodrop Spectrophotometer (Nanodrop Technologies, Montchanin, USA) using a 1.5 µL aliquot of library. A 2 µL aliquot of the library was reserved for analysis prior to size selection. The fragment size selection step was performed on a 30 µL aliquot of the library which was combined with 10 µL of size ladder L (Pippin reagents kit No.: CDF2010) and then DNA fragments of 193 – 313 bp size range were selected on a Pippin Prep (Sage Science, MA, USA). A 2 µL aliquot of the size-selected library was reserved for analysis. The pre-size selection aliquot was diluted 1:5 (6-fold dilution) and the post-size selection sample was diluted 1:4 (5-fold dilution) with water. An Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA) was used to assess the pre- and post-size selection samples and quality of each library. A standard curve of migration time versus fragment length was created from a ladder of known sizes. Sample fragment sizes were then calculated

from measured migration times, detected by fluorescence, and were translated into electropherograms (peaks) and gel-like images (bands). Each library was then sequenced in parallel on two lanes of a flow cell on an Illumina HiSeq 2500 (Illumina, San Diego, CA, USA) at Invermay Agricultural Centre (AgResearch, Mosgiel, New Zealand).

3.3.4 Single nucleotide polymorphism calling, filtering and genotyping by sequencing library control

3.3.4.1 Single nucleotide polymorphism calling

Raw data FASTQ files containing sequence reads were processed for SNP identification using the GBS analysis workflow implemented in Trait Analysis by aSSociation, Evolution and Linkage (TASSEL) v 5.0 (Glaubitz *et al.*, 2014) using default parameters except minor allele frequency (MAF) was set to 0.01. An AgResearch white clover genome assembly (version 5) was used as the reference genome (Griffiths *et al.*, 2019). Raw sequence data of 1,222 samples, 13 positive controls and 13 negative samples were analysed together. The purpose of performing the combined analysis of all pools together was to ensure commonality of SNPs nomenclature called across all samples. A summary of the bioinformatics workflow is presented in **Figure S3.1** (Appendix 2). Sequence reads were first trimmed to 64 bp and identical reads were grouped into sequence tags. The sequence tags were then aligned to the reference genome using Burrows-Wheeler Alignment (BWA) tool (Li & Durbin, 2009).

3.3.4.2 Single nucleotide polymorphism and sample filtering

After SNP calling, an investigation into influence of different filtering thresholds for 'maximum missing data' was conducted using two pools (WNZLL and FNZSL) for a range of missing genotype calls. This was examined to determine the optimal filtering threshold as there is a trade-off between retaining SNPs to provide sufficient genome coverage while ensuring quality of the SNP data. SNPs were first filtered to exclude multiallelic SNPs and retain only SNPs with a minimum and maximum read depth of 5 to 150. All filtering was performed using VCFtools v 0.1.16 (Danecek *et al.*, 2011). The dataset was split into two variant call format (VCF) files, one containing the five populations from the WNZLL pool, the other containing the five populations from the FNZSL pool. These pools were used to investigate a range of filtering thresholds (10, 30, 50, 70 and 90%) to determine the cut-off threshold for maximum percentage of missing samples per SNP. An average of the two pools was used to construct a

relationship between the percentage of SNPs surviving and the maximum percentage of missing samples allowed per SNP.

To provide the SNP marker dataset for downstream analysis, the marker set was restricted to high quality SNPs by only including biallelic SNPs with a minimum and maximum read depth range of 5 to 150, limiting missing genotype data to a maximum of 20% per SNP, and including SNPs with a MAF threshold of ≥ 0.03. Samples with a large proportion of missing data were removed from the dataset and negative control (blank) samples were removed after checking that they did not contain unduly high levels of data.

3.3.4.3 Genotyping by sequencing library quality control

A principal component analysis (PCA) was used to confirm consistency among and within GBS libraries. Positive control samples comprising a single genotype (S9), repeated in all 13 GBS libraries were used to assess SNP-calling consistency across all 13 GBS libraries. These samples are also referred to as “Control 1”. Duplicated DNA samples within a library were used to confirm SNP-calling consistency within libraries. All duplicated samples were checked but only two are presented for simplicity as they provided an example of the typical variation observed. One DNA sample from FNZSL-Low-Mid was repeated in GBS library 10, also referred to as “Control 2”; and one DNA sample from FNZSL-Low-End was repeated in GBS library 11, also referred to as “Control 3”. A VCF file was created by selecting the 13 positive (Control 1) samples and the two duplicated DNA samples from two different libraries (Controls 2 and 3) in addition to the sub set of the 1,222 individuals (*n* = 1,113) from which samples with a high proportion of missing data and duplicated samples had been excluded (see section 3.4.1). This VCF file was then filtered using the above-mentioned criteria (section 3.3.4.2) which resulted in 14,779 SNPs retained across the individuals. The VCF file was read into R using the function *read.vcfR()* and converted into a genlight object using the *vcfR2genlight()* function, both from the “*vcfR*” v 1.8.0 package (Knaus & Grünwald, 2017). Population information was added to the genlight object and the *glPca()* function from “*adegenet*” v 2.1.1 was used to perform the PCA and four principal components (PCs) were retained (Jombart, 2008). The individuals were plotted using PC1 and PC2 and were colour-coded so individual populations could be identified. The positive controls (Control 1 – 3) were highlighted in different colours on the PCA plot. As these samples were physically co-located on the biplot (see section 3.4.1), the S9 positive controls and DNA duplicates were removed from the dataset for subsequent analyses.

3.3.5 Analysis of population genetic structure

The genetic structure in the GBS SNP dataset was explored through a discriminant analysis of principal components (DAPC) (Jombart *et al.*, 2010), using the package "*adegenet*" v 2.1.1 (Jombart, 2008) for R software v 3.6.1 (R Core Team, 2019). For this analysis, 14,743 SNPs were used on 1,113 individuals (see section 3.4.1) from 24 populations as the WNZSL-Parent population was excluded due to a large amount of missing data for all samples. DAPC transforms the SNP data using PCA then performs a linear discriminant analysis (LDA) on the transformed data. The *find.clusters()* function was first used to detect the number of clusters in the populations, without prior assumptions of assignment based on knowledge of pool or population. This uses a *K*-means clustering algorithm that finds genetic clusters of individuals by maximising the variation between groups rather than the total variance of the sample. *K*-means was run sequentially from 1 to 40 genetic clusters (*K*) and the optimal clustering solution corresponded with the lowest Bayesian information criterion (BIC). Individual assignment from the *a priori* grouping to the *K*-means determined clusters was visualised and compared using the *table.value()* function from the R package "*ade4*" v 1.7-15 (Chessel, Dufour & Thioulouse, 2004).

DAPC is typically implemented using the function *dapc(),* which uses the transformed data to perform an LDA on the retained PCs. However, retaining too many PCs can lead to overfitting issues whereas retaining too few (with respect to the number of individuals) leads to losing too much genetic information with the consequence that the resulting model will not be sufficiently informative to accurately discriminate among groups. Therefore, the optimization procedure proposed by *adegenet* was used to determine the optimal number of PCs to retain for the DAPC analysis. The cross-validation method used to address this issue was the *xvalDapc()* function, and was run with 100 replicates from PC 1 to 50. The cross-validation procedure works by splitting the data into a training set (90% of the data) and a validation set (remaining 10% of the data). Members from each group, as identified by *K*-means, are selected by stratified random sampling, ensuring that at least one member of each group is represented in both the training and validation sets. DAPC is then run using the training set and the number of PCs retained is varied (i.e., 1 to 50), with DAPC repeated 100 times for each PC. The optimal number of retained PCs is determined by the accuracy of predicting the group memberships of the validation set, which is associated with the lowest root mean square error (Jombart *et al.*, 2010). DAPC aims to provide an efficient description of genetic clusters using synthetic variables, known as discriminant functions (DFs). As

part of the *dapc()* function, the number of DFs need to be specified. DFs are constructed as linear combinations of alleles that best separate the clusters (i.e., that have the largest between group variance and the smallest within group variance). This is achieved by submitting the retained PCs to an LDA, and the number of DFs that optimise the separation of individuals into pre-defined groups are determined based on the eigenvalue for each DF. Finally, a scatter plot of individuals grouped by *K*-means on DFs was created using the *scatter()* function in "*adegenet*" (Jombart, 2008), and the relationship between the inferred clusters was investigated.

Genetic variation was also assessed using analysis of molecular variance (AMOVA) and Weir and Cockerham's (1984) F-statistics. AMOVA was calculated in R using the package "*pegas*" v 0.11 to determine how genetic variation was partitioned within and between populations for all 24 populations, as well as between genetic clusters identified using the most supported *K*-value identified by DAPC (*K* = 11) with 10,000 permutations (Excoffier *et al.*, 1992; Paradis, 2010). A matrix of pairwise genetic differentiation between all population pairs from two possible population structures (the original grouping of *K* = 24 and the DAPC determined grouping of *K* = 11) was computed using the R package "*Hierfstat*" v 0.04-22 (Goudet & Jombart, 2015) using F_{ST} (Weir & Cockerham, 1984).

3.3.6 Genome-wide association study

A mixed-linear model implemented in the R package "*rrBLUP*" (Endelman, 2011) was used to perform an association analysis on a subset of individuals with both genetic and phenotypic information (605 individuals, see notes section in **Table S3.1** in Appendix 2 for a breakdown of number of individuals per pool). Markers for this analysis were filtered to retain those with a maximum of 50% missing data before the *A.mat()* function was used to impute missing values using the EM algorithm designed for GBS markers (Poland *et al.*, 2012a). This resulted in 5,757 SNPs used in the analysis. Population structure and family relatedness was accounted for with a kinship matrix calculated by rrBLUP from the genotypic data. To account for multiple testing, a Bonferroni correction was applied and markers passing the threshold at an α of 0.05 were considered statistically significant (Bonferroni, 1936). Manhattan and Quantile-Quantile (Q-Q) plots were created for each phenotypic trait: leaf area, water-soluble carbohydrate (WSC) and other nutritional attributes including soluble sugars and starch (SSS), ash, crude protein, neutral detergent fibre, acid detergent fibre and lipid content. Manhattan plots align SNPs according to their genome position and show their *p*-values for marker

associations for each trait on the *y*-axis. Q-Q plots plot expected *p*-values against the observed *p*-values, and where values deviate from the diagonal can indicate population structure and a poor genome-wide association study (GWAS) model.

3.3.7 Detection of loci under selection

Utilization of multiple methods to detect loci under divergent selection is recommended as commonality of outlier identification increases confidence in the results. Therefore, three approaches were used to analyse the 14,743 SNP dataset: PCAdapt (Luu, Bazin & Blum, 2017), BayeScan (Foll & Gaggiotti, 2008) and another F_{ST} outlier detection approach recently developed in the software package Kinship using Genotyping by sequencing with Depth adjustment (KGD, available from https://github.com/AgResearch/KGD.git) (Dodds *et al.*, 2015). Missing data for SNPs were not imputed for the following analysis methods.

Individuals used in the DAPC analysis were also split into five datasets using VCFtools (Danecek *et al.*, 2011) for the purpose of completing analyses to identify outlier loci. Outlier loci are SNPs that are responsible for differentiating high and low WSC populations. These datasets consisted of five VCF files with individuals split into their respective pools but with the Parent populations removed. Parent populations were removed so analysis methods could identify SNPs differentiating just the high and low WSC populations in each pool. All VCF files were converted into PLINK format (BED, BIM and FAM files) using PLINK v 1.9 (Purcell *et al.*, 2007). The R package "*PCAdapt*" v 4.0.3 was used to detect loci driving variation on the principal components (Luu *et al.*, 2017). The K_{PC} value (number of principal components) with the best fit to the data was determined using the scree test (Cattell, 1966), the Kaiser-Guttman criterion (Guttman, 1954; Kaiser, 1960, 1970), and interpretation of score plots with K_{PC} values higher than determined by the two formal methods (See Appendix 2, Supplementary Methods, Number of K_{PC} detection for PCAdapt analysis). Outlier SNPs based on the optimal K_{PC} value were then identified after correcting for false positives using the Bonferroni correction in each pool and outlier SNPs in common between pools were investigated (See Appendix 2, Supplementary Methods, Cut-off threshold for outlier SNP detection in PCAdapt). The "*qqman*" package v 0.1.4 (Turner, 2014) was used to create Manhattan plots of the pseudomolecule position of the SNP on the *x*-axis versus the $-\log_{10}$(*p*-value) of the Mahalanobis distance (Luu *et al.*, 2017) on the *y*-axis. The Bonferroni false-discovery thresholds, of $\alpha = 0.01$ and $\alpha = 0.05$, were plotted onto the

Manhattan plots for visualisation. Outlier SNPs are shown as “skyscrapers” – highly differentiating SNPs with high -$\log_{10}$(*p*-values) that appear above the threshold.

A complementary approach using differences in allele frequencies between populations to identify loci under selection was undertaken using both BayeScan v 2.1 and KGD-F_{ST} “*GBS-PopGen.R*” R script (Foll & Gaggiotti, 2008; Dodds *et al.*, 2015). For the BayeScan analysis, the five VCF files were converted into BayeScan format in R using the packages “*vcfR*” v 1.8.0, “*adegenet*” v 2.1.1, and “*Hierfstat*” v 0.04-22 (Jombart, 2008; Goudet & Jombart, 2015; Knaus & Grünwald, 2017). Based on the population structure analysis results (see sections 3.4.3.1 and 3.4.3.2), 11 genetic clusters were useful to describe the data, with little genetic differentiation observed between generations within the low WSC and high WSC divergent selections, respectively. For example, in the WUSLL pool, pairwise F_{ST} was 0.05 between the Low-Mid and Low-End populations. Therefore, for this analysis, the two high WSC and two low WSC populations were combined within each pool, resulting in five population pairs tested using the BayeScan and KGD-F_{ST} methods. This strategy was therefore focused on identifying SNPs that had undergone the largest change in allele frequency (i.e., between the high WSC and low WSC lines). The analysis was conducted separately for each pool so that SNPs under divergent selection could be traced back to each pool. Confirming that observed changes in allele frequencies are due to selection, rather than potentially being a consequence of random genetic drift, is important and is explained in the following section (3.3.7.1). Alpha values from BayeScan output can determine what type of selection an allele is under: neutral or diversifying. Positive alpha values suggest diversifying selection (‘adaptive’ genetic variation), while negative values suggest balancing or purifying selection. Balancing selection maintains diversity whereas purifying, or background, selection removes deleterious mutations (Fijarczyk & Babik, 2015; Cvijović, Good & Desai, 2018). BayeScan was run with default parameters (20 pilot runs with 5,000 iterations, followed by a burn-in of 50,000 iterations, and prior odds for the neutral model was 10). The *q*-value for each locus was calculated and a false discovery rate of $\alpha = 0.05$ was used to determine significant outlier loci that had positive alpha values.

In addition, a newly developed F_{ST} outlier detection method was implemented in KGD software. For this, the filtered VCF file was converted to a Reference Alternative file using the KGD *vcf2ra_ro_ao.py* python script and a separate file containing individual and population information was constructed. The *Fst.GBS.pairwise()* function was used to calculate approximate mean F_{ST} for each SNP between each population

pair, accounting for GBS read depth. The five population pairs were tested and SNPs with F_{ST} values greater than 0.3 and present in more than two pools at that threshold (both pool F_{ST} values for that SNP needs to be > 0.3) were defined as outlier SNPs. Manhattan plots were created using a modified "*qqman*" *manhattan()* script (Turner, 2014). Outlier SNPs from all three analyses (PCAdapt, BayeScan and KGD-F_{ST}) were visualised in a Venn diagram using the R package "*VennDiagram*" v 1.6.20 (Chen & Boutros, 2011).

3.3.7.1 Changes in genotypes due to selection over time

A complimentary approach to the methods described above is to observe the change in genotype frequencies from generation to generation. As an example, for a SNP under directional selection, a genotype frequency from Mid to End generations might be expected to trend upwards in the high WSC direction and downwards in the low WSC direction or vice versa. A trend such as this would imply that change is due to selection, rather than due to random genetic drift. Furthermore, it can show evidence for the type of sweep that has occurred in each pool. For example, a complete sweep will show allele fixation within a few generations. However incomplete sweeps will show beneficial alleles increasing in frequency but will not reach fixation. Each outlier SNP that was detected using two or more detection methods, in each population, was assessed in this way. Genotype proportions for each SNP were extracted using VCFtools *--extract-FORMAT-info* GT and patterns were investigated.

3.3.8 Linkage disequilibrium analysis and linked genes

The physical positions of outlier SNPs identified in two or more of the outlier analyses were used as a means of identifying potential candidate genes. For outlier SNPs located in introns and exons, the gene where they were residing was recorded as it is likely to be the best candidate. SNPs identified in coding regions of a gene were investigated further to determine if they were likely to affect protein function or structure. These included nonsense SNPs that lead to premature termination of translation and non-synonymous SNPs that lead to a change in amino acid that may affect the protein function, depending on the amino acid substitution and location. Geneious Prime 2019.1.1 (http://www.geneious.com/) was used to determine the position of the SNP in the protein and whether there was a synonymous or non-synonymous change. White clover genome annotations (Griffiths *et al.*, 2019), BLAST (Johnson *et al.*, 2008), UniProt (The UniProt Consortium, 2018) and STRING v 11.0 (Szklarczyk *et al.*, 2019) were used to identify genes and their functions.

For intergenic outlier SNPs (ioSNPs; outlier SNPs which occur outside of genes), identifying the genes with which the ioSNPs are potentially in linkage disequilibrium (LD) was determined after an analysis to estimate the global rate of LD decay. The estimated rate of LD decay was used to define an interval, inside of which candidate genes might lie. One of the limitations of the dataset in the current study was the low SNP density, which showed a mean of 16.1 (± 1.6 standard deviation) SNPs per Mbp. In addition, LD was expected to decay rapidly as white clover is a cross-pollinated species (Flint-Garcia *et al.*, 2003). This means that there would not be many SNP pairs that could be used to make LD estimates. Furthermore, while the white clover reference genome is organised into pseudomolecules, it was comprised of 22,100 scaffolds (Griffiths *et al.*, 2019). The combination of these three issues led to the development of criteria for calculating LD in the regions containing the thirteen ioSNPs. One ioSNP was used as one of a pair and LD was estimated between this and every other SNP within a prescribed physical distance. The prescribed distance used for calculating LD in these regions was determined using the criteria described below.

Firstly, the ambiguity of a SNP's physical position due to the genome being fragmented into 22,100 scaffolds (Griffiths *et al.*, 2019) needed to be addressed. LD can only be calculated with SNPs of a known physical distance apart. Therefore, a choice of window size needed to be determined. The average scaffold length for the white clover genome assembly is 38,072 bp, with the longest scaffold 734,507 bp long. As most of the ioSNPs would sit on scaffolds around the mean length, a window larger than 38,000 bp but no longer than 735,000 bp, would need to be used. Therefore, a window of 100,000 bp was chosen as it is approximately double the mean scaffold length. Secondly, in case ioSNPs were sitting on scaffolds less than 100,000 bp long, an additional criterion was added. There are gaps of unknown sequence represented by blocks of "N" in the white clover genome. Therefore, to avoid uncertainty pertaining to the physical position of SNPs, LD was only reported between SNPs if they were not separated from the ioSNP by a large gap (> 1,000 "N" blocks). LD estimates were calculated for each population within those pools where the SNP had been identified as an outlier, for each of the ioSNPs. Additionally, where LD could not be measured using the pools in which the SNP had been identified as an outlier, other pools were used instead. For example, LD for SNP 13_4850703 was calculated in populations from the FNZLL and WNZSL pools, although it was not identified as an outlier in those pools. Calculations of pairwise LD using the correlation (r^2) between an ioSNP and SNPs located within a 100,000 bp window, and an unlimited number of SNPs within the LD window, was achieved using the PLINK v 1.9 command: *plink -- bfile --r2 --ld-snp-list --*

ld-window-kb 100 *--ld-window* 99999999 *--ld-window-r2* 0 (Purcell *et al.*, 2007). Lastly, $r^2 \geq 0.25$ was used as a threshold to determine if a pair of SNPs was in LD (i.e., in non-random association).

3.4 Results

3.4.1 DNA isolation, genotyping by sequencing library evaluation and single nucleotide polymorphism filtering

Investigating genomic regions and single nucleotide polymorphism (SNP) loci under selection for foliar water-soluble carbohydrate (WSC) in white clover was facilitated by DNA extraction, genotyping by sequencing (GBS) marker genotyping and subsequent marker analyses in a series of populations selected for divergent WSC content. To this end, genomic DNA that was of high molecular weight (> 15 Kbp), free from RNA contamination and non-degraded was isolated successfully from 1,536 plants (**Figure 3.1**) with concentrations ranging from 3 to 61 ng μL^{-1}. For GBS library construction, a subset of 47 individuals were selected per population based on having sufficient DNA (> 10 ng μL^{-1}) for GBS. Evaluation of the 13 pooled GBS libraries, prior to sequencing, using the Agilent 2100 Bioanalyzer generated electropherograms for each library showing fragment sizes pre- and post-size selection (**Figure S3.2**, Appendix 2). Small adapter dimers were present in the pre-size selection libraries (88 bp) but were removed post size selection. Library fragment sizes were, therefore, successfully restricted to within the 193 – 313 bp range which focussed fragments to a range suitable for maintaining an adequate sequencing depth.

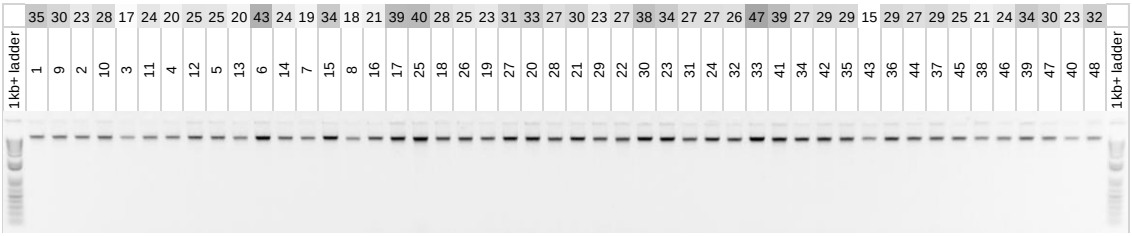


Figure 3.1 0.8% agarose (w/v) gel using lithium borate buffer containing 25 µg ethidium bromide showing DNA extracted from freeze-dried white clover leaf tissue using the 96-well plate method. Half of one plate (*n* = 48) is shown with the first and last lanes containing 5 µL of 1Kb plus DNA ladder (Invitrogen) which has a top fragment of 15 Kbp. Eight µL of 2.5x loading buffer and 2 µL of the DNA sample was loaded into each well. Concentration of each DNA sample (in ng μL^{-1}) is provided above the sample number and shaded according to high (dark grey) to low (white) concentration.

After sequencing, the data from all individuals (*n* = 1,248) in the 13 GBS libraries, including duplicates, were combined and reads were aligned to the reference genome using TASSEL v 5.0 to identify a total of 191,484 SNPs across all samples. An

investigation into the different filtering thresholds for maximum missing data was conducted using two white clover pools (WNZLL and FNZSL) and a range of 10, 30, 50, 70 and 90% missing genotype calls. Filtering to 10% missing samples per SNP retained 60% of the SNPs, 20% missing halved the number of surviving SNPs, and 90% missing left only 10% of the original 191,484 SNPs (**Figure S3.3**, Appendix 2). When using methods such as GBS there is a trade-off between the number of SNPs and the quality of SNPs. Enough SNPs need to be retained so more regions of the genome can be sampled, but the SNPs also need to meet quality requirements for downstream analyses. Analysis packages used in the current study require datasets with low levels of missing data. Therefore, only SNPs that had < 20% missing samples per SNP were retained for analysis (i.e., 80% genotyping rate), as this reduced the dataset by only half compared to more stringent filtering for missing values (**Figure S3.3**, Appendix 2). This threshold is similar to those applied in previous studies using GBS data, generally ranging from 10 to 50% missing samples per SNP (Mascher *et al.*, 2013; Alipour *et al.*, 2017; Eltaher *et al.*, 2018). Sequencing errors and low frequency SNPs are hard to distinguish, however false positives can be reduced by filtering for minor allele frequency (MAF) (Lu *et al.*, 2013). SNPs with a MAF of < 0.03 were therefore also excluded from further analysis. After filtering for depth, multiallelic, missing and MAF, a total of 14,743 SNPs were retained for 1,113 samples.

A total of 109 samples were removed across all populations. Fifteen samples were removed due to high missing data (> 80%) and one population (WNZSL-Parent) was removed completely due to a high proportion of missing data for all samples (> 80%). Positive control samples, a single genotype repeated in all 13 GBS libraries, were at first retained and checked for consistency across GBS libraries using a principal component analysis (PCA). Duplicated GBS data from 47 individuals were removed as they arose from duplicated technical replicates included for quality control and were not required for subsequent analysis. These duplicate DNA samples, within each library, were assessed prior to removal to assess consistency among replicated GBS libraries. Two examples (one DNA sample from FNZSL-Low-Mid was repeated in GBS library 10 i.e., Control 2; and one DNA sample from FNZSL-Low-End was repeated in GBS library 11, i.e., Control 3) were used to show the typical within-GBS library variation observed. The control samples were plotted as part of a large dataset and were found to cluster in close proximity (**Figure S3.4**, Appendix 2). The two samples repeated within the same library (Control 2 and Control 3) demonstrated that the same DNA within a GBS library produced minimal variation. The S9 controls (Control 1) were found to cluster together near the centre of the plot, therefore confirming GBS consistency across all 13 libraries

(**Figure S3.4**, Appendix 2). Having confirmed consistency of the GBS approach used, the S9 positive controls and the 47 duplicates of individuals were subsequently removed from further analyses.

3.4.2 Single nucleotide polymorphism distribution and density

The number of SNPs found on each pseudomolecule and the SNP density across the white clover reference genome was investigated using the 14,743 SNPs from 1,113 samples identified above. The relationship between the number of SNPs and pseudomolecule length (in Mbp) was examined. Pseudomolecules had been assigned to their relevant subgenomes, where pseudomolecules 1 to 8 belong to the *T. occidentale*-derived (Tr_{To}) subgenome, and pseudomolecules 9 to 16 belong to the *T. pallescens*-derived (Tr_{Tp}) subgenome (Griffiths *et al.*, 2019). The number of SNPs per pseudomolecule was strongly correlated to pseudomolecule length (coefficient of determination [r^2] = 0.96, **Figure S3.5**, Appendix 2), with higher numbers of SNPs on the longer pseudomolecules demonstrating that the SNPs are evenly distributed across the genome. The density of SNPs per Mbp and the distribution of SNPs across the genome was also recorded. A mean value of 16.1 (± 1.6 standard deviation) SNPs per Mbp was found across all pseudomolecules. The lowest SNP density was found on pseudomolecules 14, 10 and 6 with 12.5, 13.9 and 14.5 SNPs per Mbp, respectively. The highest densities were on pseudomolecules 5, 3 and 4 with 19.1, 18.0 and 17.7 SNPs per Mbp, respectively.

3.4.3 Population structure

Having generated the SNP dataset across the white clover populations selected for divergent WSC content, the next step was to use this resource to identify genomic regions or SNP loci that had been co-selected with the trait. This required determination of structure within and among the populations as a precursor to applying analyses to identify markers under selection.

3.4.3.1 Discriminant analysis of principal components

A preliminary assessment of population structure was undertaken using a discriminant analysis of principal components (DAPC) to determine how many clusters were described by the data and to validate the pre-defined genetic clusters (i.e., populations within pools). DAPC analysis was therefore made without any prior group assignment based on the known populations. Data were first transformed by PCA to speed up the subsequent clustering algorithm (*K*-means). To minimise the loss of information, 800

principal components (PCs) were retained, accounting for approximately 90% of the total genetic variation (**Figure S3.6**, Appendix 2). The *find.clusters()* function was then used to determine the number of clusters that best describe the data by maximising variation between clusters. The Bayesian information criterion (BIC) was used to identify the optimal number of clusters from a range of 1 to 40. The lowest BIC value (6538.781) corresponded to *K* = 11 (**Figure S3.7**, Appendix 2), therefore, this value was chosen. The assignment of individuals to the 11 clusters was compared with the *a priori* population grouping (**Figure 3.2**). The names of the 11 clusters were subsequently italicised for easy comparison with the *a priori* population names (non-italicised). There was a clear trend for the individuals in the high WSC populations within a pool to group together to form a high WSC cluster, e.g., WNZLL-High-Mid and WNZLL-High-End comprised the *WNZLL-H* cluster. The individuals from the low WSC *a priori* populations within a pool also grouped together to form a low WSC cluster, e.g., WNZLL-Low-Mid and WNZLL-Low-End comprised the *WNZLL-L* cluster. Interestingly, the individuals from all the Parent populations grouped in one cluster (*PARENT*), except for nine individuals from the WUSLL-Parent population that grouped with *WUSLL-H*. One to two samples from the FNZLL-Low-End, WUSLL-High-Mid and WUSLL-Low-End also clustered with the *PARENT* cluster. Aside from these few individuals, there was clear assignment of the two high WSC populations in each pool to one group, and the two low WSC populations in each pool to one group, with all the Parent populations grouping into one *PARENT* cluster. In addition to the assignment of individuals into clusters, DAPC can be used to provide a graphical representation of the relationship between the inferred clusters, which is described in the following paragraphs.

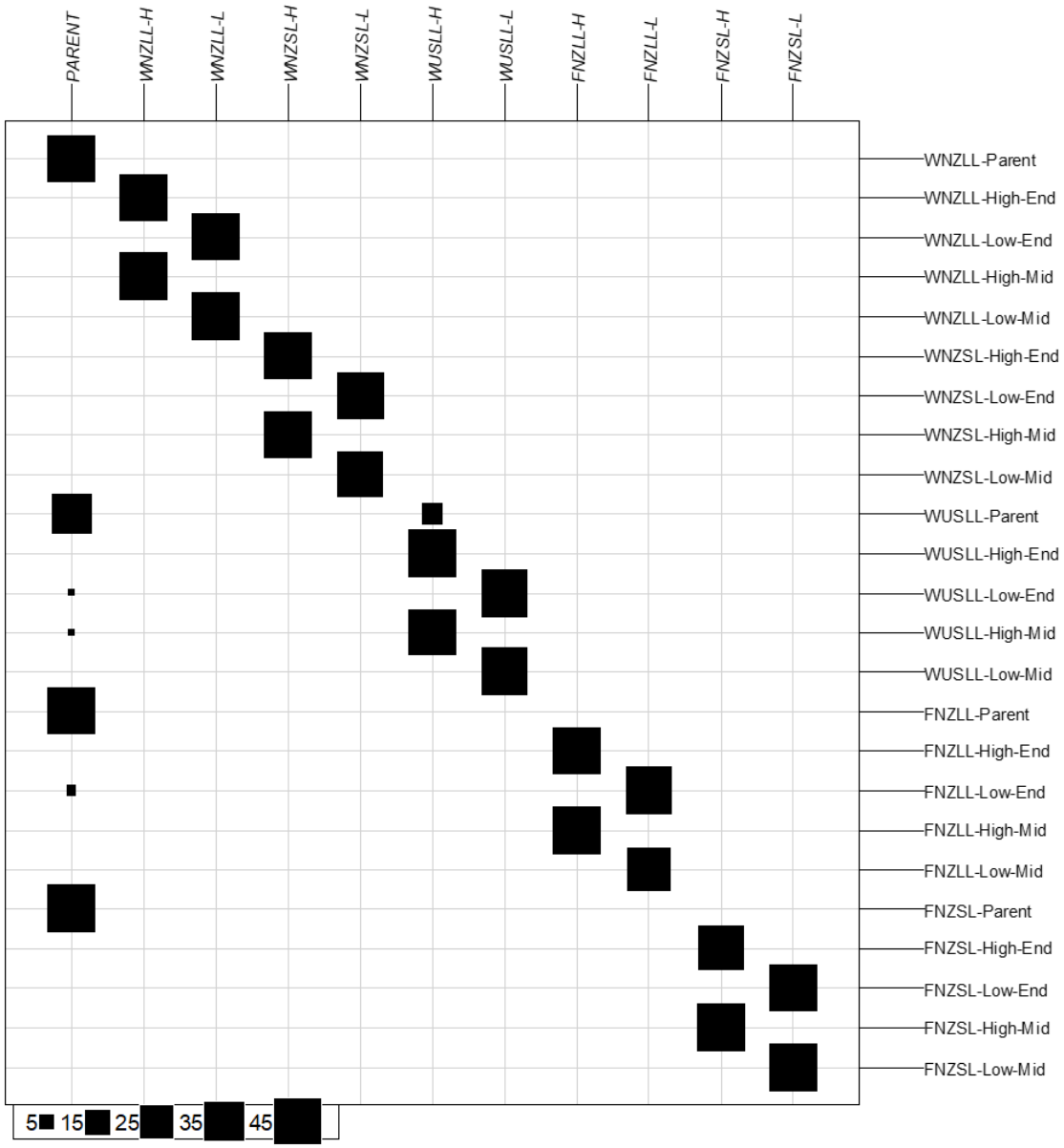


Figure 3.2 Group assignment based on the *find.cluster* function prior to discriminant analysis of principal components for 24 populations. Original populations are positioned horizontally and *K*-means determined clusters are positioned vertically. Size of black boxes represent the number of individuals assigned to the *K*-means determined cluster from the original population, with the scale presented in the bottom left-hand corner.
Note: High = high water-soluble carbohydrate (WSC), Low = low WSC, Parent = Parent generation, E = End generation, and M = Middle generation.

DAPC is typically implemented using the function *dapc()*, which performs a linear discriminant analysis (LDA) on retained PCs. However, retaining too many PCs can lead to overfitting issues and retaining too few leads to losing too much genetic information and poor discrimination. Therefore, the function *xvalDapc()* was used as a cross validation method to identify the optimal number of PCs to retain. The optimal number of PCs to retain corresponds with the maximum mean successful assignment (MSA), i.e., predictive success, and an associated root mean square error (RMSE). However, it is recommended that the optimal number of PCs to be used in the DAPC analysis should be chosen based on the lowest RMSE as sometimes the highest MSA and lowest RMSE do not correspond to the same number of optimal PCs (Jombart *et al.*, 2010). The *xvalDapc()* function determined the optimal number of PCs to retain was 43 (RMSE = 0.0033). However, both the RMSE and MSA plateaued at 6 PCs (RMSE = 0.0107 and MSA = 0.9928) with very little change thereafter (**Figure S3.8**, Appendix 2), therefore *dapc()* was run with 6 PCs retained. As part of the *dapc()* function, the number of discriminant functions (DFs) need to be specified. If some eigenvalues are much larger than others, it is recommended that only the DFs with the highest eigenvalues should be retained as they contain more information about the genetic structure of the data (Jombart *et al.*, 2010). The eigenvalues resulting from the LDA indicated that all 6 DFs captured the majority of genetic structure among the white clover populations (**Figure 3.3**, inset), therefore all six DFs were retained.

A scatter plot was drawn using the 11 clusters inferred by *K*-means, 6 PCs obtained by *xvalDapc()*, and the two axes represent the first two DFs of the DAPC analysis (**Figure 3.3**). The first DF showed a general separation of high WSC and low WSC populations with *High* clusters centred to the right of the plot, *Low* clusters centred to the left, and the *PARENT* plants clustering in the middle of the plot. The *WNZLL-H* and *FNZSL-L* clusters were clearly isolated from the bulk of the clusters. With respect to their counterpart populations (*WNZLL-L* and *FNZSL-H*), separation occurred on both the first and second DF. For *WNZSL* and *WUSLL*, Low and High population clusters showed very little separation on the first two DF and in fact clear separation was not observed on any DF (data not presented). The *FNZLL-L* and *FNZLL-H* clusters only showed clear separation on the fifth DF (**Figure S3.9**, Appendix 2).

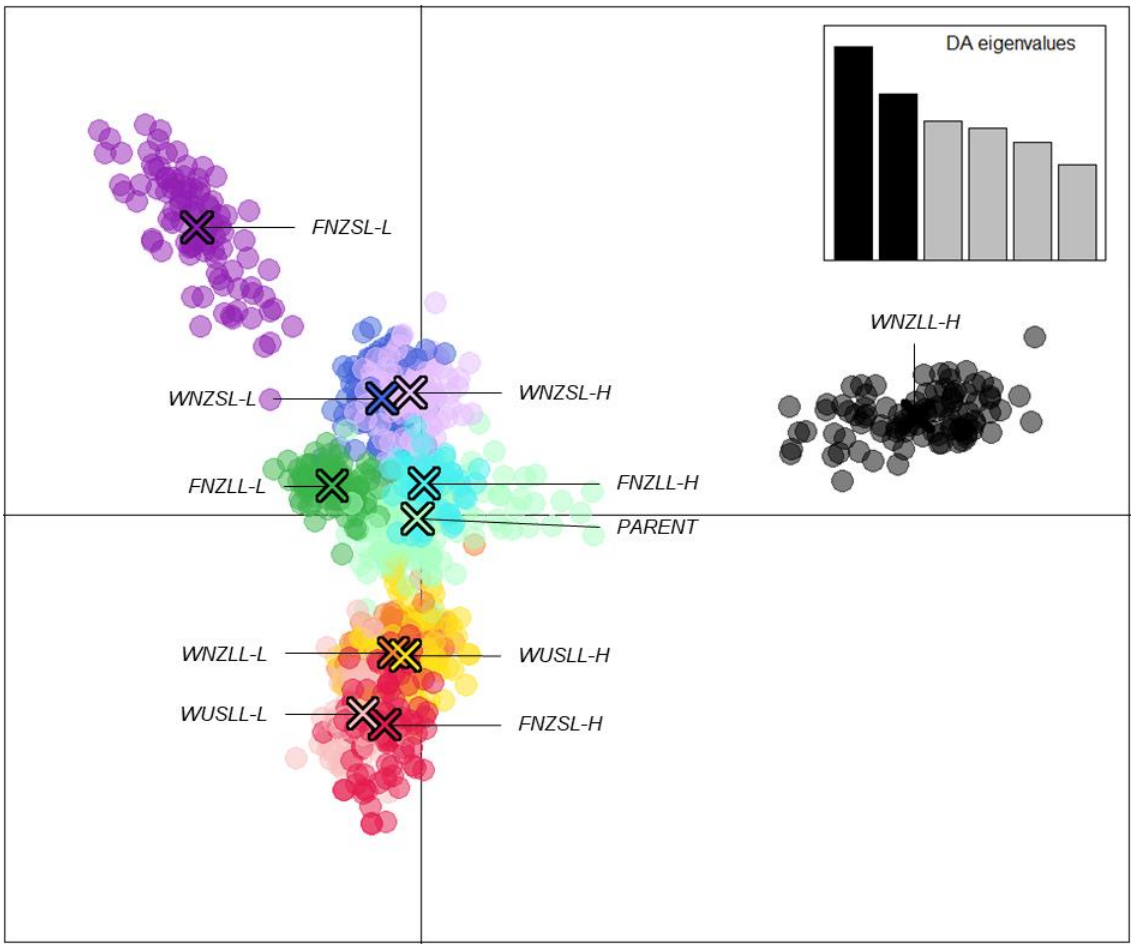


Figure 3.3 Discriminant analysis of principal components (DAPC) scatter plot of 1,113 individuals using 14,743 SNPs based on 11 assigned genetic clusters. Six principal components (**Figure S3.8**, Appendix 2) and six discriminant functions (DFs) were retained for analyses to describe the relationship between the genetic clusters. The scatter plot shows the first two DF from the DAPC analysis with the scree plot of eigenvalues of the linear discriminant analysis (LDA) shown in the inset. Populations are labelled and colour coded by *K* = 11 as determined from the *K*-means clustering algorithm. Each dot represents a single individual and the centre of each cluster, as determined by a minimum spanning tree based on the squared distances between populations, is indicated by a cross.

3.4.3.2 Pairwise F_{ST}

Pairwise F_{ST} was calculated both with no grouping other than the original populations (*K* = 24) and for the grouping identified above (*K* = 11). For biallelic marker systems (including SNPs), Wright (1978) suggests that values from 0 – 0.05 indicate little differentiation, 0.05 – 0.15 moderate differentiation, 0.15 – 0.25 great differentiation, and values above 0.25 indicate very great differentiation (Balloux & Lugon-Moulin, 2002; Hartl & Clark, 2007). For the *a priori* population grouping (*K* = 24), the WUSLL-Parent population showed moderate genetic differentiation from the NZ Parent populations

(0.06 – 0.08), while F_{ST} values were very low among the three NZ Parent populations (0.03 – 0.04). Pairwise comparisons of genetic differentiation (F_{ST}) indicated that the high WSC populations (Mid and End) within each pool were most similar to each other and the low WSC populations within each pool were most similar, with values ranging from 0.03 – 0.09. Pool generation pairs (e.g., WNZLL-Low-Mid and WNZLL-High-Mid) showed more genetic differentiation as indicated by an F_{ST} range of 0.12 – 0.22 (**Table 3.1**). Pairwise F_{ST} values based on the 11 *K*-means determined clusters showed moderate genetic differentiation between *High* and *Low* clusters within each pool, ranging between 0.13 – 0.19. Very low genetic differentiation was observed among the *PARENT* cluster and all the other clusters (0.07 – 0.10) (**Table 3.2**).

Table 3.1 Pairwise estimates of genetic differentiation among 24 white clover populations. Weir and Cockerham's (1984) F_{ST} is presented below the diagonal. Important population pairwise comparisons are underlined and in bold. F_{ST} values are shaded in a continuum of colours where a low F_{ST} (0.03) corresponds to green, an intermediate F_{ST} (0.13) corresponds to white, and a larger F_{ST} (0.23) corresponds to blue.

	WNZLL-P	WUSLL-P	FNZLL-P	FNZSL-P	WNZLL-LM	WNZLL-LE	WNZLL-HM	WNZLL-HE	WNZSL-LM	WNZSL-LE	WNZSL-HM	WNZSL-HE	WUSLL-LM	WUSLL-LE	WUSLL-HM	WUSLL-HE	FNZLL-LM	FNZLL-LE	FNZLL-HM	FNZLL-HE	FNZSL-LM	FNZSL-LE	FNZSL-HM	FNZSL-HE
WNZLL-P	-																							
WUSLL-P	0.06	-																						
FNZLL-P	0.04	0.06	-																					
FNZSL-P	0.03	0.08	0.03	-																				
WNZLL-LM	0.07	0.12	0.09	0.09	-																			
WNZLL-LE	0.10	0.15	0.12	0.11	**0.03**	-																		
WNZLL-HM	0.09	0.15	0.12	0.11	**0.17**	0.19	-																	
WNZLL-HE	0.10	0.16	0.13	0.12	0.17	**0.20**	**0.03**	-																
WNZSL-LM	0.08	0.11	0.09	0.09	0.14	0.16	0.16	0.17	-															
WNZSL-LE	0.10	0.13	0.12	0.11	0.16	0.18	0.19	0.19	**0.04**	-														
WNZSL-HM	0.08	0.10	0.10	0.09	0.14	0.16	0.17	0.17	**0.12**	0.14	-													
WNZSL-HE	0.11	0.12	0.12	0.12	0.16	0.18	0.19	0.19	0.15	**0.17**	**0.04**	-												
WUSLL-LM	0.11	0.12	0.12	0.13	0.16	0.19	0.20	0.21	0.15	0.18	0.16	0.19	-											
WUSLL-LE	0.11	0.12	0.11	0.12	0.15	0.17	0.20	0.21	0.15	0.17	0.16	0.18	**0.05**	-										
WUSLL-HM	0.08	0.07	0.08	0.10	0.14	0.16	0.17	0.17	0.13	0.15	0.12	0.15	**0.14**	0.14	-									
WUSLL-HE	0.12	0.11	0.13	0.14	0.18	0.20	0.20	0.21	0.16	0.18	0.16	0.18	0.19	**0.17**	**0.04**	-								
FNZLL-LM	0.11	0.13	0.11	0.11	0.16	0.19	0.19	0.20	0.15	0.17	0.15	0.18	0.19	0.18	0.15	0.18	-							
FNZLL-LE	0.14	0.15	0.14	0.14	0.19	0.22	0.22	0.23	0.18	0.20	0.18	0.20	0.22	0.21	0.18	0.21	**0.06**	-						
FNZLL-HM	0.10	0.10	0.08	0.11	0.15	0.18	0.18	0.19	0.15	0.17	0.15	0.16	0.18	0.17	0.14	0.17	**0.17**	0.19	-					
FNZLL-HE	0.14	0.14	0.12	0.14	0.19	0.22	0.22	0.23	0.18	0.20	0.18	0.20	0.21	0.21	0.18	0.21	0.21	**0.22**	**0.06**	-				
FNZSL-LM	0.11	0.14	0.11	0.09	0.16	0.18	0.18	0.20	0.16	0.18	0.16	0.18	0.19	0.19	0.17	0.20	0.18	0.20	0.17	0.21	-			
FNZSL-LE	0.12	0.16	0.13	0.11	0.17	0.19	0.20	0.21	0.18	0.19	0.18	0.19	0.21	0.21	0.18	0.22	0.19	0.21	0.19	0.22	**0.04**	-		
FNZSL-HM	0.11	0.14	0.11	0.09	0.16	0.19	0.18	0.20	0.16	0.19	0.16	0.19	0.20	0.19	0.16	0.20	0.19	0.21	0.17	0.21	**0.17**	0.18	-	
FNZSL-HE	0.13	0.16	0.13	0.11	0.17	0.20	0.20	0.21	0.18	0.19	0.18	0.21	0.21	0.20	0.17	0.21	0.20	0.22	0.20	0.23	0.18	**0.20**	**0.09**	-

Note: P = Parent, LM = Low-Mid, LE = Low-End, HM = High-Mid, HE = High-End. Where Low = Low water-soluble carbohydrate (WSC) and High = High WSC.

Table 3.2 Pairwise estimates of genetic differentiation among 11 white clover clusters determined from the *K*-means analysis. Weir and Cockerham's (1984) F_{ST} is presented below the diagonal. Important population pair comparisons are underlined and in bold. F_{ST} values are shaded in a continuum of colours where a low F_{ST} (0.07) corresponds to green, an intermediate F_{ST} (0.14) corresponds to white, and a larger F_{ST} (0.21) corresponds to blue.

	PARENT	*WNZLL-L*	*WNZLL-H*	*WNZSL-L*	*WNZSL-H*	*WUSLL-L*	*WUSLL-H*	*FNZLL-L*	*FNZLL-H*	*FNZSL-L*	*FNZSL-H*
PARENT	-										
WNZLL-L	0.08	-									
WNZLL-H	0.10	**0.18**	-								
WNZSL-L	0.08	0.15	0.17	-							
WNZSL-H	0.08	0.15	0.17	**0.13**	-						
WUSLL-L	0.09	0.15	0.20	0.15	0.16	-					
WUSLL-H	0.07	0.16	0.18	0.14	0.13	**0.14**	-				
FNZLL-L	0.10	0.18	0.21	0.17	0.17	0.19	0.17	-			
FNZLL-H	0.09	0.17	0.20	0.16	0.16	0.18	0.15	**0.19**	-		
FNZSL-L	0.09	0.16	0.19	0.17	0.17	0.18	0.18	0.18	0.18	-	
FNZSL-H	0.09	0.17	0.18	0.17	0.17	0.18	0.17	0.19	0.18	**0.16**	-

Note: L = low water-soluble carbohydrate (WSC) and H = high WSC.

3.4.3.3 Analysis of molecular variance

The analysis of molecular variance (AMOVA) from *K* = 24 revealed that the majority of genetic variation was partitioned within populations (77%), with the remainder of the variation partitioned among populations (23%). A hierarchical AMOVA for the 11 genetic clusters determined by *K*-means revealed that 17.5% of the variance was distributed among clusters, and only 6.6% was distributed among populations within clusters. Approximately the same amount of variance was found within populations or clusters (*K* = 24: 77%, $p < 0.001$; *K* = 11: 76%, $p < 0.001$), indicating that higher genetic variation is mainly distributed within populations (**Table 3.3**). Each of the DAPC, pairwise F_{ST} and AMOVA results indicate that genetically, the two high WSC populations for each pool can be grouped together and the two low WSC populations for each pool can be grouped together. Population groupings for subsequent outlier locus detection (section 3.4.5) were therefore based on this *K* = 10 grouping (Parent populations were excluded).

Table 3.3 Analysis of molecular variance (AMOVA) for different hierarchical levels of the 24 white clover populations using two different formulas and based on 14,743 SNPs. **A**) All 24 populations as 'population', **B**) eleven groups identified by *K*-means clustering algorithm as clusters, with the 24 populations as 'population'.

Source of variation	df	SS	MS	VC	%	*p*-value
A) *K* = 24 populations						
Among populations	23	981378.4	42668.6	859.3	23.3	<0.001
Within populations	1089	3075147	2823.83	2823.8	76.7	<0.001
B) *K* = 11 clusters						
Among clusters	10	797180.6	79718.1	651.6	17.5	<0.001
Among populations within clusters	13	184197.8	14169.1	245.2	6.6	<0.001
Within clusters	1089	3075147.2	3648.0	2823.8	75.9	<0.001

Note: df = degrees of freedom, SS = Sum of squares, MS = Mean square, VC = Variance components, % = Percentage of variation.

3.4.4 Genome-wide association study

Having completed the population structure analysis and investigated the partitioning of genetic variation via F_{ST} and AMOVA, the next step was to identify genomic regions or SNP loci that are under selection. A genome-wide association study (GWAS) was

carried out using 24 white clover populations that had both genotype and WSC phenotype data, and no markers were found to be significantly associated with WSC accumulation after correction for multiple testing (**Figure 3.4** and **Figure S3.10**, Appendix 2). However, three SNPs on pseudomolecules 1, 5 and 9 were ranked highly for both WSC and soluble sugars and starch (SSS) traits, and close to the false discovery threshold with $-\log_{10}$(*p*-values) greater than 3 for both traits (**Table S3.1**, Appendix 2). Aligning with the white clover reference genome, the SNP marker on pseudomolecule 1 was located in the coding region of a "*VPS35B*" (Vacuolar protein sorting-associated protein 35B) gene, the marker on pseudomolecule 5 was located 855 bp before a "*glgC*" (glucose-1-phosphate adenylyltransferase) gene, and the marker on pseudomolecule 9 was located in the coding region of "*UPL6*" (E3 ubiquitin-protein ligase) gene. The two SNPs located in coding regions conferred non-synonymous mutations as both SNPs altered the first base of a codon and subsequently the encoded amino acid. The SNP located in *VPS35B* caused an Ile (isoleucine) to Val (valine) change, while the SNP in *UPL6* caused a Gln (glutamine) to Glu (glutamic acid) change. The third SNP was located near a gene (*glgC*) that can be considered a prime candidate for a role in WSC accumulation, based on its inferred function. No significant SNPs were found for any other phenotypic trait, except for two SNP located on pseudomolecule 7 for Ash with a $-\log_{10}$(*p*-value) of 7.03 and 4.96. Both these SNPs were located in intergenic regions, > 10,000 bp away from genes, so functional association could not be determined.

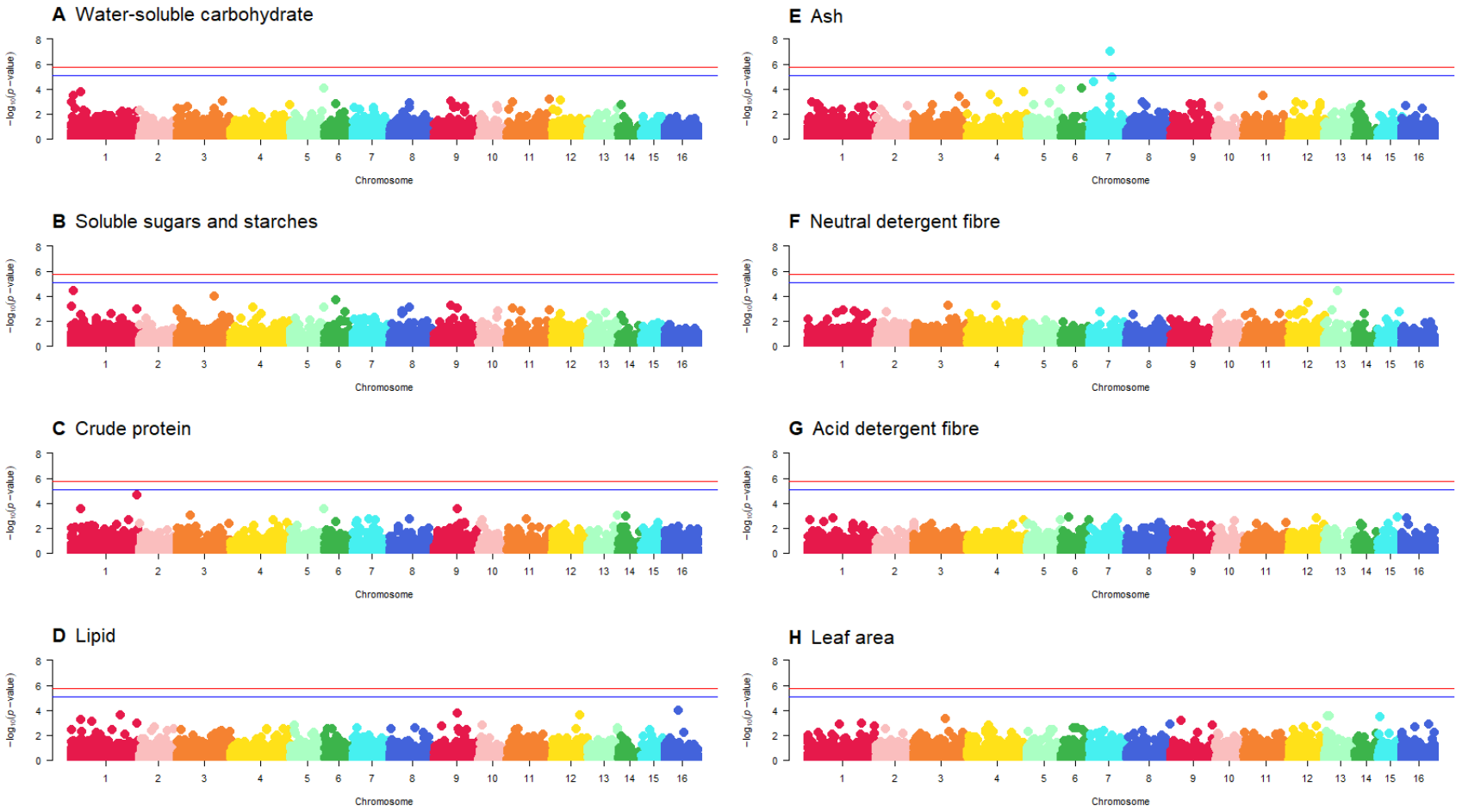


Figure 3.4 Manhattan plots from the genome-wide association study (GWAS) of eight phenotypic traits using 5,757 SNP markers and 605 individuals. -$\log_{10}$(*p*-values) are plotted against physical map position of SNPs with subgenomes of corresponding chromosomes (i.e., pseudomolecules) similarly coloured (Tr_{To} 1 – 8 and Tr_{Tp} 9 – 16). Significant loci lie above the false discovery rate thresholds as denoted by the red (α = 0.01) and blue (α = 0.05) solid lines. Quantile-Quantile plots for each trait are presented in **Figure S3.10**, Appendix 2.

3.4.5 Outlier loci detection

Another way to identify SNPs that may have been selected for during the breeding programmes is to use analyses that detect SNPs that differentiate populations. PCAdapt is used to identify SNPs corresponding to PCs that describe population structure and is a way of identifying SNPs that differentiate populations. Using several criteria, including the Kaiser-Guttman criterion (**Table S3.2**, Appendix 2), Cattell's scree test (**Figure S3.11**, Appendix 2) and interpretation of score plots, for which it was found that the second and third PCs did not ascertain population structure (**Figure S3.12**, Appendix 2), it was concluded that the K_{PC} value (number of PC's to investigate) in all four pools was K_{PC} = 1. The first PC captures the distinction between high and low WSC populations in all pools (**Figure 3.5**). Therefore, to identify SNPs related to WSC, we focused on the SNPs associated with PC1 only.

To control for false-discovery of outlier SNPs, a Bonferroni correction at α = 0.05 was used. The number of SNPs used for outlier detection in PCAdapt ranged from 10,976 to 11,479 per pool, with a mean of 11,133. The genome-wide significance thresholds were determined for each pool at α = 0.01 and α = 0.05 using their respective total number of SNPs (**Table S3.3**, Appendix 2). The average *p*-value threshold used for outlier detection at α = 0.05 was 4.49e-06, which is 5.34 on the log scale. Any SNPs associated with PC1 with $-\log_{10}$(*p*-values) larger than 5.34 were retained from each pool and identified as putative outliers (**Figure S3.13**, Appendix 2). To reduce detecting SNPs associated with population structure, SNPs identified as outliers were subjected to another criteria: they had to be present as outliers in two or more pools. Of the 643 total outlier SNPs, 36 were found in common between multiple pools based on PC1 (**Figure S3.14**, Appendix 2).

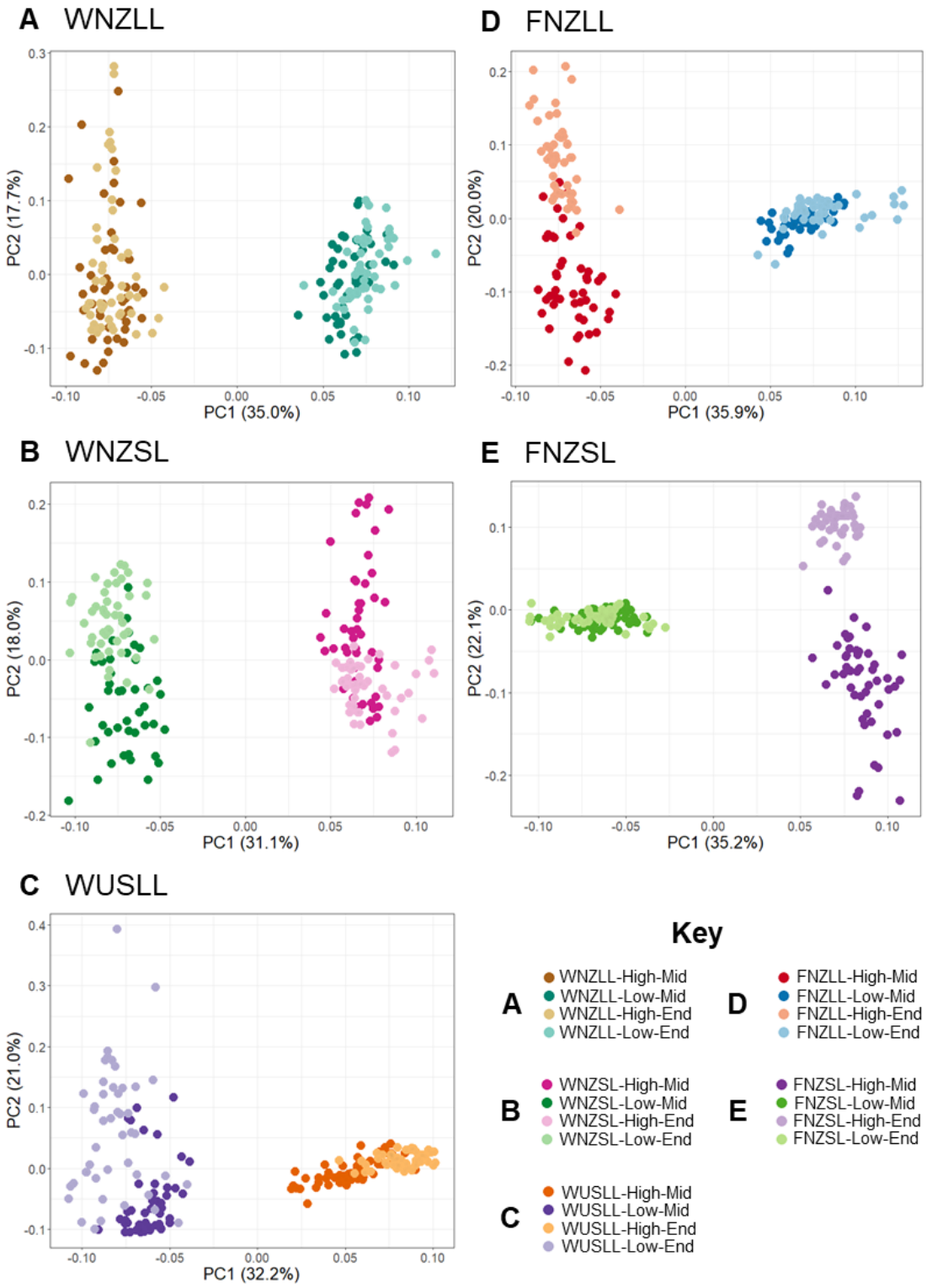


Figure 3.5 Score plots from PCAdapt analysis using the first two principal components (PC) for all five pools. Each dot represents a single individual and the colour corresponds to individuals from the same population. Each pool has four populations as the Parent populations were excluded from the analysis. Population information is displayed in the key in the bottom right corner.

A total of 329 outliers were detected using BayeScan at α = 0.05, with 27 in common between more than two pools (**Table S3.4**, Appendix 2, **Figure S3.15**, Appendix 2, and **Figure S3.16**, Appendix 2). All outliers identified by BayeScan exhibited positive alpha values, indicating that the program only detected SNPs putatively undergoing directional selection. The KGD-F_{ST} method detected the largest number of outliers including 1,188 total and 229 in common between more than two pools (**Figure S3.17**, Appendix 2 and **Figure S3.18**, Appendix 2). The strongest candidates for selection were 33 SNPs found in two or more methods of outlier detection (**Figure 3.6**). Unsurprisingly, the two F_{ST} based methods (BayeScan and KGD-F_{ST}) had the most SNPs in common (*n* = 22) whereas PCAdapt and BayeScan only had two SNPs in common and PCAdapt and KGD-F_{ST} had 13 SNPs in common.

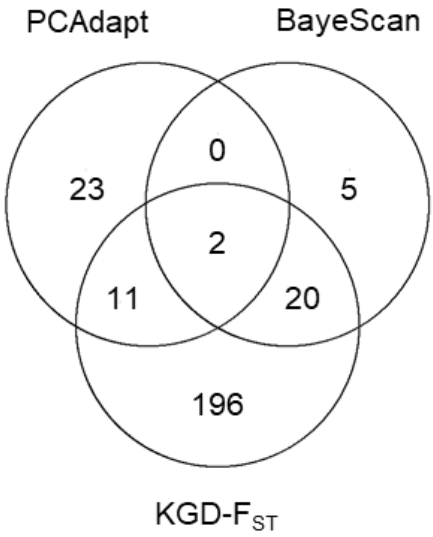


Figure 3.6 Venn diagram of the overlap between loci detected by PCAdapt, BayeScan and KGD-F_{ST}. PCAdapt and BayeScan false discovery threshold α = 0.05 was used, and SNPs in common between two or more pools at an F_{ST} > 0.3 was used for KGD-F_{ST} false discovery criteria. The 33 SNPs found as outliers in two or more analyses are considered to be strong candidates for selection.

Of the 33 candidate SNPs, five were found in exons, 15 were in introns and the remainder were either intergenic or in the promoter regions. One of the SNPs in the exons exhibited a synonymous mutation, while the other four had non-synonymous mutations leading to a change in amino acid (**Table 3.4**). A large number of the SNP-associated genes had unknown functions. To confirm SNPs were not changing due to

random genetic drift, generational changes in genotype and allele frequencies was investigated for the 33 candidates.

The vast majority of SNPs identified as outliers demonstrated a complete sweep where fixation of one allele occurred in the high WSC populations while the alternate allele became fixed in the low WSC populations within the first few generations (**Table 3.5**). This was observed for all 33 SNPs in two or more pools. Because these shifts were observed in two or more independent pools, it is less likely that the changes are due to random genetic drift than due to directional selection for WSC. Therefore, all changes in genotype frequency at all 33 SNPs were deemed to be due to selection rather than random genetic drift. There were examples where fixation was not achieved but genotype frequencies did shift across generations, for example in the WUSLL and FNZLL pools at SNP 2_6673787. Allele frequencies were fixed for the low WSC populations but there was a transition from Low-End and Low-Mid to the Parent populations and then to the High-Mid and the High-End populations whereby the alternate allele increased in frequency over the successive generations. In this case there was consistency between two different pools (WUSLL and FNZLL), indicating that selection rather than random genetic drift was the cause of the shift in allele frequencies (**Table 3.5**).

Table 3.4 Outlier SNPs detected by more than one outlier detection method (PCAdapt, BayeScan and KGD-F_{ST}), with genomic location and associated gene information. Detection method used to identify the SNP as an outlier is presented under the SNP ID in parentheses.

CHR	SNP position (bp)	SNP ID (Analysis)	Genomic region	Gene model ID and Gene annotation	Pools	Potential function of gene
1 (Tr_{To} – 1)	3,522,737	1_3522737 (BS + KGD)	Intron	chr1.jg449.t1 HIPP20, Heavy metal-associated isoprenylated plant protein 20 (*Medicago truncatula*)	WNZLL + FNZLL	Cadmium transport and detoxification
2 (Tr_{To} – 2)	6,673,787	2_6673787 (PCA + KGD)	1,246 bp from start codon 10,360 bp from start codon	chr2.jg957.t1 Unknown chr2.jg956.t1 Unknown	WUSLL + FNZLL	
2 (Tr_{To} – 2)	14,186,624 14,186,629	2_14186624 2_14186629 (BS + KGD)	Intron	chr2.jg2087.t1 OVA9, Glutamine-tRNA ligase (*M. truncatula*)	FNZLL + WUSLL	ATP binding and glutamine-tRNA ligase activity
2 (Tr_{To} – 2)	23,112,313	2_23112313 (BS + KGD)	16,487bp from start codon 14,165bp from start codon	chr2.jg3399.t1 FLA4 fasciclin-like arabinogalactan protein 4-like (*Trifolium pratense*) chr2.jg3403.t1 ribonuclease H (*T. pratense*)	WNZLL + FNZSL	Cell surface adhesion protein Intracellular protein transport
3 (Tr_{To} – 3)	28,211,144	3_28211144 (PCA + KGD)	Exon	chr3.jg4343.t1 FLA7, Fasciclin-like arabinogalactan protein 7 like (*T. pratense*)	WNZLL + FNZLL	Cell surface adhesion protein. Secondary wall biogenesis. CCC to CCT, retains Proline
4 (Tr_{To} – 4)	9,733,285	4_9733285 (PCA + KGD)	Intron	chr4.jg1351.t1 IRK, Probable LRR receptor-like serine/threonine-protein kinase IRK (*M. truncatula*)	WUSLL + FNZLL	ATP binding and protein kinase activity
4 (Tr_{To} – 4)	13,559,491	4_13559491 (BS + KGD)	Intron	chr4.jg1858.t1 FLA15 Fasciclin-like arabinogalactan protein 15 (*M. truncatula*)	WNZSL + WUSLL + FNZLL + WNZLL	Cell surface adhesion protein
4 (Tr_{To} – 4)	7,1509,072	4_71509072 (BS + KGD)	Intron	chr4.jg10116.t1 NPF4.4, Protein NRT1/ PTR FAMILY 4.4 (*Arabidopsis thaliana*)	WNZSL + FNZLL + WNZLL	Oligopeptide transmembrane transporter activity
6 (Tr_{To} – 6)	31,429,353	6_31429353 (PCA + KGD)	224 bp from start codon 674 bp from start codon	chr6.jg4668.t1 epoxide hydrolase (*T. pratense*) chr6.jg4669.t1 SPP Stromal processing peptidase (*Pisum sativum*)	WUSLL + FNZLL	Enzyme with hydrolase activity. Biosynthesis of cutin and plant defence to pathogens. Protein processing

Table 3.4 (continued)

CHR	SNP position (bp)	SNP ID (Analysis)	Genomic region	Gene model ID and Gene annotation	Pools	Potential function of gene
6 (Tr_{To} – 6)	31,429,365	6_31429365 (PCA + KGD)	212 bp from start codon 662 bp from start codon	chr6.jg4668.t1 epoxide hydrolase (*T. pratense*) chr6.jg4669.t1 SPP Stromal processing peptidase (*P. sativum*)	WUSLL + FNZLL	Enzyme with hydrolase activity. Biosynthesis of cutin and plant defence to pathogens. Protein processing
8 (Tr_{To} – 8)	40,904,996 40,905,002 40,905,003	8_40904996 8_40905002 8_40905003 (BS + KGD)	Intron	chr8.jg5802.t1 At4g03230 G-type lectin S-receptor-like serine/threonine-protein kinase At4g03230 (*A. thaliana)*	WNZLL + WNZSL	Carbohydrate binding, protein phosphorylation and plant signal under abiotic stress
9 (Tr_{Tp} – 1)	1,044,750	9_1044750 (BS + KGD)	Exon	chr9.jg173.t1 EXA1 Protein ESSENTIAL FOR POTEXVIRUS ACCUMULATION 1 (*A. thaliana)*	WNZLL + WUSLL + FNZLL + WNZSL	Translation repressor involved in defence response to pathogens T**C**T to T**A**T, swaps Ser to Tyr
11 (Tr_{Tp} – 3)	4,462,899	11_4462899 (BS + KGD)	Intron	chr11.jg696.t1 SMC5 Structural maintenance of chromosomes protein 5 (*A. thaliana)*	WUSLL + FNZLL	Protein binding and DNA repair
11 (Tr_{Tp} – 3)	9,345,053 9,345,077	11_9345053 11_9345077 (PCA + KGD)	1 bp from start codon 25 bp from start codon	chr11.jg1436.t1 Unknown	WNZLL + FNZLL + WNZSL	
11 (Tr_{Tp} – 3)	17,480,539	11_17480539 (BS + KGD)	Exon	chr11.jg2695.t1 Unknown	WNZSL + FNZLL	**G**AA to **C**AA swaps Phe to Gln
11 (Tr_{Tp} – 3)	21,109,404 21,109,408	11_21109404 11_21109408 (BS + KGD)	Intron	chr11.jg3295.t1 ROPGEF5 Rop guanine nucleotide exchange factor 5 (*A. thaliana*)	WNZSL + WUSLL + FNZSL	Guanine-nucleotide exchange factor involved in development
11 (Tr_{Tp} – 3)	46,932,936	11_46932936 (BS + KGD)	4,519 bp from stop codon	chr11.jg7041.t1 BRI1 Protein BRASSINOSTEROID INSENSITIVE 1 (*A. thaliana*)	WNZLL + FNZLL	Brassinosteroid signalling pathway, mediator of cell expansion.
11 (Tr_{Tp} – 3)	63,153,863	11_63153863 (BS + KGD)	Exon	chr11.jg9352.t1 FIB2 Mediator of RNA polymerase II transcription subunit 36a (*A. thaliana*)	WNZLL + FNZLL	Transcription regulation C**C**T to C**A**T, swaps Pro to His
12 (Tr_{Tp} – 4)	3,437,942	12_3437942 (PCA + KGD)	Intron	chr12.jg513.t1 Unknown	FNZLL + FNZSL + WNZSL	
12 (Tr_{Tp} – 4)	16,032,423	12_16032423 (PCA + BS + KGD)	Intron	chr12.jg2376.t1 RANB2 zinc finger Ran-binding domain-containing protein 2 isoform X2 (*Abrus precatorius*)	WNZSL + FNZLL	mRNA processing and RNA splicing

Table 3.4 (continued)

CHR	SNP position (bp)	SNP ID (Analysis)	Genomic region	Gene model ID and Gene annotation	Pools	Potential function of gene
13 (Tr_{Tp} – 5)	4,850,703	13_4850703 (PCA + BS + KGD)	2,119 bp from stop codon 7,115 bp from start codon	chr13.jg751.t1 Unknown chr13.jg750.t1 BHLH Basic helix-loop-helix protein (*T. repens*, *P. sativum*)	WNZLL + FNZSL	 Flavonoid biosynthetic process Anthocyanin synthesis
13 (Tr_{Tp} – 5)	23,675,463	13_23675463 (BS + KGD)	2,811 bp from start codon 9,488 bp from start codon	chr13.jg3560.t1 MIND1 Putative septum site-determining protein minD homolog, chloroplastic (*A. thaliana*) chr13.jg3559.t1 PPPDE thiol peptidase family protein, putative (*M. truncatula*)	WUSLL + FNZLL	chloroplast fission and negative regulation of cell division Unknown
13 (Tr_{Tp} – 5)	23,675,501	13_23675501 (BS + KGD)	1,713 bp from stop codon	chr13.jg3560.t1 MIND1 Putative septum site-determining protein minD homolog, chloroplastic (*A. thaliana*)	WNZLL + FNZLL	Cell division and chloroplast fission
14 (Tr_{Tp} – 6)	7,499,208	14_7499208 (BS + KGD)	Exon	chr14.jg1087.t1 DTX46 Protein DETOXIFICATION 46, chloroplastic (*A. thaliana*)	WNZLL + FNZLL	Xenobiotic transmembrane transporter activity, enhances tolerance to drought, salinity and cold stress. A**A**G to A**T**G, swaps Lys to Met
15 (Tr_{Tp} – 7)	6,515,376	15_6515376 15_6515380 (PCA + KGD)	101 bp from start codon 105 bp from start codon	chr15.jg971.t1 Unknown	WNZLL + WNZSL	
15 (Tr_{Tp} – 7)	17,315,253	15_17315253 (PCA + KGD)	348 bp from start codon	chr15.jg2597.t1 disease resistance protein rga3-like, partial (*T. pratense*)	FNZLL + FNZSL	Disease resistance
16 (Tr_{Tp} – 8)	32,428,574	16_32428574 (BS + KGD)	Intron	chr16.jg4564.t1 At1g19450 Sugar transporter ERD6-like 4 (*A. thaliana*)	WNZLL + FNZSL	Carbohydrate (glucose) proton symporter activity

Note: CHR = chromosome (i.e., pseudomolecule), Tr_{To} = white clover *Trifolium occidentale*-derived subgenome, Tr_{Tp} = white clover *T. pallescens*-derived subgenome, bp = base pairs, SNP ID = pseudomolecule number and bp position of the SNP, PCA = PCAdapt, BS = BayeScan, KGD = KGD-F_{ST}, Pools = specific pools that SNPs were identified as significant are listed.

Table 3.5 Genotype percentages for 33 outlier SNPs detected by more than one analysis method (PCAdapt, BayeScan and KGD-F_{ST}) in 24 populations from five pools. Genotypes are presented as AA, Aa and aa, where AA = homozygote for reference allele, Aa = heterozygote, aa = homozygote for alternate allele. Genotypes are colour coded on a continuum where a higher percentage (100%) corresponds to blue and a lower percentage corresponds to green (0%), with intermediate percentages (50%) as white. SNPs that were not identified as outliers in specific pools are coloured in grey. Detection method used to identify the SNP as an outlier is presented under the SNP name in parentheses.

SNP ID (Analysis)	Pool	WNZLL					WNZSL					WUSLL					FNZLL					FNZSL				
	Pop	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE
	AA	100	100	75	0	16.7	100	100	ND	100	100	ND	100	100	100	100	ND	0	100	100	100	ND	100	88.9	100	100
1_3522737	Aa	0	0	25	16.7	0	0	0	ND	0	0	ND	0	0	0	0	ND	0	0	0	0	ND	0	0	0	0
(BS + KGD)	aa	0	0	0	83.3	83.3	0	0	ND	0	0	ND	0	0	0	0	ND	100	0	0	0	ND	0	11.1	0	0
	n	16	26	8	6	12	11	15	ND	9	11	ND	1	14	8	6	ND	4	11	12	18	ND	9	18	12	1
	AA	100	100	54.2	90	100	84	74.1	ND	52.9	50	100	100	48.1	28.6	26.5	100	100	81	26.1	14.8	100	100	66.7	100	100
2_6673787	Aa	0	0	8.3	5	0	3.1	11.1	ND	20.6	26.5	0	0	25.9	28.6	17.6	0	0	14.3	21.7	11.1	0	0	6.7	0	0
(PCA + KGD)	aa	0	0	37.5	5	0	13	14.8	ND	26.5	23.5	0	0	25.9	42.9	55.9	0	0	4.8	52.2	74.1	0	0	26.7	0	0
	n	30	33	24	20	16	32	27	ND	34	34	30	30	27	28	34	27	26	21	23	27	28	22	15	9	16
	AA	8.3	4.8	60	0	0	100	100	ND	41.7	50	0	6.7	44.4	100	100	100	100	42.9	0	0	28.6	45.5	80	100	100
2_14186624 2_14186629	Aa	19.4	19	10	0	0	0	0	ND	8.3	10	21	6.7	0	0	0	0	0	14.3	0	0	21.4	27.3	10	0	0
(BS + KGD)	aa	72.2	76.2	30	100	100	0	0	ND	50	40	79	86.7	55.6	0	0	0	0	42.9	100	100	50	27.3	10	0	0
	n	36	21	10	6	4	3	1	ND	12	10	28	15	9	16	10	10	9	14	1	3	14	22	10	13	14
	AA	100	100	60	0	0	100	100	ND	65.2	100	100	100	ND	87.5	78.9	ND	ND	ND	100	100	100	100	0	0	0
2_23112313	Aa	0	0	0	0	0	0	0	ND	0	0	0	0	ND	0	0	ND	ND	ND	0	0	0	0	0	0	0
(BS + KGD)	aa	0	0	40	100	100	0	0	ND	34.8	0	0	0	ND	12.5	21.1	ND	ND	ND	0	0	0	0	100	100	100
	n	16	6	10	16	11	14	9	ND	23	3	24	12	ND	16	19	ND	ND	ND	4	1	7	4	1	20	16
	AA	100	100	75.9	10.3	2.2	58	60	ND	100	100	86	50	97.1	84.4	97.9	5.6	12.1	0	97.1	100	34.1	46.7	9.5	41.2	65
3_28211144	Aa	0	0	3.4	20.5	13.3	35	17.1	ND	0	0	2.8	0	2.9	11.1	2.1	16.7	9.1	0	2.9	0	27.3	33.3	0	32.4	20
(PCA + KGD)	aa	0	0	20.7	69.2	84.4	7.5	22.9	ND	0	0	11	50	0	4.4	0	77.8	78.8	100	0	0	38.6	20	90.5	26.5	15
	n	17	31	29	39	45	40	35	ND	37	42	36	14	34	45	47	36	33	7	35	41	44	45	21	34	40

Table 3.5 (continued)

SNP ID (Analysis)	Pool	WNZLL					WNZSL					WUSLL					FNZLL					FNZSL				
	Pop	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE
	AA	44.2	30.4	71.4	45.7	80.9	100	92.9	ND	48.9	68.4	100	88.9	67.9	14.3	22.7	71.4	62.5	64.3	7	0	38.9	36.7	40.9	69.2	85
4_9733285	Aa	34.9	50	16.7	47.8	19.1	0	0	ND	24.4	18.4	0	0	10.7	2.9	13.6	11.4	21.9	19	9.3	2.2	25	13.3	34.1	25.6	10
(PCA + KGD)	aa	20.9	19.6	11.9	6.5	0	0	7.1	ND	26.7	13.2	0	11.1	21.4	82.9	63.6	17.1	15.6	16.7	83.7	97.8	36.1	50	25	5.1	5
	n	43	46	42	46	47	10	14	ND	45	38	31	27	28	35	22	35	32	42	43	45	36	30	44	39	40
	AA	0	4.8	90.9	100	100	100	100	ND	0	0	100	100	46.7	8.7	0	100	100	60	27.3	20	100	100	50	100	100
4_13559491	Aa	0	0	0	0	0	0	0	ND	0	0	0	0	6.7	8.7	0	0	0	20	0	0	0	0	16.7	0	0
(BS + KGD)	aa	100	95.2	9.1	0	0	0	0	ND	100	100	0	0	46.7	82.6	100	0	0	20	72.7	80	0	0	33.3	0	0
	n	21	21	11	5	9	14	9	ND	2	9	33	33	15	23	19	2	15	15	11	5	6	3	6	5	3
	AA	ND	0	81.8	100	ND	100	84.6	ND	0	0	100	100	100	94.7	84	100	100	88.9	0	3.7	0	53.3	73.3	ND	ND
4_71509072	Aa	ND	0	0	0	ND	0	0	ND	0	0	0	0	0	0	0	0	0	0	0	0	0	13.3	13.3	ND	ND
(BS + KGD)	aa	ND	100	18.2	0	ND	0	15.4	ND	100	100	0	0	0	5.3	16	0	0	11.1	100	96.3	100	33.3	13.3	ND	ND
	n	ND	5	11	3	ND	13	13	ND	11	14	16	10	14	19	25	5	10	9	9	27	3	15	15	ND	ND
	AA	16.7	31.3	40.7	47.4	45.5	81	35	ND	32.4	27.8	81	90.3	25	21.1	7.7	0	12.5	73.7	83.3	100	100	100	46.2	0	ND
6_31429353	Aa	19	12.5	18.5	5.3	0	0	10	ND	8.1	13.9	0	6.5	15	10.5	2.6	0	6.3	5.3	0	0	0	0	26.9	0	ND
(PCA + KGD)	aa	64.3	56.3	40.7	47.4	54.5	19	55	ND	59.5	58.3	19	3.2	60	68.4	89.7	100	81.3	21.1	16.7	0	0	0	26.9	100	ND
	n	42	32	27	19	22	21	20	ND	37	36	16	31	20	38	39	14	16	19	18	22	23	34	26	2	ND
	AA	16.7	31.3	85.2	63.2	45.5	81	40	ND	83.8	52.8	81	90.3	30	28.9	7.7	0	12.5	94.7	83.3	100	100	100	76.9	100	ND
6_31429365	Aa	19	12.5	7.4	5.3	0	0	10	ND	5.4	33.3	0	6.5	20	13.2	2.6	0	6.3	5.3	0	0	0	0	11.5	0	ND
(PCA + KGD)	aa	64.3	56.3	7.4	31.6	54.5	19	50	ND	10.8	13.9	19	3.2	50	57.9	89.7	100	81.3	0	16.7	0	0	0	11.5	0	ND
	n	42	32	27	19	22	21	20	ND	37	36	16	31	20	38	39	14	16	19	18	22	23	34	26	2	ND
8_40904996	AA	100	100	73.3	8.3	0	100	100	ND	0	0	88	90.9	8.3	100	100	100	100	100	100	ND	ND	100	72.7	100	100
8_40905002	Aa	0	0	0	0	0	0	0	ND	7.7	0	4	0	0	0	0	0	0	0	0	ND	ND	0	0	0	0
8_40905003	aa	0	0	26.7	91.7	100	0	0	ND	92.3	100	8	9.1	91.7	0	0	0	0	0	0	ND	ND	0	27.3	0	0
(BS + KGD)	*n*	8	3	15	12	11	30	29	ND	13	7	25	22	12	5	11	2	9	8	5	ND	ND	3	11	11	32

Table 3.5 (continued)

SNP ID (Analysis)	Pool	WNZLL					WNZSL					WUSLL					FNZLL					FNZSL				
	Pop	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE
	AA	0	0	38.7	100	100	85	100	ND	33.3	9.7	0	0	66.7	95.8	100	10.5	15.4	36.7	75.9	88.5	56.5	21.4	52.2	19	63.6
9_1044750	Aa	0	0	6.5	0	0	5	0	ND	11.1	9.7	0	0	0	0	0	10.5	0	13.3	17.2	7.7	13	10.7	4.3	14.3	4.5
(BS + KGD)	aa	100	100	54.8	0	0	10	0	ND	55.6	80.6	100	100	33.3	4.2	0	78.9	84.6	50	6.9	3.8	30.4	67.9	43.5	66.7	31.8
	n	32	17	31	28	34	20	13	ND	27	31	4	4	12	24	21	19	13	30	29	26	23	28	23	21	22
	AA	100	100	100	100	100	100	100	ND	100	100	0	13.3	85.7	100	ND	0	0	100	100	100	28.6	55.6	100	100	100
11_4462899	Aa	0	0	0	0	0	0	0	ND	0	0	0	0	0	0	ND	0	0	0	0	0	0	0	0	0	0
(BS + KGD)	aa	0	0	0	0	0	0	0	ND	0	0	100	86.7	14.3	0	ND	100	100	0	0	0	71.4	44.4	0	0	0
	n	10	16	19	19	16	9	2	ND	7	14	11	15	7	13	ND	12	5	19	26	24	7	9	12	12	10
	AA	100	100	85.7	33.3	21.1	0	0	ND	100	100	100	100	88.5	100	100	100	100	77.3	3.6	16.7	100	100	95.7	100	100
11_9345053 11_9345077	Aa	0	0	4.8	6.7	10.5	0	0	ND	0	0	0	0	0	0	0	0	0	4.5	14.3	23.3	0	0	0	0	0
(PCA + KGD)	aa	0	0	9.5	60	68.4	100	100	ND	0	0	0	0	11.5	0	0	0	0	18.2	82.1	60	0	0	4.3	0	0
	n	5	10	21	15	19	19	9	ND	31	18	4	11	26	36	30	29	19	22	28	30	1	11	23	29	14
	AA	12.5	61.5	92.9	ND	0	100	100	ND	15.8	0	100	100	100	100	ND	0	21.1	100	100	100	100	100	72.7	ND	ND
11_17480539	Aa	0	7.7	0	ND	0	0	0	ND	0	0	0	0	0	0	ND	0	0	0	0	0	0	0	0	ND	ND
(BS + KGD)	aa	87.5	30.8	7.1	ND	100	0	0	ND	84.2	100	0	0	0	0	ND	100	78.9	0	0	0	0	0	27.3	ND	ND
	n	8	13	14	ND	3	17	15	ND	19	8	5	19	8	3	ND	27	19	9	20	8	7	5	11	ND	ND
	AA	100	100	75	100	100	68	66.7	ND	7.7	0	100	100	0	0	0	87.5	15.4	66.7	ND	ND	ND	0	100	100	100
11_21109404 11_21109408	Aa	0	0	0	0	0	12	4.2	ND	0	0	0	0	0	0	0	0	0	0	ND	ND	ND	0	0	0	0
(BS + KGD)	aa	0	0	25	0	0	21	29.2	ND	92.3	100	0	0	100	100	100	12.5	84.6	33.3	ND	ND	ND	100	0	0	0
	n	18	18	16	13	21	34	24	ND	13	2	9	3	4	3	3	8	13	3	ND	ND	ND	2	10	13	15
	AA	0	0	68.4	100	100	67	34.8	ND	ND	ND	0	0	83.3	33.3	42.9	100	100	21.4	0	0	ND	ND	84.6	100	100
11_46932936	Aa	0	0	0	0	0	6.1	8.7	ND	ND	ND	0	0	0	0	0	0	0	7.1	0	0	ND	ND	3.8	0	0
(BS + KGD)	aa	100	100	31.6	0	0	27	56.5	ND	ND	ND	100	100	16.7	66.7	57.1	0	0	71.4	100	100	ND	ND	11.5	0	0
	n	1	6	19	29	22	33	23	ND	ND	ND	2	8	6	9	7	3	6	14	16	12	ND	ND	26	25	19

Table 3.5 (continued)

SNP ID (Analysis)	Pool	WNZLL					WNZSL					WUSLL					FNZLL					FNZSL				
	Pop	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE
	AA	0	0	67.9	100	100	89	73.9	ND	100	100	100	ND	33.3	100	100	100	100	28.6	0	0	57.1	80	57.9	90.9	66.7
11_63153863	Aa	0	0	3.6	0	0	0	4.3	ND	0	0	0	ND	0	0	0	0	0	0	0	0	4.8	0	21.1	0	0
(BS + KGD)	aa	100	100	28.6	0	0	11	21.7	ND	0	0	0	ND	66.7	0	0	0	0	71.4	100	100	38.1	20	21.1	9.1	33.3
	n	2	6	28	15	4	18	23	ND	20	10	1	ND	6	18	18	6	8	7	13	9	21	10	19	22	18
	AA	11.1	50	16.7	0	0	33	64.7	ND	81.3	75	64	55.6	50	56.3	70.6	86.2	33.3	30	4	9.5	100	89.3	25	11.1	20
12_3437942	Aa	0	8.3	0	0	0	0	0	ND	0	0	7.1	0	0	0	17.6	13.8	0	15	0	0	0	7.1	6.3	11.1	0
(PCA + KGD)	aa	88.9	41.7	83.3	100	100	67	35.3	ND	18.8	25	29	44.4	50	43.8	11.8	0	66.7	55	96	90.5	0	3.6	68.8	77.8	80
	n	27	24	24	14	10	3	17	ND	16	16	14	18	2	16	17	29	6	20	25	21	35	28	16	18	10
	AA	90.9	95	83.3	93.2	100	4	66.7	ND	100	100	88	92.1	83.9	71	100	2.3	5.7	96.6	100	100	ND	44.4	73.1	66.7	96.7
12_16032423	Aa	4.5	2.5	5.6	6.8	0	4	4.2	ND	0	0	9.4	5.3	6.5	12.9	0	6.8	8.6	0	0	0	ND	0	11.5	9.5	0
(PCA + BS + KGD)	aa	4.5	2.5	11.1	0	0	92	29.2	ND	0	0	3.1	2.6	9.7	16.1	0	90.9	85.7	3.4	0	0	ND	55.6	15.4	23.8	3.3
	n	44	40	36	44	44	25	24	ND	31	40	32	38	31	31	28	44	35	29	20	30	ND	9	26	21	30
	AA	100	100	56.3	0	0	0	23.5	ND	33.3	80	100	ND	60	0	0	100	100	53.8	60.7	78.6	100	100	35.3	0	0
13_4850703	Aa	0	0	0	0	0	0	0	ND	0	20	0	ND	0	0	0	0	0	15.4	10.7	3.6	0	0	0	0	0
(PCA + BS + KGD)	aa	0	0	43.8	100	100	100	76.5	ND	66.7	0	0	ND	40	100	100	0	0	30.8	28.6	17.9	0	0	64.7	100	100
	n	30	24	16	12	9	3	17	ND	3	5	2	ND	10	13	23	12	6	13	28	28	28	15	17	12	11
	AA	61.5	61.9	55.6	ND	100	100	100	ND	86.7	88.9	100	100	81.3	11.1	11.1	100	ND	50	0	0	100	100	66.7	100	100
13_23675463	Aa	7.7	9.5	11.1	ND	0	0	0	ND	6.7	11.1	0	0	0	0	0	0	ND	0	0	0	0	0	0	0	0
(BS + KGD)	aa	30.8	28.6	33.3	ND	0	0	0	ND	6.7	0	0	0	18.8	88.9	88.9	0	ND	50	100	100	0	0	33.3	0	0
	n	13	21	9	ND	5	4	5	ND	15	27	17	3	16	9	9	2	ND	8	6	26	17	14	6	2	1
	AA	0	0	33.3	ND	100	100	20	ND	60	63	53	0	75	0	0	100	ND	50	0	0	82.4	100	50	0	100
13_23675501	Aa	0	0	0	ND	0	0	20	ND	26.7	22.2	0	33.3	6.3	0	0	0	ND	0	0	0	5.9	0	0	0	0
(BS + KGD)	aa	100	100	66.7	ND	0	0	60	ND	13.3	14.8	47	66.7	18.8	100	100	0	ND	50	100	100	11.8	0	50	100	0
	n	13	21	9	ND	5	4	5	ND	15	27	17	3	16	9	9	2	ND	8	6	26	17	14	6	2	1

Table 3.5 (continued)

SNP ID (Analysis)	Pool	WNZLL					WNZSL					WUSLL					FNZLL					FNZSL				
	Pop	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE	LE	LM	P	HM	HE
	AA	100	100	59.1	0	0	32	20	ND	40.9	0	71	73	50	63.6	44.4	ND	0	78.6	100	100	100	100	66.7	100	100
14_7499208	Aa	0	0	13.6	0	0	11	10	ND	13.6	0	0	13.5	0	13.6	0	ND	0	14.3	0	0	0	0	7.4	0	0
(BS + KGD)	aa	0	0	27.3	100	100	58	70	ND	45.5	100	29	13.5	50	22.7	55.6	ND	100	7.1	0	0	0	0	25.9	0	0
	n	6	15	22	23	16	19	10	ND	22	2	24	37	12	22	18	ND	3	14	13	20	14	21	27	24	19
	AA	27.5	25.6	52.5	100	100	94	63	ND	8.5	6.4	50	30.2	24.2	29.7	33.3	32.4	21.6	36.8	11.9	0	66.7	73	41	61.8	27.8
15_6515376 15_6515380	Aa	35	38.5	17.5	0	0	0	7.4	ND	25.5	14.9	20	34.9	33.3	27	16.7	35.3	48.6	21.1	9.5	4.5	24.2	10.8	25.6	29.4	30.6
(PCA + KGD)	aa	37.5	35.9	30	0	0	5.7	29.6	ND	66	78.7	30	34.9	42.4	43.2	50	32.4	29.7	42.1	78.6	95.5	9.1	16.2	33.3	8.8	41.7
	n	40	39	40	33	33	35	27	ND	47	47	40	43	33	37	36	34	37	38	42	44	33	37	39	34	36
	AA	29	19.4	51.6	20.7	41.4	26	65.5	ND	74.2	67.7	27	35.5	36	42.4	31	0	13.8	53.3	54.8	64.5	9.7	3.3	51.6	66.7	100
15_17315253	Aa	54.8	38.7	35.5	55.2	41.4	41	31	ND	19.4	32.3	58	48.4	32	51.5	44.8	3.2	37.9	30	38.7	35.5	32.3	30	38.7	29.2	0
(PCA + KGD)	aa	16.1	41.9	12.9	24.1	17.2	33	3.4	ND	6.5	0	15	16.1	32	6.1	24.1	96.8	48.3	16.7	6.5	0	58.1	66.7	9.7	4.2	0
	n	31	31	31	29	29	27	29	ND	31	31	33	31	25	33	29	31	29	30	31	31	31	30	31	24	31
	AA	0	0	90.9	100	100	ND	100	ND	100	100	0	0	27.3	0	0	ND	0	44.4	7.1	11.5	100	100	100	0	0
16_32428574	Aa	0	0	0	0	0	ND	0	ND	0	0	0	0	18.2	0	0	ND	100	0	14.3	11.5	0	0	0	0	0
(BS + KGD)	aa	100	100	9.1	0	0	ND	0	ND	0	0	100	100	54.5	100	100	ND	0	55.6	78.6	76.9	0	0	0	100	100
	n	4	1	11	21	17	ND	7	ND	13	13	1	8	11	22	26	ND	1	9	28	26	33	29	11	15	1

Note: Pop = Population, LE = Low-End, LM = Low-Mid, P = Parent, HM = High-Mid, HE = High-End, SNP ID = pseudomolecule number and bp position of the SNP, PCA = PCAdapt, BS = BayeScan, KGD = KGD-F_{ST}, *n* = number of individuals with data that were used to calculate genotype percentages, ND = No Data.

3.4.6 Linkage disequilibrium analysis and linked genes

After completing the outlier detection analyses, linkage disequilibrium (LD) analysis was performed to identify potential candidate genes. After applying criteria (section 3.3.8) for calculating LD for the 13 intergenic outlier SNPs (ioSNPs), LD could only be successfully calculated for four of the ioSNPs. This was not due to SNP density, as the nine remaining ioSNPs did have neighbouring SNPs within 100,000 bp, rather it was due to SNP allele frequencies in the populations under selection. Because LD is calculated based on allele frequency, LD cannot be calculated for SNPs that are homozygous in the population (allele frequency = 0 or 1). Because many of these ioSNPs exhibited complete sweeps, and therefore showed allelic fixation, LD could not be calculated. However, from the four ioSNPs where LD could be calculated, the maximum distance at which LD did not decay past $r^2 = 0.25$, was 10,303 bp (10 Kbp) (**Table S3.5**, Appendix 2). Therefore, for all ioSNPs, genes within a maximum distance of 10 Kbp either side of each ioSNP were recorded. When considering potential candidate genes, the upstream and downstream regulatory elements of the gene, including the promoter region, were also considered. This is because regulatory proteins (repressors and activators) can bind to distal regulatory sequences that are usually located within 1 Kbp of the transcription initiation site (Taiz & Zeiger, 2010a). If ioSNPs were located less than 1 Kbp away from the start codon of a gene, they were also classified as putatively in LD with the gene due to the proximity to promoter. The validity of the ioSNP was further demonstrated by observing the $-\log_{10}(p\text{-values})$ obtained in the PCAdapt analysis and comparing them to the LD measurements for neighbouring SNPs. This showed a common peak for the ioSNPs with both the LD and *p*-values but no peaks for the neighbouring SNPs that were not identified as outliers (an example is presented in **Figure S3.19**, Appendix 2). A total of 13 candidate genes were identified for 12 of the ioSNPs. One ioSNP (2_23112313) had two genes on either side but they were greater than 10 Kbp away (**Table 3.4**).

3.5 Discussion

3.5.1. Single nucleotide polymorphism discovery workflow from genotyping by sequencing data

Multiple factors influence the single nucleotide polymorphisms (SNPs) detected in genotyping by sequencing (GBS) data. These include enzyme choice, laboratory processes, GBS data analysis workflow and SNP caller which all affect the SNPs detected in a given sequence dataset. SNP-calling workflow and SNP distribution are discussed in greater detail below.

After GBS library preparation and sequencing, SNP-calling using TASSEL identified 191,484 SNPs across all 1,113 samples. It has been demonstrated that SNPs obtained from different workflows using the same raw dataset, have limited overlap (Torkamaneh, Laroche & Belzile, 2016; Wickland *et al.*, 2017; Ashby, 2019). This suggests that, in this study, only a subset of the available SNP variation in the GBS dataset is likely to have been utilised and is not fully comprehensive. It must be acknowledged that with reduced representation sequencing, typically less than 10% of the genome is normally sampled, therefore these methods will not sample SNPs in linkage disequilibrium (LD) with all of the genes under selection. This is further compounded by the issue of missing data and data quality. Although steps were taken to balance sequencing depth and coverage by utilizing a double restriction enzyme digest, and sequencing over two Illumina HiSeq lanes, there were still substantial missing data. Within the TASSEL workflow, reads are removed if they align to multiple positions (i.e., both sub-genomes) in the genome. After quality control (only including biallelic SNPs, a minimum and maximum read depth range of 5 to 150, limiting missing genotype data to a maximum of 20% per SNP, and including SNPs with a minor allele frequency threshold of ≥ 0.03) 14,743 high quality SNPs were retained, an approximate 10-fold reduction in the original SNP dataset. Thus, while the SNP-calling workflow and filtering results in an increase in the robustness of the data as ambiguous assignment of reads and poor-quality SNPs are removed, there is a marked reduction in the number of retained SNPs. Even after these filtering steps, there were missing data (51.8%), which limited available analyses to a few methodologies (PCAdapt, BayeScan and KGD-F_{ST}) that are able to manage some missing data.

For the genome-wide association study (GWAS) however, imputation was required as GWAS algorithms require complete datasets. Therefore, the current preferred imputation method (EM algorithm) for GBS data was used (Endelman, 2011; Poland *et al.*, 2012a). The EM algorithm is designed for GBS markers and produces low imputation error (ca. 30%) compared to an imputation method based on the mean (ca. 70%) (Poland *et al.*, 2012a). Although a random forest regression imputation method was slightly more accurate (imputation error of ca. 20%), the computational time required for imputation was substantially longer (22 hours) compared to the EM algorithm (3 minutes) (Poland *et al.*, 2012a). Hence, the EM algorithm is both accurate and time-efficient. This highlights the lack of population genomic analyses available that can utilise low coverage sequence-based data. GBS datasets, therefore, require onerous filtering or imputation in preparation for use with existing analysis packages. This, in turn, markedly reduces the number of available markers. Recently, in response

to the emergence of reduced representation sequencing methods such as GBS, increased emphasis has been placed on developing methodologies to address this challenge, with LD and genomic relationship analyses developed specifically for low-depth sequencing data (Dodds *et al.*, 2015; Bilton *et al.*, 2018). However, more analyses that accommodate missing data and focus on detecting selection need to be developed. Despite some limitations of GBS datasets, they are time and cost-efficient compared to existing marker platforms. For example, a white clover microsatellite (SSR) marker resource of approximately 1,000 SSRs is available (Griffiths *et al.*, 2013), but the marker density is insufficient for the purposes of the current study, and the time and effort to generate SSR genotype data on over 1,000 individuals with 1,000 SSRs is significant. Another common marker platform is a SNP chip, a resource that has yet to be developed for white clover. SNP chips array development requires sequencing and SNP calling using a large number of individuals and populations to produce a panel of SNPs, which is a large upfront development cost. Additionally, SNP chip data suffer from ascertainment bias due to the SNP panel of individuals used to select the SNPs (Heslot *et al.*, 2013). This ascertainment bias causes a bias towards common alleles in the data which can skew measures of population structure and relatedness among individuals (Nielsen, Hubisz & Clark, 2004; Albrechtsen, Nielsen & Nielsen, 2010). The GBS platform, therefore, is the marker of choice available to white clover that provides sufficient density across a range of populations for the analyses utilised in the current study.

3.5.1.1 Single nucleotide polymorphism distribution using the white clover reference genome

It has been shown that using a high-quality reference genome can provide numerous benefits to downstream genetic analyses (Benevenuto *et al.*, 2019). The current white clover genome assembly is fragmented with an N_{50} (length of the contig, when ordered from largest to smallest, at the point where 50% of the sequence data have been accounted) of 122 Kbp and 22,100 scaffolds aligned into pseudomolecules based on linkage mapping data (Griffiths *et al.*, 2019). Further data are required, including from long-read sequencing technologies, to fill the gaps and refine the assembly, and resolve any potential misplacement of scaffolds within sub-genomes. The data presented in the current study showed the fewest SNPs were observed on the smallest pseudomolecule (Tr_{Tp} 6 is 32.3 Mbp; 404 SNPs identified), while the largest pseudomolecule had the most SNPs (Tr_{To} 1 is 95.9 Mbp; 1,549 SNPs identified). A highly positive correlation was observed between the number of SNPs found on a pseudomolecule and the size of the

pseudomolecule (coefficient of determination [r^2] = 0.96). Studies using GBS data in other polyploid species (e.g., wheat) and other species from the legume family (e.g., chickpea and pea) have shown similarly high positive correlations (Poland *et al.*, 2012b; Alipour *et al.*, 2017; Ma *et al.*, 2017; Alipour *et al.*, 2019; Deokar, Sagi & Tar'an, 2019). Furthermore, the SNP assignment to pseudomolecules showed no bias towards one subgenome over the other. Previous studies using GBS data have been able to detect genes under selection with a similar marker density compared to the current study. For example, Biazzi *et al.* (2017) utilised 8,494 SNP markers to identify loci associated with forage quality traits in alfalfa (*Medicago sativa*). These marker positions were based on the reference genome of *M. truncatula* (barrel clover) which has a genome size of ca. 500 Mbp (Bennett & Leitch, 2011; Branca *et al.*, 2011), approximately half the size of white clover (1093 Mbp) (Bennett & Leitch, 2011), and resulted in a SNP density of ca. 17 SNPs per Mbp. Similarly, Inostroza *et al.* (2018) utilised 8,324 SNPs aligned to the *M. truncatula* genome as a reference to identify SNPs associated with cold tolerance in white clover and a SNP density of ca. 16.6 SNPs per Mbp was observed. As the SNP density observed in the current study (16.1 SNPs per Mbp) is comparable to previous studies that successfully identified loci associated with traits of interest, the marker set is expected to have sufficient coverage of the white clover genome to identify loci associated with water-soluble carbohydrate (WSC). Loci associated with WSC are further discussed in sections 3.5.6 and 3.5.7.

3.5.2 Population structure

A preliminary assessment of population structure is required for GWAS and outlier detection approaches as it can be a confounding factor in the analysis. This is because the association of a SNP with a trait may be a product of an underlying, unrecognised population genetic structure. By accounting for population structure, power to identify associations is increased and false positive associations are reduced. To investigate population structure and partitioning of genetic variation, three analyses were used. An analysis of molecular variance (AMOVA) revealed that the genetic variation within the 24 white clover populations accounted for about 77% of the total (**Table 3.3**). This indicates that there is little population structure among these populations as the majority of variance was detected within populations. Populations of outcrossing perennial plants tend to be genetically diverse and have less genetic differentiation among populations (Hamrick & Godt, 1996; Nybom, 2004). The variance among white clover populations identified by AMOVA (23%) is in accordance with this and aligns with earlier studies reporting higher variation within, rather than among, white clover populations and

cultivars. For example, 19% variation was partitioned among a diverse set of white clover populations (Collins *et al.*, 2012), 24.3% variation was partitioned among populations within three north-eastern states of the USA (Gustine & Huff, 1999), and 21.2% variation was partitioned among three white clover varieties from the UK, Netherlands and Denmark (Khanlou *et al.*, 2011). Higher levels of intra-population diversity, as observed here, can be attributed to obligate outcrossing (Hamrick & Godt, 1996), breeding system (Annicchiarico & Piano, 1995), and very recent human-associated range expansion (Zeven, 1991).

The *K*-means clustering algorithm implemented in the discriminant analysis of principal components (DAPC) analysis determined that the 24 population SNP dataset described 11 genetic clusters. The makeup of the clusters was supported by AMOVA and pairwise F_{ST} analyses. To assign 24 populations derived from five original pools to 11 clusters, the two high WSC populations (High-Mid and High-End) within each pool grouped together as a single cluster, as did the two low WSC populations (Low-Mid and Low-End), while all Parent populations co-located in a single cluster. AMOVA showed greater variation occurred among clusters (17.5%) than among populations within clusters (6.6%), indicating that populations within clusters were very similar. A common measure of population differentiation is the fixation index (F_{ST}). In this study genetic differentiation among the 24 recognised populations was low to moderate (F_{ST} = 0.03 – 0.23) according to the scale proposed by Wright (1978). High WSC populations within each pool had a low level of genetic differentiation (F_{ST} range: 0.03 – 0.09), as did the low WSC populations within each pool (F_{ST} range of: 0.03 – 0.06), whereas F_{ST} values among divergent populations (e.g., High-Mid vs Low-Mid and High-End vs Low-End) were much higher (F_{ST} range of: 0.12 – 0.22). This indicates that selection was successful in driving genetic separation between the high and low WSC populations. These observations suggest that selection within the first 2 – 3 generations for low and high WSC individuals (i.e., from Parent to Mid) produced the largest genetic changes, with fewer genetic changes occurring in the next 2 – 3 generations (i.e., from Mid to End). These observations are reflected in the observed phenotypes (**Figure 2.4**, Chapter 2): there was a mean difference in WSC among the Low-Mid and High-Mid populations of 54 grams per kilogram dry matter (g kg^{-1} DM), whereas there was a mean difference in WSC among the High-Mid and High-End populations of 16 g kg^{-1} DM, and a mean difference among the Low-Mid and Low-End populations of 8 g kg^{-1} DM (**Table 2.2**, Chapter 2). Furthermore, this pattern was also observed for the changes in genotype frequencies in the outlier SNPs, as the majority of these SNPs exhibited fixation for one genotype in the Mid generations and no changes were observed

between the Mid and End populations (**Table 3.5**). Thus, selection for high WSC white clover individuals can be achieved in a short time frame of 2 – 3 generations and for the purposes of identifying SNPs under selection for divergent foliar WSC, the population grouping that showed the most genetic differentiation (*K* = 11) should be used.

The DAPC result demonstrated that all the Parent populations grouped into one cluster, suggesting a lack of population structure at the parental source material level. However, the pairwise F_{ST} analysis showed the United States of America (US) material from the Widdup US large leaf (WUSLL) pool was slightly more genetically distinct from the New Zealand/Aotearoa (NZ) cultivars (F_{ST} 0.06 – 0.08), than the NZ Parent populations were among themselves (F_{ST} 0.03 – 0.04) (**Table 3.2**). It has been demonstrated that white clover has extremely large effective population sizes worldwide and exhibits negligible population structure on continental and global scales (George *et al.*, 2006; Olsen, Sutherland & Small, 2007; Kooyers & Olsen, 2012, 2013), with pairwise F_{ST} values < 0.03 in previous studies (Wright *et al.*, 2017; Inostroza *et al.*, 2018). Therefore, it is not surprising that, although F_{ST} values were slightly higher between different countries of origin, they were still very low. Furthermore, cultivars used in the WUSLL-Parent population included a mixture of US (e.g., Tillman and SRVR; Widdup *et al.*, 2015) and NZ material (e.g., Huia and Ranger; (Williams, 1983; Caradus *et al.*, 1995; Widdup *et al.*, 2010)). As the pedigrees of the Parent populations had an overlap in cultivars, it may be expected that the Parent populations should not show a great level of genetic differentiation, which is supported by the population structure analyses. All three methods suggest minimal population structure was observed in the 24 white clover populations. However, divergent selection has created a structure that differentiates high and low WSC populations and within the divergent lines there was low genetic variation.

3.5.2.1 BayeScan artificially inflated F_{ST} values due to hierarchical structure

After assessing population structure, a combination of three genome scan methods identified SNPs associated significantly with WSC levels that may also be linked to genes influencing this trait in white clover. However, the number of outlier SNPs (outlier SNPs are loci that are responsible for differentiating high and low WSC populations) varied among the methods with few significant SNPs in common (**Figure 3.6**). Unsurprisingly, the strongest overlap between methods was between the two F_{ST}-derived methods: BayeScan and KGD-F_{ST}. However, BayeScan F_{ST} values were higher than those estimated by KGD-F_{ST}, for all pairwise comparisons. For example, using

BayeScan in the Widdup NZ large leaf (WNZLL) pool, the minimum F_{ST} value determined for a SNP locus was 0.21 with a maximum of 0.69 and mean of 0.23. In the equivalent KGD-F_{ST} analysis the minimum F_{ST} value was 0.0, maximum of 0.99, and the mean was 0.04. Both sets of F_{ST} values fitted χ^2 distributions (data not presented), but the mean BayeScan F_{ST} values were higher.

BayeScan is relatively robust against confounding demographic processes, but strong selection, hierarchical structure, population bottlenecks and recent migration can impact this method and artificially inflate F_{ST} values (Hermisson, 2009; Narum & Hess, 2011; Lotterhos & Whitlock, 2014, 2015). Given the contrasting outcomes from BayeScan and KGD-F_{ST}, it was important to assess the potential for this having occurred with the current dataset and to make a determination as to the reliability of the outcomes. To test if F_{ST} values were inflated due to population structure, BayeScan was re-run using the WNZLL pool, this time including all five associated populations without combining the two high WSC populations together and without combining the two low WSC populations together (**Figure S3.20**, Appendix 2). F_{ST} values from 0 – 0.05 indicate little differentiation, 0.05 – 0.15 moderate differentiation, 0.15 – 0.25 great differentiation, and values above 0.25 indicate very great differentiation (Wright, 1978; Balloux & Lugon-Moulin, 2002; Hartl & Clark, 2007). Applying the BayeScan analysis to the WNZLL pool structured in this way altered the F_{ST} values to: the lowest F_{ST} value of 0.05, a mean F_{ST} of 0.096, and the highest F_{ST} of 0.65. Therefore, adding in the Parent population, and splitting the *WNZLL-H* and *WNZLL-L* clusters into two populations each, reduced the F_{ST} values dramatically (c.f. lowest F_{ST} = 0.21, mean F_{ST} = 0.23 and highest F_{ST} = 0.69). This indicates that analysing populations that are closely related causes an inflation of F_{ST}. Inflation of F_{ST} values due to population structure and a similar evolutionary history has been documented (Excoffier *et al.*, 2009; Eckert *et al.*, 2010b). It is also not unusual for minimum F_{ST} values to sit around 0.2 when strong selection is occurring (Narum & Hess, 2011). For example, Reinert *et al.* (2019) presented F_{ST} values from a BayeScan analysis to identify loci influenced by natural and artificial selection in barley that never dropped below 0.1. One of the benefits to BayeScan is that it runs two models for each locus: a neutral and a selection model. The posterior probability for both models is calculated and the posterior odds is used to provide more evidence for one model compared to the other. Log_{10}(posterior odds) ≤ 2 indicates decisive evidence for selection and corresponds to large positive alpha values. Therefore, not only do the SNPs need to have high F_{ST} values, but also large alpha values to be indicative of putative adaptive selection. Although the inflated F_{ST} values technically indicate all the SNPs are under selection, as F_{ST} values greater than 0.25

indicate very great genetic differentiation, the type of selection they are under can be determined from the alpha value. BayeScan calculates *q*-values for each locus, which is a test statistic directly related to the false discovery rate (FDR) and should be used to make decisions (Foll & Gaggiotti, 2008). Therefore, the F_{ST} values calculated by BayeScan and presented in this study should be disregarded and focus should instead be placed on the alpha values of SNPs above the FDR threshold as indicated by the *q*-value, which was implemented in the current study.

3.5.2.2 Mitigation of confounding population structure in outlier detection analyses and genome-wide association studies.

In outlier detection analyses, population structure can be controlled by a covariance matrix (Günther & Coop, 2013) or latent factors (Frichot *et al.*, 2013; Duforet-Frebourg, Bazin & Blum, 2014). When using PCAdapt, individuals are not sorted into predefined populations, instead population structure is ascertained using principal component analysis (PCA), which is similar to latent factors (Lotterhos & Whitlock, 2015), and determines population structure and outlier loci concurrently (Duforet-Frebourg *et al.*, 2014). Population structure in GWAS analysis methods are often controlled for by PCA (Price *et al.*, 2006), genomic control (Devlin & Roeder, 1999) or linear mixed models (Kang *et al.*, 2010). Methods based on linear mixed models that incorporate pairwise relatedness between individuals have been shown to capture population structure more effectively than other methods (Kang *et al.*, 2010; Sul & Eskin, 2013). For example, rrBLUP controls for population structure using a kinship matrix based on genetic diversity parameters from molecular markers, which avoids spurious associations arising by precisely modelling genetic relatedness among individuals (Endelman, 2011). To confirm population structure adjustment was successful in the GWAS, the distribution of $-\log_{10}$(*p*-values) from the association between each genotyped SNP and the phenotype of interest were assessed on the Q-Q plots. If population structure is successfully controlled, the $-\log_{10}$(*p*-values) will lie on the diagonal line which is the expected distribution under a model of no association (null hypothesis). As the $-\log_{10}$(*p*-values) did not deviate from the expected distribution (**Figure S3.10**, Appendix 2), population structure was appropriately controlled for in the GWAS. As mentioned above (section 3.5.2.1), population structure in the BayeScan analysis inflated F_{ST} values due to violated assumptions of evolutionary independence, however SNPs under selection were still successfully identified due to testing for local adaptation, i.e., locus *q*-values were used to determine whether a locus is under selection (Storey & Tibshirani, 2003). In the current study, additional criteria were imposed to reduce the number of outlier

SNPs detected that may be artefacts of population structure (see section 3.3.7). These included identifying outlier SNPs in common among multiple pools, assessing genotype changes among populations and comparing outliers among multiple analytical methods. The importance of utilising multiple outlier methods is outlined in the following section (section 3.5.3).

3.5.3 Importance of using multiple analytical methods to detect signatures of selection

It was noted that most SNPs identified as outliers demonstrated a complete sweep. That is, fixation of one allele occurred in the high WSC populations and the alternate allele was fixed in the low WSC populations. This was often achieved within a few generations as the Mid generations (i.e., generation two or three) showed fixation or near fixation for one allele in the High-Mid populations and fixation or near fixation occurred for the alternate allele in the Low-Mid populations. Because these SNPs exhibit clear changes in allele frequencies, multiple outlier detection methods were able to detect these strong differences. BayeScan was however unable to detect subtle changes in allele frequency such as an incomplete sweep. This was demonstrated in multiple pools for SNPs 2_6673787, 4_9733285 and 12_3437942, as examples. These SNPs each showed gradual increases in genotype AA (homozygote for reference allele) in the low WSC populations and genotype aa (homozygote for alternate allele) in the high WSC populations from the Mid to End generations. KGD-F_{ST} and PCAdapt detected these SNPs as outliers but BayeScan did not, highlighting a potential inability of BayeScan to be able to detect subtle changes in allele frequency. Under simulated strong selection, BayeScan can correctly identify all markers under selection. However under weak selection, only two of five markers were correctly identified (Narum & Hess, 2011), supporting the observation of BayeScan's inability to detect loci under weak selection. Because of this limitation, population pairwise comparisons within a pool could not be reliably undertaken using BayeScan. Therefore, in order to detect both types of selection (strong and weak e.g., complete sweep and incomplete sweep), comparing among populations that exhibit the largest differences in allele frequencies will be the most successful approach for identifying SNPs under selection. As the population structure results demonstrated that there was very little genetic differentiation observed between generations within a line (e.g., Low-Mid and Low-End), the largest changes in allele frequency would most likely be observed when comparing the high and low WSC lines (with the High-Mid and High-End populations grouped together and Low-Mid and Low-End populations were grouped together).

3.5.4 Genotype frequency changes provide evidence for selection rather than random genetic drift

An observation is that for 19 of the 33 outlier SNPs identified, fixation of the "A" allele (reference) is occurring in the high WSC populations for one pool, but fixation of the "a" allele (alternate) is occurring in the high WSC populations in a different pool. This is most likely due to phasing differences, where the SNP is either in coupling or repulsion with regards to the gene of interest. For example, a gene variant that is linked to increasing WSC in the high WSC populations in one pool (FNZSL) is linked to the "A" allele (e.g., SNP 16_32428574, **Table 3.5**), but in another pool (WNZLL) the "A" allele is linked but in the opposite phase (**Figure S3.21**, Appendix 2). Another observation is that the 33 outlier SNPs showed a shift from one homozygous state to the alternate homozygous state when moving from high to low WSC populations, with no SNPs exhibiting selection for the heterozygous form. This indicates that there was no heterozygote advantage detected for the high or low WSC trait. Another observation is that Parental populations showed fixation for one genotype (e.g., AA) but the Mid and End populations showed fixation for the alternate genotype (e.g., aa) with no heterozygotes or transition observed. This pattern (fixation for one allele in the Parent population but fixation of the alternate allele in one of the high WSC or low WSC populations) occurred at eight loci (FNZLL for 1_3522737 and 11_4462899; FNZSL for 2_23112313, 11_21109404, 11_21109408, and 16_32428574; WUSLL for 11_21109404 and 11_21109408). The most likely explanation for these observations is that some of the populations were under sampled, meaning that the population allele/genotype frequencies could not be reliably estimated. In three of these cases there was a very low number of samples representing the Parent populations (1 – 4 individuals) and a low number of samples in the Mid populations for seven of these cases (average of 3.3. samples). Therefore, detection of heterozygotes could simply have been missed due to under sampling. This may also result in possible sampling bias. It may be that if different individuals were sampled, the alternate allele would be observed in the population. This is possible as numerous populations had low sample sizes and rarer alleles may not have been captured.

An alternative explanation is heterozygote under calling, which is a potential weakness of GBS when there is low coverage and depth (Glaubitz *et al.*, 2014; Swarts *et al.*, 2014), as it is possible to observe only one of the two alleles for an individual, meaning that a heterozygous individual is observed as homozygous. To minimise this impact, a double restriction enzyme digest was chosen for the complexity reduction step,

which increases the genome coverage relative to a frequent cutting single enzyme digest, and the libraries were sequenced over two lanes to increase read depth (Poland *et al.*, 2012b). Furthermore, low depth SNPs were removed prior to analyses. Overall therefore, heterozygote under-calling should be a minor issue in this experiment and unlikely to be the underlying reason for homozygous genotypes observed in the Parental generations.

No outlier SNPs were found to change genotype frequencies due to random genetic drift (changes in allele frequencies in a population from one generation to another due to chance fluctuations, typically occurring in populations with small sample sizes), as confirmed by checking genotype frequencies. Therefore, candidate genes associated with outlier SNPs were investigated using physical position and proximity to genes as well as applying a linkage disequilibrium analysis for intergenic outlier SNPs. Functional annotations of genes were used to identify genes that are most likely associated with WSC and are discussed in the following sections (3.5.6 and 3.5.7).

3.5.5 Linking single nucleotide polymorphisms with candidate genes

Linkage disequilibrium (LD), the non-random association between alleles of adjacent loci, was used to determine an estimated distance for associating SNPs identified from GWAS and outlier analyses with candidate genes in the white clover populations. LD typically decays at short distances to r^2 (correlation between a pair of loci) = 0.2 – 0.25 in outcrossing plant species such as maize (*Zea mays* L., 100 – 1,500 bp) (Remington *et al.*, 2001; Tenaillon *et al.*, 2001) and perennial ryegrass (*Lolium perenne L*., 174 – 1,750 bp) (Faville *et al.*, 2018). The pattern of LD decay is linked to the genetic variability of a population, which results from the number and relatedness of the parent plants used to generate a breeding population. For example, in alfalfa (*Medicago sativa* L.), Li *et al.* (2014) found that LD decayed more rapidly in populations generated from 25 wild genotypes compared to breeding populations generated from 30 genotypes from modern cultivars. Furthermore, in perennial ryegrass, Faville *et al.* (2018) demonstrated that LD decayed less rapidly (r^2 = 0.2 in 1,506 bp) in a population that was generated from six pair crosses between randomly selected individuals from two cultivars. Alternatively, another population created by a larger number and a greater diversity of parents (a polycross of 80 randomly selected individuals from six different cultivars) exhibited more rapid LD decay (r^2 = 0.2 in 366 bp) (Faville *et al.*, 2018). Similarly, Auzanneau *et al.* (2007) demonstrated that LD decayed less rapidly in a population generated from six individuals from a narrow-based perennial ryegrass cultivar (Aberavon, 1,600 Kbp) compared to a population generated from 336 individuals from

a core collection (174 Kbp). These studies suggest that using genotypes from modern cultivars to create a breeding population can reduce LD decay, as LD decays more rapidly when the number of unrelated parents increases. The number of parent plants used to create the first generation in a breeding pool and their relatedness determines the initial LD, which then decreases as the number of generations increases (Auzanneau *et al.*, 2007). Therefore, populations developed from fewer founders and undergo fewer generations of selection will require a lower marker density for GWAS due to the slow LD decay. For the current study, 20 – 30 individuals from multiple white clover cultivars were used for crossing, ranging from four to six cultivars per pool. Given that the material used was exclusively from modern breeding cultivars, the F_{ST} values demonstrated that the material was not excessively diverse (pairwise F_{ST} range from 0.03 to 0.08 between the Parent populations), and there were only four or six generations created, we would expect slow LD decay in these populations. There is a lack of LD estimates published for white clover populations to date, with only one in the past two years in which LD was estimated using SNP markers. Inostroza *et al.* (2018) determined that LD decayed to r = 0.2 in 134 Kbp, using 192 genotypes from six naturalised white clover populations. This serves as a baseline example of "rapid" LD decay in white clover although the authors note that it is likely an overestimation due to the reference genome used (*Medicago truncatula*) which is smaller than the white clover genome. Therefore, for the material used in this experiment, we would expect LD to hold at distances equal to or longer than 134 Kbp.

Due to limitations of the current dataset, namely the low SNP density, LD decay was unable to be determined for each population and an alternative approach was implemented. This involved focusing on identified intergenic outlier SNPs (ioSNPs) and calculating LD between these SNPs and all other SNPs within a 100 Kbp window. Using this approach LD could be calculated for four of the 13 ioSNPs. The maximum distance at which LD did not decay past $r^2 = 0.25$, was 10,303 bp (10 Kbp). Information from nine of the ioSNPs that could not be used because they exhibited a complete sweep which fixed the allele frequencies for SNPs under selection meaning that LD with neighbouring SNPs could not be calculated. Although there were different alleles at the locus among the populations (both A and a), within the populations there is only one allele present (e.g., A), so r^2 is indeterminable. LD is calculated based on allele and haplotype frequencies (**Figure S3.22**, Appendix 2). For example, in a given population of 20 individuals, the following allele frequencies were observed: A = 1, a = 0, B = 0.3 and b = 0.7 (where A = reference allele at locus 1, a = alternate allele at locus 1, B = reference allele at locus 2 and b = alternate allele at locus 2); and the haplotype frequencies were

observed: AB = 0.3, aB = 0, Ab = 0.7 and ab = 0. Using the equations in **Figure S3.22** (Appendix 2), calculation of r^2 results in an error (Hill & Robertson, 1968). Hence, SNPs with fixed genotypes within a population cannot be used to calculate LD and detect linked loci.

Three of the four remaining ioSNPs exhibited incomplete sweeps, with non-fixed allele frequencies at the end point generations, and LD was able to be calculated in the pools where the SNP was shown under selection. By contrast, LD estimation for the fourth ioSNP, 13_4850703 required the use of data from pools in which the SNP was not under selection. The estimated rate of LD decay shown here is likely to be an under estimate of the actual LD decay in white clover due to the low number of SNPs used to calculate LD but a greater SNP density dataset would need to be used to investigate this further.

3.5.6 One single nucleotide polymorphism associated with water-soluble carbohydrate accumulation identified by outlier detection methodology

Sugars in plants play important roles as nutrients, signal molecules and solutes during stress. The vacuole is an organelle that stores solutes as nutrient reservoirs and also plays an important role in adaptation to salt stress and drought (Rizhsky *et al.*, 2004), as well as cold stress (Wormit *et al.*, 2006; Schulze *et al.*, 2012). Various abiotic stresses lead to the accumulation of sugars, in particular glucose and fructose (Wormit *et al.*, 2006), thus suggesting that vacuole sugar transporters play a role in these situations (Pertl-Obermeyer *et al.*, 2016).

In the past two decades, vacuole sugar transporters have been characterised in detail and are mainly membrane proteins that belong to the major facilitator superfamily (*MFS*), which has two subfamilies known as sucrose transporters (*SUTs*) and monosaccharide transporters (*MSTs*) (Pertl-Obermeyer *et al.*, 2016). Within the monosaccharide transporters, there is a family of vacuolar sugar transporters called early responsive to dehydration (*ERD*) which contain importers and exporters of vacuolar sugars (Büttner, 2007). *ERD* six-like transporters are involved in energy-independent sugar efflux from the vacuole (Wei *et al.*, 2014) and are induced by dehydration and cold treatment (Kiyosue *et al.*, 1998). There are 19 known *ERD6*-like genes in *Arabidopsis* and expression has been shown to vary from homologue to homologue in *A. thaliana* (Büttner, 2007). For example, expression of *ERD6* in *A. thaliana* has been shown to decrease in leaves after exposure to high salinity and ABA

treatment, however *ERD6-like 1* (*ESL1*) was only induced by ABA treatment and not salt stress (Yamada *et al.*, 2010).

Based on LD estimates a white clover homologue of *ERD6-like 4* (At1g19450) was found to be physically associated with one outlier SNP, detected in both the WNZLL and FNZSL pools. Endler *et al.* (2006) identified *ERD6-like 4* as being most homologous to the sugar beet (*Beta vulgaris*) hexose transporter *U43629* (Chiou & Bush, 1996). This sugar transporter has been hypothesised to catalyse facilitated diffusion (substrate transport down a concentration gradient without the additional input of energy) of glucose across the vacuole membrane (Chiou & Bush, 1996). In plants it is uncommon for monosaccharide transporters to function as facilitators (Büttner, 2007; Taiz & Zeiger, 2010b). However, in the last decade Sugars Will Eventually be Exported Transporters (*SWEET*) proteins were discovered (Chen *et al.*, 2010), which function as energy-independent uniporters of sucrose and glucose at the plasma membrane (Chen *et al.*, 2010; Chen *et al.*, 2012), and fructose on the tonoplast membrane (Chardon *et al.*, 2013). The above-mentioned evidence suggests a role for *ERD6-like 4* in WSC accumulation. It is possible that *ERD6* homologs are involved in movement of sugars out of the vacuole during circumstances where carbohydrate reallocation is important (Büttner, 2007). Although there is compelling evidence to suggest *that ERD6-like 4* plays a role in WSC accumulation, possibly in the form of osmotic adjustment, further experiments need to be done to test this hypothesis for *ERD6-like 4* protein expression leading to an increase or decrease in WSC content in white clover plants.

3.5.7 Putative candidate genes under selection identified from genome-wide association study

GWAS based on a subset of the 24 white clover populations failed to identify SNPs significantly associated with WSC levels, when accounting for false discovery testing, although three SNPs were highly ranked in both the WSC and soluble sugars and starch (SSS) trait analyses. The false discovery rate applied was controlled by Bonferroni's multiple testing correction method, which has been suggested to be too stringent (Hirschhorn & Daly, 2005; Wang *et al.*, 2005) and applying conservative thresholds has shown to exclude real associations (Yang *et al.*, 2010). Other studies have instead used a conservative *p*-value ($p < 0.001$) approach to reduce Type I error, whereby SNPs that are the highest ranking and have $-\log_{10}(p\text{-value}) \geq 3.0$ should be reported (Kang *et al.*, 2015; Biazzi *et al.*, 2017; Sakiroglu & Brummer, 2017). Applying conservative thresholds may exclude real associations (Yang *et al.*, 2010), therefore the biological significance

of the gene the SNP is located near should also be investigated (Zhang, 2016). As previously mentioned, the Q-Q plots produced from the GWAS in this study indicated population structure was successfully controlled as most *p*-values were similar to the expected *p*-values, with a few SNPs in both the SSS and WSC plots exhibiting deviation from the line with $-\log_{10}(p\text{-values}) > 3.0$. Candidate genes in close proximity to SNPs that were in common to both traits (WSC and SSS) and highly ranked were investigated. The first SNP 1_2338028 was in the exon of *VPS35B*, which plays a role in vesicular protein sorting (Yamazaki *et al.*, 2008) and is involved in plant growth and leaf senescence.

Another SNP (5_47903593) may be important for WSC accumulation based on the gene annotation. This SNP was located 855 bp away (upstream of the promoter region) from the start codon of *glgC* (glucose-1-phosphate adenylyltransferase) which encodes the small subunit of ADP-glucose pyrophosphorylase (ADPG-Ppase) (Weber *et al.*, 1998). This subunit is involved in starch biosynthesis, which in turn may affect glucose and sucrose concentrations (Burgess *et al.*, 1997). As the SNP was located in the promoter region, we cannot determine whether *glgC* transcription is initiated, reduced or inhibited; nor if the SNP is linked to a functional part of the gene. However, there is an expectation that higher *glgC* activity confers greater levels of starch (Ballicora, Iglesias & Preiss, 2004). It is speculated that increased levels of sucrose may be responsible for inducing expression of this gene as expression was highest in tissues with a high carbohydrate content, e.g., seeds in *Pisum sativum* (Muller-Rober *et al.*, 1990; Burgess *et al.*, 1997). Two transcripts encode the small subunit of ADPG-Ppase in both *Vicia faba* (Weber *et al.*, 1995) and *P. sativum* (Burgess *et al.*, 1997) and based on genome annotations (Griffiths *et al.*, 2019), this is also the case in white clover. Other dicots such as *Arabidopsis*, potato, spinach and sugar beet have only one small subunit, suggesting that the duplication occurred after legumes diverged from other dicots (Morell *et al.*, 1987; Muller-Rober *et al.*, 1990; Nakata *et al.*, 1991; Villand, Olsen & Kleczkowski, 1993; Müller-Röber, Nast & Willmitzer, 1995; Burgess *et al.*, 1997). *Psagps2* in *P. sativum*, the orthologue of white clover *glgC*, had variable expression across a range of tissues, with the highest expression levels observed in seeds (Burgess *et al.*, 1997). *VfAGPC* in *V. faba* is expressed only in the cotyledons, with no transcripts detected in leaves or flowers (Weber *et al.*, 1995). Given the lack of *VfAGPC* expression in leaves in these species, expression of *glgC* in white clover leaves may be hard to detect. This may impact the ability to detect this protein in the transcriptomic and proteomic study presented in Chapter 4. Another study (Zhang *et al.*, 2015) found the large subunit of *glgC* (Glucose-1-phosphate adenylyltransferase large subunit) to be

associated with drought resistance in *Medicago sativa*. This suggests that *glgC* may also play a role in osmotic adjustment or osmoprotection, as osmotic adjustment is the capacity of cells to accumulate solutes and use them to lower water potential during periods of osmotic stress. Low molecular weight sugars are often involved in osmotic adjustment which adds to the confidence of associating *glgC* to WSC accumulation in white clover.

The last SNP (9_23070656) was associated with the *UPL6* gene, which is involved in protein post-translational modification, mediating the addition of ubiquitin groups to target proteins and the proteasomal degradation of target proteins. Increased protein degradation due to environmental stress has been observed in plants as a way to mobilize nitrogen or eliminate damaged proteins (Eckert *et al.*, 2010a). This may have a role in the concentration of protein to sugars in the cells as other studies have found evidence of induced protein turnover (enhancement of ubiquitin ligase) in response to water deficit and cold stress (Kiyosue *et al.*, 1996; Abernethy & McManus, 1999; Schulze *et al.*, 2003; Kim & Kim, 2013; Patel *et al.*, 2015). However, expression of these genes will need to be investigated further to identify their role in WSC accumulation in white clover.

Only one study has investigated the genetic control of WSC in white clover and revealed that there were four candidate genes that control stolon WSC (Inostroza *et al.*, 2018). However there are stark differences between the findings in Inostroza *et al.* (2018) and the current study. Firstly, their experimental unit was the stolon, typically a carbon sink (importers of assimilates), whereas leaves, assayed in the current study, are carbon sources (exporters of assimilates) and the control of WSC in both tissues are likely to be under different mechanisms (Ballicora *et al.*, 2004). Secondly, the degradation of WSC was investigated in Inostroza *et al.* (2018), whereas in the current study WSC accumulation was the focus. Additionally, identification of genes involved in stolon WSC degradation was via GWAS using a dataset of a greater SNP density (16.6 SNPs per Mbp) compared to the current study (6 SNPs per Mbp as 5,757 SNPs were used for the GWAS), so all the genes important for WSC accumulation may not have been detected. Inostroza *et al.* (2018) identified four candidate genes associated with WSC degradation rate, including a prolyl 4-hydroxylase alpha-like protein, a putative RING-finger E3 ubiquitin ligase, plant invertase/pectin methylesterase inhibitor and peptide/nitrate transporter. None of these genes were in common with the genes associated with outlier SNPs or detected by GWAS. The genetic control of foliar WSC accumulation requires further elucidation and is investigated further in Chapter 4.

3.6 Conclusions

Past research has investigated the genetics of WSC in white clover stolons in response to cold-stress but no studies have examined the genetics underpinning increased foliar WSC in white clover. In this study, five breeding pools that had undergone selection for divergent foliar WSC were utilised to identify candidate genes associated with this trait. GBS SNP markers from outlier analyses or GWAS were found to differentiate low and high WSC populations, and from this, two candidate genes were identified: *ERD6-like 4* and *glgC*. SNPs associated with a range of other candidate genes were also identified, with involvement in numerous aspects of plant development, membrane transport, post-translational processing, cell division and pathogen response. LD analysis showed relatively rapid LD decay in these populations, which is not uncommon in outcrossing species. It demonstrated that the marker density was insufficient to confidently link candidate genes to intergenic SNPs more than 10,000 bp away. However, clear phenotypic separation of the high and low WSC populations provide a platform for further genetic investigation through the use of transcriptomics and proteomics.

3.7 Acknowledgements

I would like to thank AgResearch colleagues: Narsaa Na for help collecting leaf material for DNA extraction. Craig Anderson and Won Hong for DNA extraction guidance. Won Hong, Anna Larking and Rachel Tan for GBS library guidance and preparation. Ruy Jauregui for bioinformatic analysis. Rachael Ashby, Ken Dodds and Paul Maclean for statistical support.

CHAPTER 4

Transcriptomic and proteomic analysis of foliar water-soluble carbohydrate accumulation in white clover (*Trifolium repens* L.)

4.1 Abstract

Mitigating the environmental impact of pastoral agriculture systems in New Zealand/Aotearoa is required to improve the quality of waterways. Increasing water-soluble carbohydrate (WSC) levels in white clover leaves is one way to reduce nitrogen produced by ruminants. Determining the genetic control of foliar WSC is important to enable efficient breeding for high WSC white clover cultivars. Variation in genes can provide insight into the genetic changes that underlie artificial selection for phenotypes. Four populations from two pools previously used in the phenotyping and genotyping chapters (Chapters 2 and 3) were utilised for transcriptomic and proteomic analyses. The aims of the present study were to identify carbohydrate pathway genes involved in WSC accumulation and determine whether convergent evolution in WSC levels was achieved through artificial selection sorting the same allelic variants of pathway genes into high and low WSC populations. Fourteen gene models from seven gene families of carbohydrate genes (*glgC*, *WAXY*, *glgA*, *glgB*, *BAM*, *AMY* and *ISA3*) were identified as being under selection. Three gene families (*glgC*, *WAXY* and *BAM*) in the starch synthesis and degradation pathway were found to be under selection in both high WSC populations, while an additional four gene families were under selection in at least one of the high WSC populations. Patterns of single nucleotide polymorphism (SNP) variation in chr8.jg3312.t1 (*AMY*), chr5.jg7106.t1 (*glgC*) and chr3.jg7615.t1 (*WAXY*) separated low and high WSC individuals. Sequencing of additional individuals is needed to determine whether allelic variants for the other 11 gene models contribute significantly to WSC levels, and whether allelic variation of these genes is important in breeding high WSC white clover individuals. Allelic variation of carbohydrate pathway genes may be necessary but insufficient to fully understand WSC accumulation as a genome-wide association study (GWAS) identified a transcription factor (*GTE9*) linked to WSC levels, suggesting the potential contribution of cis-acting regulation in WSC accumulation.

4.2 Introduction

Pasture systems for ruminants in New Zealand/Aotearoa (NZ) consist mainly of perennial ryegrass and white clover mixed swards (Barrett *et al.*, 2015). White clover (*Trifolium repens* L.) plays an important role in the environmental sustainability and nutritional quality of these swards. The nitrogen-fixing ability of white clover significantly reduces nitrogen fertiliser applied to pastures in NZ and the high protein content increases the nutritional quality of the feed (Ledgard & Steele, 1992; Harris *et al.*, 1998). Mitigating the environmental impact of pasture systems in NZ is required to improve the quality of waterways. Nitrogen and phosphorous runoff cause excessive algae growth, disrupting the natural river ecosystem (McDowell, Larned & Houlbrooke, 2009). Whilst significant improvement in the reduction of nitrogen in NZ's rivers has occurred in the past two decades, there is strong evidence that nutrient levels in rivers increase in proportion to the levels of agricultural activity in river catchments (Larned *et al.*, 2016). Reducing nitrogen runoff from pastures can be achieved by reducing the nitrogen that is produced by the animals. Previous studies have shown that increased levels of water-soluble carbohydrates (WSC) in perennial ryegrass have reduced nitrogen loss through urine and dung from ruminants (Miller *et al.*, 2000; Miller *et al.*, 2001; Moorby *et al.*, 2006). Therefore, it is thought that increasing the overall WSC of the pasture system, including that of white clover, will decrease the loss of nitrogen to the land from ruminants (Edwards *et al.*, 2007; Widdup *et al.*, 2010).

Previous breeding efforts have focused on increasing WSC content in white clover populations (Widdup *et al.*, 2010). In response to this artificial selection, the populations showed significant shifts in their levels of WSC, as measured by near infra-red reflectance spectroscopy (NIRS) (Chapter 2). A subset of these populations, two high WSC and two low WSC populations from two pools (WNZLL and FNZLL) were used in the current study to identify genes involved in WSC accumulation. Transcriptome and proteome profiling provide an additional tool to genetic analyses (outlier detection and genome-wide association study, GWAS) using genotyping by sequencing (GBS) single nucleotide polymorphism (SNP) data to identify candidate genes responding to selection and for identifying processes important in adaptation and plasticity (Voelckel *et al.*, 2017). RNA-Seq was used to obtain information on transcript differential expression and sequential window acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH MS) profiling was used to obtain information on protein differential abundance between high and low WSC populations. This comparative data allowed me to address questions about the potential and limitations

of different experimental approaches for identifying candidate genes responding to artificial selection. It also allowed me to test a prediction from the hypothesis that standing genetic variation provides a means for adaptation and rapid morphological/physiological evolution (Lai *et al.*, 2019). More specifically, I investigated whether the same artificial selection regimes for high and low levels of WSC applied to different genetic pools of white clover would select the same allelic variants controlling expression levels of carbohydrate pathway genes and in so, doing alter levels of WSC produced in the leaves of the plant. If so, this would provide a model system for investigating physiological and biochemical processes of convergent evolution (Sackton & Clark, 2019). In any event, the experiment had the potential to identify candidate gene targets for selection in future breeding programmes that seek to increase WSC content in white clover populations.

One weakness of comparative transcriptome and proteome profiling is that underlying population structure and possible underlying confounders, such as stochastic and systematic errors, can contribute to long lists of differential expression where only some will be informative. In the current study, the historic relationships of four populations of white clover was reconstructed to provide a phylogenetic framework for studying differential expression patterns that could be linked to artificial selection for foliar WSC levels. GBS-derived SNP data for these four populations were obtained in Chapter 3 and were used to reconstruct the evolutionary relationship of the four populations using Neighbour Net analyses (Bryant & Moulton, 2004). Phylogenetic relationships of the four populations using 183 individuals and 222 high quality filtered GBS SNPs were reconstructed to produce a phylogenetic tree that reflected the known breeding pedigree of the four populations. Patterns of transcript expression and protein abundance were superimposed onto this phylogeny (**Figure 4.1**). Of most interest were patterns of differential expression that could and could not be mapped to branches of the phylogenetic tree and in particular, patterns of expression that appeared to be convergent between populations from different pools subjected to the same direction of artificial selection. SNP variation from non-coding and coding regions of 14 of the candidate carbohydrate metabolism gene model IDs identified as responding to selection were examined. Doing this allowed the question to be asked: whether convergent phenotypes (high or low levels of WSC) might have arisen from similar sorting of ancestral allelic variation or whether different physiological or metabolic processes generated similar phenotypes in the two pools? This study was motivated by a recent discussion of whether phylogenetic modelling of expression data may provide a more efficient means of identifying candidate genes under selection than standard

approaches of pairwise comparison (Voelckel *et al.*, 2017). Although, phylogenetic modelling was not undertaken in the present work, the implications of my findings for this approach have been discussed, as is evidence for the sorting of allelic variation that may have significance for rapid morphological evolution.

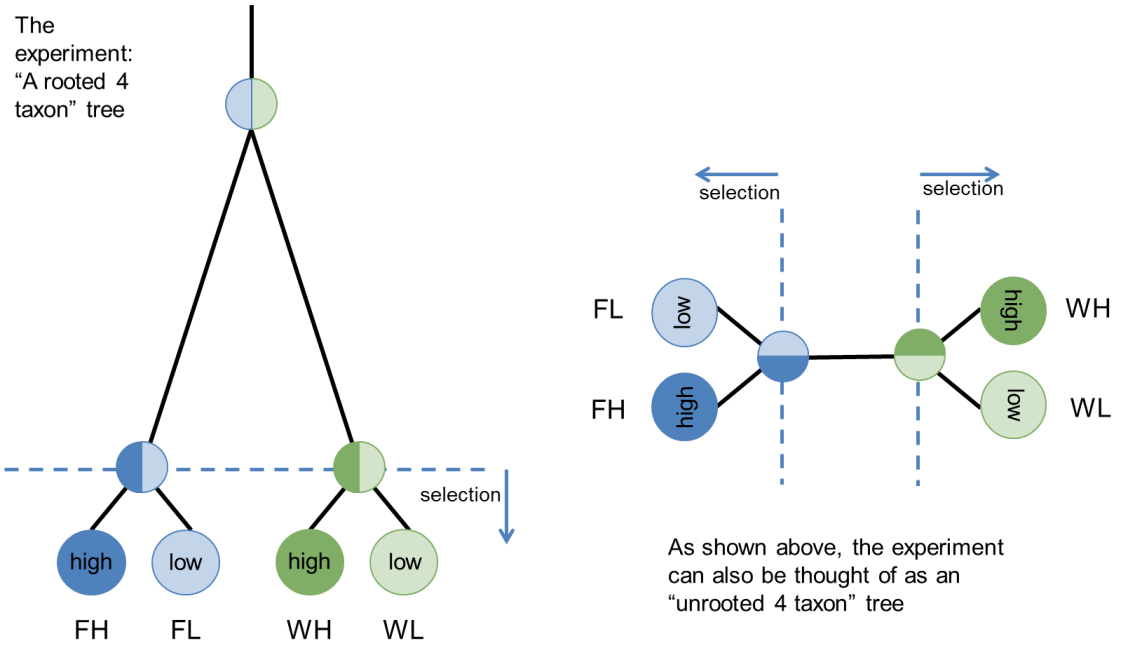


Figure 4.1 Experimental design as a four-taxon tree. The tree on the left depicts a rooted four-taxon tree and the figure on the right depicts an unrooted taxon tree. FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

4.3 Materials and methods

4.3.1 Experimental design

Four populations from two pools used in Chapters 2 and 3 were utilised for transcriptomic and proteomic analyses. These populations were: WNZLL-High-End, WNZLL-Low-End, FNZLL-High-End and FNZLL-Low-End; hereafter referred to as WH, WL, FH and FL, respectively. Five individuals from each population were used for transcriptomic analysis and six individuals from each population were used for proteomic analysis (**Table S4.1**, Appendix 3). Three samples per population were common between two analyses.

4.3.2 Phylogenetic tree constructed with genotyping by sequencing data

Hierarchical clustering of genotyping by sequencing (GBS) single nucleotide polymorphism (SNP) data was used to determine whether SNP variation across the white clover genomes distinguished the High-End and Low-End populations of the FNZLL and WNZLL pools. One hundred and eighty-three individuals from the four populations were used to construct a phylogenetic tree using SNP data obtained from Chapter 3 (section 3.3.4.1). The unfiltered variant call format (VCF) file used in Chapter 3 was first filtered to only retain individuals from the FH, FL, WH and WL populations, which included 48, 48, 47 and 47 individuals per population, respectively. The new VCF file, containing 190 individuals, was then filtered to only include biallelic SNPs, a minimum and maximum read depth range of 5 to 150, limiting missing genotype data to a maximum of 20% per SNP, and including SNPs with a minor allele frequency (MAF) threshold of ≥ 0.03. The above filtering scheme was performed using VCFtools v 0.1.16 (Danecek *et al.*, 2011). After this filtering, the VCF file was imported into R software v 3.6.1 (R Core Team, 2019) using the function *read.vcfR()* from the "*vcfR*" v 1.8.0 package (Knaus & Grünwald, 2017). Individuals with more than 80% missing data were removed using the function *is.na()* in conjunction with *apply()*, which reduced the number of individuals down to 183. The data were then filtered to remove SNPs with a high degree of missing genotype calls using *is.na()* and *apply()* by limiting missing genotype data to a maximum of 80% per sample. This filtering resulted in less than 5% missing data overall and 222 SNPs retained. A hierarchical cluster analysis was then performed using the *hclust()* function, using the Euclidean distance calculated using the *dist()* function, to produce a phylogenetic tree. Populations were identified manually from the phylogenetic tree and borders were placed around individuals belonging to each population.

4.3.3 Water-soluble carbohydrate content of individuals used in transcriptomic and proteomic analyses

To confirm the individuals used in the transcriptomic and proteomic analyses had divergent water-soluble carbohydrate (WSC) phenotypes, the WSC content of the 32 samples (8 samples per population) was used to create boxplots displaying the WSC content for each population. Soluble sugars and starch content, as determined by near infra-red reflectance spectroscopy (NIRS) from Chapter 2 (section 2.3.3), was plotted in RStudio using the function *ggplot()* from the "*ggplot2*" v 3.3.0 package (Wickham, 2016).

4.3.4 RNA isolation, quality control and sequencing

Total RNA was extracted from frozen leaf material following the ISOLATE II Plant Kit bench-top protocol (Bioline, London UK). For sample homogenization, leaf tissue was ground to a fine powder using a mortar and pestle placed in a liquid nitrogen (N) bath to ensure samples remained frozen to avoid RNA degradation. A maximum of 100 mg of powder was transferred to a 1.5 mL Eppendorf tube also placed in a liquid N bath. To the tube, 350 µL of Lysis buffer (RLY) and 3.5 µL of β-mercaptoethanol was added and then vortexed quickly to thoroughly mix. The remaining RNA extraction steps were performed following the manufacturers protocol.

After extraction, RNA samples were stored on ice and 2 µL was measured on a nanodrop Spectrophotometer (Nanodrop Technologies, Montchanin, USA) for quality. RNA was then quantified with a Qubit RNA high sensitivity kit (Life Technologies, Thermo Fisher Scientific Inc.) and a Qubit 2.0 Fluorometer (Life Technologies, Thermo Fisher Scientific Inc.) with the Qubit working solution made to the manufacturer's instructions. If samples were of high quality and concentration, 6 µL of each sample was sent to the Massey Genome Service (Massey University, Palmerston North, New Zealand) and an RNA quality score (RQS) was determined for each sample using the RNA Standard LabChip Assay. The RQS correlates well with Agilent's RNA Integrity Number (RIN) and follows the same 0 – 10 scale rating, with values closer to 10 indicating very high-quality RNA. The remaining extracted RNA was stored at -80°C until needed.

Samples that had 260/280 and 260/230 ratios above 1.8, concentrations above 50 ng μL^{-1}, and RQS above 7.5 were chosen for sequencing. Twenty samples were sent on dry ice to Novogene (Hong Kong, China) for RNA-Seq paired-end sequencing. Quality control using Agarose gel electrophoresis, Agilent 2100 and Nanodrop were performed by Novogene prior to library preparation and sequencing. RNA preparation (mRNA enrichment using oligo(dT) beads, random fragmentation using fragmentation buffer, double-stranded cDNA synthesis, addition of A overhang and adaptor ligation, fragment size selection, PCR amplification, and library quality test) and RNA-Seq (Illumina sequencing) were performed by Novogene. All 20 samples were sequenced on a single lane, returning 145.1 G raw data with a mean of 24,185,978 raw reads per sample.

4.3.4.1 Homeolog-specific transcriptional profiling and differential expression analysis

The software HyLiTE v 1.6.2 was used to obtain read count matrices for differential expression analysis (Duchemin *et al.*, 2015). The HyLiTE analysis was run based on the approach in the HyLiTE manual, which does not include any QC filtering https://hylite.sourceforge.io/tutorial.html#tutorial. SAM files were first created by mapping paired end mRNA reads with Bowtie 2 (Langmead & Salzberg, 2012) against *T. occidentale* reference sequences. HyLiTE then identified polymorphisms indicative of parental origin (*T. occidentale* and *T. pallescens*). The resulting output from HyLiTE comprised tables with read counts for parent species to gene models (counts for orthologs) and read counts of unambiguously assignable reads for white clover for each homeolog. The read counts for each homeolog were combined to give one count value for the transcript. A total of 39,602 transcripts were identified. The workflow for running HyLiTE can be found in the HyLiTE folder at https://github.com/MarniTausen/CloverAnalysisPipeline.

Count data for the 39,602 transcripts were then imported into RStudio. Basic filtering to remove transcripts with total counts less than 10 across the full dataset was performed, resulting in 34,078 transcripts retained. Data were then analysed in six pairwise comparisons (WH vs WL, FH vs FL, WH vs FL, FH vs WL, WH vs FH and WL vs WL). Each pairwise (PWC) comparison was filtered to remove transcripts with total counts less than 10. The R package "*DESeq2*" v 1.24.0 was used to perform differential expression analysis (Love, Huber & Anders, 2014). An accurate way of estimating $\log_2$ fold changes (LFCs) can be obtained by allowing for shrinkage of LFC estimates towards zero when the information for a transcript is low. Shrinkage identifies the largest LFCs that are not due to low counts and then uses these to inform a prior distribution. The large LFC from transcripts with lots of statistical information are not shrunk, but the imprecise LFC are shrunk. Hence, shrinkage can be used to reduce potential false positives that can result from underestimates of dispersion. Shrinkage using approximate posterior estimation for general linear model (apeglm) was first performed to increase accuracy of estimating LFCs using the function *lfcShrink()* (Zhu, Ibrahim & Love, 2018). The adjusted *p*-value (Benjamini & Hochberg, 1995) and LFC were plotted to create volcano plots for each PWC. Transcripts that had both an adjusted *p*-value less than 0.05 and a LFC cut-off at ± 2 were classified as differentially expressed.

4.3.5 Protein isolation and determination

Leaf samples from 6 individuals per population (WH, WL, FH and FL) were previously collected at the 2017 harvest (Chapter 3), frozen in liquid N and stored at -80°C. These 24 samples were then freeze-dried and transported on silica gel to Australia where samples were prepared for quantitative proteomics using sequential window acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH MS) analysis by the Australian Proteomic Analysis Facility (see supplementary methods, "Protein extraction and sequential window acquisition of all theoretical fragment ion spectra" in Appendix 3). Proteomics aims to ascertain and quantify the most accurate representation of proteins within a sample at a given point in time. SWATH MS is an approach that uses data independent acquisition to quantify proteins without prior knowledge of the proteins in the sample. This approach quantifies a large set of proteins across multiple samples with consistency, accuracy, sensitivity and high reproducibility (Gillet *et al.*, 2012). A FASTA file containing a list of 68,557 *Trifolium repens* proteins was supplied from AgResearch (Griffiths *et al.*, 2019) as a reference for protein identification.

4.3.5.1 Differential expression analysis of protein data

The SWATH MS protein abundance data were normalised to the total intensity of the respective sample. Normalised protein abundance data were then imported into R and Wilcoxon tests were performed using the function *wilcox.test()* from the "*stats*" package v 3.6.1 (R Core Team, 2019) on each pairwise comparison (PWC). If a protein had a *p*-value smaller than 0.05 and a $\log_2$ fold change (LFC) ± 1.5 (as specified by the Australian Proteomic Analysis Facility), it was highlighted as a differentially abundant protein (DAP). The *p*-value and LFC were plotted to visualise DAPs for each PWC.

4.3.6 Gene ontology enrichment analysis and Kyoto Encyclopedia of Genes and Genomes pathway analysis

Gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) IDs were assigned to each transcript and protein in the *T. occidentale* and *T. repens* genomes, respectively, using translated protein sequences. Protein sequences from the *T. occidentale* and *T. repens* genomes were aligned to the NCBI "nr" database using the BLASTP function of diamond v 0.9.24 (Buchfink, Xie & Huson, 2015) and had their domains searched using InterProScan v 5.36-75.0 (Jones *et al.*, 2014). The diamond and InterProScan XML result files were imported into OmicsBox v 1.2.4 (Götz *et al.*, 2008) where GO terms and annotation were assigned with default settings. In addition,

the eggNOG-Mapper v 1.0.3 within OmicsBox was run with EggNOG 5.0.0 (Huerta-Cepas *et al.*, 2019). Default settings were used except for the specification of XML output. Biological significance of differentially expressed transcripts (DETs) and differentially abundant proteins (DAPs) was explored by GO term enrichment analysis including biological process (BP) and molecular function (MF) using singular enrichment analysis (SEA) by GO Analysis Toolkit and Database for Agricultural Community (AgriGO) v 2 (Tian *et al.*, 2017) available at http://systemsbiology.cau.edu.cn/agriGOv2/c_SEA.php. For each pairwise comparison, a list of DETs and DAPs was created with their corresponding GO terms and was used as the input list with GO annotations for AgriGO. A list of reference transcripts and proteins with their corresponding GO terms was created using the 34,078 reference transcripts and 6,577 reference proteins and used as the customised annotated reference input. Fisher's statistical test was applied with a significance level at 0.05. No multiple adjustment method was applied but Plant GO slim was selected for the gene ontology type to specify plant related GO term output.

In parallel, KEGG mapper pathway analysis on DETs and DAPs was performed using the KEGG mapper search and colour pathway (Kanehisa & Sato, 2020). The KEGG maps with hits of more than five transcripts or proteins for the WH-WL and FH-FL PWCs were recorded. As the starch and sucrose metabolism map had more than five hits for both PWCs for both transcriptomic and proteomic datasets, it was selected as a good candidate to identify genes involved in WSC accumulation. The starch and sucrose metabolism reference pathway was visualised for both PWCs.

4.3.7 Identification of matching gene model IDs between the *Trifolium repens* and *T. occidentale* genomes

Each transcript and protein has an identifier called a gene model ID. These gene model IDs are unique to each of the reference genomes (*T. occidentale* and *T. repens*), with the *T. occidentale* gene model IDs beginning with "jg" and the *T. repens* gene model IDs beginning with "chrX.jg", where X specifies the chromosome number (1 to 16). Identical gene model IDs needed to be identified between the two reference genomes to find commonality between the two datasets. Identification of matching gene model IDs was performed by blasting all translated protein sequences from the *T. repens* genome against all translated protein sequences from the *T. occidentale* genome using BLASTP. A maximum number of three hits (gene model IDs from *T. repens*) for each query (gene model ID from *T. occidentale*) were allowed. The gene model ID with the smallest E-

value (≤ 1e-10), largest per cent identity match (> 85%) and a similar length to the query (match length/query length > 0.5) was determined as a good candidate for a match. If there were two very good matches (both > 95% identity match), then both gene model IDs were recorded. Gene function for the query and hit gene model IDs was also used to verify that the gene model IDs were a good match. Gene model ID match between the two datasets was also important for identifying SNP and gene variants in candidate genes which is described in the following section (4.3.8).

4.3.8 Single nucleotide polymorphism variation in candidate genes identified from RNA-Seq data

SNP variation in non-coding and coding regions of candidate carbohydrate metabolism genes responding to selection were examined. FASTA reads (obtained from RNA-Seq) were mapped to the *T. repens* genome using Spliced Transcripts Alignment to a Reference (STAR) v 2.5.2b, limiting each read to be mapped only once to avoid subgenome ambiguous reads. The reads were prepared using Split'N'Trim in Genome Analysis Toolkit (GATK) v 3.6 and then variant calling in GATK was done using Haplotypecaller. Variants were filtered using settings: "MQ>30.00 && DP>20 && QUAL>20.00" resulting in 2,694,398 SNPs retained post filtering. For more detail please see "Variant calling using STAR and GATK" in Supplementary methods in Appendix 3.

4.3.8.1 Discriminant analysis of principal components

Discriminant analysis of principal components (DAPC) was used to identify SNP variants for the 14 candidate genes that separate high and low WSC individuals. This was performed by first subsetting the VCF file obtained in section 4.3.8 into 14 VCF files, each one including SNPs identified in coding and non coding regions for one of the candidate genes. This was performed using VCFtools using the following settings:

vcftools --recode --recode-INFO-all --positions --vcf --stdout > X.vcf

Where: X is the name of the gene model ID.

Each of the 14 VCF files were then imported into R using *read.vcfR()* and converted to genind objects using *vcfR2genind()*, both from the "*vcfR*" package v 1.10.0 (Knaus & Grünwald, 2017). Population data were assigned, where the WH and FH individuals were grouped into one High population, and the WL and FL individuals were grouped into one Low population. The *xvalDapc()* function from the "*adegenet*" package v 2.1.2 (Jombart, 2008) was run with 100 replicates from PC 1 to 20. Individual density

plots were then created using the function *scatter()* for each candidate gene. A loading plot was generated for each gene model ID to identify which SNPs differentiate the High and Low individuals using the function *loadingplot()*. SNPs with loadings above 0.03 were selected as contributing the most to the observed separation of individuals on the first discriminant function. Genotype calls for each SNP were extracted using VCFtools --*extract-FORMAT-info* GT and converted to binary: 0/0 = 0, 0/1 = 1, 1/1 = 2. Where: 0 = homozygous for the reference allele, 1 = heterozygous, and 2 = homozygous for the alternate allele. A trend where all the individuals from the high WSC populations have the same genotype would suggest there is evidence that a particular SNP is potentially functionally associated with higher WSC content. A summary of the analysis workflow is presented in **Figure 4.2**, including the populations used, the transcriptomic and proteomic workflow and SNP variation.

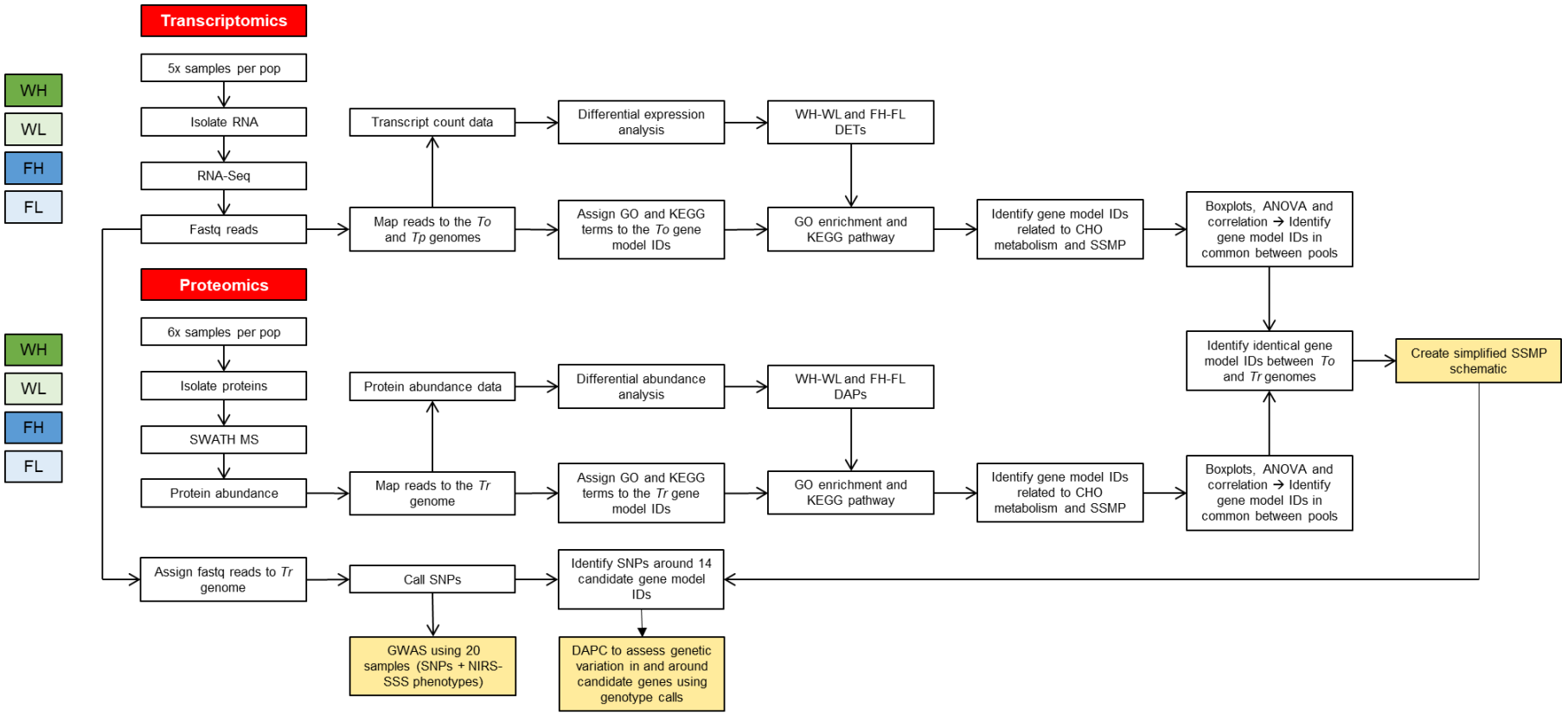


Figure 4.2 Summary of transcriptomic and proteomic analysis workflow. Main 'outcome' boxes are coloured in yellow and distinguished from the 'process' boxes (white). CHO = carbohydrate, DAP = differentially abundant proteins, DET = differentially expressed transcripts, GO = gene ontology, GWAS = genome-wide association study, KEGG = Kyoto Encyclopedia of Genes and Genomes, NIRS = near infra-red reflectance spectroscopy, pop = population, SSMP = starch and sucrose metabolism pathway, SSS = Soluble sugars and starch, SWATH MS = sequential window acquisition of all theoretical fragment ion spectra mass spectrometry, *To* = *T. occidentale*, *Tp* = *T. pallescens*, *Tr* = *T. repens*, FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

4.4 Results

4.4.1 Water-soluble carbohydrate content of individuals used in transcriptomic and proteomic analyses

Water-soluble carbohydrate (WSC) content of individuals for each population was plotted and statistically significantly differences were observed between the high and low WSC populations (**Figure 4.3**). A greater difference in the mean WSC content was observed between the WH and WL populations (approximately 100 grams per kilogram dry matter, g kg^{-1} DM) than between the FH and FL populations (approximately 75 g kg^{-1} DM). As will be discussed, this is an observation perhaps relevant for understanding the significance of differences in patterns of differential expression for carbohydrate metabolism genes observed for the WNZLL and FNZLL pools.

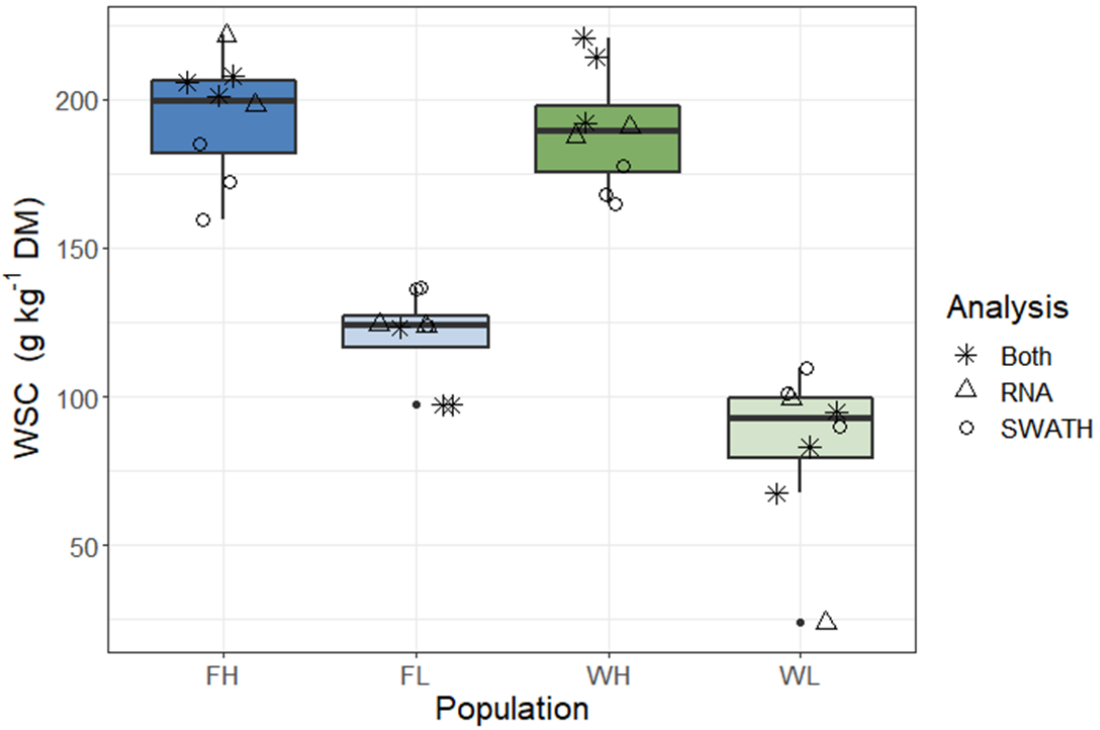


Figure 4.3 Boxplots of water-soluble carbohydrate (WSC) content (g kg^{-1} dry matter) grouped by population. Data points used to construct boxplots are displayed and shape coded to represent which analysis (SWATH proteomics, RNA transcriptomics or both) the samples were used in. WSC was measured by near infra-red spectroscopy.
Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

4.4.2 Phylogenetic relatedness the four white clover populations as determined by genotyping by sequencing single nucleotide polymorphism data

A phylogenetic tree produced from hierarchical clustering of Euclidean distances derived from 222 genotyping by sequencing (GBS) single nucleotide polymorphisms (SNPs) identified populations from the same pool as being more closely related (i.e., WH and WL; FH and FL) than populations with similar WSC content (i.e., WH and FH; WL and FL) (**Figure 4.4**). The 222 SNPs were distributed across the genome with at least one SNP present in every chromosome.

4.4.3 RNA isolation and differentially expressed transcripts and proteins

4.4.3.1 RNA isolation and data quality

Extracted RNA was high quality with nanodrop ratios above 1.99 for all samples and RNA quality scores (RQS) ranging from 7.7 – 9.1 (**Table S4.2**, Appendix 3). Quality control at Novogene prior to sequencing revealed similar sample quality (**Table S4.3**, Appendix 3) with only one sample showing a slight decline in RNA quality (72-28, RIN score of 7.5) between RQS testing and arrival in Hong Kong. The transcriptomes of 20 leaf samples were sequenced. After removing the low-quality reads of raw data, 474.5 million clean reads were obtained, with 23.7 million reads per sample on average. In addition, the Q30 score of each sample ranged from 92.6 to 93.4% (**Table S4.3**, Appendix 3).

4.4.3.2 Differentially expressed transcripts

A total of 34,078 transcripts (post-filtering to remove transcripts with less than 10 counts in total) were identified from the leaf tissue. To identify differentially expressed transcripts (DETs) related to WSC accumulation in the two pools, differential expression analysis was performed by comparing the transcript expression profiles between populations with high WSC content and populations with low WSC content. Using a *p*-value (Benjamini-Hochberg adjusted) at a significance threshold of $\alpha = 0.05$; 2,646 DETs were observed in the WH-WL pairwise comparison (PWC) and 1,354 DETs were observed in the FH-FL PWC. Among these DETs, 162 genes were in common between WH-WL and FH-FL PWC (**Figure S4.1**, Appendix 3).

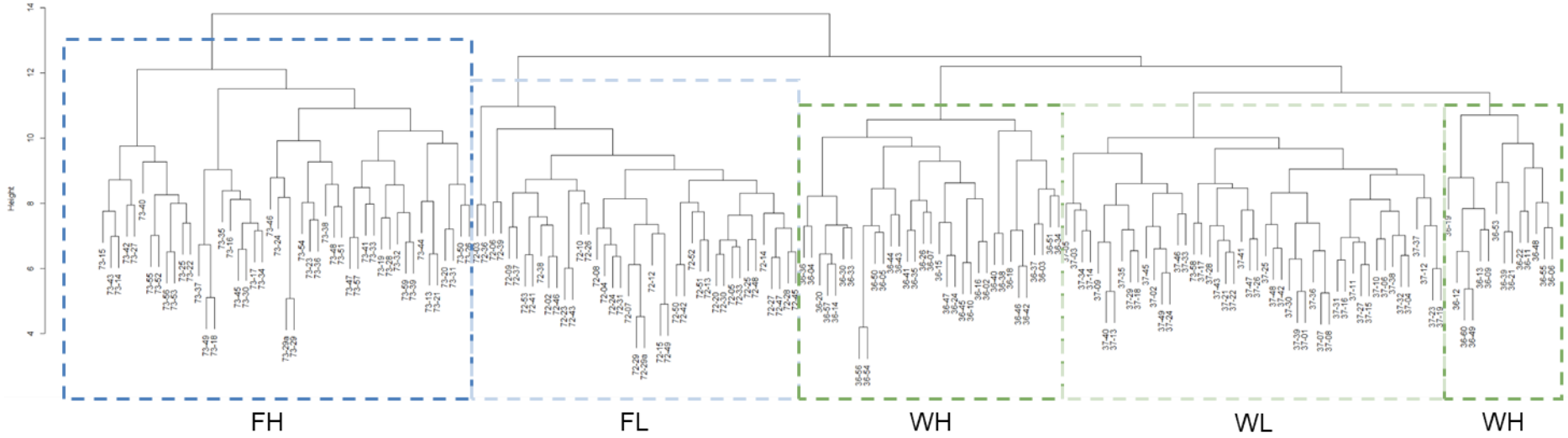


Figure 4.4 Phylogenetic tree of 183 samples from four populations created using 222 genotyping by sequencing SNP markers with maximum missing data of 5%. FH individuals are surrounded by dark blue box, FL individuals are surrounded by light blue box, WH individuals are surrounded by dark green boxes, WL individuals are surrounded by light green box. Height is Euclidean distance.
Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

4.4.3.3 Differentially abundant proteins

Sequential window acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH MS) proteomics identified 6,577 proteins from leaf tissue. To identify differentially abundant proteins (DAPs) related to WSC accumulation in the two pools, differential abundance analysis was performed by comparing the protein abundance profiles between populations with high WSC content and populations with low WSC content. Using a *p*-value significance threshold of $\alpha = 0.05$ with no false discovery rate (FDR) adjustment; 294 DAPs were observed in the WH-WL PWC and 676 DAPs were observed in the FH-FL PWC. Among these DAPs, 40 genes were in common between WH-WL and FH-FL (**Figure S4.1**, Appendix 3).

4.4.4 Gene ontology enrichment analysis of transcriptome and proteome data

With a *p*-value significance threshold of $\alpha = 0.05$ and no FDR adjustment, gene ontology (GO) enrichment analysis identified a total of 32 GO terms in the transcriptome dataset and 3 GO terms enriched in the protein dataset for the WH-WL PWC. A total of 5 GO terms in the transcriptome dataset and 21 GO terms in the protein dataset were enriched in the FH-FL PWC. Carbohydrate metabolic process was enriched only in the WH-WL PWC for both the transcriptomic and proteomic datasets at the significance threshold applied. There were 151 genes related to carbohydrate metabolic process identified in the transcriptome DET dataset for WH-WL (**Table 4.1**). There were 42 proteins related to carbohydrate metabolic process identified in the proteome DAP dataset for WH-WL (**Table 4.2**). No enrichment for carbohydrate metabolic process was inferred in the FH-FL PWC for either the transcriptomic or proteomic datasets at the significance threshold applied. However, as described below (section 4.4.4.1) and later discussed, analysis of individual genes and proteins linked to “carbohydrate metabolic process” in the FH-FL comparison showed similar patterns of up- and down-regulation for some of these 151 transcripts and 42 proteins in the WH-WL comparison.

Table 4.1 Gene ontology (GO) enrichment for differentially expressed transcripts with GO information available in six pairwise comparisons using transcriptomic data. Differentially expressed transcripts were significant at a Benjamini-Hochberg adjusted *p*-value of less than 0.05 for each pairwise comparison. GO terms enriched for biological processes and biological functions are reported, with important GO terms related to water-soluble carbohydrate accumulation underlined and in bold.

PWC	*n* DET	*n* REF	*n* GO terms	GO category	GO term	Description	*n* in DET	*n* in REF	*p*-value
WH-WL	2326	29639	32	P	GO:0009875	pollen-pistil interaction	19	139	0.013
				P	GO:0005975	**carbohydrate metabolic process**	**151**	1603	0.015
				P	GO:0006412	translation	103	1056	0.016
				P	GO:0007049	cell cycle	101	1037	0.017
				P	GO:0051704	multi-organism process	184	2003	0.019
				P	GO:0009790	embryo development	65	645	0.026
				P	GO:0010467	gene expression	463	5370	0.03
				P	GO:0022414	reproductive process	196	2212	0.05
				F	GO:0005198	structural molecule activity	79	612	1.40e-05
				F	GO:0030246	carbohydrate binding	85	520	0.0051
				F	GO:0003682	chromatin binding	33	290	0.021
				F	GO:0000166	nucleotide binding	404	4679	0.035
FH-FL	1126	29639	5	P	GO:0006259	DNA metabolic process	118	2451	0.0086
				P	GO:0007049	cell cycle	52	1037	0.031
				F	GO:0000166	nucleotide binding	227	4679	0.00046
				F	GO:0003774	motor activity	11	154	0.034
WH-FL	986	29639	1	P	GO:0006259	DNA metabolic process	139	2451	1.20e-08
FH-WL	1212	29639	10	P	GO:0006259	DNA metabolic process	128	2451	0.0051
				P	GO:0019725	cellular homeostasis	30	469	0.012
				P	GO:0065008	regulation of biological quality	97	1873	0.015

Table 4.1 (continued)

PWC	*n* DET	*n* REF	*n* GO terms	GO category	GO term	Description	*n* in DET	*n* in REF	*p*-value
				P	GO:0008361	regulation of cell size	5	40	0.023
				P	GO:0042592	homeostatic process	52	979	0.039
				F	GO:0016787	hydrolase activity	356	7074	0.0003
FH-WL	1212	29639	10	F	GO:0000166	nucleotide binding	238	4679	0.0012
				F	GO:0030246	carbohydrate binding	36	520	0.002
				F	GO:0016788	hydrolase activity, acting on ester bonds	145	2819	0.0053
				F	GO:0003824	catalytic activity	723	16198	0.03
				P	GO:0006259	DNA metabolic process	89	2451	0.039
				P	GO:0042592	homeostatic process	39	979	0.045
WH-FH	879	29639	5	F	GO:0000166	nucleotide binding	165	4679	0.023
				F	GO:0030246	carbohydrate binding	23	520	0.042
				F	GO:0003677	DNA binding	110	3106	0.044
				P	GO:0019748	secondary metabolic process	64	508	3.00e-05
				P	GO:0008219	cell death	46	366	0.00036
				P	GO:0009607	response to biotic stimulus	138	1456	0.0024
				P	GO:0006950	response to stress	384	4559	0.0083
				P	GO:0009605	response to external stimulus	175	1971	0.0097
WL-FL	2191	29639	20	P	GO:0051704	multi-organism process	177	2003	0.011
				F	GO:0030246	carbohydrate binding	59	520	0.00088
				F	GO:0000166	nucleotide binding	401	4679	0.0029
				F	GO:0003824	catalytic activity	1290	16198	0.014
				F	GO:0016301	kinase activity	244	2853	0.015
				F	GO:0016740	transferase activity	581	7216	0.031

Table 4.1 (continued)

PWC	*n* DET	*n* REF	*n* GO terms	GO category	GO term	Description	*n* in DET	*n* in REF	*p*-value
WL-FL	2191	29639	20	F	GO:0060089	molecular transducer activity	46	480	0.046
				F	GO:0004872	receptor activity	46	480	0.046

Note: PWC = Pairwise comparison, *n* DET = number of differentially expressed transcripts, *n* REF = number of transcripts in the reference list, *n* GO terms = number of GO terms enriched, GO category = either biological process (P) or molecular function (F), *n* in DET = number of transcripts in the GO category that were differentially expressed, *n* in REF = number of transcripts in the GO category that were in the reference list, *p*-value = *p*-value for the GO enrichment analysis for the GO term, FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

Table 4.2 Gene ontology (GO) enrichment for differentially abundant proteins with GO information available in six pairwise comparisons using proteomic data. Differentially abundant proteins were significant at a *p*-value of less than 0.05 for each pairwise comparison. GO terms enriched for biological processes and biological functions are reported, with important GO terms related to water-soluble carbohydrate accumulation underlined and in bold.

PWC	*n* DAP	*n* REF	*n* GO terms	GO category	GO term	Description	*n* in DAP	*n* in REF	*p*-value
WH-WL	289	6424	3	P	GO:0005975	**Carbohydrate metabolic process**	**42**	612	0.0076
				P	GO:0009991	Response to extracellular stimulus	11	111	0.013
				F	GO:0016787	Hydrolase activity	95	1703	0.038
FH-FL	663	6424	21	P	GO:0006950	Response to stress	2204	1394	4.00e-06
				P	GO:0009607	Response to biotic stimulus	81	461	4.00e-06
				P	GO:0051704	Multi-organism process	92	547	6.40e-06
				P	GO:0009605	Response to external stimulus	94	573	1.40e-06
				P	GO:0019725	Cellular homeostasis	32	175	0.0012
				P	GO:0050896	Response to stimulus	270	2159	0.0030
				P	GO:0009719	Response to endogenous stimulus	82	577	0.0031
				P	GO:0009628	Response to abiotic stimulus	114	851	0.0046
				P	GO:0065008	Regulation of biological quality	68	485	0.0082
				P	GO:0006091	Generation of precursor metabolites and energy	57	398	0.0093
				P	GO:0009056	Catabolic process	123	952	0.0100
				P	GO:0042592	Homeostatic process	37	243	0.0120
				P	GO:0040029	Regulation of gene expression, epigenetic	12	62	0.0240
				F	GO:0003824	Catalytic activity	510	4256	0.004

Table 4.2 (continued)

PWC	*n* DAP	*n* REF	*n* GO terms	GO category	GO term	Description	*n* in DAP	*n* in REF	*p*-value
WH-FL	523	6424	24	P	GO:0051704	Multi-organism process	72	547	9.40e-05
				P	GO:0009607	Response to biotic stimulus	60	461	0.00041
				P	GO:0009605	Response to external stimulus	71	573	0.00056
				P	GO:0006810	Transport	106	931	0.00085
				P	GO:0051234	Establishment of localization	106	945	0.0013
				P	GO:0051179	Localization	110	992	0.0016
				P	GO:0030154	Cell differentiation	26	182	0.0042
				P	GO:0050896	Response to stimulus	212	2159	0.0096
				P	GO:0048869	Cellular developmental process	32	253	0.01
				P	GO:0065008	Regulation of biological quality	54	485	0.016
				P	GO:0019725	Cellular homeostasis	23	175	0.017
				P	GO:0040007	Growth	23	179	0.022
				P	GO:0007154	Cell communication	55	512	0.028
				P	GO:0042592	Homeostatic process	29	243	0.028
				P	GO:0007165	Signal transduction	45	145	0.036
				F	GO:0008289	Lipid binding	16	99	0.0068
				F	GO:0004871	Signal transducer activity	12	80	0.029
FH-WL	459	6424	9	P	GO:0019725	Cellular homeostasis	28	175	7.00e-05
				P	GO:0042592	Homeostatic process	34	243	0.0002
				P	GO:0050896	Response to stimulus	183	2159	0.025
				P	GO:0006950	Response to stress	119	1394	0.042
				P	GO:0006091	Generation of precursor metabolites and energy	38	398	0.049

Table 4.2 (continued)

PWC	*n* DAP	*n* REF	*n* GO terms	GO category	GO term	Description	*n* in DAP	*n* in REF	*p*-value
FH-WL	459	6424	9	F	GO:0008289	Lipid binding	13	99	0.025
WH-FH	373	6424	9	P	GO:0006950	Response to stress	102	1394	0.021
				P	GO:0009605	Response to external stimulus	44	573	0.046
				P	GO:0051704	Multi-organism process	42	547	0.05
				F	GO:0003824	Catalytic activity	282	4256	0.046
WL-FL	435	6424	12	P	GO:0009056	Catabolic process	94	952	0.00054
				P	GO:0005975	Carbohydrate metabolic process	59	612	0.0066
				P	GO:0051704	Multi-organism process	53	547	0.0085
				P	GO:0019748	Secondary metabolic process	19	156	0.011
				P	GO:0009607	Response to biotic stimulus	43	461	0.027
				F	GO:0003824	Catalytic activity	334	4256	0.02
				F	GO:0016740	Transferase activity	111	1358	0.039

Note: PWC = Pairwise comparison, *n* DAP = number of differentially abundant proteins, *n* REF = number of proteins in the reference list, *n* GO terms = number of GO terms enriched, GO category = either biological process (P) or molecular function (F), *n* in DAP = number of proteins in the GO category that were differentially abundant, *n* in REF = number of proteins in the GO category that were in the reference list, *p*-value = *p*-value for the GO enrichment analysis for the GO term, FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

4.4.4.1 Differential expression of transcripts and proteins linked to carbohydrate metabolism

Boxplots of the 151 transcripts linked to carbohydrate metabolic process were created for each population using the normalised transcript expression values for each sample. Binary patterns of differential expression (up- or down-regulation) could be mapped against the phylogenetic tree produced from SNP data for the four populations (**Figure 4.1** and **Figure 4.4**). Of particular interest were transcripts where both WH and FH had a significantly higher level of expression than WL or FL, or where WH and FH had a lower expression than in both WL and FL. While the ontology analysis did not identify enrichment of carbohydrate metabolic process in the FH-FL PWC, 49 of the 151 transcripts linked to carbohydrate metabolic process showed similar expression patterns for FH and WH populations (albeit not significantly similar at $p = 0.05$) compared to FL and WL populations (**Figure 4.5**). All 49 transcripts were significant at the Benjamini-Hochberg adjusted *p*-value for WH-WL but only five were significant at the adjusted *p*-value for FH-FL (**Table S4.4**, Appendix 3). Furthermore, 13 of the 42 proteins linked to carbohydrate metabolic process showed similar abundance patterns for FH and WH populations compared to FL and WL populations (**Figure 4.6**). Of these 13 proteins, none were significant at the *p*-value for FH-FL, but 5 were significant if the α threshold was relaxed to 0.1 (**Table S4.5**, Appendix 3). The patterns of similarity were further assessed below.

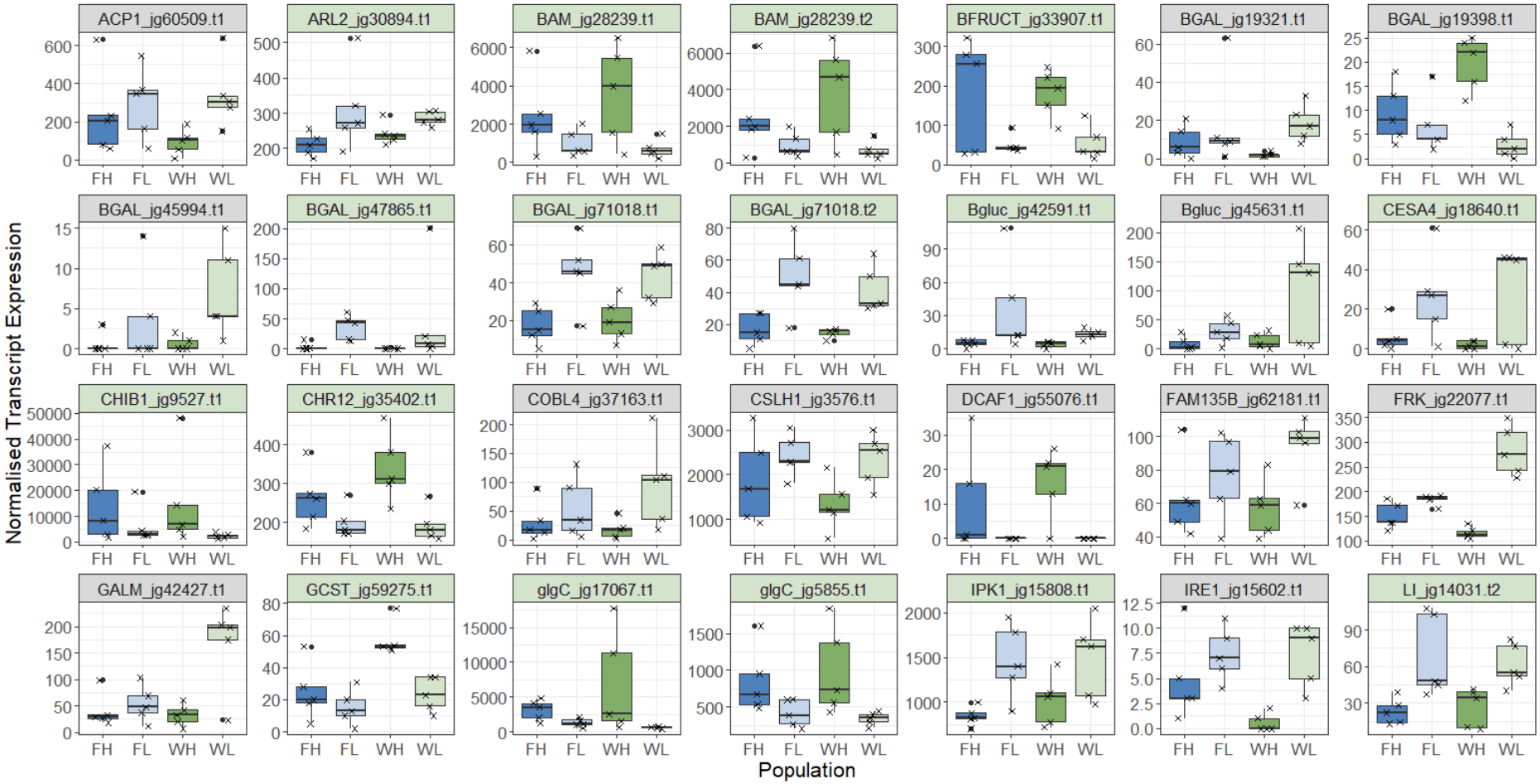


Figure 4.5 (Figure legend on next page)

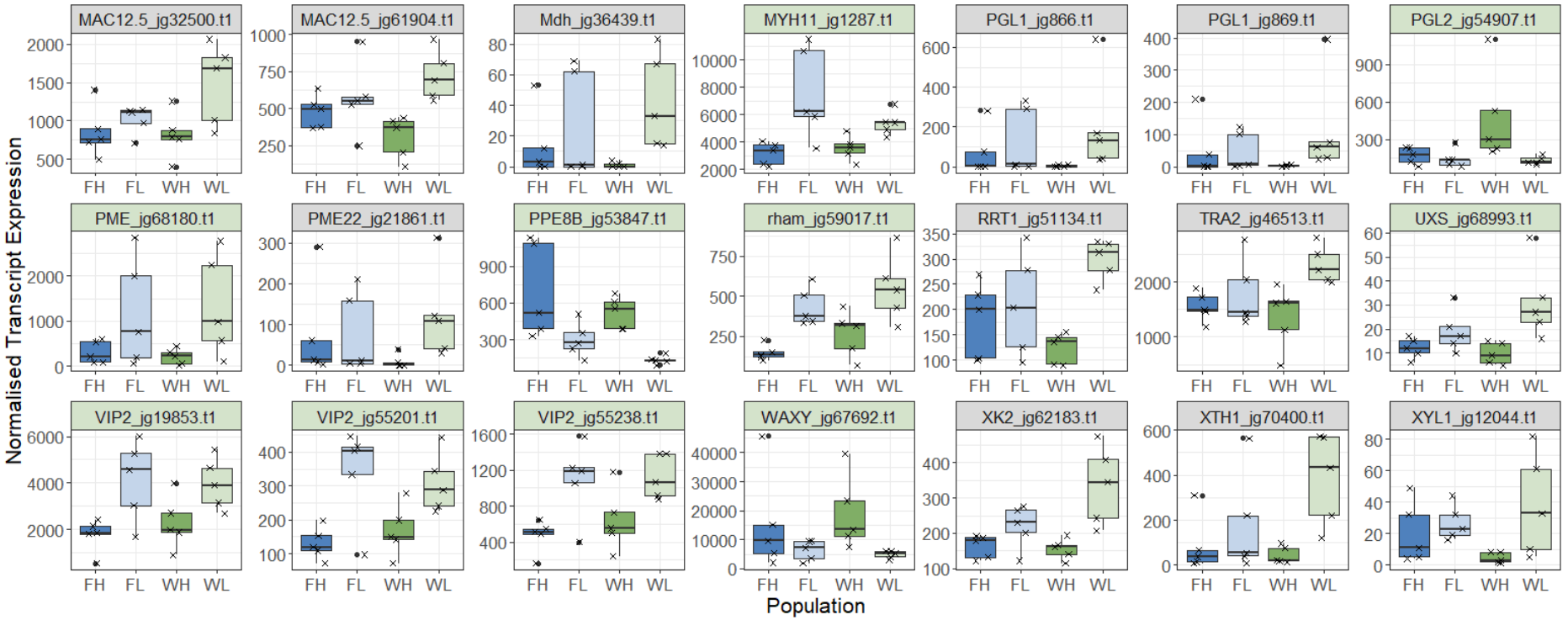


Figure 4.5 Boxplots of 49 transcripts linked to carbohydrate metabolic process using normalised transcript expression values for each population. Data points are overlaid on the boxplots for each population. Transcripts have gene names assigned, which are alphabetically ordered and transcripts coloured in light green represent the 26 candidate transcripts under selection as identified in the correlation analysis. See **Table S4.4** (Appendix 3) for corresponding fold changes and *p*-values for the 26 candidate transcripts. Note that the *y*-axis scale is different between transcripts. *ACP1* = acyl carrier protein 1, *ARL2* = ADP-ribosylation factor-like protein 2, *BAM* = beta-amylase, *BFRUCT* = beta-fructofuranosidase, *BGAL* = beta-galactosidase, *Bgluc* = glucan endo-1,3-beta-glucosidase-like, *CESA4* = cellulose synthase A catalytic subunit 4, *CHIB1* = acidic endochitinase, *CHR12* = probable ATP-dependent DNA helicase CHR12 isoform X1, *COBL4* = COBRA-like protein 4, *CSLH1* = cellulose synthase, *FAM135B* = protein FAM135B isoform X1, *FRK* = probable fructokinase-6, *GALM* = aldose 1-epimerase, *GCST* = aminomethyltransferase, *glgC* = glucose-1-phosphate adenylyltransferase large subunit 1, *IPPK* = inositol-pentakisphosphate 2-kinase, *IRE1* = serine/threonine-protein kinase/endoribonuclease IRE1a-like isoform X1, *LI* = cyanogenic beta-glucosidase, *MAC12.5* = probable alpha-

mannosidase At5g13980, *MDH* = malate dehydrogenase, *MYH11* = myosin-11 isoform X1, *PGL1* = probable 6-phosphogluconolactonase 4, *PGL2* = probable 6-phosphogluconolactonase 2, *PME* = pectin methylesterase, *PME22* =putative pectinesterase/pectinesterase inhibitor 22, *PPE8B* = pectinesterase/pectinesterase inhibitor, *rham* = probable rhamnogalacturonate lyase B isoform X1, *RRT1* = rhamnogalacturonan I rhamnosyltransferase 1-like, *TRA2* = AES90954.1 Transaldolase 2, *UXS* = UDP-glucuronic acid decarboxylase 2-like, *VIP2* = inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase VIP2 isoform X2, *WAXY* = granule-bound starch synthase 1, *XK2* = xylulose kinase 2, *XTH1* =xyloglucan endotransglucosylase/hydrolase 1, *XYL1* = alpha-xylosidase 1.

Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

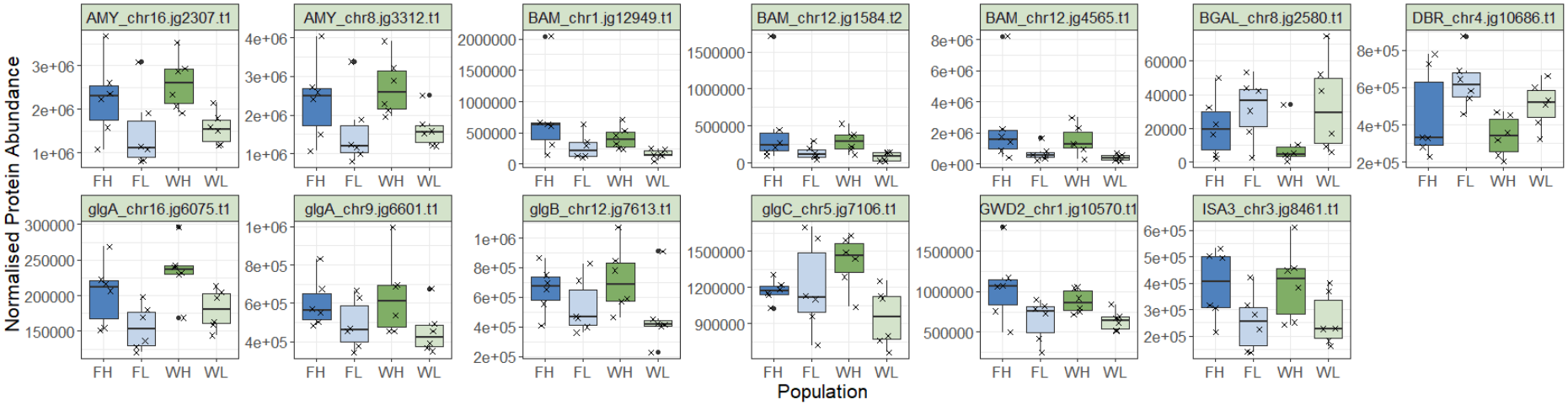


Figure 4.6 Boxplots of 13 proteins linked to carbohydrate metabolic process using normalised protein expression values for each population. Data points are overlaid on the boxplots for each population. Proteins have gene names assigned, which are alphabetically ordered and proteins coloured in light green represent the 13 candidate proteins under selection as identified in the correlation analysis. See **Table S4.5** (Appendix 3) for corresponding fold changes and *p*-values for all 13 candidate proteins. Note that the *y*-axis scale is different between proteins. *AMY* = alpha-amylase 3, *BAM* = beta-amylase, *BGAL* = beta-galactosidase 1, *DBR* = 2-alkenal reductase (NADP(+)-dependent), *glgA* = starch synthase, *glgB* = 1,4-alpha-glucan-branching enzyme 1, *glgC* = glucose-1-phosphate adenylyltransferase small subunit 2, *GWD2* = alpha-glucan water dikinase, *ISA3* = isoamylase 3.

Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

4.4.4.2 Pattern assessment analyses using permutation analysis of variance and correlation

Although the expression differences within the FNZLL populations (FH and FL) was not statistically significant for the majority of the DET and DAP identified as enriched from the WH-WL GO analysis, they did show a similar pattern to the WH-WL expression profile (**Figure 4.5** and **Figure 4.6**), this very obvious trend required further investigation. The analyses using PWC expression differences were focused on within-pool difference, but it does not consider the gene effects that are in common to both pools and therefore potentially universal for the species. Therefore, to determine if the trend found in the WNZLL pool was consistent with the FNZLL pool, two tests were performed in parallel. The first test was a permutation analysis of variance (ANOVA) that was performed with 1e+7 permutations on each of the 49 transcripts and 13 proteins with pool (W, F) and selection (H, L) as separate factors and an interaction term (see supplementary methods, "Permutation analysis of variance and correlation tests" in Appendix 3). All 49 transcripts and all 13 proteins were significant for the selection factor (**Table S4.6**, Appendix 3 and **Table S4.7**, Appendix 3), indicating that the effect of the selection was consistent on the 49 transcripts and 13 proteins between the two pools.

The second analysis involved three correlation tests performed in R for the relative expression levels of carbohydrate metabolism transcripts and proteins in the high and low WSC populations, for both the WNZLL and FNZLL pools. For this test, transcripts and proteins were first filtered to only include those with $\log_2$ fold change (LFC) at least ± 0.1; this resulted in 26 transcripts and 13 proteins retained. The first correlation examined the LFCs between the high and low populations in both pools for the 26 transcripts, with WH-WL PWC and FH-FL PWC as the two variables. The second correlation examined the $\log_2$ expression means for the high WSC populations in both the WNZLL and FNZLL pools for the 26 transcripts, with WH and FH as the two variables. The third correlation examined the $\log_2$ expression means for the low WSC populations in both the WNZLL and FNZLL pools for the 26 transcripts, with WH and FH as the two variables (see supplementary methods, "Permutation analysis of variance and correlation tests" in Appendix 3). These three correlations were then repeated using the 13 proteins identified from the proteomic dataset. For two of the three correlations using the transcriptomic dataset (26 transcripts), and the three correlations using the proteomic dataset (13 proteins), Pearson correlation coefficients (r) were in excess of 0.9 and associated *p*-values were less than 1e-4 (**Figure S4.2**, Appendix 3). There was

only one correlation (the LFC using the transcriptomic dataset) that had a slightly lower r value of 0.74 (p = 1.655e-05), suggesting a slightly weaker (although still high) relationship between the LFC of the two PWC. This relationship was weakened due to a large LFC between WH and WL, while the LFC between FH and FL was substantially smaller. As will be discussed, this observation reflects the observed phenotypes. The strong and highly significant relationships indicated that selection had similar effect on transcript and protein expression in the two pools. All 26 transcripts and 13 proteins were therefore identified as candidate genes that had responded to selection in the two pools (**Table S4.4**, Appendix 3 and **Table S4.5**, Appendix 3).

4.4.5 Kyoto Encyclopedia of Genes and Genomes analysis of transcriptome and proteome data

Kyoto Encyclopedia of Genes and Genomes (KEGG) identifiers (IDs) were assigned to 1,230 transcripts (out of 2,646) for the WH-WL PWC and 584 transcripts (out of 1,354) for the FH-FL PWC. These transcripts had a Benjamini-Hochberg adjusted *p*-value less than 0.05 but there was no LFC threshold applied. KEGG-IDs were also assigned to 200 proteins (out of 294) for the WH-WL PWC and 476 (out of 676) proteins for the FH-FL PWC. These proteins had a *p*-value less than 0.05 but there was no LFC threshold applied. A total of 44 maps were enriched for genes and proteins to 17 KEGG pathways with a hit of more than five genes or proteins (**Table S4.8**, Appendix 3). A KEGG pathway map is a molecular interaction network diagram that uses experimental evidence from well-studied organisms to represent the current knowledge of molecular interactions, reaction and relation networks. One map in the carbohydrate metabolism pathway, starch and sucrose metabolism, was identified as having more than five genes and proteins for each PWC. Maps were visualised for each PWC and dataset (transcriptomic and proteomic) with transcripts and proteins assigned a colour based on the LFC. For the transcriptomic and proteomic data, the right-hand side of the pathway was highlighted as up-regulated for the WH-WL PWC (**Figure S4.3**, Appendix 3 and **Figure S4.4**, Appendix 3), while the left-hand side of the pathway was highlighted as up-regulated for the FH-FL PWC (**Figure S4.5**, Appendix 3 and **Figure S4.6**, Appendix 3). This analysis appears to suggest that different parts of the starch and sucrose metabolism reference pathway are active in WH and FH populations.

4.4.5.1 Simplified schematic of starch and sucrose metabolism pathway to identify similarities between the two datasets

Using the gene model IDs identified from the KEGG analyses that were involved in the starch and sucrose metabolism pathway (SSMP), and the gene model IDs identified from the GO enrichment that had a function in the SSMP, a simplified schematic of the SSMP was generated to identify which gene model IDs were up- and down-regulated for each PWC and dataset in the pathway; and to identify if there was similarity between the datasets (transcriptomic and proteomic) for the gene model IDs. The gene model IDs have different names for the two datasets, therefore, to make direct comparisons about the expression and abundance profiles between them, identification of matching gene model IDs was determined. The following nomenclature is used below for matching gene model IDs between the transcriptomic and proteomic datasets: the matching gene model IDs are presented with a "/" separating the *T. repens* gene model IDs (beginning with "chrX.jg", where X specifies the chromosome number [1 to 16]) and the *T. occidentale* gene model IDs (beginning with "jg"), with the per cent identity match (IDM) presented after the two gene model IDs. Gene model IDs for each of the two datasets were matched and gene model IDs identified in the SSMP were used in a simplified schematic (**Figure 4.7**). The genes included in this pathway were related to WSC metabolism, including glucose biosynthesis, sucrose biosynthesis, starch synthesis, degradation of starch to maltose and dextrose, and cell wall component metabolism, such as degradation of cellulose into glucose.

At the assigned *p*-values of significance, there was little overlap between the transcriptome and proteome datasets for both PWCs. Only three gene families (*BAM*, *glgC* and *WAXY*) consisting of five gene model IDs (*BAM*: chr12.jg1584.t2 and chr12.jg4565.t1, *glgC*: chr4.jg5408.t1, *WAXY*: chr1.jg8555.t1 and chr3.jg7615.t1) showed corroboration between the transcriptome and proteome datasets for at least one PWC (**Figure 4.7**). The gene model IDs (chr12.jg1584.t2/jg.28239.t1 [98.8% identity match (IDM)] and chr12.jg4565.t1/jg28239.t2 [99.8% IDM]) for *BAM* (beta-amylase), were significantly up-regulated in WH for both the transcriptomic (both LFC > 1.7 and Benjamini-Hochberg adjusted (adj) *p*-values < 0.05) (**Table S4.4**, Appendix 3) and proteomic (both LFC > 3.4 and *p*-values < 0.05) (**Table S4.5**, Appendix 3) datasets; were significantly up-regulated in FH for the transcriptomic dataset (both LFC > 0.2 and *p*-values < 0.05) (**Table S4.4**, Appendix 3); and although not significant as determined by the Wilcoxon tests (both LFC > 2.5 and *p*-values = 0.093 and 0.065 for chr12.jg1584.t2 and chr12.jg4565.t1, respectively), the two gene model IDs were

identified as significant in FH for the proteomic dataset in the ANOVA (**Table S4.7**, Appendix 3) and correlation analyses (**Figure S4.2**, Appendix 3).

The gene model ID (chr4.jg5408.t1/jg17067.t1 [95.8% IDM]) for *glgC* (glucose-1-phosphate adenylyltransferase), was significantly up-regulated in both WH and FH for the transcriptomic dataset (both LFC > 1.2 and adj *p*-values < 0.05) (**Table S4.4**, Appendix 3); was significantly up-regulated in FH for the proteomic dataset (LFC = 2.54 and *p*-value = 0.026); and although not significant in the proteomic dataset for WH as determined by the Wilcoxon test (LFC = 1.75 and *p*-value = 0.18), followed the same expression profile as FH. As this gene model ID had been found to be significant in the FH-FL comparison only it was not one of the proteins that was included in the ANOVA and correlation analyses, as those proteins were determined from the WH-WL PWC only.

The gene model IDs (chr1.jg8555.t1/jg67692.t1 [96.9% IDM] and chr3.jg7615.t1/jg.67692.t1 [98.4% IDM], note the gene model ID for the transcriptomic dataset was matched to both gene model IDs for the proteomic dataset) for *WAXY* (granule bound starch synthase) were significantly up-regulated in WH and FH for the transcriptomic dataset (LFC > 1.5 and adj *p*-value < 0.01 for WH; and LFC > 0.2 and *p*-value < 0.05 for FH) (**Table S4.4**, Appendix 3); were significantly up-regulated in FH for the proteomic dataset (chr1.jg8555.t1: LFC = 3.35 and *p*-value = 0.015; and chr3.jg7615.t1: LFC = 3.41 and *p*-value = 0.0087); and although not significant for WH as determined by the Wilcoxon test (chr1.jg8555.t1: LFC = 2.17 and *p*-value = 0.13; and chr3.jg7615.t1: LFC = 2.29 and *p*-value = 0.24), followed the same expression profile as FH. As with the gene model ID for *glgC* (previous paragraph), these gene model IDs were found to be significant in the FH-FL comparison only, and therefore were not included in the ANOVA and correlation analyses.

Twelve gene model IDs showed the same expression profile for both WH-WL and FH-FL in one of the datasets only (transcriptomic or proteomic). These gene model IDs belonged to: *glgC* (chr5.jg7106.t1 and jg5855.t1/chr11.jg7807.t1 [91.7% IDM]), *glgA* (chr16.jg6075.t1 and chr9.jg6601.t1), *glgB* (chr12.jg7613.t1), *BAM* (chr1.jg12949.t1), *AMY* (chr16.jg2307.t1 and chr8.jg3312.t1), *ISA3* (chr3.jg8461.t1), *BGAL* (jg71018.t1 and jg71018.t2) and *LI* (jg14031.t2). For the gene model IDs identified in the transcriptomic dataset, the *LI* and the two *BGAL* gene model IDs were down-regulated and significant for WH-WL at the Benjamini-Hochberg adjusted (adj) *p*-value threshold (α = 0.05) but only significant at the *p*-value threshold (α = 0.05) for FH (**Table S4.4**,

Appendix 3). However, the ANOVA analysis showed a significant response for selection group (L and H) for all three gene model IDs (**Table S4.6**, Appendix 3). The other gene model ID from the transcriptomic dataset (jg5855.t1) was significantly up-regulated at the adj *p*-value for both PWCs (**Table S4.4**, Appendix 3). For the gene model IDs identified in the proteomic dataset, all eight were identified as significantly up-regulated in the WH-WL PWC (all LFC > 1.3 and *p*-values < 0.05) but not significantly up-regulated in the FH-FL PWC (all LFC > 1 and *p*-values > 0.05) (**Table S4.5**, Appendix 3). However, the ANOVA analysis showed a significant response for selection group (L and H) for all eight gene model IDs (**Table S4.6**, Appendix 3).

An increase in soluble sugars and starch content in the WH and FH populations was observed from phenotype data (**Figure 4.3**). Up-regulation of starch synthesis from α-D-Glucose-1P and ADP-glucose; and starch degradation into maltose and dextrin is inferred from both transcriptomic and proteomic datasets in both WH and FH. Therefore, the genes involved in this part of the SSMP (*glgC*, *WAXY* and *BAM*) were identified as candidate genes undergoing selection in both pools. The other gene model IDs involved in the starch synthesis and degradation pathways (*AMY*, *glgA*, *glgB* and *ISA3*), encompassing a total of 14 gene model IDs, were also investigated for SNPs that may be linked to high or low WSC phenotypes. This is reported in the following section (4.4.6).

4.4.6 Discriminant analysis of principal components to identify single nucleotide polymorphisms driving separation between high and low water-soluble carbohydrate individuals in 14 candidate genes.

For the 14 candidate genes, 33 to 124 SNPs (mean = 92.5) were used for each DAPC analysis. Clear separation of high WSC (High) and low WSC (Low) individuals on the first discriminant function (DF) was observed for four gene model IDs (e.g., chr8.jg3312.t1 *AMY*), but for other genes no clear separation between High and Low individuals was observed on the first DF (e.g., chr1.jg12949.t1 *BAM*). Amongst the 4 gene model IDs that did show clear High and Low separation, 1 to 5 SNPs, with a mean of 2.8 SNPs, were identified as separating High and Low individuals. There were three gene model IDs (chr8.jg3312.t1 [*AMY*], chr5.jg7106.t1 [*glgC*] and chr3.jg7615.t1 [*WAXY*]) that had SNPs that showed clear differences in genotypes between the High and Low individuals. For example, all individuals from the high populations were heterozygous at SNP 24305522 in chr8.jg3312.t1 (*AMY*), while eight out of ten individuals from the low populations were homozygous for the reference allele at this

SNP. The other 11 gene model IDs did not show distinct differences like the above-mentioned example (**Figure S4.7**, Appendix 3).

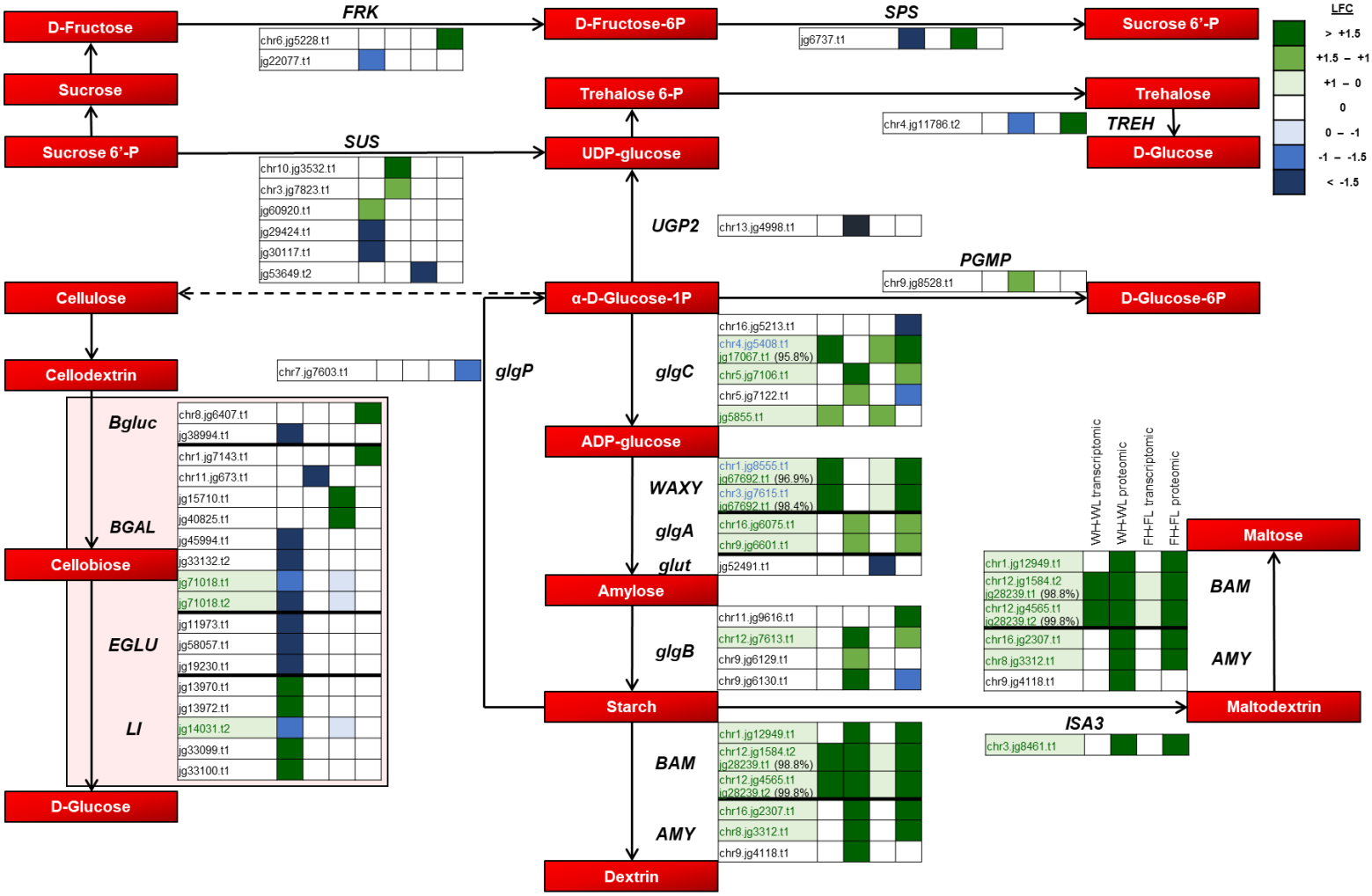


Figure 4.7 (Figure legend on next page)

Figure 4.7 Simplified schematic of genes involved in the starch and sucrose metabolism pathway. *AMY* = alpha-amylase, *BAM* = beta-amylase, *BGAL* = beta-glucosidase, *Bgluc* = glucan endo-1,3-beta-glucosidase, *EGLU* = endoglucanase, *FRK* = fructokinase, *glgA* = starch synthase, *glgB* = 1,4-alpha-glucan-branching enzyme, *glgC* = glucose-1-phosphate adenylyltransferase, *glgP* = glycogen phosphorylase, *glut* = glucuronosyltransferase, *ISA3* = isoamylase 3, *LI* = cyanogenic beta-glucosidase, *PGMP* = phosphoglucomutase, *SPS* = sucrose-phosphate-synthase, *SUS* = sucrose synthase, *TREH* = trehalase, *UGP2* = UTP-glucose-1-phosphate uridylyltransferase, and *WAXY* = granule bound starch synthase.

The dashed arrow indicates there is one intermediate sugar (GDP-Glucose) between α-D-Glucose-1P and cellulose. Next to each pathway is a heatmap list of gene model IDs. A thick black line is present on the heatmap when there is more than one gene involved in the conversion from one carbohydrate to the next, with the corresponding gene name present to the left of the gene model IDs. The colours in the heatmaps represent the level of $\log_2$ fold change (LFC) for WH-WL and FH-FL, where: < -1.5 = dark blue, -1.5 to -1 = medium blue, -1 to 0 = light blue, 0 = white (no significant change), 0 to 1 = light green, 1 to 1.5 = medium green and > 1.5 = dark green. The columns in the heatmaps, from left to right, correspond to WH-WL transcriptomic, WH-WL proteomic, FH-FL transcriptomic and FH-FL proteomic. Colour is only added to the heatmap for transcripts with a significant LFC or if gene model IDs were identified as significant from the permutation ANOVA and correlation analyses and that were one of the 26 genes or 13 proteins.

Gene model IDs (left-most column in each heatmap) highlighted in light green indicate that there was expression profile commonality between WH-WL and FH-FL in either the transcriptomic dataset, proteomic dataset or both datasets. For three of these gene model IDs in *glgC* and *WAXY* a blue transcript is highlighted and this indicates that the LFC for the WH-WL proteomic dataset was above +1.5 and followed the same expression profile pattern as FH-FL, but the *p*-value was not significant for the PWC. Percentage next to gene model ID indicates the per cent identity match between the *T. repens* and *T. occidentale* transcripts.

Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

4.5 Discussion

4.5.1 Reading the transcriptome and proteome

In the present study, transcriptomic and proteomic analyses were undertaken to identify carbohydrate metabolism genes in white clover populations responding to artificial selection for high or low levels of water-soluble carbohydrate (WSC). This study identified a number of candidate genes involved in starch synthesis and starch degradation. The commonality between findings from the transcriptome and proteome, and issues for interpretation of these data, are discussed below.

Identifying genes that underpin phenotypes can be achieved through genomics, transcriptomics, proteomics, or a combination of the three. However, concordance between the transcriptome and proteome is often moderate, with Pearson correlation coefficients (r) typically in the range of r = 0.30 – 0.52 (de Sousa Abreu *et al.*, 2009; Maier, Güell & Serrano, 2009; Voelckel *et al.*, 2010; Ghazalpour *et al.*, 2011; Vogel & Marcotte, 2012). Downstream processes such as post-transcriptional regulation and protein degradation may cause discrepancies between transcriptome and proteome profiles (Cox, Kislinger & Emili, 2005; Ghazalpour *et al.*, 2011). Protein profiles may predict phenotypes more reliably than transcript profiles due to the closer link between traits and protein levels (Voelckel *et al.*, 2017). Furthermore, proteomic data provide protein-level validation of gene expression data. However, the number of proteins surveyed in proteomic studies is far fewer than the number of gene transcripts analysed in RNA-Seq studies. Therefore, a combination of the two approaches is suggested to accurately and efficiently identify genes under selection (Voelckel *et al.*, 2017).

Data analyses were targeted to a defined subset of a biologically relevant pathway and gene families that were consistently observed across two genetic pools. Hence, the discussion is focused on the carbohydrate metabolism for both gene ontology (GO) term and Kyoto Encyclopedia of Genes and Genomes (KEGG). To my knowledge, this is the first evaluation of the white clover transcriptome and proteome for populations with divergent WSC levels and the patterns observed in these data now provide a foundational basis for investigating genes that determine high WSC in white clover leaves.

4.5.1.1 Gene ontology enrichment analysis of transcriptome and proteome data

In the present study, GO enrichment showed 151 carbohydrate transcripts and 42 carbohydrate proteins in the WH-WL pairwise comparison (PWC) were enriched. Interestingly, in these analyses, carbohydrate metabolism was not enriched for the FH-FL PWC, for either dataset. However, when assessing transcript expression and protein abundance for all four populations, 26 transcripts and 13 proteins followed a common trend in the two pools (**Figure 4.5** and **Figure 4.6**). This trend consisted of a pattern of either a higher or lower level of expression in WH and FH when compared to WL and FL. Permutation analysis of variance (ANOVA) tests and correlation analyses emphasised that the gene model IDs are expressed in a similar manner, with all gene model IDs showing a significant difference between high and low WSC populations (**Table S4.6**, Appendix 3 and **Table S4.7**, Appendix 3).

Based on the differential expression analysis using DESeq 2 for the 26 transcripts, five gene model IDs were significant for both WH-WL and FH-FL at the adjusted *p*-value threshold of 0.05 and an additional three gene model IDs were significant in both WNZLL and FNZLL pools if the adjusted *p*-value was relaxed to 0.1 in FH-FL. These eight gene model IDs belonged to seven genes: *BFRUCT1*, *CHR12*, *glgC* (two gene model IDs), *LI*, *MYH11*, *PPE8B* and *rham* (**Table S4.4**, Appendix 3). Nineteen of the 26 gene model IDs were significant at the *p*-value threshold of 0.05 for both PWCs although the $\log_2$ fold change (LFC) was above the ±2 threshold for only three gene model IDs (**Table S4.4**, Appendix 3).

Based on the differential protein abundance analysis, using Wilcoxon tests, for the 13 proteins no gene model IDs were significant at the threshold of 0.05 for FH-FL, but five were significant if the threshold was relaxed to 0.1. Three of these gene model IDs belong to *BAM*, one was an alpha-glucan water dikinase gene (*GWD2*) and the last was a starch synthase gene (*glgA*) (**Table S4.5**, Appendix 3). As there were only five gene model IDs that were significantly differentially expressed in the FH-FL PWC for the transcriptomic dataset, there was not a large enough number of carbohydrate related genes for "carbohydrate metabolic process" to be identified by GO enrichment. Similarly, as the 13 gene model IDs in the proteomic dataset were not significantly differentially expressed in the FH-FL PWC, it is not surprising that no enrichment was observed for "carbohydrate metabolic process" as the GO enrichment analysis only uses significantly differentially expressed transcripts (DETs) and differentially abundant proteins (DAPs) for the analysis. To test whether relaxing the α threshold would show enrichment for

"carbohydrate metabolic process" in the FH-FL PWC, GO enrichment was re-run using a less stringent α threshold of 0.1 for DETs and DAPs. However, when this was performed, carbohydrate metabolic process was still not enriched in either the transcriptomic or proteomic datasets. This indicates that, even at a lower threshold, the number of carbohydrate related genes that are differentially expressed in FH-FL was still too low to show any enrichment in the GO analysis.

In comparison to WH–WL there was less phenotypic divergence between the FH and FL populations. There was an approximately 100 g kg^{-1} dry matter (DM) difference between WH and WL compared with approximately 75 g kg^{-1} DM between the FH and FL populations. This could explain why fewer WSC related transcripts and proteins were identified as being significantly differentially expressed and differentially abundant compared to the WH-WL PWC – a greater LFC was observed in the WH-WL PWC than the FH-FL PWC for 21 out of the 26 transcripts and 10 out of the 13 proteins (**Table S4.4**, Appendix 3 and **Table S4.5**, Appendix 3). It is unlikely that the difference in WSC content between the WNZLL and FNZLL pools was due to environmental or other confounding factors. Spatial positioning of plants within the Latin square was also examined, as was the health status and abiotic stress exposure of individuals used in the transcriptomic and proteomic analyses. No individuals used in this chapter were recorded as having been exposed to soil moisture deficit (where the sprinkler did not reach and hand watering was required), nor obviously affected by common diseases of white clover, such as powdery mildew and pepper spot. Instead it is more likely that breeding for low WSC was more successful in WL than in FL. It should be noted, though, that phenotypic WSC content was only measured at one daily time point and it is possible that the starch and sugar level in the FL individuals did not deplete overnight as much as it did for the WL individuals. Future studies should sample WSC content at multiple time points during the day (e.g., 8am, 10am, 12pm, 2pm, 4pm and 6pm) to gather a more accurate representation of the WSC phenotype for each plant. Samples taken at these time points could also be used in either transcriptomic or proteomic studies.

4.5.1.2 Corroboration of transcriptome and proteome data to identify genes undergoing selection

In the present study, corroboration between expression patterns in the transcriptome and proteome for three genes: beta-amylase (*BAM*), glucose-1-phosphate adenylyltransferase (*glgC*) and granule bound starch synthase (*WAXY*) was observed.

For *BAM*, up-regulation in both datasets was observed in the WH and FH populations for two gene model IDs (chr12.jg4565.t1 and chr12.jg1584.t2). For *glgC*, up-regulation in the FH population for both the transcriptomic and proteomic datasets and for the WH transcriptomic dataset was observed for one gene model ID (chr4.jg5408.t1). Two gene model IDs (chr1.jg8555.t1 and chr3.jg7615.t1) for *WAXY* showed up-regulation in the WH transcriptomic dataset and in the FH transcriptomic and proteomic datasets. These three genes play a direct role in the synthesis and degradation of starch. *glgC* is involved in the conversion of α-D-glucose-1P into ADP-glucose, *WAXY* is involved in converting ADP-glucose into amylose (a precursor of starch), and *BAM* is involved in the degradation of starch into maltose and dextrin.

In addition to the five *glgC*, *WAXY* and *BAM* gene model IDs, there were nine other gene model IDs that showed a similar pattern of expression in both FNZLL and WNZLL pools for at least one of the datasets (transcriptomic or proteomic) in the starch biosynthesis and degradation pathway. One gene model ID showed similar up-regulation in both WH and FH for the transcriptomic data alone: *glgC* (jg5855.t1/chr11.jg7807.t1 [91.7% IDM]). The remaining eight gene model IDs showed similar up-regulation in both WH and FH for the proteomic data: *glgC* (chr5.jg7106.t1), *glgA* (chr16.jg6075.t1 and chr9.jg6601.t1), *glgB* (chr12.jg7613.t1), *BAM* (chr1.jg12949.t1), *AMY* (chr16.jg2307.t1 and chr8.jg3312.t1), and *ISA3* (chr3.jg8461.t1). *glgA*, or starch synthase, converts ADP-glucose into amylose. *glgB*, or 1,4-alpha-glucan-branching enzyme, converts amylose into starch. *AMY*, or alpha amylase, converts starch into maltose and dextrin. *ISA3*, or isoamylase 3, converts starch into maltodextrin. Single nucleotide polymorphism (SNP) variation for all 14 of these gene model IDs was investigated and is discussed in the following section (4.5.2).

4.5.2 Investigating single nucleotide polymorphism variation linked to changes in carbohydrate metabolism

SNP variation as an explanation for individuals with differing WSC content was assessed for the 14 candidate gene model IDs using multiple DAPC. For three of the genes (*AMY*, *glgC* and *WAXY*), different genotype frequencies were observed for the high and low WSC individuals. These three genes play a role in synthesising starch and degrading starch into maltose and dextrin (high molecular weight to low molecular weight WSC). The remaining 11 gene model IDs had no SNPs with different genotype frequencies. As one of these gene model IDs had differences in transcript levels but no SNP variation was found, it is possible that this was driven by selection for a transcription

factor (TF) affecting expression of this gene. There is some evidence for the role of selection on transcription factors as a consequence of phenotypic selection for divergent WSC content. In the current dataset, one TF (*GTE9-like*, chr5.jg2741.t1) linked to carbohydrate metabolism was identified from a genome-wide association study (GWAS) using SNPs identified from the transcriptome (see supplementary methods, "Genome-wide association study" in Appendix 3; **Figure S4.8**, Appendix 3). There are few studies that have studied TF *GTE9*. However, one recent study by Misra, McKnight and Mandadi (2018) provides evidence that *GTE9* mediates responses to sugar signals in *Arabidopsis thaliana* which supports *GTE9* involvement in sugar metabolism.

Additionally, as there were eight gene model IDs that showed differences in protein levels between low and high WSC groups, without concurrent changes in transcript levels, it is also possible that some form of post-translational modification is occurring. The flow of genetic information from DNA to protein is initiated by transcription. Thus, it has been thought that the number of transcripts is the main determinant of the amount of protein (Vogel & Marcotte, 2012). However, many proteins are present at levels that cannot be predicted from the level of mRNA, including most transcription factors; and the occurrence of post-transcriptional regulation makes mRNA an imperfect indicator of protein regulation and modification (Ghaemmaghami *et al.*, 2003). The main determinant of protein level in a cell is post-transcriptional regulation, which encompasses post-translational modification, modulation of protein half-life and delay of protein synthesis (Liu, Beyer & Aebersold, 2016). As these processes affect protein formation, they could be responsible for discrepancies in the relationship between the level of mRNA and the level of corresponding protein (Lee *et al.*, 2011; Ponnala *et al.*, 2014).

Expression changes in transcripts and proteins in the starch and sucrose metabolism pathway have been observed as potential responses to selection for WSC levels, but at present it is unclear what drives those differences. Are there mutations in the genes that might act in a concerted fashion (Carciofi *et al.*, 2012; Li *et al.*, 2016)? Does a transcription factor or microRNA bind to some or all the promoters of all the genes involved in this pathway (Sun *et al.*, 2003; Fu & Xue, 2010; Van Harsselaar *et al.*, 2017; López-González *et al.*, 2019)? Is there a trans-acting factor, cis-acting element, or post translational modification occurring (Tiessen *et al.*, 2002; Kolbe *et al.*, 2005; Li *et al.*, 2017a)? It is important to note that very few individuals were used for the DAPC and GWAS analyses (20 in total). Future studies including more individuals with known phenotype to examine sequence variation in the 14 candidate gene model IDs would

likely be informative. Designing primers to sequence a small portion (allele-specific assays designed around the SNPs identified in this study) of each of the 14 candidate gene model IDs would be a cost-effective and rapid way to identify homozygotes and heterozygotes. While this study has revealed genes that potentially underpin improved foliar WSC accumulation in white clover, further investigation is still required to more fully understand the genetic determinants of WSC accumulation.

4.5.3 Accounting for evolutionary history when identifying genes under selection

The design of the present study meant that patterns of differential expression amongst transcripts and proteins could be mapped onto branches of the underlying phylogenetic tree that describes the relationship of the four populations (**Figure 4.1**). While assignment is influenced by assumed levels of significance, for a specified *p*-value we could infer expression level changes that occurred on different branches of the phylogenetic tree, as well as changes that did not fit onto this phylogenetic tree. This mapping identified changes that could be explained most parsimoniously as a convergent response to selection (up-regulated patterns of expression shared by WH and FH populations and/or down-regulated patterns of expression shared by WL and FL populations; Pattern 6 on **Figure S4.9** and **Figure S4.10**), changes more likely explained by correlation with phylogenetic structure (patterns of differential expression differing between the WNZLL and FNZLL pools; Pattern 5 on **Figure S4.9** and **Figure S4.10**), and those that represent undetermined stochastic/systematic noise (i.e., could not be attributed to either of the two other causes; Other patterns on **Figure S4.9** and **Figure S4.10**).

For the 151 transcripts and 42 proteins that were identified from the WH-WL GO enrichment analysis (carbohydrate related gene model IDs), the expression patterns were mapped to branches of the phylogenetic tree (**Figure S4.9**, Appendix 3). The percentage of the transcripts and proteins that mapped to a convergent selection response (Pattern 6) was consistent between both datasets (32.5% and 31% for transcriptome and proteome, respectively). As the pattern assignment was focused on the 151 transcripts and 42 proteins that were identified by GO enrichment in the W pool, it was not surprising to find expression patterns that mapped uniquely to W populations (e.g., 17 – 19% were mapped to WH and 14 – 21% were mapped to WL). Furthermore, as expression of transcripts and proteins had to differ between WH and WL to be considered differentially expressed, it was not surprising that no gene model IDs exhibited expression Pattern 5.

To obtain a better estimate of the number of differentially expressed transcripts and proteins that could be mapped to all the above-mentioned patterns, an alternative methodology called a weighted gene correlation network analysis (WGCNA) was implemented (performed by Australian Proteomic Analysis Facility). This analysis identifies “hubs” of highly correlated gene model IDs (showing the same expression profiles) using hierarchical clustering from expression data, and summarises each hub using representative eigenvalues (Langfelder & Horvath, 2008; Ficklin *et al.*, 2017). For this, ANOVAs were run on all transcripts and proteins, and transcripts and proteins that were significantly different in at least two populations were classified as differentially expressed (3,792 transcripts and 633 proteins were identified as differentially expressed). The transcripts and proteins with similar expression profiles were grouped into a hub. The representative expression profile of the hubs was then assigned to one of the pattern types (**Figure S4.10**, Appendix 3). The mapping suggested that many of the genes exhibiting differential expression in the experiment were not genes whose response was linked to selection (only 16.6% transcripts and 8.7% proteins were mapped to Pattern 6). This is of interest because in transcriptome and proteome profiling studies long lists of differentially expressed genes and proteins are often identified. In the present study at least some of these patterns could be explained. It is encouraging that many genes involved in carbohydrate metabolism were mapped as a selection response indicating the potential of differential expression studies for identifying candidate genes responding to selection. An interesting question is whether a more formal approach for phylogenetic modelling of transcriptome and proteome profiles such as recently discussed (Rohlfs & Nielsen, 2015; Voelckel *et al.*, 2017; Cope, O'Meara & Gilchrist, 2020) would also highlight the carbohydrate genes identified by this analysis, as well as other genes exhibiting differential expression. Perhaps one limitation of gene and protein expression profiling is the very large number of differentially expressed genes that can be identified in these studies. This is because understanding and assessing all gene expression changes is a cumbersome and time-consuming task. Phylogenetic modelling provides one approach that might help to more efficiently identify candidate genes of potential interest. With the data obtained in the present study an interesting question would be whether phylogenetic modelling would identify carbohydrate metabolism genes in both the WNZLL and FNZLL pools that have responded to artificial selection. If so, this would be significant for downstream analyses. GO enrichment and KEGG analyses are limited by the significance of a transcript or protein from PWC differential analysis. This was seen in the current study, where the FNZLL pool had fewer significant WSC gene model IDs identified from the PWC for both

datasets, and the GO enrichment and KEGG analyses did not identify many carbohydrate-related transcripts and proteins as under selection. Furthermore, different parts of the starch and sucrose metabolism pathway were highlighted. When significance criteria are relaxed or ANOVA and correlation are used, there are similarities between the two pools.

4.5.4 Developing a metabolic model for carbohydrate metabolism in white clover

Developing a metabolic model for carbohydrate metabolism in white clover would likely be of great value in achieving more sustainable agriculture with white clover (Cañas *et al.*, 2017; Fang, Luo & Wang, 2019). Such a model would require understanding the contribution of different regulatory processes that determine flux through the sucrose synthesis/degradation pathway. With my work to date, some progress has been made towards understanding the metabolic process that contribute to the WSC phenotypes. However, more detailed studies are needed to evaluate the relative contribution of cis- and trans-acting transcriptional, and post-translational modification in determining this metabolic flux.

4.6 Conclusions

Transcriptome sequencing and proteomics are often undertaken in parallel with genomics to add an additional layer of evidence to the genomic data. Standard pairwise comparison (PWC) analyses for differential transcript expression and protein abundance was able to identify significant changes in expression of genes related to carbohydrate metabolism for the WNZLL pool. However, PWC analysis for the FNZLL pool failed to show the same enrichment for carbohydrate metabolism genes. This is correlated to an apparently weaker selection response in the FNZLL populations, with a lower level of phenotypic divergence observed between low and high populations. Using PWC we were not able to identify a convergent expression response in WNZLL and FNZLL high water-soluble carbohydrate (WSC) populations. However, by accounting for phylogenetic relationships I was able to identify expression patterns most parsimoniously explained as a response to artificial selection. Fourteen carbohydrate genes were shown by analysis of variance (ANOVA) and correlation analysis to have similar patterns of expression in response to selection. Increasing WSC content in white clover leaves appears to be strongly affected by regulation of genes involved in the starch biosynthetic and degradation pathways. There was strong evidence that sorting of allelic variants for gene model IDs belonging to *glgC*, *WAXY* and *AMY* gene families correlated to the high WSC phenotype. There was strong evidence for an additional 11

gene model IDs responding to selection, but the significance of the observed single nucleotide polymorphism (SNP) variation is unclear. Future work involving sequencing the 14 gene regions in more population samples may help to identify allelic variants that can be used to efficiently breed for high WSC white clover individuals.

4.7 Acknowledgements

I would like to thank Anna Larking (AgResearch) for help collecting leaf material and RNA extraction guidance. Simon Zhang (Custom Science) for organising RNA sample shipment. Mehdi Mirzaei, Dana Pascovivi and Jemma Wu (all Macquarie University, Australia) for protein extraction, proteomics and WGCNA analyses. Abdul Baten (AgResearch) for extracting protein sequences from the *T. occidentale* and *T. repens* genomes. Marni Tausen (Aarhus University, Denmark) for performing the HyLiTE analysis and SNP calling. Paul Maclean (AgResearch) for statistical support, including help with differential transcript and protein expression analyses, GO enrichment, KEGG analyses, correlation, ANOVA and identification of matching gene model IDs.

CHAPTER 5

Local adaptation in white clover (*Trifolium repens* L.) populations associated with climate variation

5.1 Abstract

New Zealand's agricultural sector is dependent on rainfall for its productivity and so periods of soil moisture deficit (SMD) can impose adverse conditions on pastoral systems. The aim of the current study was to identify regions of the genome associated with tolerance to high SMD. The approach taken involved assessing 17 naturalised white clover populations likely to have undergone local adaptation in climatically contrasting environments, namely areas of high and low SMD in the South Island/Te Waipounamu of New Zealand/Aotearoa (NZ). Each population was represented by approximately 40 individuals sampled from a single paddock at each of the 17 locations (n = 674) that had not been resown or oversown with white clover in the previous 20 years. Population structure assessment was performed using genotype data obtained by genotyping by sequencing (GBS) on each individual from each population to minimise confounding effects. Analysis based on the resulting 15,120 single nucleotide polymorphisms (SNPs) identified significant population structure. Three to five genetic clusters were observed from ADMIXTURE and sNMF analyses. Using this same dataset, 64 candidate SNPs associated with variation in SMD were identified by at least two outlier detection analyses (OutFLANK) and environmental association analyses (*LEA*, *lfmm* and BayeScEnv). Mapping these SNPs to the white clover reference genome suggested a subset were associated with genes involved in carbohydrate metabolism and root morphology based on gene ontology analysis. Variation for five SNPs suggests directional selection for a common set of allelic variants has occurred in a subset of the populations from high SMD environments. However, a more comprehensive suite of experiments is suggested to elucidate the genetic basis of adaptation of white clover populations to SMD. This is the first in depth assessment of population structure of naturalised white clover populations in NZ and provides a starting resource for future studies assessing local adaptation, investigating genetic and physiological controls of response to SMD, and breeding for SMD tolerance in white clover.

5.2 Introduction

Plants are sessile in nature and are exposed to changing environments including various abiotic stresses. These abiotic stresses can include temperature extremes, flooding, water scarcity, high light intensity, salinity and nutrient deficiency/excess that either individually, or in combination, negatively affect plant growth, development and productivity. Abiotic stress contributes substantially to reduction in global crop yields (Acquaah, 2012). Management of abiotic stress, for example through irrigation, is required to ensure crop productivity in extreme environmental conditions. Improving plant abiotic stress tolerance is therefore a key objective for breeders to ensure the longevity and reliability of the world's global food supply. Increasing our understanding of the genetic basis of plant traits is fundamental for improvement of crops. The most limiting factor in crop production, especially in agriculture, is water (Acquaah, 2012). Soil moisture deficit (SMD) is the number of days per year when the soil moisture in the root zone is less than half of the soil moisture holding capacity (van Ham *et al.*, 2016), and is calculated as the incoming daily rainfall, outgoing daily potential evapotranspiration and fixed available water capacity. New Zealand's pastoral agriculture systems are dependent on rain water for productivity and so periods of SMD can impose adverse growth conditions and reduce feed availability from pastures. Extreme weather events, such as the frequency and duration of droughts that influence SMD, are expected to increase in frequency due to the increased rate of the changing climate (Reisinger *et al.*, 2010). A plant's ability to react to changing environmental conditions depends on a combination of adaptation, gene flow and phenotypic plasticity (Anderson, Panetta & Mitchell-Olds, 2012). Adaptation via changes in allele frequencies or genotypic recombination, i.e., microevolution, may be the key process for species longevity (Bell & Gonzalez, 2009). Hence, genetic variability underpins a species' vegetative persistence in its current distribution and has the potential to mitigate the adverse effects of rapidly progressing climate change (Pauls *et al.*, 2013).

In New Zealand/Aotearoa (NZ), the occurrence of prolonged periods of SMD are becoming more frequent and have substantial negative economic effects (Kamber *et al.*, 2013). As SMD imposes strong selective pressure on natural plant populations, investigating local adaptation to high SMD regions in naturalised populations can be informative for efforts to produce SMD tolerant populations (Mickelbart *et al.*, 2015). Studying patterns of local adaptation at the genomic level is a way to uncover signatures of selection and to identify genes, pathways and environmental factors driving the evolutionary process (Barrett & Schluter, 2008). Landscape genomics is a recent

research field that combines landscape genetics and population genomics to link environmental factors to present-day adaptive genetic variation and the gene variants that drive local adaptation (Rellstab *et al.*, 2015). The Southern Alps/Kā Tiritiri o te Moana in the South Island/Te Waipounamu of NZ create a distinct rain shadow effect, where regions on the Eastern side are typically drier and regions on the Western side are typically wetter (Sturman & Wanner, 2001). The environmental heterogeneity represented make it possible to investigate adaptation to SMD in plant populations across this gradient. A previous study by van Ham *et al.* (2016) identified 26 white clover populations distributed across East and West sides of the South Island/Te Waipounamu that were present in "naturalised" pasture. Naturalised pasture was arbitrarily defined by the authors as more than 20 years uninterrupted pasture (i.e., white clover was not resown or oversown in that timeframe). Due to the hypothesised strong environmental selective pressure, it is expected that local adaptation has driven the production of SMD tolerant white clover populations – by long term survival of better-adapted plants and elimination of inferior plants from the pasture and, assuming flowering has been able to occur under farm grazing managements, sexual reproduction among adapted individuals followed by seed set and seedling recruitment in the pasture. Hence, assessment of genetic variation among white clover populations across contrasting SMD environments may support identification of regions of the genome associated with SMD tolerance.

The principal objective of this study was to identify regions of the genome associated with SMD tolerance in white clover. A preliminary experiment to determine the minimum number of samples required from each pasture population to adequately capture population genetic diversity was performed as described in Appendix 4, Chapter 5 Supplementary Material, Supplementary experiment – Sample size determination. This underpinned the main experiment in which white clover populations from contrasting SMD zones were sampled, single nucleotide polymorphism (SNP) genotype data were generated for each population using genotyping by sequencing (GBS), and a landscape genomics approach, through environmental association analyses (EAAs), was used to identify potential adaptive SNPs. Furthermore, outlier detection analyses were used to identify SNPs differentiating "Dry" and "Wet" environment populations, categorised based on SMD. The commonality between SNPs identified from the two approaches were assessed. As population structure is often a confounder of EAAs, the population structure of a subset of these white clover populations identified from van Ham *et al.* (2016) also needed to be assessed.

5.3 Materials and methods

5.3.1 Site information and sample collection

Based on long-term soil moisture deficit (SMD) data from the National Institute of Water and Atmospheric Research (NIWA) and known 'naturalised' pasture sites (arbitrarily defined as pastures that had not been resown or over-sown for > 20 years) from van Ham *et al.* (2016), a total of 18 pastures were selected across the South Island/Te Waipounamu of New Zealand/Aotearoa (NZ) to represent contrasting SMD environments (**Figure 5.1**). These paddocks are privately owned land and consent to remove samples was provided from land owners at the time of collection (September 2018 to December 2018). Only 17 pastures were ultimately used to identify regions of the genome associated with SMD tolerance in white clover as, after collecting the material, further discussions with the landowner revealed that one population was in fact a more recent sowing (see section 5.4.1) and was consequently excluded. Due to the large rain shadow effect caused by the Southern Alps/Kā Tiritiri o te Moana, the wettest areas were found on the West Coast/Te Tai Poutini of the South Island/Te Waipounamu (populations Cape Foulwind [CF], Rahu Saddle [RS], Kumara Junction [KJ], Whataroa [WR], and Haast [HA]) and included three populations from the Tasman/Te Tai o Aorere (Lower Takaka [LT]), Marlborough/Tauihu (Rai Valley [RV]) and Otago/Otakou regions (Makarora [MR]). The driest areas were the Canterbury Plains/Waitaha (populations Clarence [CL], Kaikoura [KK], Waipara [WP], Waikuku [WK] and Southbridge [SB]), Otago/Otakou (populations Arrowtown [AT], Fruitlands [FL], Middlemarch [MM] and Omarama [OM]) and the Marlborough/Tauihu region (Awatere Valley [AV]) (**Figure 5.1**).

At each of the 18 sites, 50 white clover plants were sampled in linear transects across the field, giving a total of 850 individuals. Various field studies have reported clonal patches of white clover can occupy varying areas from a few cm up to 3 m^2 (Harberd, 1963; Cahn & Harper, 1976; Gustine & Sanderson, 2001). Therefore, to avoid collecting plants from the same clonal patch, plants were collected at 10 m intervals with a 10 m distance between transects. Leaf material was collected fresh and placed in silica gel to desiccate and store the material until the return to the laboratory for DNA extraction. A preliminary study (See Chapter 5 Supplementary Material, Supplementary experiment – Sample size determination in Appendix 4) was performed to determine the minimum sample size required to adequately represent the genetic variation within the pasture population. Briefly, a set of individuals (n = 91) from different cultivars were screened with microsatellite (simple sequence repeat, SSR) markers and assayed for

genetic diversity as the sample size was increased in increments of 10. This study found that a minimum of 30 individuals were required to represent the genetic variation within a cultivar. Therefore, a sample of 50 individuals from each pasture population was considered sufficient and provided for additional samples to be available in case DNA extraction failed. Forty samples per population were subsequently genotyped (section 5.3.2), to allow for samples with poor quality GBS data to be excluded. Allowing for redundancy at DNA extraction and genotyping stages meant that a minimum of 30 samples should be available for subsequent analyses.

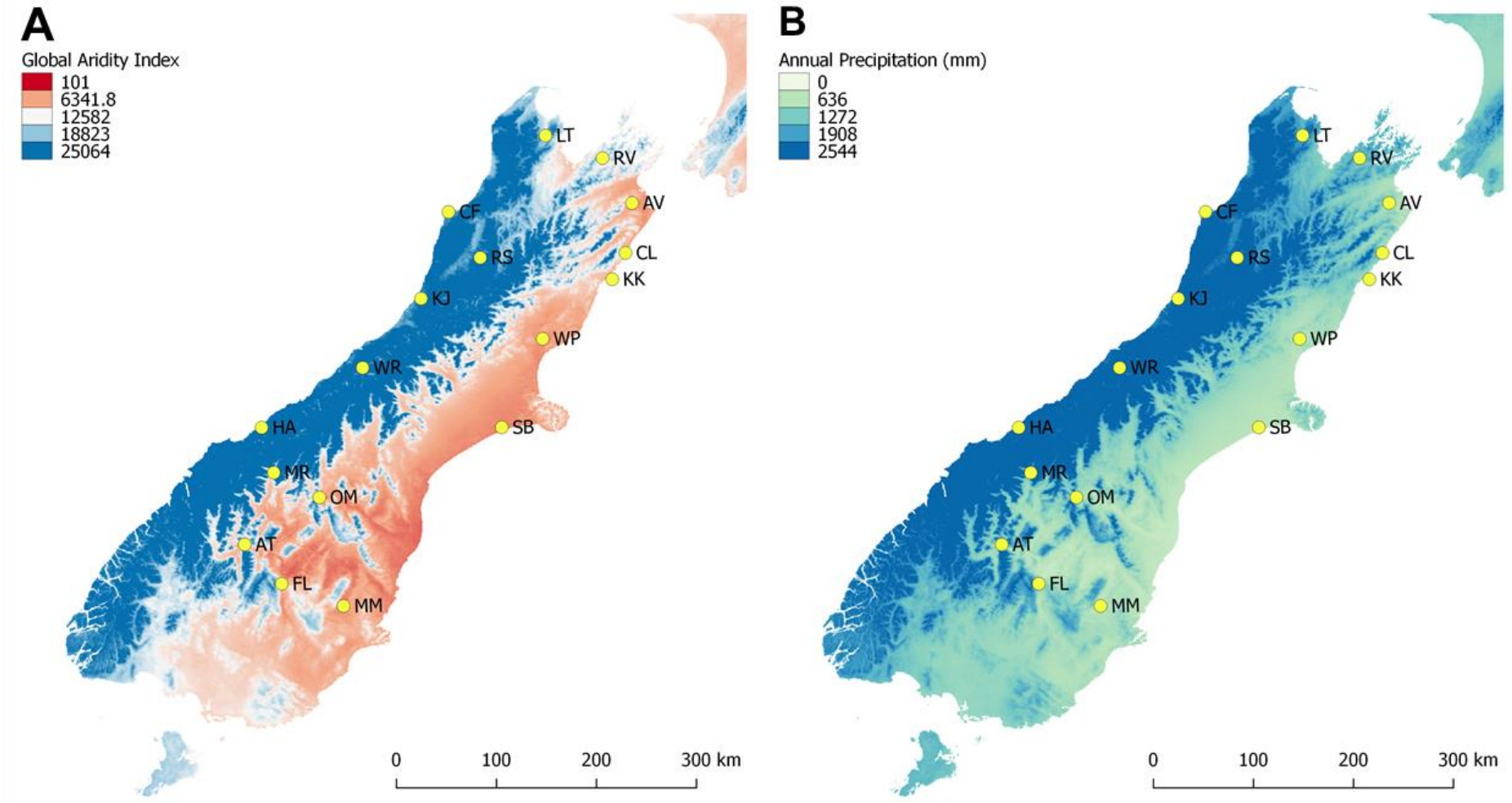


Figure 5.1 Geographic locations of 17 white clover populations sampled in the South Island/Te Waipounamu of New Zealand/Aotearoa. Global aridity index (**A**) and annual precipitation (**B**) were strongly correlated with principal component 1, which explains 47.6% of the climatic variation

across the 17 populations. Maps were created in QGIS v 2.18.28 and are shaded according to the climatic variables downloaded from the WorldClim Global Climate Data (http://WorldClim.org). The colour scale in **A** represents the global aridity index (AI). Lower AI values and red colours indicate areas are more arid, while higher AI values and blue colours indicate areas are more humid. The colour scale in **B** represents the annual precipitation (AP). Lower AP and green colours represent less rainfall, and higher AP and blue represent more rainfall. AT = Arrowtown, AV = Awatere Valley, CF = Cape Foulwind, CL = Clarence, FL = Fruitlands, HA = Haast, KJ = Kumara Junction, KK = Kaikoura, LT = Lower Takaka, MM = Middlemarch, MR = Makarora, OM = Omarama, RS = Rahu Saddle, RV = Rai Valley, SB = Southbridge, WP = Waipara and WR = Whataroa. Waikuku (WK) is removed from maps as it was not used in analyses.

Slope details were determined at the time of collection using a compass and clinometer for each site. Slope has two components: the gradient (the angle that the plane makes with a horizontal surface) and the aspect (the direction of the plane with respect to some arbitrary zero, in the current experiment, North). Each site was scored for gradient into one of seven categories: flat (0 – 3°), undulating (4 – 7°), rolling (8 – 15°), strongly rolling (16 – 20°), moderately steep (21 – 25°), steep (26 – 35°) and very steep (> 35°), with the aspect relative to North recorded. Where the paddock had areas of different gradients (e.g., flat and rolling), samples were only taken from the flat area.

At each site, 15 individual soil core samples (3 cm diameter x 7.5 cm deep) were taken from random points across the field, in a zig-zag pattern, with a minimum distance of 15 m between samples. Areas such as around fences, troughs, manure piles and gates were avoided. Samples from each site were bulked together and kept at 4^{o}C until sent to Eurofins (Auckland, NZ) to analyse soil physiochemical properties. Analysis of soils included: pH, Olsen P, Sulphate sulphur (SO_4-S), Calcium (Ca), Magnesium (Mg), Potassium (K), Sodium (Na), CEC (cation exchange capacity) and base saturation.

Nineteen bioclimatic variables including minimum, mean and maximum air temperature (°C) and precipitation (mm) data from years 1970 – 2000 were downloaded from WorldClim v 2.0 using the highest spatial resolution (~ 1 km^2) (Fick & Hijmans, 2017). An additional three variables including the average solar radiation (kJ m^{-2} day^{-1}), wind speed (m s^{-1}) and water vapour pressure (kPa) for each month of the year (January – December) for years 1970 – 2000, were also downloaded from WorldClim v 2 using the highest spatial resolution (Fick & Hijmans, 2017). Annual potential evapotranspiration (PET) for the 1970 – 2000 period (mm/year) and aridity index (AI) for the same period (AI = Mean annual precipitation divided by the mean annual potential evapotranspiration, where higher AI indicate more humid conditions and lower AI indicate higher aridity) calculated from WorldClim v 2.0 Global Climate Data were also downloaded (Trabucco & Zomer, 2019). All data were selected for downloading using the latitude and longitude for each population (**Table S5.1**, Appendix 4).

5.3.2 DNA isolation and genotyping

For the 18 populations, genomic DNA was extracted for genotyping by sequencing (GBS) in a 96-well plate format from a minimum of 48 individuals per population (9 plates in total), following the freeze-dried/silica gel-dried tissue protocol of Anderson *et al.* (2018). Samples were stored at -20°C until needed. All DNA samples were resolved and

visualised by electrophoresis to assess quality, and quantified using the methodology outlined in Chapter 3 (section 3.3.2). Samples that did not produce a single clear high molecular weight band on an electrophoresis gel were re-extracted. If the second extraction did not yield good quality DNA, then another individual was chosen for extraction.

From the sampled 900 individuals, a total of 720 were chosen for GBS, and consisted of 40 individuals from each of the 18 populations. Each GBS library can contain a maximum of 96 individuals, but one negative and one positive control was included for each library, leaving 94 positions for samples. The negative control (a water control) was used to assess cross contamination in the libraries and the positive control (inbred white clover S_9 individual (Cousins & Woodfield, 2006; Griffiths *et al.*, 2019)) was used to confirm consistency of GBS and single nucleotide polymorphism (SNP) capture across all libraries. Eight libraries were generated and 32 spare wells were filled with duplicated DNA samples chosen at random from across all populations. The duplicated samples included duplication of samples within a GBS library and duplication of samples among GBS libraries. The duplicated DNA samples within a GBS library and the positive S9 control samples were used to confirm consistency both within and between all GBS libraries, as explained in Chapter 3, section 3.3.4.3. Eight GBS libraries were constructed following the protocol of Poland *et al.* (2012b) with some modifications. See Chapter 3, section 3.3.3 for details of GBS library construction.

5.3.3 Single nucleotide polymorphism calling, filtering and genotyping by sequencing library control

5.3.3.1 Single nucleotide polymorphism calling

After sequencing (two lanes of data per plate), raw data FASTQ files containing sequence reads were processed for SNP identification using the GBS analysis work flow implemented in Trait Analysis by aSSociation, Evolution and Linkage (TASSEL) v 5.0 (Glaubitz *et al.*, 2014) using default parameters except minor allele frequency (MAF) was set to 0.01. An AgResearch white clover genome assembly was used as the reference genome (Griffiths *et al.*, 2019). Raw sequence data of 752 samples, 8 positive controls and 8 negative samples were analysed together. The purpose of performing the combined analysis of all populations together was for commonality and consistency of SNP nomenclature across all populations. A summary of the bioinformatics workflow is presented in **Figure S3.1** (Appendix 2). Sequence reads were first trimmed to 64 bp and identical reads were grouped into sequence tags. The sequence tags were then

aligned to the reference genome using Burrows-Wheeler Alignment tool (Li & Durbin, 2009).

5.3.3.2 Single nucleotide polymorphism and sample filtering

After SNP calling, the marker set was restricted to high quality SNPs by only including biallelic SNPs, a minimum and maximum read depth range of 5 to 150, limiting missing genotype data to a maximum of 20% per SNP, and including SNPs with a MAF threshold of ≥ 0.03. All filtering was performed using VCFtools v 0.1.16 (Danecek *et al.*, 2011). Samples with a large proportion of missing genotype data (> 80%) were removed from the dataset and negative control samples were removed after confirming that they did not contain unduly high levels of data.

5.3.3.3 Genotyping by sequencing library quality control

The R package "*PCAdapt*" v 4.3.1 was used to confirm consistency among and within GBS libraries. Positive control samples of a single genotype (S9), repeated in all 8 GBS libraries were used to check consistency (co-location on principal component analysis, PCA, biplots) across all 8 GBS libraries (**Figure S5.1 A**, Appendix 4). Duplicated DNA samples within a library were similarly used to confirm consistency within libraries. All duplicated samples, blank negative and S9 positive controls were checked: duplicates (data not presented) and S9 positive controls (**Figure S5.1 A**, Appendix 4) all physically co-located on biplots, while negative controls had minimal read data. All were then removed from the dataset prior to subsequent population analyses. Six samples (CF-11A, KK-10C, KK-8B, WP-3E, WP-8C and AT-9) were also removed as they contained high levels of missing data (> 80%). The remaining samples (from 18 populations) were run through PCAdapt and score plots were produced (see section 5.3.4.3 for protocol details).

5.3.4 Statistical analyses

5.3.4.1 Assessing population genetic structure

Two complimentary genetic clustering approaches were used to assess population structure: ADMIXTURE (Alexander & Lange, 2011) and the sNMF function from the "*LEA*" package v 2.6.0 (Frichot & François, 2015). ADMIXTURE v 1.3.0 was used to perform a model-based clustering analysis on the dataset, comprising 674 individuals from 17 populations, to assess population structure and determine the putative number of ancestral populations (K_P). This program uses maximum likelihood modelling to

estimate ancestry, which is much faster than Bayesian modelled ancestry implemented in STRUCTURE (Pritchard *et al.*, 2000). The ADMIXTURE analysis was run under default parameters from K_P = 1 to K_P = 20, with 20 iterations for each K_P. The cross-validation flag (--cv) was used to determine the K_P-value with the lowest cross-validation error. The K_P-value with the lowest cross-validation error was determined as the optimal K_P. ADMIXTURE results using meanQ files were displayed in the web version of pophelper (https://roymf.shinyapps.io/structure/) and the average proportion of each population assignment to K_P = 3 clusters was plotted on a distribution map using QGIS v 2.18.28.

The sNMF model uses a fast and accurate algorithm similar to Bayesian clustering programmes such as STRUCTURE (Pritchard *et al.*, 2000; François & Durand, 2010), and uses least-squares estimates of ancestry proportions rather than maximum likelihood estimates (Frichot *et al.*, 2014). Cross-validation was implemented through an entropy criterion to evaluate the quality of fit of the model to the data and to aid determination of the number of putative ancestral populations that best explain the genotypic data (Alexander & Lange, 2011). The *snmf()* function was run with K_P between 1 and 10 with cross-entropy using 20 repetitions for each K_P (**Figure S5.2 C**, Appendix 4). The most likely K_P was chosen based on minimized cross-entropy.

5.3.4.2 Assessing population genetic variation

Quantification of genetic variation within and among populations was assessed using analysis of molecular variance (AMOVA) and Weir and Cockerham's (1984) F-statistics.

Analysis of molecular variance

AMOVA was calculated in R using the package "*pegas*" v 0.11 (Paradis, 2010) to determine partitioning of genetic variation within and among populations, for three defined hierarchical partitions with 10,000 permutations for each AMOVA (Excoffier *et al.*, 1992; Paradis, 2010). The first hierarchical partition was the *a priori* grouping of 17 populations. The second was regional, where populations were categorised as "Dry" (AT, AV, CL, FL, KK, MM, OM, SB, and WP) or "Wet" (CF, HA, KJ, LT, MR, RS, RV, and WR), as indicated by soil moisture deficit. The third was between genetic clusters identified using the most supported K_P-value identified by ADMIXTURE (K_P = 3). Populations were assigned to one of the three genetic clusters based on which cluster was present at the highest proportion within the population (White-coloured cluster =

AT, FL, OM, CL and RV; Green-coloured cluster = AV, CF, HA, MM, MR, SB and WP; Black-coloured cluster = KJ, KK, LT, RS and WR).

F-statistics

For the F-statistics analysis, a matrix of pairwise genetic differentiation (F_{ST}; Weir & Cockerham, 1984) between all population pairs from the *a priori* population structure (the original grouping of K_P = 17) was computed using the R package "*Hierfstat*" v 0.04-22 (Goudet & Jombart, 2015).

5.3.4.3 Outlier detection analyses

Utilization of multiple methods to detect loci under divergent selection is recommended because when more than one line of evidence identifies the same outlier there is increased confidence in the results. Two outlier detection approaches were used to analyse the 15,120 SNP dataset of the 17 populations and 674 individuals: PCAdapt (Luu *et al.*, 2017) and OutFLANK (Whitlock & Lotterhos, 2015). Missing genotype data were not imputed for these methods.

PCAdapt

PCAdapt ascertains population structure using PCA and identifies markers under selection as those that are excessively correlated with population structure, hence individuals are not organized into predefined populations (Luu *et al.*, 2017). The variant call format (VCF) file (from section 5.4.1) was converted into PLINK format (BED, BIM and FAM files) using PLINK v 1.9 (Purcell *et al.*, 2007). The R package "*PCAdapt*" v 4.3.1 was then used to detect loci driving variation on the principal components (Luu *et al.*, 2017). The BED file was imported into R using the *read.pcadapt()* function with the "bed" type specified. PCAdapt was run using the *pcadapt()* function with a maximum "K_{PC}" of 20 and a "*min.maf*" of 0.03 specified. The K_{PC} value (number of principal components to investigate) with the best fit to the data was determined using the scree test (Cattell, 1966) implemented by running the *scree_plot()* function (**Figure S5.3**, Appendix 4). The *a priori* population grouping of K_P = 17, based on the initial populations, was used to visualise populations on the score plot which was obtained using a modified *score_plot()* function. The clustering of individuals on score plots can be used to visualise population structure captured by principal components (PCs). Outlier SNPs were identified as those responsible for differentiating "Dry" and "Wet" site populations. To obtain *p*-values for each of the outlier SNPs on the second PC, *pcadapt()* was re-run with the "componentwise" method specified. Outlier SNPs were corrected for false

positives using a *q*-value threshold of 0.05 from the "*qqman*" package v 2.16.0, which is considered a false discovery rate (FDR) of 5%. Outlier SNPs differentiating populations based on PC2 were identified using the *get.pc()* function.

OutFLANK

OutFLANK can identify F_{ST} outliers by inferring a neutral F_{ST} distribution using likelihood on a trimmed distribution of F_{ST} values. This approach has shown lower false positive rates compared to other F_{ST} outlier methods (BayeScan, Fdist2 and FLK) based on simulation studies (Whitlock & Lotterhos, 2015). The VCF file (from section 5.4.1) was imported into R using the *read.vcfR()* function from the "vcfR" v 1.10.0 package (Knaus & Grünwald, 2017). In R, loci with more than 60% missing data were removed, resulting in 8,795 remaining loci. SNP data were converted into 0, 1, 2 format using the *revalue()* function from the "*plyr*" v 1.8.6 package (Wickham, 2011) with missing data encoded as "9". Samples were classified as either "Dry" or "Wet" for population grouping. Genotype data, SNP loci and population information were then used to create an F_{ST} matrix using the function *MakeDiploidFSTMat()*. The function *OutFLANK()* was run using default parameters. Outlier SNPs were corrected for false positives using a *q*-value threshold of 0.05 from the "*qqman*" package using the *qvalue()* function.

5.3.4.4 Environmental association analyses to detect adaptive loci

Three environmental association analyses (EAAs) were used to identify putative adaptive loci. For the first two EAAs, latent factor mixed models (LFMMs) were implemented in two R packages (*LEA* and *lfmm*) to identify loci showing significant correlations with environmental variables while accounting for neutral population structure (Frichot *et al.*, 2013; Frichot & François, 2015). BayeScEnv was used in the last EAA to detect associations between environmental factors and genetic differentiation.

Environmental variables used in three environmental association analyses

Important environmental variables can be extracted from a multivariate dataset and may be expressed as a new set of two or three variables called principal components (PCs). These PCs reduce the dimensionality of a multivariate dataset as the PCs correspond to a linear combination of the original variables. As many of the available 79 environmental variables in the current dataset are related (e.g., Annual Mean Temperature, Max Temperature of Warmest Month etc), PCA was performed to identify PCs that best summarise the range of environmental variation. Environmental variables

included were: soil moisture deficit (SMD), aridity index (AI), potential evapotranspiration (PET), 19 standard bioclimatic variables (Annual Mean Temperature, Mean Diurnal Range, Isothermality, Temperature Seasonality, Max Temperature of Warmest Month, Min Temperature of Coldest Month, Temperature Annual Range, Mean Temperature of Wettest Quarter, Mean Temperature of Driest Quarter, Mean Temperature of Warmest Quarter, Mean Temperature of Coldest Quarter, Annual Precipitation, Precipitation of Wettest Month, Precipitation of Driest Month, Precipitation Seasonality, Precipitation of Wettest Quarter, Precipitation of Driest Quarter, Precipitation of Warmest Quarter, Precipitation of Coldest Quarter), solar radiation (srad) from each month of the year (1 = Jan – 12 = Dec), water vapour (vapr) from each month of the year (1 = Jan – 12 = Dec), wind speed (wind) from each month of the year (1 = Jan – 12 = Dec), Altitude, NZSC soil order, Steepness of Slope, Aspect of slope, soil pH (pH), cation exchange capacity (CEC), volume weight (VW), Olsen Phosphorus (Olsen P), Sulphate sulphur (SO_4-S), Calcium (Ca), Exchangeable Calcium (Eca), Magnesium (Mg), Exchangeable Magnesium (Emg), Potassium (K), Exchangeable Potassium (EK), Sodium (Na), Exchangeable Sodium (Ena), Calcium base saturation (CaBS), Magnesium base saturation (MgBS), Potassium base saturation (KBS) and Sodium base saturation (NaBS) (**Table S5.2**, Appendix 4).

To identify the environmental variables that best summarise the range of environmental variation, three separate PCAs were performed using the *PCA()* function from the "*FactoMineR*" v 2.3 package (Lê, Josse & Husson, 2008). The first PCA comprised 22 important environmental variables relating to precipitation and temperature: SMD, AI, PET and the 19 standard bioclimatic variables. Four environmental variables (SMD, AI, PET and PC1 co-ordinates, see section 5.4.4) were imported into R and converted into the ".env" format using the *write.env()* function. These four variables were used in subsequent *LEA*, *lfmm* and BayeScEnv analyses.

Two additional PCAs were investigated to determine if the environmental variables included contributed to the "Dry" and "Wet" separation. The second PCA assessed the monthly srad, vapr and wind variables. These 36 variables were reduced to two PCs explaining a total of 87.5% of the original variability across the 17 white clover populations (**Figure S5.4**, Appendix 4). The third PCA assessed 21 soil and site variables: Altitude, NZSC soil order, Steepness of Slope, Aspect of slope, pH, CEC, VW, Olsen P, SO_4-S, Ca, Eca, Mg, Emg, K, EK, Na, Ena, CaBS, MgBS, KBS and NaBS. These 21 variables were reduced to two PCs explaining a total of 59.6% of the original variability for the 17 white clover populations (**Figure S5.4**, Appendix 4). The second

and third PCAs did not show clear separation of “Dry” vs “Wet” site populations for any one PC, hence they were excluded from subsequent analyses.

Environmental Association Analysis 1:* LEA – *Genotype data conversion

The first LFMM analysis to identify putative adaptive loci was run using the r package “*LEA*” v 2.6.0 (Frichot & François, 2015). For this analysis, population structure analysis and data imputation had to be performed prior to the environmental association test. Data were first converted to appropriate formats as outlined below.

The VCF file (from section 5.4.1) was imported using the *read.vcfR()*function from the *vcfR* v 1.10.0 package (Knaus & Grünwald, 2017). Genotypes were converted into 0, 1, 2 format using the *revalue()* function from the “*plyr*” v 1.8.6 package (Wickham, 2011). Six loci were removed from the dataset as there was no variation present at each of those loci (i.e., monomorphic throughout the 17 populations). Thus 15,114 SNPs were used for *LEA* analysis. Genotype data was then converted into “.lfmm” and “.geno” format using the *write.lfmm()* and *write.geno()* functions respectively, from the “*LEA*” package.

LEA – *Population structure to determine number of latent factors*

For the population structure analysis, PCA was used to determine the number of significant components (i.e., latent factors; K_E) to be evaluated using the function *pca()*. The *tracy.widom()* function was then used to compute Tracy-Widom tests for each eigenvalue (Patterson, Price & Reich, 2006). *p*-values from the Tracy-Widom tests determined 33 PCs should be retained (**Figure S5.2 A**, Appendix 4), however as seen on the scree plot (**Figure S5.2 B**, Appendix 4), there were only five PCs that explained a high percentage of variation. In parallel, the function *snmf()* was also used to infer individual admixture coefficients (see section 5.3.4.1). Both PCA and sNMF determined K_E of 5 best represented the number of major genetic clusters in the data (**Figure S5.2**, Appendix 4), hence a K_E of 5 was chosen as the number of putative ancestral populations to use to impute missing genotype data.

LEA – *Data imputation and LFMM*

The repetition with the lowest cross entropy score for a K_E of 5 was identified using the *which.min()* function and was then incorporated into the *impute()* function to impute missing genotypes. Using the imputed genotype datafile and the four environmental variables, genome-wide association analysis based on latent factor mixed models using

the "*lfmm()*" function was performed. A total of three models were run, each varying the number of latent factors (K_E) to best control for population structure. This included the optimal K_E determined from the PCA and sNMF, plus two other K_E-values that also had similarly low cross-entropy (K_E = 3 – 5). All three models were run with default parameters, except 10 repetitions (replicate runs), a 10,000 burn-in period and 20,000 cycles were performed. The *z*-scores from the 10 runs were combined using the median value. These median *z*-scores were then recalibrated using the genomic inflation factor (λ) as a rescaling factor to avoid false discoveries. This λ value is used to modify the baseline null hypothesis in order to limit inflation related to population structure and other confounding factors (François *et al.*, 2016). λ was calculated for each of the four environmental variables using:

lambda <- median(zs.median^2)/qchisq(0.5, df = 1)

Where: zs.median is the combined z-scores for one environmental variable using the median.

p-values for each of the four environmental variables were then computed from the combined z-scores using:

adj.p.values <- pchisq(zs.median^2/lambda, df = 1, lower = FALSE)

The histograms of corrected *p*-values were then inspected to ensure they exhibited a uniform distribution (as expected under the null hypothesis of selective neutrality for most loci) but also that a peak close to zero (potentially selected loci) was present (**Figure S5.5**, Appendix 4). This adjustment of *p*-values increases the power of the LFMM test (Frichot & François, 2015). A list of candidate SNPs was then generated after correcting for false positives using a *q*-value threshold of 1e-05, which is a common value used in the literature. *LEA* LFMM outliers considered potentially adaptive were detected as those SNPs with a significant *q*-value (< 1e-05) across at least two values of K_E for each of the environmental variables.

Environmental Association Analysis 2:* lfmm – *Genotype data conversion

In the second LFMM analysis, the R package "*lfmm*" v 1.0 (Caye *et al.*, 2019) was used. For this analysis, the four environmental variables (PC1, SMD, AI and PET) were analysed separately for the 17 white clover populations. The VCF file (from section 5.4.1) was imported using the *read.vcfR()* function from the "*vcfR*" v 1.10.0 package. Genotypes were converted into 0, 1, 2 format using the *revalue()* function from the "*plyr*"

v 1.8.6 package, with missing data encoded as “9”. Genotype data was not imputed for this analysis.

lfmm – *Population structure to determine number of latent factors*

The number of latent factors (K_E) for *lfmm* analysis was determined by maximum variance explained by PCA using the *prcomp()* function from the R package “*stats*” v 3.6.1 (R Core Team, 2019). The variance explained by each PC was plotted (**Figure S5.6**, Appendix 4) and three latent factors were retained.

lfmm – *LFMM*

Regularized least squares estimates for latent factor mixed models using a ridge penalty was implemented using the function *lfmm_ridge()* for K_E = 3 for each of the four variables. *lfmm_test()* was then used to perform association testing using each of the models (four in total). The *q*-value for each locus was calculated using the R package “*qvalue*” and a false discovery rate of α = 0.05 was used to determine significant adaptive loci.

Environmental Association Analysis 3: BayeScEnv

BayeScEnv v 1.1 was implemented to determine whether genetic differentiation (measure by F_{ST}) is related to environmental differentiation (de Villemereuil & Gaggiotti, 2015). This method is based on the Bayesian approach proposed by Beaumont and Balding (2004) and extended by Foll and Gaggiotti (2008). The VCF file (from section 5.4.1) was converted into BayeScan format (.bsc) in R using the packages “*vcfR*” v 1.8.0, “*adegenet*” v 2.1.1, and “*Hierfstat*” v 0.04-22 (Jombart, 2008; Goudet & Jombart, 2015; Knaus & Grünwald, 2017). In order to determine which SNPs were significant for each of the environmental variables, BayeScEnv analysis was performed four times with each analysis assessing one of the four environmental variables. Four environmental text files were created for each of the environmental variables and saved as tab delimited. Within an environmental text file, 17 environmental values were present in one row, with the columns corresponding to each of the 17 populations. BayeScEnv was run for each environmental variable separately with default parameters (20 pilot runs with 5,000 iterations, followed by a burn-in of 50,000 iterations). The *q*-value for each locus was calculated and a false discovery rate of α = 0.05 was used to determine significant outlier loci that had positive alpha values.

5.3.4.5 Detecting candidate gene model IDs associated with outlier and adaptive single nucleotide polymorphisms

Candidate SNPs under selection were those identified as either an outlier SNP or adaptive SNP in common between at least two of the methods (OutFLANK, *LEA*, *lfmm* and BayeScEnv). For candidate SNPs located in exons and introns, the gene where they were residing was recorded as it is likely to be the best candidate. For candidate SNPs which occurred outside of genes, the gene model ID that was closest to the SNP (< 2,000 bp away) was recorded. White clover genome annotations (Griffiths *et al.*, 2019), BLAST (Johnson *et al.*, 2008), UniProt (The UniProt Consortium, 2018) and STRING v 11.0 (Szklarczyk *et al.*, 2019) were used to identify genes and their functions.

5.3.4.6 Single nucleotide polymorphism variation for a subset of candidate gene model IDs

Variation of genotypes of a subset of the candidate SNPs was examined. SNPs were chosen based on two criteria: either a SNP was located near a gene model ID that had biologically relevant gene function for soil moisture stress; or the SNP was identified as significant for at least three of the methods. Genotype proportions for each of the 11 candidate SNPs were extracted from the VCF file (from section 5.4.1) using VCFtools --*extract-FORMAT-info* GT. Genotype patterns were then assessed for each SNP to determine what genotypes distinguished “Dry” and “Wet” individuals. Simple analysis of variance (ANOVA) tests were performed in RStudio using the function *anova()* from the “*stats*” package to determine if the means for each genotype (AA, Aa or aa) were different at $\alpha = 0.01$ for each of the 11 candidate SNPs with genotype percentages categorised as either “Dry” or “Wet”. ANOVAs were performed for each of the 33 candidate SNP/genotype combinations using the following code:

anova(aov(xx_y~SMD grouping, data = data))

Where: xx is either AA, Aa or aa for each SNP (y).

Residuals were checked for each test and no patterns were observed.

5.4 Results

5.4.1 Quality control of single nucleotide polymorphism data

A total of 162,676 single nucleotide polymorphisms (SNPs) were identified pre filtering for 768 individuals. One population, Waikuku (WK), was removed from the dataset. After

collecting the material for the WK population, further discussions with the landowner identified that it was a more recent sowing, only 5 – 6 years prior to sampling (**Table S5.1**, Appendix 4). The WK samples were included for genotyping to determine if time of sowing had any impact on genetic variation and structure. From the score plot produced in PCAdapt, WK clustered with a large group on the principal component (PC)1-PC2 plot but on PC3 it separated as a discrete group (**Figure S5.1**, Appendix 4). Given this knowledge of the WK population's history and the fact that the source genetics was so different to other material, WK samples were unsuitable for further study. After filtering for depth, multiallelic, missing and minor allele frequency, a total of 15,120 SNPs were retained for 674 samples from 17 populations. The mean SNP depth per population was 37 ± 5.3 standard deviation (SD). The lowest SNP depth was found in Middlemarch (MM; 30.3) and the highest SNP depth was found in Fruitlands (FL; 46.2).

5.4.2 Determination of minimum number of individuals to represent a population

Using seven single locus SSRs, 91 white clover individuals from two cultivars ('Grasslands Huia' and 'Crau') were assessed for genetic variation estimates for datasets of different sample sizes (10, 20, 30 and 40). A typical trend observed was the reduction in standard error with increasing dataset size. Although there was no significant change in H_O and H_E values from 20 – 40 individuals, the standard error was minimal for sample sizes above 30. Similarly, this trend was observed for mean pairwise F_{ST} (**Figure 5.2**). Hence a minimum sample size of 30 individuals was assessed as sufficient for accurate allele frequency estimates (see Chapter 5 Supplementary Material, Supplementary experiment – Sample size determination in Appendix 4).

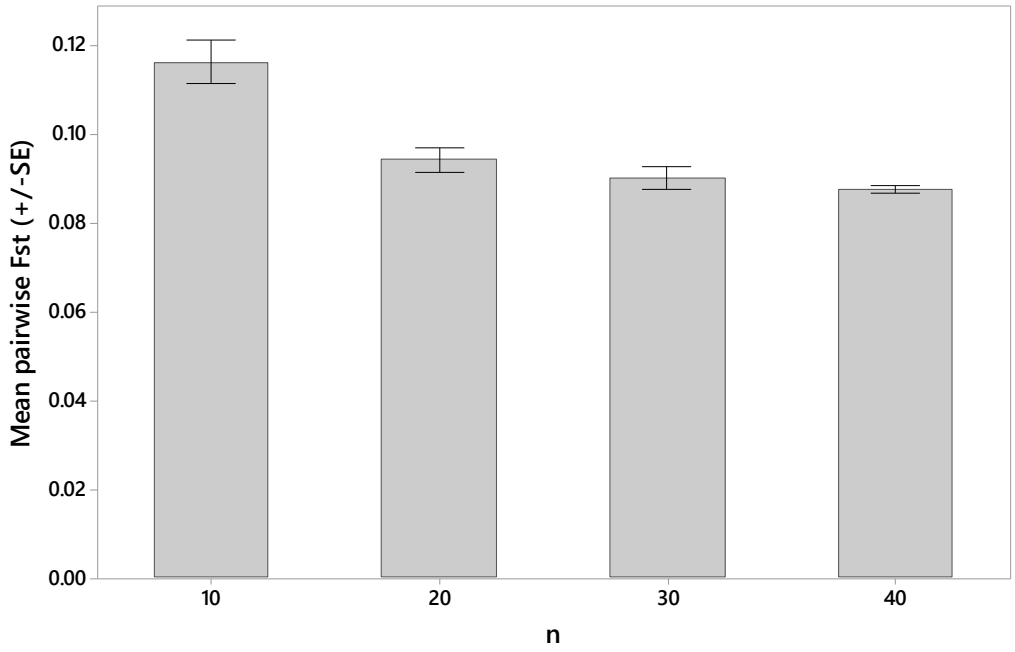


Figure 5.2 Mean pairwise F_{ST} between the 10 random replications for the white clover dataset at each sample size (n). Error bars are the standard error (SE).

5.4.3 Population structure, variation and differentiation

5.4.3.1 Population structure

Utilisation of ADMIXTURE with the dataset representing white clover populations from 17 locations and a total of 674 individuals identified three genetic clusters (putative number of ancestral populations, K_P) based on the lowest cross-validation error (**Figure 5.3**). A structure defined by two genetic clusters was identified by PCAdapt, where individuals that had a portion of their genetic material belonging to the ADMIXURE-determined White-coloured cluster from $K_P = 3$ separated from the rest of the individuals (**Figure S5.7**, Appendix 4). sNMF also indicated three to five genetic clusters from the STRUCTURE-like cross-entropy validation (**Figure S5.2**, Appendix 4). No distinct population structure that aligned with “Dry” or “Wet” sites was observed from any of the three population structure analyses.

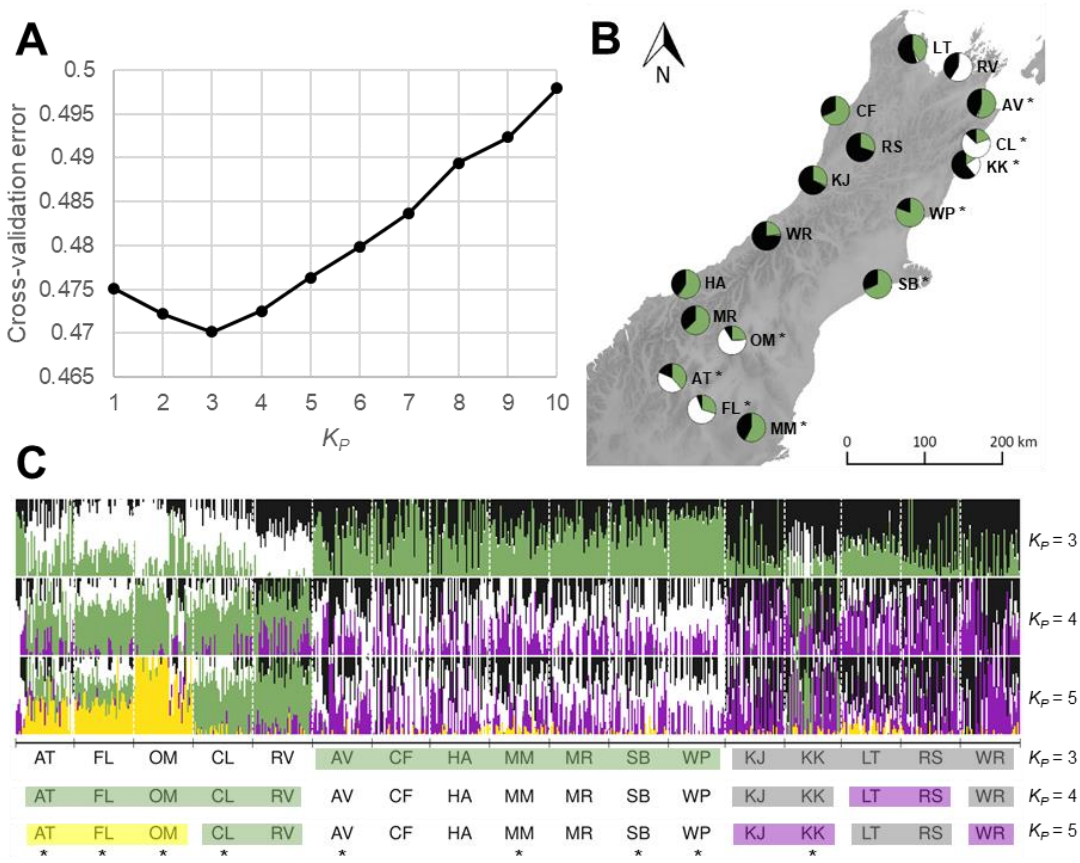


Figure 5.3 Genetic clustering of 17 white clover populations inferred using ADMIXTURE. **A**) Cross-validation error plot of putative ancestral populations (K_P) sub-population values from K_P = 1 through K_P = 10. K_P = 3 was selected for further analyses as it was the value that minimised cross-validation error. **B**) Map of study area overlaid with pie charts displaying the mean cluster (White-, Green- and Black-coloured) proportions inferred by ADMIXTURE when the K_P-value was set as three. **C**) Stacked bar plots for 674 white clover individuals grouped into their respective population for K_P-values of 3 – 5. Each bar represents the genetic data from one individual and the bars are filled by colours representing the assignment probability to each of the inferred clusters. The colour code for K_P = 3 aligns with **B**. K_P = 4 and K_P = 5 were included as the sNMF STRUCTURE-like cross-entropy validation indicated 3 – 5 putative ancestral populations were present based from the sNMF population analysis (**Figure S5.2**, Appendix 4). Colour along *x*-axis over population codes correspond with the groupings at K_P = 3 – 5. *Note*: colours among the three graphs in **C** do not correspond to the same cluster across all three. * indicates populations have high soil moisture deficit and are categorised as "Dry". AT = Arrowtown, AV = Awatere Valley, CF = Cape Foulwind, CL = Clarence, FL = Fruitlands, HA = Haast, KJ = Kumara Junction, KK = Kaikoura, LT = Lower Takaka, MM = Middlemarch, MR = Makarora, OM = Omarama, RS = Rahu Saddle, RV = Rai Valley, SB = Southbridge, WP = Waipara and WR = Whataroa.

5.4.3.2 Analysis of Molecular Variance

The Analysis of Molecular Variance (AMOVA) from K_P = 17 revealed that the majority of genetic variation was partitioned within populations (96.1%), with minimal variation

partitioned among populations (3.9%). A hierarchical AMOVA based on the two major regions ("Dry" vs "Wet") indicated very little variation occurred between regions (0.3%), and only 3.8% of genetic variation was partitioned among populations. Finally, a third AMOVA based on the three genetic clusters indicated minimal variation between the three clusters (2.1%) and only 2.5% variation partitioned among populations within the clusters. For all three AMOVAs, most of the genetic variation was partitioned within populations (95.4 – 96.1%; **Table S5.3**, Appendix 4).

5.4.3.3 Pairwise F_{ST}

Pairwise estimates of genetic differentiation, F_{ST} (**Table 5.1**), among the 17 white clover populations were low overall (0.001 – 0.047) and consistent with population structure. FL, OM, CL and RV were genetically distinct from the rest of the populations as indicated by larger F_{ST} values (**Table 5.1**). Additionally, AT exhibited intermediate (~0.023) F_{ST} values (F_{ST} = 0.015 – 0.027) with populations belonging to the Green- and Black-coloured clusters (**Figure 5.3**). KK and WR (both Black-coloured cluster) were also both more genetically distinct from MM and WP (both Green-coloured cluster) (F_{ST} = 0.028 – 0.033). SB and WP were almost indistinguishable from each other (F_{ST} = 0.001).

Table 5.1 Pairwise estimates of genetic differentiation among 17 white clover populations located in the South Island/Te Waipounamu of New Zealand/Aotearoa. Weir and Cockerham's (1984) F_{ST} is presented below the diagonal. F_{ST} values are shaded in a continuum of colours where a low F_{ST} (0.001) corresponds to green an intermediate F_{ST} (0.023) corresponds to white, and a larger F_{ST} (0.047) corresponds to blue. Populations are ordered the same as **Figure 5.3 C** and solid black lines reflect distinctions between population grouping based on K_P (putative number of ancestral populations) = 3.

	AT	FL	OM	CL	RV	AV	CF	HA	MM	MR	SB	WP	KJ	KK	LT	RS	WR
AT	-																
FL	0.012	-															
OM	0.021	0.020	-														
CL	0.017	0.027	0.035	-													
RV	0.025	0.038	0.043	0.011	-												
AV	0.018	0.036	0.039	0.021	0.025	-											
CF	0.018	0.033	0.039	0.025	0.036	0.011	-										
HA	0.015	0.030	0.037	0.025	0.032	0.008	0.007	-									
MM	0.025	0.033	0.037	0.040	0.045	0.018	0.018	0.019	-								
MR	0.017	0.031	0.037	0.028	0.034	0.009	0.008	0.003	0.016	-							
SB	0.016	0.032	0.034	0.024	0.036	0.008	0.006	0.004	0.013	0.006	-						
WP	0.019	0.032	0.036	0.027	0.044	0.012	0.008	0.009	0.014	0.009	0.001	-					
KJ	0.025	0.041	0.045	0.030	0.023	0.010	0.017	0.014	0.022	0.015	0.017	0.022	-				
KK	0.020	0.035	0.039	0.016	0.009	0.013	0.024	0.023	0.032	0.024	0.024	0.033	0.015	-			
LT	0.023	0.036	0.035	0.036	0.033	0.014	0.020	0.020	0.012	0.022	0.016	0.020	0.022	0.025	-		
RS	0.022	0.039	0.039	0.029	0.023	0.014	0.016	0.012	0.018	0.016	0.009	0.018	0.008	0.014	0.014	-	
WR	0.027	0.043	0.047	0.036	0.027	0.015	0.021	0.011	0.028	0.016	0.022	0.030	0.010	0.016	0.024	0.012	-

Note: AT = Arrowtown, AV = Awatere Valley, CF = Cape Foulwind, CL = Clarence, FL = Fruitlands, HA = Haast, KJ = Kumara Junction, KK = Kaikoura, LT = Lower Takaka, MM = Middlemarch, MR = Makarora, OM = Omarama, RS = Rahu Saddle, RV = Rai Valley, SB = Southbridge, WP = Waipara and WR = Whataroa.

5.4.4 Determination of environmental variables used in environmental association analyses

Using only environmental variables (no genetic data) a principal component analysis (PCA) reduced 22 environmental variables to two principal components (PCs) accounting for a total of 75.9% of the original variability across the 17 locations (**Figure 5.4**). The first of these PCs explained 47.6% of the total variance and the second explained 28.3% of the total variance (**Figure 5.4**). Precipitation related variables strongly influenced PC1, while temperature related variables influenced PC2. Precipitation and Temperature variables were not correlated. Soil moisture deficit (SMD), potential evapotranspiration (PET), Isothermality and Precipitation Seasonality were positively correlated but only SMD and PET contributed to PC1. Aridity Index (AI), Annual Precipitation, Precipitation of Wettest Month, Precipitation of Driest Month, Precipitation of Wettest Quarter, Precipitation of Driest Quarter, Precipitation of Warmest Quarter and Precipitation of Coldest Quarter were all positivity correlated and contributed to PC1. SMD and PET were negatively correlated with the above-mentioned eight precipitation variables (**Figure 5.4**).

Annual Mean Temperature, Min Temperature of Coldest Month, Mean Temperature of Driest Quarter, Mean Temperature of Warmest Quarter and Mean Temperature of Coldest Quarter were positively correlated but only Mean Temperature of Warmest Quarter did not strongly contribute to PC2. Mean Diurnal Range, Temperature Seasonality, Max Temperature of Warmest Month, Temperature Annual Range and Mean Temperature of Wettest Quarter were positively correlated but only Max Temperature of Warmest Month and Mean Temperature of Wettest Quarter did not strongly contribute to PC2 (**Figure 5.4**).

There was clear separation between the “Dry” and “Wet” clusters based on PC1 with secondary spread based on temperature within the “Dry” cluster i.e., the “Dry” cluster is separated from the “Wet” cluster by SMD and PET but the “Dry” cluster also shows separation within by temperature: Otago/Otakou (greater temperature fluctuations) top-left; Canterbury Plains/Waitaha and Marlborough/Tauihu bottom-right; and West Coast/Te Tai o Poutini, Tasman/Te Tai o Aorere and Marlborough/Tauihu right-hand side (**Figure 5.4**). As PC1 contributed the most to the “Dry” vs “Wet” environmental split, the PC1 co-ordinates were retained as one of the environmental variables to assess in further analyses. In addition to the PC1 co-ordinates, the SMD, AI and PET variables were also investigated as discrete traits as they exhibited a large

contribution to the first PC and were of strong interest in this study. These four environmental variables were utilised for the subsequent environmental association analyses, reported in the following section (5.4.6).

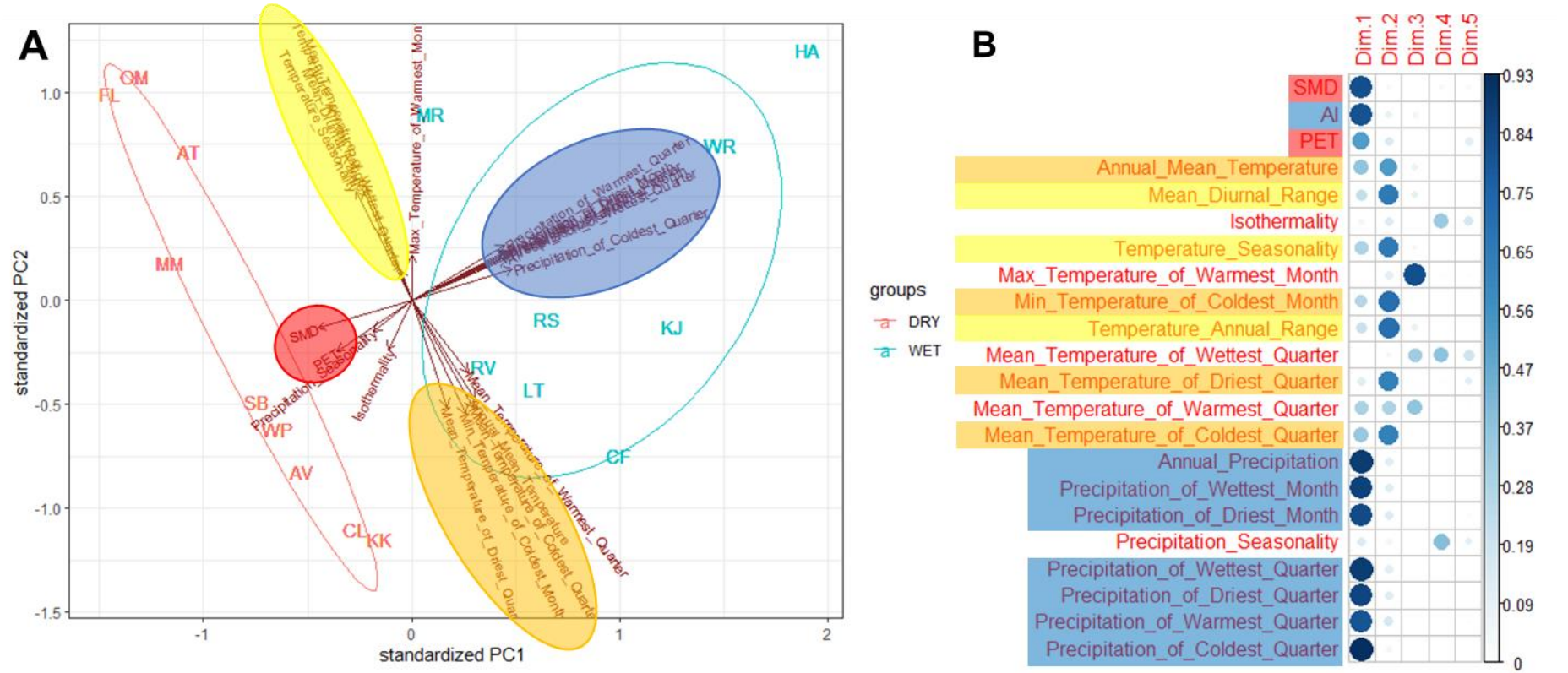


Figure 5.4 Principal component analysis of 22 environmental variables. **A**) Biplot displaying populations grouped as either “Dry” (red) or “Wet” (blue) and the environmental variables. Percentage of explained variance for principal component (PC)1 = 47.6% and PC2 = 28.3%. **B**) List of the 22 environmental variables and their contribution to the first five PCs. The contribution of each environmental variable is displayed in a continuum where large dark blue circles indicate a greater contribution to the respective PC (Dim 1 – 5), while a smaller circle and pale blue indicates less contribution to the respective PCs, and no circle indicates the variable had no contribution to the respective PC. Coloured circles in **A** highlight variables contributing to PC1 and PC2, and correspond with coloured boxes in **B**. AT = Arrowtown, AV = Awatere Valley, CF = Cape Foulwind, CL = Clarence, FL = Fruitlands, HA = Haast, KJ = Kumara Junction, KK = Kaikoura, LT = Lower Takaka, MM = Middlemarch, MR = Makarora, OM = Omarama, RS = Rahu Saddle, RV = Rai Valley, SB = Southbridge, WP = Waipara and WR = Whataroa.

5.4.5 Outlier single nucleotide polymorphism detection

Two different methods were employed to identify SNPs putatively under selection, PCAdapt and OutFLANK (Whitlock & Lotterhos, 2015; Luu *et al.*, 2017). PCAdapt determines population structure based on PCs as an initial step. Using Cattell's scree test (Cattell, 1966) (**Figure S5.3**, Appendix 4), two PCs were identified as explaining the greatest proportion of variance (25.3% combined) and were used subsequently in PCAdapt (**Figure S5.7**, Appendix 4). The first PC showed no clear distinction of individuals. The second PC captured the distinction between a group containing a subset of the "Dry" site populations (AT, CL, FL, KK and OM) plus one "Wet" site population (RV), and both the majority of the "Wet" site populations (CF, HA, KJ, LT, MR, RS and WR) and the remaining "Dry" site populations (AV, MM, SB and WP) (**Figure S5.7**, Appendix 4). This grouping reflects the population structure detected by ADMIXTURE (**Figure 5.3**), where individuals in the White-coloured genetic cluster are physically located in the upper cluster on the PCAdapt score plot (**Figure S5.7**, Appendix 4). PCAdapt identified a total of 355 outlier SNPs separating the populations, based on PC2. However, PCAdapt is not an appropriate tool for the objective of this study as it can only be used to identify SNPs associated with the detected population structure and cannot be used to identify SNPs that are environmentally related (i.e. discriminating populations from "Dry" vs "Wet" environments).

OutFLANK on the other hand was used to group populations *a priori* into the two environments ("Dry" and "Wet") and identify SNPs that distinguish the two genetic groups based on F_{ST}. OutFLANK, the most conservative of the two packages used, had the lowest number of outlier SNPs detected, with 176 identified across the 15,120 SNP loci. A total of 70 SNPs were identified as outliers in common among the two approaches, however all outlier SNPs detected by OutFLANK were used to compare with adaptive SNPs identified in the following section (5.4.6).

5.4.6 Detection of single nucleotide polymorphisms associated with adaptation

Three methods were employed to identify SNPs putatively associated with adaptation to environmental variables. Latent factor mixed models (LFMMs) were implemented in two different R packages (*LEA* and *lfmm*) (Frichot & François, 2015; Caye *et al.*, 2019) and an F_{ST}-based genome scan method was implemented in BayeScEnv (de Villemereuil & Gaggiotti, 2015). For *LEA*, multiple latent factors (K_E) were used to control for population structure, including $K_E = 3 – 5$, while a K_E of 3 was used for *lfmm*. A total

of 375 adaptive SNPs were identified as being significantly associated with at least one of the four environmental variables using the *LEA* method. These environmental variables comprised the aridity index (AI), soil moisture deficit (SMD), potential evapotranspiration (PET) and the PC1 co-ordinates from PCA analysis of environmental variables including SMD, AI, PET and the standard 19 bioclimatic variables described above (section 5.4.4) Of these 375 SNPs, 106 (28%) were common to between more than one of the environmental variables (**Figure S5.8**, Appendix 4). Similarly, using the *lfmm* method, a total of 342 adaptive SNPs were determined as being significantly associated with at least one of the environmental variables described above. Of these, 201 (59%) were common to more than one of the environmental variables (**Figure S5.8**, Appendix 4). By contrast, the F_{ST}-based BayeScEnv method identified 101 adaptive SNPs significantly associated with at least one of the environmental variables described above. Of these 101 SNPs, however, none were associated with AI or PET and only 9 (9%) were common to multiple environmental variables, in this case PC1 and SMD (**Figure S5.8**, Appendix 4). All putative adaptive SNPs identified by *LEA*, *lfmm* and BayeScEnv, regardless of which environmental variable they were significant for, were compared with the outlier SNPs detected by OutFLANK.

5.4.7 Candidate loci and functional annotation of gene model IDs

To identify candidate loci under selection and associated with soil moisture deficit stress, the commonality of outlier and adaptive SNPs was determined. SNPs identified by multiple methods provide confidence that these SNPs are likely positives. A total of 64 SNPs were found in common between at least two of the OutFLANK, *LEA*, *lfmm* and BayeScEnv methods, with only one common to three methodologies and none to all four (**Figure 5.5**). Of these 64 candidate SNPs, 17 were found in exons, 13 in introns and the remainder were either intergenic or found < 2,000 bp away from gene model IDs. A total of 46 gene model IDs were identified as associated with 49 SNPs, based on physical position and proximity to genes (**Table 5.2**). There was an element of redundancy as in some cases multiple SNPs were located in the same gene model ID (hence 46 gene model IDs from 49 SNPs).

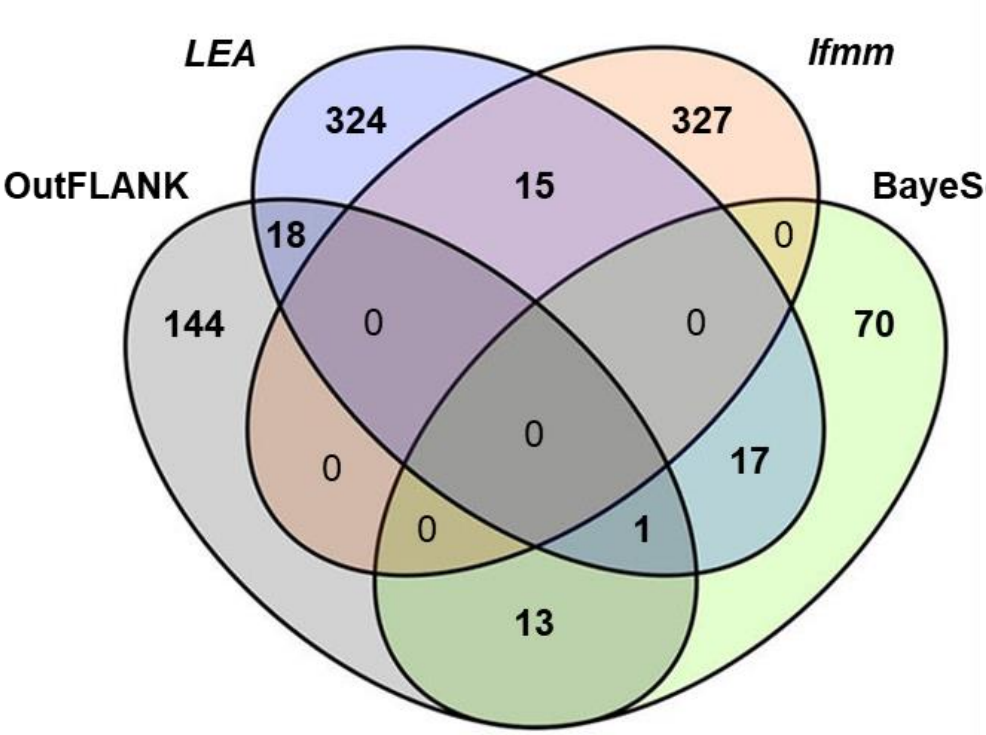


Figure 5.5 Venn diagram comparing the total number of outlier and adaptive SNPs detected by OutFLANK, *LEA*, *lfmm* and BayeScEnv methods. The OutFLANK dataset included 8,795 SNPs and 674 individuals that were assigned to two groups: “Dry” or “Wet”. Outlier SNPs were detected as those that differentiated the “Dry” and “Wet” individuals and environmental variables were not included in this analysis. The *LEA* dataset included 15,114 SNPs and 674 individuals that were assigned to the *a priori* grouping of 17 populations. *Lfmm* and BayeScEnv datasets included 15,120 and 674 individuals that were assigned to the *a priori* grouping of 17 populations. *LEA*, *lfmm* and BayeScEnv identified adaptive SNPs as those that were associated with at least one of the four environmental variables – (1) PC1 co-ordinates from a PCA of environmental variables including soil moisture deficit (SMD), aridity index (AI), potential evapotranspiration (PET) and the standard 19 bioclimatic variables; (2) SMD; (3) AI; and (4) PET. All outlier and adaptive SNPs were corrected for false positives using *q*-values. Any SNP identified by more than one method was termed a candidate SNP. A total of 64 candidate SNPs were detected by at least two methods, with one SNP detected by three methods.

Table 5.2 Candidate SNPs under selection identified by more than one outlier or adaptive SNP detection methodology. *Q*-values are reported for each environmental variable that the SNP was significant for, for each of the detection methods. *Q*-values for OutFLANK were determined using the "Dry" vs "Wet" site grouping, which is not an environmental variable (in contrast to PC1, SMD, AI and PET). Genomic location and associated gene information are presented.

			***q*-value for environmental variable-associated SNP †**							
SNP	**CHR**	**Methods**	**PC1**	**SMD**	**AI**	**PET**	**"Dry" vs "Wet"**	**Gene Model ID and genomic region**	**Annotation**	**Potential function**
1_42534748	1	BayeScEnv	0.0027					INTERGENIC		
		LEA	3.5e-09							
1_42534782	1	BayeScEnv	0.0048					INTERGENIC		
		LEA	3.3e-08							
1_66422932	1	BayeScEnv	0.0081	0.0035				chr1.jg9666.t1 (INTRON)	Poly [ADP-ribose] polymerase 2	Double-strand break repair
		LEA	2.7e-09							
2_46150696	2	BayeScEnv	0.036					chr2.jg6917.t1 (EXON)	Heparanase-like protein 3	Metabolic process
		LEA	7.2e-10	0.0001	0.00026					
3_42141745	3	BayeScEnv	0.046					chr3.jg6353.t1 (EXON)	Inositol transporter 1-like	Carbohydrate transport
		LEA	3.6e-15							
3_56065966	3	BayeScEnv	0.023					733 bp from start codon chr3.jg8565.t1	Cytochrome P450	RNA phosphodiester bond hydrolysis, endonucleolytic
		LEA	2.7e-08							
3_75145668	3	BayeScEnv	0.0002					chr3.jg11334.t1 (INTRON)	RNA polymerase sigma factor sigB	Regulation of transcription, DNA-templated
		LEA	5.4e-46		7.4e-08					
4_55784676	4	BayeScEnv	0.022					INTERGENIC		
		LEA	5.4e-12							
5_20886874	5	BayeScEnv	0.012					chr5.jg3036.t1 (INTRON)	Phosphate transporter PHO1 homolog 10 isoform X1	Phosphate ion transport
		LEA	5.4e-06		1.1e-09					
5_20886881	5	BayeScEnv	0.015					chr5.jg3036.t1 (INTRON)	Phosphate transporter PHO1 homolog 10 isoform X1	Phosphate ion transport
		LEA			4.9e-09					
7_1040038	7	BayeScEnv	0.021					chr7.jg162.t1 (EXON)	Probable inactive histone-lysine N-methyltransferase SUVR2	Gene silencing by RNA-directed DNA methylation
		LEA			4.6e-13					

Table 5.2 (continued)

SNP	CHR	Methods	PC1	SMD	AI	PET	"Dry" vs "Wet"	Gene Model ID and genomic region	Annotation	Potential function
7_1302527	7	BayeScEnv	0.048					974 bp from stop codon chr7.jg205.t1	Polyadenylate-binding protein 2	RNA binding
		LEA	3.5e-07					1,291 bp from stop codon chr7.jg206.t1	Protein exordium-like 2	Cellular response to hypoxia
7_50249174	7	BayeScEnv	0.027					chr7.jg7992.t1 (EXON)	Mannosylglycoprotein endo-beta-mannosidase	Carbohydrate metabolic process
		LEA		1.2e-14	5.5e-21					
8_29159561	8	BayeScEnv	0.0048					chr8.jg3996.t1 (EXON)	Transportin MOS14	Protein import into nucleus and from nucleus
		LEA	5.4e-09							
9_8446930	9	BayeScEnv	0.017					INTERGENIC		
		LEA	3.9e-16		4.3e-06					
13_2073758	13	BayeScEnv	0.03					50 bp from stop codon chr13.jg338.t1	Proline-rich receptor-like protein kinase PERK8	Protein phosphorylation
		LEA	5.6e-10							
13_32341386	13	BayeScEnv	0.033					chr13.jg4904.t1 (EXON)	Heat shock cognate 70 kDa protein	Stress response
		LEA		7.7e-59	2.1e-31					
16_21540893	16	BayeScEnv	0.034					chr16.jg2981.t1 (EXON)	Isoleucine-tRNA ligase, chloroplastic/mitochondrial	Isoleucyl-tRNA aminoacylation
		LEA	6.0e-06		4.5e-09	4.5e-13				
		OutFLANK					6.3e-06			
1_4206355	1	BayeScEnv	0.0024					chr1.jg554.t1 (INTRON)	E3 ubiquitin-protein ligase WAV3-like	Root development and gravitropism
		OutFLANK					1.3e-06			
2_30295355	2	BayeScEnv	0.032					INTERGENIC		
		OutFLANK					0.0020			
2_47257509	2	BayeScEnv	0.0077					1,084 bp from stop codon chr2.jg7076.t1	Pentatricopeptide repeat-containing protein	Membrane
		OutFLANK					0.041			
2_47257510	2	BayeScEnv	0.00049					1,085 bp from stop codon chr2.jg7076.t1	Pentatricopeptide repeat-containing protein	Membrane
		OutFLANK					0.00072			
2_47257537	2	BayeScEnv	0.00033					1,112 bp from stop codon chr2.jg7076.t1	Pentatricopeptide repeat-containing protein	Membrane
		OutFLANK					0.00072			

Table 5.2 (continued)

SNP	CHR	Methods	PC1	SMD	AI	PET	"Dry" vs "Wet"	Gene Model ID and genomic region	Annotation	Potential function
3_12109714	3	BayeScEnv	0.021	0.0028				chr3.jg1808.t1 (EXON)	Putative complex I intermediate-associated protein 30	mitochondrial respiratory chain complex I assembly
		OutFLANK					0.019			
5_45296487	5	BayeScEnv	0.039					647 bp from start codon chr5.jg6699.t1	Disease resistance protein RGA2	Defence response
		OutFLANK					0.0022			
6_36819169	6	BayeScEnv	0.01					chr6.jg5430.t1 (EXON)	Protein ANTI-SILENCING 1 isoform X1	Negative regulation of chromatin silencing
		OutFLANK					0.0034			
7_22676114	7	BayeScEnv	0.0073					chr7.jg3614.t1 (INTRON)	O-glucosyltransferase rumi homolog	Transferase activity, transferring glycosyl groups
		OutFLANK					1.3e-06			
10_31293855	10	BayeScEnv	0.025					chr10.jg4731.t1 (EXON)	Probable inactive leucine-rich repeat receptor-like protein kinase At3g03770	Protein autophosphorylation
		OutFLANK					0.029			
11_20297161	11	BayeScEnv	0.028					chr11.jg3155.t1 (EXON)	Cationic amino acid transporter 2, vacuolar	Amino-acid transport
		OutFLANK					0.0022			
11_26982613	11	BayeScEnv	0.047					744 bp from stop codon chr11.jg4109.t1	Hypothetical protein	
		OutFLANK					0.0028			
14_21277692	14	BayeScEnv	0.027					chr14.jg3151.t1 (EXON)	CLP protease regulatory subunit CLPX1, mitochondrial	Protein folding
		OutFLANK					0.029			
1_64955755	1	*LEA*	3.1e-11	1.3e-08	5.2e-14	5.4e-08		1,527 bp from start codon chr1.jg9441.t1	Lysine-rich arabinogalactan protein 19	Differentiation, cell-cell recognition, embryogenesis and programmed cell death
		lfmm		0.027	0.038	0.0043				
2_19798297	2	*LEA*		7.0e-07				129 bp from stop codon chr2.jg2931.t1	SWI/SNF complex subunit SWI3B	DNA binding
		lfmm				0.023				
4_18643366	4	*LEA*	1.0e-07	4.6e-06	7.0e-11	1.8e-15		1,289 bp from stop codon chr4.jg2548.t1	Nodule Cysteine-Rich (NCR) secreted peptide	Nodule morphogenesis
		lfmm	1.9e-06	1.2e-07	1.3e-06	4.1e-08				
4_28301182	4	*LEA*	7.9e-13					chr4.jg3860.t1 (INTRON)	Protein MAK16 homolog	Maturation of 5.8s RNA
		lfmm		0.044						

Table 5.2 (continued)

SNP	CHR	Methods	PC1	SMD	AI	PET	"Dry" vs "Wet"	Gene Model ID and genomic region	Annotation	Potential function
4_57492619	4	*LEA*	6.4e-06					INTERGENIC		
		lfmm	0.0024	0.033	0.040					
4_57492635	4	*LEA*	1.9e-07					INTERGENIC		
		lfmm	0.0023	0.020	0.026	0.027				
7_38154411	7	*LEA*		0.041				691 bp from start codon chr7.jg6155.t1	Tyrosine decarboxylase 1	Cellular amino acid metabolic process
		lfmm	0.021	2.4e-07	9.1e-05	7.5e-08				
8_19017512	8	*LEA*	2.5e-13					INTERGENIC		
		lfmm	0.031							
8_32566913	8	*LEA*	1.2e-13	1.4e-07	7.0e-13	7.3e-21		65 bp from stop codon chr8.jg4503.t1	F-box protein skip16-like	Protein ubiquitination
		lfmm	0.015	0.038	0.049					
10_31116406	10	*LEA*	1.2e-07	6.2e-13	5.6e-16	2.0e-07		1,154 bp from stop codon chr10.jg4701.t1	Protein QUIRKY	Transferase activity, transferring glycosyl groups
		lfmm		0.015	0.022	0.0063		1,942 bp from stop codon chr10.jg4702.t1	beta-glucuronosyltransferase GlcAT14A	Transferase activity
11_15851586	11	*LEA*		1.2e-08	5.2e-10	9.2e-10		chr11.jg2437.t1 (INTRON)	Mitochondrial fission protein ELM1	Mitochondrial fission
		lfmm	1.1e-08	3.6e-08	1.2e-07	5.6e-09				
11_54897968	11	*LEA*		4.3e-07	5.9e-07	7.3e-09		chr11.jg8152.t1 (EXON)	Pentatricopeptide repeat-containing protein At4g31850, chloroplastic	Response to auxin
		lfmm		0.0022	0.0058					
11_54898004	11	*LEA*			6.8e-06	8.2e-06		chr11.jg8152.t1 (EXON)	Pentatricopeptide repeat-containing protein At4g31850, chloroplastic	Response to auxin
		lfmm		0.0022	0.0058					
12_46261080	12	*LEA*	2.9e-08		1.2e-07			INTERGENIC		
		lfmm		0.0035	0.0058					
15_31597018	15	*LEA*		4.0e-22	8.6e-25	4.3e-19		chr15.jg4786.t1 (INTRON)	ABC transporter C family member 5	Response to salt stress
		lfmm		3.4e-07	0.00013	9.7e-07				
1_48939405	1	*LEA*	2.3e-09		2.9e-09	2.8e-10		INTERGENIC		
		OutFLANK					0.0022			

Table 5.2 (continued)

SNP	CHR	Methods	PC1	SMD	AI	PET	"Dry" vs "Wet"	Gene Model ID and genomic region	Annotation	Potential function
1_48939412	1	*LEA*		2.0e-11	1.9e-15	6.5e-13		INTERGENIC		
		OutFLANK					0.0020			
1_89696968	1	*LEA*		7.9e-06	1.1e-05	2.0e-07		2,018 bp from stop codon chr1.jg13031.t1	Phosphoglucosamine mutase	Carbohydrate metabolic process
		OutFLANK					0.00012			
2_7592081	2	*LEA*		1.2e-15				chr2.jg1086.t1 (INTRON)	GATA transcription factor 24	Regulation of transcription, DNA-templated
		OutFLANK					0.00073			
2_17981686	2	*LEA*	3.6e-10	1.7e-07	1.7e-10			542 bp from stop codon chr2.jg2660.t1	Heat shock cognate 70 kDa protein	Stress response
		OutFLANK					0.00060			
2_17981688	2	*LEA*	2.0e-07		3.0e-06			544 bp from stop codon chr2.jg2660.t1	Heat shock cognate 70 kDa protein	Stress response
		OutFLANK					0.00060			
4_22587619	4	*LEA*		2.0e-07	1.6e-11			890 bp from start codon chr4.jg3074.t1	Protein FAM133	RNA binding
		OutFLANK					4.3e-05			
4_59635492	4	*LEA*		5.7e-05		1.5e-06		INTERGENIC		
		OutFLANK					0.0028			
4_59635493	4	*LEA*	9.3e-14					INTERGENIC		
		OutFLANK					0.0089			
5_5063027	5	*LEA*		1.3e-14	1.2e-18	8.7e-18		chr5.jg742.t1 (INTRON)	Hypothetical protein	
		OutFLANK					0.00012			
6_15437952	6	*LEA*			1.6e-06			INTERGENIC		
		OutFLANK					0.0023			
8_2464855	8	*LEA*	5.0e-15					chr8.jg302.t1 (EXON)	Exportin-2	Protein import into nucleus and from nucleus
		OutFLANK					0.015			
9_34153967	9	*LEA*		8.9e-17	3.8e-14			1,196 bp from stop codon chr9.jg5048.t1	E3 ubiquitin-protein ligase SGR9, amyloplastic	Gravitropism
		OutFLANK					0.023			
9_42641599	9	*LEA*		2.3e-07	1.5e-10	3.8e-06		chr9.jg6321.t1 (EXON)	Laminin subunit alpha-2	Membrane
		OutFLANK					0.039			

Table 5.2 (continued)

SNP	CHR	Methods	PC1	SMD	AI	PET	"Dry" vs "Wet"	Gene Model ID and genomic region	Annotation	Potential function
9_45089027	9	*LEA*			0.00031			chr9.jg6685.t1 (INTRON)	Calreticulin	Vesicle-mediated transport
		OutFLANK					0.0046			
10_15009932	10	*LEA*		3.6e-34	1.0e-23	9.2e-10		chr10.jg2169.t1 (INTRON)	Regulatory protein NPR1	Plant defence
		OutFLANK					0.0024			
13_6228069	13	*LEA*		4.7e-14	1.7e-10			chr13.jg958.t1 (EXON)	VAN3-binding protein	Integral component of membrane
		OutFLANK					0.0033			
15_29906485	15	*LEA*		8.9e-05	9.6e-06			INTERGENIC		
		OutFLANK					0.011			

Note: SNP = chromosome_position in base pairs, CHR = Chromosome, PC1 = PC1 co-ordinates from PCA of environmental variables including soil moisture deficit (SMD), aridity index (AI), potential evapotranspiration (PET) and 19 standard bioclimatic variables, and bp = base pairs.
† *LEA q*-values from the three latent factors models are reported.
Shading – groups SNPs together based on gene model ID. Ordered by methods in common, with solid black line between.

5.4.8 Single nucleotide polymorphism variation for a subset of candidate gene model IDs

Genotype frequencies of 11 candidate SNP were estimated in each of the 17 populations (**Table 5.3**) to determine if populations from “Dry” and “Wet” sites were distinguished by genotype at these loci. As described below, only one SNP (1_89696968) demonstrated a statistically significant difference (p = 0.0001) in mean genotype frequency when comparing populations in the “Dry” and “Wet” categories, with the AA genotype (homozygous for reference allele) significantly higher in the “Dry” group at an $\alpha = 0.01$. While no clear differences in genotype frequencies between “Dry” and “Wet” site populations were observed at any of the other SNPs (e.g., all “Dry” site populations homozygous for one genotype and all “Wet” site populations homozygous for the other), there were trends indicative of selection at a further four of the SNPs. These are described below, alongside SNP 1_89696968.

For the first SNP (1_89696968), all nine “Dry” site populations (SMD = 81.8 – 127.8) and three “Wet” site populations exhibited high genotype percentages (> 70%) for the AA genotype. These three “Wet” site populations included Lower Takaka (LT), Makarora (MR) and Rai Valley (RV), all of which had moderate (25.7 – 45.2) soil moisture deficit (SMD) values. The remaining five “Wet” site populations with low SMD (0 – 7.2) had a lower proportion of the AA genotype (< 70%). For the second and third SNP loci (2_17981686 and 2_17981688), which are close physical neighbours, four “Wet” site populations comprising Cape Foulwind (CF), Haast (HA), Kumara Junction (KJ) and Whataroa (WR) displayed lower AA genotype percentages (≤ 55%) when compared to all other populations (all > 58%). Three of these populations were from sites with the lowest SMD values (all 0) and the fourth, CF, also had a very low SMD of 7.2. For the fourth SNP (3_42141745), four populations from the “Dry” grouping and located around the Otago/Otakou region (Arrowtown [AT], Fruitlands [FL], Middlemarch [MM] and Omarama [OM]) displayed higher (> 70%) aa (homozygous for alternate allele) genotype frequencies, while all other populations displayed higher (> 70%) heterozygote frequencies. The fifth SNP (9_34153967) displayed a similar Otago/Otakou focussed trend to 3_42141745, with AT, FL, MM and OM displaying higher (> 53%) aa genotype frequencies, whereas the remaining populations were relatively admixed between AA, Aa and aa genotypes (**Table 5.3**).

Table 5.3 Genotype percentages for 11 candidate SNPs detected by more than one analysis method (OutFLANK, LEA, lfmm and BayeScEnv) in 17 white clover populations located in the South Island/Te Waipounamu of New Zealand/Aotearoa, categorised as "Dry" or "Wet". Soil moisture deficit (SMD) values are presented below their respective population and colour coded on a continuum where a higher SMD corresponds to red and a lower SMD corresponds to yellow, with intermediate SMD as orange. Genotypes are presented as AA, Aa and aa, where AA = homozygote for reference allele, Aa = heterozygote, aa = homozygote for alternate allele. Genotypes are colour coded on a continuum where a higher percentage (100%) corresponds to blue and a lower percentage corresponds to green (0%), with intermediate percentages (50%) as white. Detection methods used to identify the SNP putatively associated with adaptation are presented under the SNP name in parentheses. Associated gene model ID and its annotation are presented for each SNP. Important genotype trends are surrounded by a box, underlined and in bold. Mean genotype percentages and number of individuals are presented for both "Dry" and "Wet" sites.

Gene model ID (annotation)	SNP ID (Analysis)		"Dry" sites										"Wet" sites								
		Pop	AT	AV	CL	FL	KK	MM	OM	SB	WP	Mean	CF	HA	KJ	LT	MR	RS	RV	WR	Mean
		SMD	81.8	125.5	108.1	124.4	117.1	127.8	110.5	116.3	95.6	111.9	7.2	0	0	25.7	28.4	2.6	45.2	0	13.6
chr1.jg13031.t1 (Phosphoglucosamine mutase)	**1_89696968** (*LEA* + OutFLANK)	AA	**84.4**	**77.4**	**89.5**	**89.5**	**81**	**92.3**	**94.9**	**86.5**	**100**	88.4*	68.6	65.4	54.5	**78.4**	**78.8**	66.7	**71.4**	70	69.2*
		Aa	9.4	16.1	5.3	7.9	9.5	2.6	2.6	10.8	0	7.1	22.9	11.5	33.3	13.5	12.1	20	0	26.7	17.5
		aa	6.3	6.5	5.3	2.6	9.5	5.1	2.6	2.7	0	4.5	8.6	23.1	12.1	8.1	9.1	13.3	28.6	3.3	13.3
		n	32	31	19	38	21	39	39	37	28	31.5	35	26	33	37	33	30	14	30	29.8
chr2.jg2660.t1 (Heat shock cognate 70 kDa protein)	**2_17981686** (*LEA* + OutFLANK)	AA	75	76	82.4	58.3	75	72.4	61.5	81.3	84.2	74	**46.7**	**55**	**34.8**	77.3	78.9	68.8	69.2	**30**	57.6
		Aa	0	4	0	8.3	0	17.2	7.7	0	0	4.1	20	10	4.3	9.1	0	0	0	5	6.1
		aa	25	20	17.6	33.3	25	10.3	30.8	18.8	15.8	21.9	33.3	35	60.9	13.6	21.1	31.3	30.8	65	36.3
		n	12	25	17	12	12	29	13	16	19	17.2	15	20	23	22	19	16	13	20	18.5
chr2.jg2660.t1 (Heat shock cognate 70 kDa protein)	**2_17981688** (*LEA* + OutFLANK)	AA	75	76	82.4	58.3	75	72.4	61.5	81.3	84.2	74	**46.7**	**55**	**34.8**	77.3	78.9	68.8	69.2	**30**	57.6
		Aa	0	4.0	0	8.3	0	17.2	7.7	0	0	4.1	20	10	4.3	9.1	0	0	0	5	6
		aa	25	20	17.6	33.3	25	10.3	30.8	18.8	15.8	21.9	33.3	35	60.9	13.6	21.1	31.3	30.8	65	36.4
		n	12	25	17	12	12	29	13	16	19	17.2	15	20	23	22	19	16.0	13	20	18.5
chr3.jg6353.t1 (Inositol transporter 1-like)	**3_42141745** (BayeScEnv + *LEA*)	AA	0	15.4	12.9	0	26.9	0	0	0	0	6.1	0	0	7.7	0	0	0	9.1	0	2.1
		Aa	26.9	80.8	87.1	11.1	73.1	30	14.3	72	79.2	52.7	70.8	84	88.5	50	100.0	60.9	87.9	96.2	79.8
		aa	**73.1**	3.8	0	**88.9**	0	**70**	**85.7**	28	20.8	41.2	29.2	16	3.8	50	0	39.1	3	3.8	18.1
		n	26	26	31	27	26	10	14	25	24	23.2	24	25	26	10	23	23	33	26	23.8

Table 5.3 (continued)

Gene model ID (annotation)	SNP ID (Analysis)	Pop	"Dry" sites AT	AV	CL	FL	KK	MM	OM	SB	WP	Mean	"Wet" sites CF	HA	KJ	LT	MR	RS	RV	WR	Mean
		SMD	81.8	125.5	108.1	124.4	117.1	127.8	110.5	116.3	95.6	111.9	7.2	0	0	25.7	28.4	2.6	45.2	0	13.6
chr9.jg5048.t1 (E3 ubiquitin-protein ligase SGR9, amyloplastic)	**9_34153967** (*LEA* + OutFLANK)	AA	11.5	28	24.1	16.7	36	9.5	16.7	34.8	30	23	29.6	27.6	52.4	60	31.8	38.1	43.5	16	37.4
		Aa	34.6	28	27.6	23.3	32	14.3	20.8	17.4	30	25.3	18.5	27.6	33.3	15	27.3	33.3	21.7	48	28.1
		aa	**53.8**	44	48.3	**60**	32	**76.2**	**62.5**	47.8	40	51.7	51.9	44.8	14.3	25	40.9	28.6	34.8	36	34.5
		n	26	25	29	30	25	21	24	23	20	24.8	27	29	21	20	22	21	23	25	23.5
chr1.jg554.t1 (E3 ubiquitin-protein ligase WAV3-like)	**1_4206355** (BayeScEnv + OutFLANK)	AA	50	100	71.4	16.1	85.2	75.7	36.4	92.5	97.2	69.4	94.9	100	100	74.4	100	97.4	78.9	100	93.2
		Aa	43.8	0	28.6	51.6	11.1	24.3	30.3	7.5	2.8	22.2	5.1	0	0	23.1	0	2.6	21.1	0	6.5
		aa	6.3	0	0	32.3	3.7	0	33.3	0	0	8.4	0	0	0	2.6	0	0	0	0	0.3
		n	32	38	21	31	27	37	33	40	36	32.8	39	39	38	39	40	39	19	38	36.4
chr4.jg2548.t1 (Nodule Cysteine-Rich (NCR) secreted peptide)	**4_18643366** (*LEA* + *lfmm*)	AA	60	51.5	58.8	20	70.4	64.9	62.5	55	44.7	54.2	62.5	25	100	55.3	57.1	78.9	67.6	100	68.3
		Aa	20	33.3	35.3	60	22.2	32.4	37.5	30	42.1	34.8	29.2	25	0	34.2	28.6	21.1	26.5	0	20.6
		aa	20	15.2	5.9	20	7.4	2.7	0	15	13.2	11	8.3	50	0	10.5	14.3	0	5.9	0	11.1
		n	5	33	34	5	27	37	16	40	38	26.1	24	4	2	38	7	38	34	2	18.6
chr7.jg7992.t1 (Mannosylglycoprotein endo-beta-mannosidase)	**7_50249174** (BayeScEnv + *LEA*)	AA	72.7	62.5	40	66.7	42.9	58.3	56.3	45.5	33.3	53.1	28.6	0	66.7	16.7	0	70	71.4	0	31.7
		Aa	18.2	0	0	0	0	16.7	12.5	27.3	55.6	14.5	35.7	0	0	33.3	0	10	7.1	0	10.8
		aa	9.1	37.5	60	33.3	57.1	25	31.3	27.3	11.1	32.4	35.7	100	33.3	50	100	20	21.4	100	57.5
		n	11	8	10	9	7	12	16	11	9	10.3	14	5	3	12	3	10	14	5	8.3
chr13.jg4904.t1 (Heat shock cognate 70 kDa protein)	**13_32341386** (BayeScEnv + *LEA*)	AA	50	0	50	76.2	23.1	0	41.2	0	0	26.8	0	ND	0	0	0	0	47.6	0	6.8
		Aa	37.5	100	44.4	23.8	38.5	0	47.1	33.3	0	36	0	ND	0	0	0	0	42.9	0	6.1
		aa	12.5	0	5.6	0	38.5	100	11.8	66.7	100	37.2	100	ND	100	100	100	100	9.5	100	87.1
		n	16	1	18	21	13	5	17	9	5	11.7	3	ND	4	5	4	4	21	4	6.4

Table 5.3 (continued)

Gene model ID (annotation)	SNP ID (Analysis)	Pop	"Dry" sites									"Wet" sites								
			AT	AV	CL	FL	KK	MM	OM	SB	WP	Mean	CF	HA	KJ	LT	MR	RS	RV	WR	Mean
		SMD	81.8	125.5	108.1	124.4	117.1	127.8	110.5	116.3	95.6	111.9	7.2	0	0	25.7	28.4	2.6	45.2	0	13.6
chr15.jg4786.t1	**15_31597018**	AA	65.4	100	74.1	81.0	72.7	78.3	73.9	69.2	57.7	74.7	40.0	32.5	61.8	82.8	47.5	60.6	90	46.2	57.6
		Aa	34.6	0	25.9	19.0	27.3	21.7	26.1	30.8	38.5	24.9	57.1	62.5	35.3	17.2	45.0	36.4	10	53.8	39.7
(ABC transporter C family member 5)	(*LEA* + *lfmm*)	aa	0	0	0	0	0	0	0	0	3.8	0.4	2.9	5	2.9	0	7.5	3	0	0	2.7
		n	26	2	27	21	11	23	23	13	26	19.1	35	40	34	29	40	33	30	39	35
chr16.jg2981.t1	**16_21540893**	AA	78.6	66.7	61.1	60	50	41.2	78.6	45.5	25	56.3	72.7	82.4	70.6	60.9	57.1	72.2	73.9	96	73.2
		Aa	0	6.7	0	0	0	0	0	0	5.0	1.3	9.1	0	0	4.3	4.8	11.1	4.3	0.0	4.2
(Isoleucine-tRNA ligase, chloroplastic/ mitochondrial)	(BayeScEnv + *LEA* + OutFLANK)	aa	21.4	26.7	38.9	40	50	58.8	21.4	54.5	70	42.4	18.2	17.6	29.4	34.8	38.1	16.7	21.7	4	22.6
		n	14	15	18	15	12	17	14	22	20	16.3	22	17	17	23	21	18.0	23	25	20.8

Note: Pop = Population, SNP ID = chromosome number and base pair position of the SNP, *n* = number of individuals with data that were used to calculate genotype percentages, ND = No Data, AT = Arrowtown, AV = Awatere Valley, CF = Cape Foulwind, CL = Clarence, FL = Fruitlands, HA = Haast, KJ = Kumara Junction, KK = Kaikoura, LT = Lower Takaka, MM = Middlemarch, MR = Makarora, OM = Omarama, RS = Rahu Saddle, RV = Rai Valley, SB = Southbridge, WP = Waipara and WR = Whataroa.

* Indicates *p*-values from analysis of variance (ANOVA) < 0.01 at α = 0.01.

5.5 Discussion

5.5.1 Genetic variation and structure of white clover populations

Assessment of population structure is a critical step as unaccounted structure can confound outlier and adaptive single nucleotide polymorphism (SNP) methodologies. Determining such structure using population genetic variation analyses, which included analysis of molecular variance (AMOVA) and pairwise F_{ST}, indicated that the 17 white clover populations were weakly differentiated. This negligible population differentiation, with low pairwise F_{ST} values (<0.05), aligns with previous findings from analysis of white clover populations sampled across North America (F_{ST} < 0.03) (Wright *et al.*, 2017). Similarly, the majority of genetic variation was partitioned within populations (96.1%) as identified by AMOVA in the current study. Studies assessing neutral marker variation of white clover sampled from across North American States also observed a partitioning of variation to within-population (92.3 – 96.4%) (Kooyers & Olsen, 2012, 2013). Furthermore, Kooyers and Olsen (2013) identified a high degree of molecular variance (95%) partitioned within New Zealand/Aotearoa (NZ) white clover populations located around Christchurch and Oamaru. Similarly, Annicchiarico and Carelli (2014) observed the highest proportion (90.4%) of genetic variability was explained by the within-population component in Italian white clover populations. The higher levels of intrapopulation diversity observed here correlates with certain life history traits such as an outcrossing breeding system and a widespread geographic range, which is observed in white clover (Hamrick & Godt, 1996; Nybom, 2004).

The weak population differentiation was also supported by ADMIXTURE and sNMF population structure results, where individuals were highly admixed between three to five genetic clusters. The negligible population structure identified here reinforces previous findings that white clover displays minimal structure on a regional and continental scale (George *et al.*, 2006; Olsen *et al.*, 2007; Kooyers & Olsen, 2012, 2013; Wright *et al.*, 2017). In the current study, three clusters were identified (**Figure 5.3**) across the range of "Dry" and "Wet" environments. The distribution of the individuals belonging to the larger clusters (Green- and Black-coloured; **Figure 5.3**) had no apparent relationship with geographic region or degree of soil moisture deficit (SMD). Similarly, the origin of individuals in the White-coloured cluster (**Figure 5.3**) also exhibited no clear relationship with geographic region but did include individuals from three Otago/Otakou populations (Arrowtown [AT], Fruitlands [FL] and Omarama [OM]; SMD = 81.8 – 124.4), two Canterbury Plains/Waitaha (Clarence [CL] and Kaikoura [KK];

SMD = 108.1 – 117.1) and one Marlborough/Tauihu population (Rai Valley [RV]; SMD = 45.2), all of which had moderate to high SMD values. However, as all populations from "Dry" sites were not included in this White-coloured cluster it is unlikely that the population structure observed is due to SMD variation. It is possible that the structure observed is influenced by cultivar origins with, for example, the three clusters representing three different cultivars. At a putative number of ancestral populations (K_P) of three, CL and RV, in the upper South Island/Te Waipounamu, contain individuals of a similar ancestral genetic composition to the three Central Otago/Otakou populations, AT, FL and OM. This may be a consequence of convergent adaptation but may also be a result of a common cultivar having been originally sown at these locations. Unfortunately, cultivar sowing records could not be acquired for any of the farms, and so any influence of cultivar origin on structure could not be determined here. Future studies could integrate the GBS data from the 17 populations with new data acquired from individuals sampled from cultivars prevalent in the NZ market 20 – 50 years ago (**Table S5.4**, Appendix 4). The relative positioning of survivor populations and cultivars on a phylogenetic tree could aid identification of the cultivar(s) that the 17 white clover populations are most closely related to and whether the observed population structure is cultivar-related. The land class (e.g. flat or hill), the type of stock (e.g. sheep or beef), in conjunction with the cultivars' year of release (targeting the approximate time the pastures were originally sown, i.e. 20 – 50 years ago), could be used to help determine which candidate cultivars to include in such a study (**Table S5.4**, Appendix 4).

Performing this study could be beneficial for two additional reasons. Firstly, it could identify cultivar(s) that are potential sources of genetic variation providing SMD tolerant phenotypes. For example, if a cultivar was found to be closely related to populations from the White-coloured cluster (**Figure 5.3**), which included many "Dry" site populations, that this cultivar could possess genetic variation at candidate gene(s) that could be selected for in a breeding program. Equally, this may be due to coincidence as a certain cultivar may have been sown at these sites, and these sites just so happened to be located in dry areas. In this case, there is no functional association between the cultivar and genetic variation. As discussed above, a critical piece of information that would have supported a more robust interpretation of the data is having the original cultivar information, from farmer sowing records. Secondly, it could help in determining how many generations the population has undergone since sowing. This is of importance as it can aid in determining how long white clover can persist in the field. This is a concern of the current study as different aged cohorts, due to buried seed, migration (both discussed in section 5.5.3) and clones persisting, can result in

overlapping generations within the pasture. In white clover there are two levels of population structure, the number of genetic individuals and the number of ramets (clones) per genetic individual. The latter is due to the white clover growth phases, i.e. the taproot phase and clonal phase (see Chapter 1, section 1.1.1.3). All white clover plants transition from the taproot phase to the clonal phase within two and a half years, with clonal plants forming 1 – 2.5 years after sowing (Brock *et al.*, 2000; Brock & Hay, 2001). Anecdotally, stolons are expected to persist for 5 – 6 years however there is a lack of data published to support this. The longest period of study assessing stolon persistence in the field has been three years (Brock *et al.*, 2000; Brock & Hay, 2001) but Harberd (1963) speculates clones could last up to 20 years. Pests and diseases such as clover root weevil, white clover leaf mosaic virus and slugs can impose major stress on the white clover plants and so clones surviving for more than ten years is unlikely (Brock & Hay, 2001). A sophisticated modelling approach to determine generation time based on genetic distance or mutation rate could be performed to obtain an accurate estimate. Sequence variation from the most closely related cultivar and the population of interest could be used to estimate a mutation rate and an estimated number of generations. If there are few mutations then it is likely that few generations have passed but if the mutation rate is higher then it is more likely that more generations have passed due to recombination. It is theorised that due to white clovers perennial nature, that some individuals may have persisted and could be examples of the material that was originally sown, while other individuals may be derived from successive generations. Hence determining the number of generations could also estimate how long white clover can persist in the field.

The known breeding history suggests a number of these “older” cultivars were used to generate some of the “newer” cultivars (Caradus, Hay & Woodfield, 1996; Caradus *et al.*, 1997; Caradus & Woodfield, 1997; Woodfield *et al.*, 2001; Woodfield *et al.*, 2003; Widdup & Barrett, 2011) which could result in phylogenies with low resolution (reduced clarity) among cultivars and the 17 populations. However, studies using microsatellite/simple sequence repeat markers (SSRs, see Appendix 4, Chapter 5 Supplementary Material, Supplementary experiment – Sample size determination, Introduction) and GBS data have shown genetic separation of NZ and international white clover cultivars (Jahufer *et al.*, 2003; George *et al.*, 2006; Faville *et al.*, 2020b), suggesting distinction between cultivars is achievable (this is discussed further in Chapter 6).

No obvious “Dry” vs “Wet” site population structure was observed from ADMIXTURE, sNMF and PCAdapt (**Figure S5.7**, Appendix 4; **Figure 5.3**) which would have been potentially indicative of different material having been originally sown in the two mega environments (e.g., one cultivar sown on the East/”Dry” side and another on the West/”Wet” side). Hence the lack of population structure observed facilitates the characterisation of the genetics of local adaptation as population structure can be a major confounder (false positive associations – SNPs associated with population structure rather than with adaptation to environments) in distinguishing between neutral alleles and locally-adapted alleles (De Mita *et al.*, 2013; de Villemereuil & Gaggiotti, 2015). The outlier and adaptive SNPs are discussed in detail in the following section (5.5.2).

5.5.2 Single nucleotide polymorphisms putatively under natural selection

Different strategies to manage soil moisture deficit (SMD) stress have been adopted by different plant groups, ranging from physical alterations to minimise water loss to the adjustment of osmotic chemical attributes to withstand periods of reduced water. Compatible solutes, also known as osmoprotectants, have been studied extensively in the context of in SMD stress tolerance, because osmotic adjustment aids in the maintenance of turgor pressure under water stress conditions (Pilon-Smits *et al.*, 1999; Le & McQueen-Mason, 2006; Ashraf & Foolad, 2007; Parida *et al.*, 2008; Li *et al.*, 2012; Filippou *et al.*, 2014). Osmoprotectants are molecules with a low molecular weight, are highly soluble, and electrically neutral and non-toxic at molar concentrations. Osmoprotectants include amino acids (ectoine and proline), ammonium compounds (b-alanine betaine, choline-O-sulfate, dimethyl-sulfonio propionate, glycinebetaine and polyamines) and sugars and sugar alcohols (D-ononitol, fructan, mannitol, sorbitol and trehalose) (Chen & Murata, 2002; Rontein, Basset & Hanson, 2002). Manipulation of osmoprotectant genes to increase osmoprotectant accumulation is one method to improve SMD stress tolerance in plants (Reguera, Peleg & Blumwald, 2012; Prakash, Kumar & Mikawlrawng, 2014). White clover is known to increase the concentration of water-soluble carbohydrates under SMD (Turner, 1990; Lee *et al.*, 2008; Li, Shi & Peng, 2013). In the current study, one SNP identified as potentially under selection was associated (in the exon) with the gene model ID chr3.jg6353.t1 (inositol transporter 1-like). The *Arabidopsis thaliana INOSITOL TRANSPORTER1* (*INT1*) is a member of a small gene family that encodes a tonoplast-localized H^+/inositol symporter that mediates the efflux of inositol from the vacuole (Schneider *et al.*, 2008). Based on the gene annotation, this SNP may be associated with response to SMD. Although no studies

have identified a link between *INT1* and improved drought tolerance specifically, one study in tobacco identified that overexpression of a plasma membrane inositol transporter *MfINT-like*, identified from *Medicago falcata*, resulted in greater tolerance to freezing temperatures and plants were also able to develop more fresh weight when grown under water stress (Sambe *et al.*, 2015). Findings from previous reports by Zhuo *et al.* (2013), Tan *et al.* (2013) and Sambe *et al.* (2015) also indicate that *MfINT-like* expression is induced by cold and salt stress, in both tobacco and *Medicago*. Interestingly, allele variation at this SNP indicated a higher proportion of homozygotes for the alternate allele (aa) in four populations located in the Otago/Otakou region including: Arrowtown (AT), Fruitlands (FL), Middlemarch (MM) and Omarama (OM). This variation is unlikely to be related to population structure because although AT, FL and OM mostly contained individuals belonging to the Yellow cluster at K_P = 5 (**Figure 5.3**), MM aligned to a different cluster at each of K_P 3 – 5 (**Figure 5.3**). Hence no obvious trend exists between SNP variation and neutral genetic structure. However, as this allelic variation was only present for a subset of all of the "Dry" site populations, it may have been selected for only at a regional scale rather than in all populations as a general response to high SMD conditions. Alternatively, all four populations were identified as grouping together on the environmental principal component analysis (PCA) biplot (**Figure 5.4**) and the allelic variation may be instead due to fluctuating temperatures, i.e. cold winter, hot summer and the variation between high temperature and low temperature that occurs during the same day (Mean Diurnal Range). Therefore, the efficacy of this SNP and the gene's role in response to SMD, and potentially prevention of freezing damage, requires further investigation.

One SNP, 1_89696968, located approximately 2,000 bp from the end of a phosphoglucosamine mutase gene (chr1.jg13031.t1) may also play a role in osmoprotection, based on its inferred function. Phosphoglucosamine mutase (*GlmM*) is an enzyme that catalyses the interconversion of glucosamine 6-phosphate to glucosamine 1-phosphate and is involved in the regulation of osmolyte homeostasis (Jolly *et al.*, 2000). However, experimental evidence for this function has only been examined in bacteria and its involvement in plant cell osmolyte homeostasis remains to be determined. Genotype variation at this SNP locus showed that all "Dry" site populations (SMD = 81.8 – 127.8) have a high frequency of the homozygous reference genotype (AA) with percentages ranging from 77.4 – 100% AA (mean = 88.4%). Populations with intermediate SMD values (LT, MR and RV; 25.7 – 45.2) had slightly lower AA percentages (71.4 – 78.8%, mean = 76.2%) and populations with low SMD values (CF, HA, KJ, RS and WR; 0 – 7.2) had the lowest AA frequency (54.5 – 70%,

mean = 65%) (**Table 5.3** and **Figure 5.6**). This trend is suggestive of positive selection for the AA genotype at this SNP at the “Dry” sites.

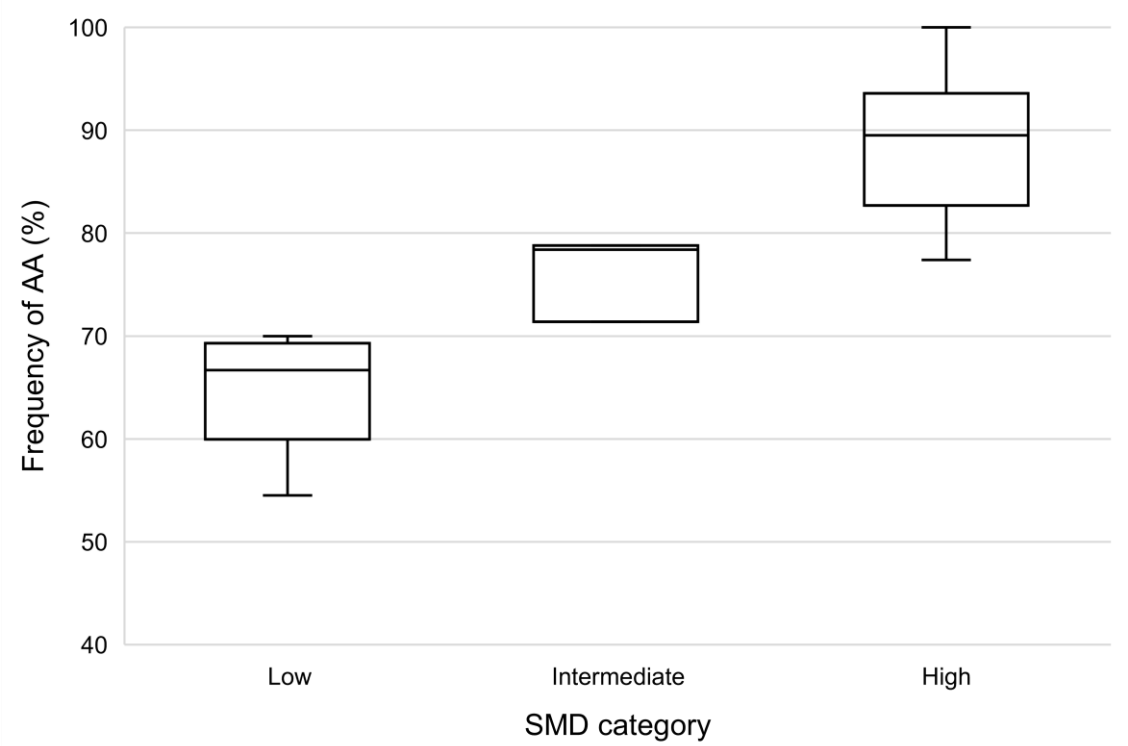


Figure 5.6 Genotype frequencies of the homozygous reference allele (AA) at SNP 1_89696968 for each of the 17 white clover populations categorised by soil moisture deficit (SMD). Number of samples (*n*) = 5 for low SMD populations (SMD = 0 – 7.2), *n* = 3 for intermediate SMD populations (SMD = 25.7 – 45.2) and *n* = 9 for high SMD populations (SMD = 81.8 – 127.8).

Another way plants can mitigate SMD stress is through the modification of their root system. Root system characteristics are fundamentally important for soil exploration to acquire resources, and are therefore strongly related to plant adaptation to environmental stresses such as high SMD (Ludlow & Muchow, 1990). Poor SMD tolerance has been associated with weakly tap rooted and shallow rooted white clover root systems (Caradus & Woodfield, 1998). In the current study, three SNPs were associated with root development genes. The first SNP, 15_31597018, was located in the intron of an ABC transporter C family member 5 (*ABCC5*) gene (chr15.jg4786.t1). An *Arabidopsis thaliana ABCC5*, also known as *AtMRP5*, is involved in root auxin regulation controlling the transition from primary root elongation to lateral root formation (Gaedeke *et al.*, 2001). Furthermore, *AtMRP5* has been shown to confer partial drought insensitivity by regulating stomatal aperture through control of the plasma membrane anion channels of guard cells (Klein *et al.*, 2003; Nagy *et al.*, 2009). The second SNP,

9_34153967, was located 1,196 bp from the end of an E3 ubiquitin-protein ligase SHOOT GRAVITROPISM9 (*SGR9*) gene (chr9.jg5048.t1). *SGR9* is localised in amyloplasts (starch containing plastids) and modulates amyloplast dynamics, hence plays a role in gravitropism (Nakamura *et al.*, 2011). Gravitropism is an important factor in determining root growth angle – a measure of downward growth. Uga *et al.* (2013) demonstrated in rice that increasing the root growth angle results in deeper rooting, and thereby improved drought avoidance. Similar to *INT1* (see above), populations AT, FL, MM and OM all exhibited a higher frequency of the aa genotype (53.8 – 76.2%) at 9_34153967 compared to the remaining populations (14.3 – 51.9%) (**Table 5.3**). The third SNP, 1_4206355, was located in the intron of a E3 ubiquitin-protein ligase wavy growth 3 (*WAV3*) gene (chr1.jg554.t1). *WAV3* also plays a role in root gravitropism (Sakai *et al.*, 2012).

Finally, SNP 4_18643366 was located 1,289 bp from the end of a Nodule Cysteine-Rich (*NCR*) secreted peptide gene (chr4.jg2548.t1). *NCR* potentially plays a role in *Rhizobium*-host symbiosis compatibility (Durgo *et al.*, 2015). van Ham *et al.* (2016) identified strong evidence for selection for desiccation tolerant *Rhizobia* phenotypes in high SMD environments.

Although no major SNP variation trend was observed for *ABCC5*, *WAV3* and *NCR* (**Table 5.3**), the functions of these genes suggest further investigation into root phenotypes and *Rhizobium* strains colonising nodules would be of value. These experiments could utilise experimental systems such as rhizoboxes, with plants grown in soil and RGB and hyperspectral root imaging are used to visualise root growth over time. These rhizoboxes can also be used to vary water content available to the plant, therefore plant root growth can be assessed under water stress (Bodner *et al.*, 2017). A subset of plants utilised for the common garden experiment described in section 5.5.3 could also be assessed for root architecture. If significant differences in root architecture were observed then measuring the expression of these genes in roots could help to elucidate how and whether root systems and the *Rhizobia*/clover symbiotic relationship have responded to local adaptation. To address the question of environmental adaptation of Rhizobium, stolon cuttings including roots, or whole plants with nodules, could be collected from the pastures utilised in the current study. Bacteria extracted from the Rhizobia root nodules could be assessed for genetic diversity following the protocol described in van Ham *et al.* (2016).

5.5.3 Assumptions and future studies

Within this chapter there has been an underlying assumption that the white clover individuals sampled are locally adapted to environmentally-contrasting habitats. Individuals may not be locally adapted if they are the result of contamination from another local source, for example, a more recent sowing compared to the naturalised population, and therefore may not have been present long enough to have undergone adaptation to the environment. The assumption that the populations' samples are exclusively derived from the originally-sown seed could be violated. The samples could include off-type volunteers/contaminants that are not from the sown cultivar but have emerged either as volunteers from the seedbank; or been deposited in the pasture from another source (e.g., a recent sowing) by livestock transfer; or via off-type pollination (assuming flowering).

A white clover seedbank can develop in the soil when white clover is allowed to flower and set seed (Roberts, 1981). The amount of total white clover seed within the soil seedbank can range from 2 kg ha^{-1} up to approximately 25 kg ha^{-1} depending on the region sampled, management, soil fertility and land classification i.e. lowland or hill pastures (Suckling & Charlton, 1978; Lancashire, Ralston & Scott, 1985; Wedderburn *et al.*, 1996). This is much more than the typical 3 – 4 kg ha^{-1} white clover seed over-sowing rate. Typically, pasture soils grazed by sheep contain fewer seeds in the soil seedbank than pasture soils grazed by cattle (Suckling & Charlton, 1978). An even number of sheep and cattle pastures were sampled in the current experiment (**Table S5.1**, Appendix 4), hence it is possible that a large soil seedbank could be present in some of the pastures sampled. However, the viability of white clover seed within a soil seedbank is generally very low. Furthermore, successful germination of white clover seeds from the soil seed bank is generally very low. Lewis (1973) determined white clover seed buried in mineral soil (pH 6.5 and organic carbon 4.6%) at discrete depths of 13, 26 or 39 cm and left undisturbed showed reduced germination with depth. Successful germination ranged from 6 – 23% germination after 1 year, 5 – 7% germination after 4 years but regardless of depth, germination was 1% after 20 years. Similarly, in peat soil (pH 4.2 and organic carbon 36.2%) at a depth of 26 cm, white clover seeds retained 7, 6 and 2% germination after 1, 4 and 20 years, respectively (Lewis, 1973). A factor influencing interpretation of these data is that in one trial white clover seed buried in soil did not germinate after five years, but it was noted that the germination conditions were likely unsuitable for overcoming dormancy (Crocker, 1938). White clover seed requires scarification and stratification to promote germination,

therefore if seeds remain undisturbed, they are unlikely to germinate. Although a large amount of white clover seed could be present in the soil seedbank of the sampled pastures, it is unlikely that buried seed would contribute to a large number of seedbank individuals being sampled as part of the target populations, as all pastures have remained undisturbed for at least a decade (**Table S5.1**, Appendix 4).

Secondly, because white clover seed can be easily transported via animals (Suckling & Charlton, 1978), the individuals sampled could be derived from white clover plants from neighbouring pastures. White clover seed can be spread by livestock through internal seed dispersal. Mature seed heads are consumed by livestock, the seeds get scarified through the digestive tract, and are then deposited in faeces. Legume seeds extracted from cattle and sheep dung have been shown to retain their germination (Suckling, 1950; Yamada & Kawaguchi, 1972). Although sampling in the current experiment was conducted to avoid collecting plants on and near visible animal tracks and dung piles and only sampling from mature plants, not seedlings; there is no way to completely rule this out unless detailed information about grazing practice and livestock movement is obtained from landowners. This is only an issue if neighbouring pastures have been recently resown and the gene flow from these recent sowing could reduce the extent of local adaptation. Unfortunately, there were a number of populations used in the current experiment that were only one of a few pastures that had not been resown, with the majority of the surrounding paddocks having been resown recently. However, correct grazing management to reduce white clover flowering in the resown paddocks had been described by the farmers.

Finally, gene flow by pollen from resown pastures could also reduce the ability to detect a local adaptation signal, through dilution of adaptive gene genotypes. White clover is primarily bee pollinated. A bee's typical home range is 2 – 5 km which suggests that pollen could move to and from the resown and naturalised pastures easily (Newstrom-Lloyd, 2013). However, Marshall, Michaelson-Yeates and Williams (1999) demonstrated that although pollen transfer is high over short distances (less than 1 meter), low levels of pollen transfer were observed when plants were more than 2 m apart. Woodfield *et al.* (1995) also showed that the majority of gene flow from a donor source occurred within the first 4 m and at distances further than 24 m, it was negligible. Furthermore, low rates of gene flow were measured at 100 – 250 m from the donor plot (Woodfield *et al.*, 1995). Honey bees have been shown to deposit pollen onto the first 15 – 20 inflorescences they visited and only sporadic pollen deposition was observed up to the 50th inflorescence (Marshall *et al.*, 1999). Hence, gene flow from resown to

naturalised pastures is likely to be minimal. As the spatial location of each individual plant sample was recorded, the current dataset could be used to investigate the spatial genetic variability. This would be of interest to determine how closely related individuals are across the paddock and if, for example, individuals on one side of the paddock are more closely related compared to individuals on the other side of the paddock, or if no population structure exists within a paddock.

The three scenarios discussed above are strongly centred around ambiguity of whether the individuals sampled are locally adapted to the environmentally contrasting habitats. One simple method to test this would be to perform a common garden experiment. Mature white clover seed or stolon cuttings could be gathered from a subset of the populations. Clones from these individuals could then be placed in environmentally contrasting conditions, for example, high soil moisture deficit (SMD) and low SMD. An ideal control for this experiment would be a random sample of seed from the cultivar known to be originally sown at the site. However as mentioned in section 5.5.1, this information is unknown. Hence, individuals sampled from a range of cultivars, ideally closely genetically related (identified from the positioning of naturalised populations and cultivars on a phylogenetic tree, as previously suggested) could provide adequate controls. If individuals from the “Dry” sites survived in the high SMD conditions, and vice versa, when compared to cultivar material that has not been exposed to the environment, then this would be evidence of adaptation. The successful identification of candidate genes for water-soluble carbohydrate accumulation, through transcriptomics and proteomics approaches (See Chapter 4), suggests that one or both methodologies could be employed to further investigate the SMD tolerance candidates identified here and to establish if SNP variation in those genes has contributed to high or low SMD adaptation. The role of the inositol transporter and the role of sugars in response to SMD as osmoprotectants (Singh *et al.*, 2015) could be examined through a targeted metabolomics study of the material used in the common garden experiment described above. Furthermore, phenotyping using a more sophisticated experimental approach, such as the use of rhizoboxes, coupled with transcriptomics of root material may aid in elucidating which genes have responded to local adaptation for root systems and whether a correlated change in root morphology occurs.

Genetic variation alone may be insufficient to identify determinants of local adaptation in white clover. Epigenetic changes accompanying environmental change can occur at a faster rate than the sorting of allelic variation and mutational change (Miryeganeh & Saze, 2020; Rey *et al.*, 2020). Gene expression is controlled through

chemical modification of DNA, RNA and proteins in response to environmental signals, known as epigenetics (Donohue, 2014; Junaid *et al.*, 2018). One form of epigenetic change is methylation, and is the most extensively studied mechanism for epigenetic gene regulation (Portela & Esteller, 2010). Changes in the methylome (regions of nucleic acid methylation modifications in the genome) could be responsible for adaptation to high SMD. In plants, repetitive and transposable elements are heavily methylated and cytosine methylation of promoter regions can inhibit transcription of genes (Zhang *et al.*, 2010a). Identification of differentially-methylated genomic regions could therefore be used as a complementary approach to discovering genes associated with high SMD response. The majority of gene model IDs identified in the current experiment play a role in transcription or post-translational modification, which could be missed by a transcriptomic and proteomic study. Expression levels of transcription factors are generally very low and so identifying significant log fold expression changes are not usually observed (Ghaemmaghami *et al.*, 2003). Hence, the impact of epigenetic modifications can be assessed through a methylation genotyping study. Reduced representation sequencing protocols such as bsRADseq, DREAM and EPiGBS can be used to capture a subset of the entire epigenome through characterisation of partial methylation profiles (Jelinek & Madzo, 2016; Trucchi *et al.*, 2016; van Gurp *et al.*, 2016). Locus-specific bisulphite sequencing can target regions of the genome of specific interest (Hernández *et al.*, 2013; Lam *et al.*, 2020). If the cost of whole-genome sequencing is not an issue then whole genome bisulphite sequencing (ca. USD$1000 per sample) (Lister *et al.*, 2009; Hansen, Langmead & Irizarry, 2012) or assay for transposase-accessible chromatin using sequencing (ATAC-seq) can be used. The latter maps genome-wide chromatin accessibility, which is tightly linked to gene expression (Miskimen, Chan & Haines, 2017). Methylation factors could explain how some white clover populations are more fit in one habitat than another. These methylation differences could be assessed in the common garden experiment mentioned above and address the questions: is there epigenetic variation within and among naturalized white clover populations? What are the methylation patterns for putatively adaptive genes? Future studies could aim to compare genomic and epigenetic diversity to transcriptome variation and to how the resulting variation relates to local adaption for SMD tolerance.

5.6 Conclusions

In the current study population structure and genetic variation was assessed in naturalised white clover populations from contrasting New Zealand/Aotearoa (NZ)

environments, using 15,120 SNP markers. Three to five genetic clusters were identified by ADMIXTURE and sNMF, but minimal population differentiation was observed using pairwise F_{ST}, and AMOVA revealed the majority of genetic variation was partitioned within populations. Corroboration between at least two outlier detection (OutFLANK) and environmental association analyses (*LEA*, *lfmm* and BayeScEnv) identified 64 single nucleotide polymorphisms (SNPs) as significantly associated with environmental variation, when a *q*-value FDR was applied. Five of the 64 SNPs exhibited differences in SNP genotype frequency between contrasting environments and were potentially associated with genes involved in carbohydrate metabolism and root system characteristics. One SNP, associated with the gene phosphoglucosamine mutase, showed a statistically-significant difference in AA genotype frequency between “Dry” sites and “Wet” site populations. This study presents the first in-depth population structure assessment of naturalised white clover populations in NZ. Further studies using a common garden experimental approach are recommended to confirm local adaptation of white clover to contrasting soil moisture deficit environments.

5.7 Acknowledgements

I would like to thank Craig Anderson (AgResearch), Kulatunga Tennakoon (Lincoln University) and Assoc. Prof Hayley Ridgeway (Lincoln University/Plant and Food Research) for help in collecting missions. Anna Larking (AgResearch) for GBS library construction. Ruy Jauregui (AgResearch) for SNP calling. Limei Zhang (Nanchang University, Jiangxi, China) for statistical support and Prof Ken Olsen for advice and assistance (Washington University, St Louis, Missouri, USA).

CHAPTER 6

General discussion

The studies reported in this thesis sought to identify loci associated with two white clover traits, foliar water-soluble carbohydrate (WSC) accumulation and soil moisture deficit (SMD) tolerance, with the goals of generating new knowledge on the genetic control of these traits and identifying candidate loci for use in marker-assisted breeding. In summary, a phenotyping study in Chapter 2 characterised divergent WSC phenotypes in artificially selected breeding populations. These populations were utilised to investigate underlying genetic features influencing this trait through multiple studies including genomic analyses (Chapter 3) and transcriptomic and proteomic analyses (Chapter 4). Chapter 5 moved away from artificial controlled selection to investigating natural selection in response to environmental conditions, in this case, SMD. This study focussed on analyses of single nucleotide polymorphism (SNP) variation in genotyping by sequencing (GBS) data from naturalised white clover populations collected from 17 sites representing contrasting SMD across the South Island/Te Waipounamu, New Zealand/Aotearoa. The findings reported in Chapters 2 – 5 make a significant contribution to our understanding of the genetic determinants of WSC and SMD phenotypes in white clover. The key findings, limitations of the studies, future directions and potential applications for the forage seed and pastoral agriculture industries are discussed below.

6.1 Key findings

Previously generated divergent WSC lines derived from five breeding pools were characterised for foliar WSC, leaf size and other nutrient measures. The pools with low and high WSC lines showed significant divergence in foliar WSC content. Despite a previously noted correlation between WSC content and leaf size, little correlation was observed between the two, indicating that these two variables are independent in these breeding programmes. These outcomes suggest that breeding for increased WSC content itself rather than being solely a function of leaf size, can be achieved in both large and small leaf size classes of white clover in as few as 2 – 3 generations.

Genome, transcriptome and proteome analyses in conjunction with the phenotype data were combined in white clover for the first time, to identify gene candidates responsible for variation in foliar WSC levels – a critical trait for improving the environmental sustainability and productivity of New Zealand/Aotearoa (NZ) pastures. These multiple lines of evidence implicated involvement of the *glgC* gene in foliar WSC accumulation as it is involved in the first step in the starch biosynthesis pathway. A genome-wide association study (GWAS) to investigate the relationship

between SNP markers distributed across the genome and WSC content, detailed in Chapter 3, identified *glgC* (gene model ID chr5.jg7106.t1, glucose-1-phosphate adenylyltransferase small subunit 2) as highly ranked with $-\log_{10}(p\text{-values})$ greater than 3 for both the WSC and soluble sugars and starches (SSS) traits. Similarly, gene model IDs for *glgC* were also differentially expressed in the transcriptome and proteome analyses reported in Chapter 4. For example, this same *glgC* gene model ID described above was also found to be significantly upregulated in the proteome dataset for both the Widdup New Zealand large leaf (WNZLL) High WSC-End generation (WH) and Ford New Zealand large leaf (FNZLL) High WSC-End generation (FH) populations (Chapter 4). Interestingly, SNP variation from RNA-Seq data for this gene model ID identified eight SNPs discriminating between high and low WSC individuals. Two other gene model IDs for *glgC* (both glucose-1-phosphate adenylyltransferase large subunit 1) were identified as being significantly upregulated in the WH and FH populations: one (jg5855.t1) was significantly upregulated in the transcriptome dataset for both populations; and the other (chr4.jg5408.t1) was significantly upregulated in both the transcriptome and proteome datasets for both WH and FH populations. Overall, multiple lines of evidence corroborate the importance of *glgC* for increasing foliar WSC accumulation in white clover.

Gene model IDs from six additional gene families (*AMY*, *BAM*, *glgA*, *glgB*, *ISA3* and *WAXY*) in the starch biosynthesis and degradation pathway showed significant upregulation in both the WH and FH (high WSC) populations for at least one of the transcriptome or proteome datasets. Interestingly, SNP variation from RNA-Seq data for two of these gene model IDs (*AMY* and *WAXY*) identified one and two SNPs, respectively, discriminating between high and low WSC individuals in the form of distinct genotype differences.

The outlier analyses described in Chapter 3 identified a large number of genes as potentially under selection during breeding for divergent foliar WSC. For the most part, the transcripts and proteins of these genes were not investigated as there was little overall corroboration between the genome, transcriptome and proteome. In the one case where the transcripts and proteins for a gene (chr16.jg4564.t1, At1g19450 Sugar transporter *ERD6-like 4*) identified in Chapter 3 were briefly investigated. The gene was not identified by sequential window acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH MS) proteomics, and the closest match in the *T. occidentale* genome for transcriptomics (jg42295.t1 [83.9% IDM]) had very low transcript counts (min = 0 and max = 8.8 read counts) and was not significantly differentially

expressed. The lack of corroboration for the sugar transporter *ERD6-like 4* gene from the transcriptome and proteome analyses, may suggest this outlier analysis result was a false positive.

A landscape genomics study described in Chapter 5, assessed 17 populations likely to have undergone local adaptation in climatically contrasting “Dry” and “Wet” environments, namely areas of high and low soil moisture deficit (SMD) in the South Island/Te Waipounamu of NZ. This is the first study to identify potentially adaptive SNPs to high SMD in white clover utilising naturalised populations. Sixty-four SNPs were identified as significantly associated with environmental variation when a *q*-value false discovery rate was applied. Of these SNPs, five exhibited differences in genotype frequency when compared between subsets of the 17 populations defined as being from “Dry” and “Wet” environments. One SNP, associated with a phosphoglucosamine mutase gene (chr1.jg13031.t1), exhibited a statistically significant difference in the frequency of the reference allele homozygous genotype between the contrasting environments. The remaining four SNPs also showed genotype frequency differences between “Dry” and “Wet” but these were not statistically significant. The five SNPs were associated with genes involved in carbohydrate metabolism and root system characteristics. This finding and those described above implicate sorting of allelic variation as important in the evolutionary response of white clover to both natural and artificial selection.

6.2 Comparison of genetic variation and structure between the two white clover datasets (foliar water-soluble carbohydrate and soil moisture deficit)

Population structure results indicated that stronger population structure was observed in the artificial selection populations (WSC) in contrast to the weaker population structure observed in the naturalised populations (SMD). Comparing pairwise F_{ST} values among populations in the respective studies, F_{ST} was found to be larger among the WSC populations (0.03 – 0.23; **Table 3.1**) in contrast to the SMD populations (0.001 – 0.047; **Table 5.1**). Similarly, analysis of molecular variance (AMOVA) identified a higher proportion of genetic variation was partitioned among populations in the WSC populations (23.3%; **Table 3.3**) compared to among the SMD populations (3.9%; **Table S5.3**, Appendix 4).

In the WSC study, both pairwise F_{ST} and discriminant analysis of principal components (DAPC) analyses suggest that the founder (Parent) populations in the breeding programmes were not strongly differentiated. These Parent populations were

composed of individuals taken from different cultivars/breeding populations which were subsequently crossed together. Parent populations from among NZ pools showed little to no detectable differentiation (F_{ST} 0.03 – 0.04), however, when comparing Parent populations from the NZ and United Sates of America (US) pools, there was slightly more differentiation (F_{ST} 0.06 – 0.08). This suggests slight differentiation of cultivars from different countries of origin. The *K*-means clustering algorithm implemented in DAPC population structure analysis assigned all individuals from the Parent populations into one cluster. However, assessment of population structure of the Parent populations solely (i.e. analysed in the absence of the Mid and End generation populations of the divergent lines) was not performed. It may be that when all populations were combined, the stronger differentiation of the divergent WSC populations obscured distinctions among the Parent populations from the different pools in the DAPC analysis. Therefore, it may be that if the Parent populations were assessed independently of the rest of the selected populations, distinction between the cultivars that comprised the Parent populations could possibly be observed. This hypothesis is supported by the observation that when the Parent population was included in a preliminary PCAdapt analysis for the FNZLL pool, there was distinction among the different cultivars of which Parent population was comprised (**Figure 6.1**). Although there were only seven to eight individuals from each of the cultivars (Aran, Grasslands Kopu, Grasslands Kopu II, Grasslands Pitau and Grasslands Sustain) and breeding line (G23 selection), this plot highlights that white clover cultivars can be distinguished from each other using GBS SNP data.

As mentioned above, the cultivar components of the WSC FNZLL Parental pool were able to be differentiated (**Figure 6.1**), an observation consistent with previous studies suggesting that white clover cultivars can be genetically distinguished. Jahufer *et al.* (2003) determined the genetic relatedness of 32 white clover cultivars using 39 microsatellite/SSR markers. Cluster analysis assigned the 32 cultivars into 14 clusters, and there was a strong correlation between geographic origin of the germplasm and genetic relationships based on pedigree. Similarly, Kölliker *et al.* (2001) identified separation among 52 white clover cultivars and accessions using AFLP markers. Less evidence for genetic distinction was obtained in a microsatellite study by George *et al.* (2006), in which 15 SSRs were used to characterise a range of white cultivars from nine countries (mainly in Europe and Australasia) to assess population diversity. Limited clustering of individuals from the same cultivar and substantial overlap between each cultivar was observed, suggesting that population structure among white clover cultivars was limited (George *et al.*, 2006). A more recent study by Faville *et al.* (2020b), used

pooled GBS SNP data to assess diversity from extant naturalised (from wild grasslands and mature farm systems) white clover populations in countries where white clover was introduced (NZ, Australia and the US). These populations were compared to the putative source populations in western Europe – i.e., those assessed by George *et al.* (2006). Given the similarity in the geographic origins and potential pedigree of the cultivars studied by George *et al.* (2006) and naturalised populations studied by Faville *et al.* (2020b) the different findings of these authors are best explained by differences in the resolution of the molecular marker systems that were used.

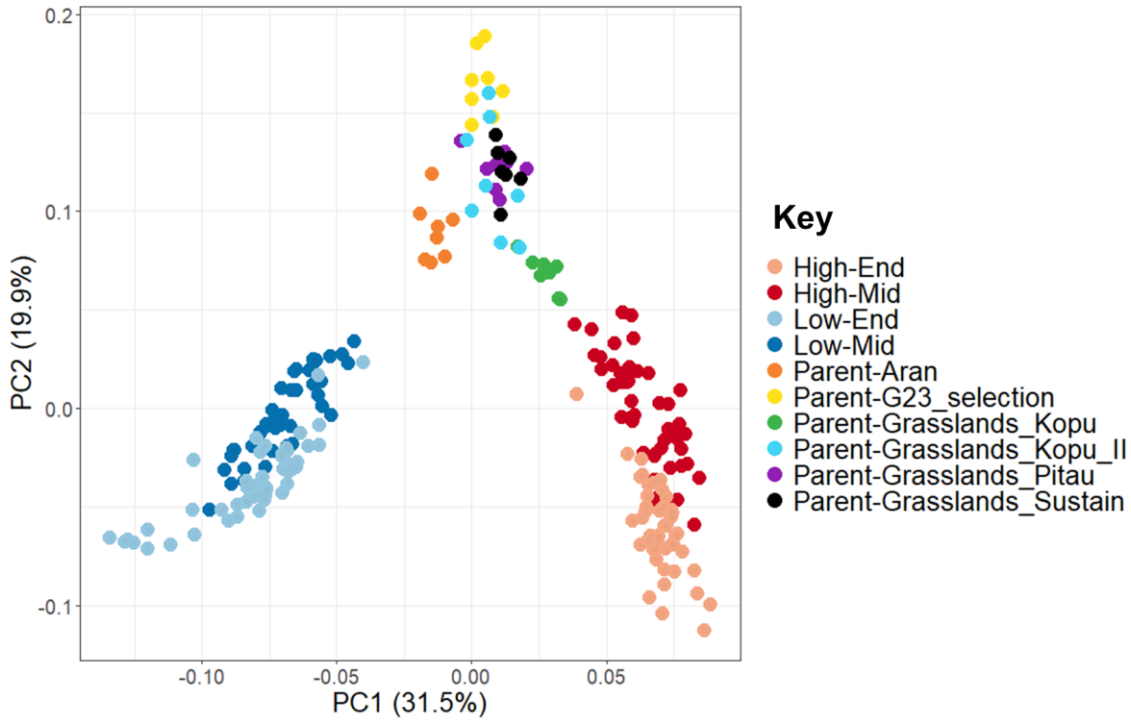


Figure 6.1 Score plot from a PCAdapt analysis using the first two principal components (PCs) and the five populations from the FNZLL pool: High-End, High-Mid, Low-End, Low-Mid and Parent. The individuals from the Parent population are separated into cultivars and breeding lines and colour coded. Population information is displayed in the key. High = high water-soluble carbohydrate (WSC), Low = low WSC, Mid = middle generation, End = end generation.

The ability to genetically differentiate cultivars provides an exciting opportunity to resolve the population structure observed in the SMD study. As described in Chapter 5, there were three genetic clusters identified by the ADMIXTURE analysis. One of the clusters, referred to as “White-coloured”, included individuals from the Otago/Otakou region but also from Marlborough/Tauihu and Canterbury Plains/Waitaha regions. In the absence of sowing records, in order to determine whether the observed population

structure was a consequence of the cultivars that were originally sown in these pastures, it was suggested in Chapter 5 that a range of white clover cultivars could be genotyped (using GBS derived SNPs) and analysed in conjunction with the pasture populations. Using genotypes from a range of white clover cultivars encompassing the time period during which the pastures were known to have been sown (and likely to have been commercially available at the time of sowing), along with the 17 white clover populations, phylogenetic analyses may help elucidate whether the observed population structure resulted from persistence of the original cultivars that were sown. If, for example, the individuals from the White-coloured cluster were closely related to one or two cultivars it would provide evidence for their recent common origin. Alternatively, a simple analysis that uses data already available from the current studies could be used to identify individuals from the WSC Parent populations that belong to cultivars of interest e.g. Aran, Grasslands Kopu, Grasslands Kopu II, Grasslands Pitau and Grasslands Sustain from FNZLL and Grasslands Huia, Grasslands Prestige, Demand and Prop from Ford New Zealand small leaf (FNZSL) and co-analyse these data with the 17 SMD populations. Observation of alignment between cultivars and SMD populations (via principal component analysis or a distance-based dendrogram) could be used to infer the cultivar origins of the sown pastures. Such an analysis would be limited by the range of cultivars it assesses, but if significant differentiation were observed then it would warrant a larger genotyping study on a wider variety of cultivars to be performed in the near future. Future expansion of the white clover dataset described by Faville *et al.* (2020b) will include the 17 SMD populations and cultivars likely to have been commercially available at the time of sowing (Dr Andrew Griffiths, pers comm October 2020) and may serve to resolve this issue.

6.3 Limitations of the methodology used and recommendations for future research

6.3.1 Chapter 2 – Phenotyping water-soluble carbohydrate populations

The moderate correlation observed between near infra-red reflectance spectroscopy (NIRS), which was the WSC phenotype used for breeding the WSC populations (Widdup *et al.*, 2010), and the time-consuming but accurate anthrone wet chemistry measurement of WSC suggests that there may be some inaccuracy associated with the current NIRS calibration. Greater progress might be made using artificial selection with a more accurate phenotyping method. This might be achieved through either developing a NIRS calibration specifically designed for white clover or shifting to a hyperspectral or

multi-spectral imaging approach. However, in the present study, the approximation of WSC phenotype was still sufficient to differentiate divergently selected populations and identify genetic determinants of the trait. For example, the genome-wide association study (GWAS) in Chapter 3 used both genotype and phenotype information to identify a gene (*glgC*) that was associated with the WSC trait as measured by NIRS. There was additional support for this gene as a candidate from both transcriptomic and proteomic analyses, as mentioned above.

6.3.2 Chapter 3 – Genomic analysis of water-soluble carbohydrate populations

Two of the major limitations for the GBS dataset from the WSC populations was the amount of missing data for each SNP, and the SNP density across the genome. The extent of missing data meant that data analysis was limited to protocols that are not affected by the inclusion of some missing data. For the GWAS, genotype data had to be imputed as GWAS algorithms require complete datasets (Poland *et al.*, 2012a). But for accurate imputation, a very reduced set of SNPs (*n* = 5,757) were used. GBS libraries were sequenced on two lanes of an Illumina flow cell, in order to increase sequencing depth, and this served to reduce missing data. Therefore, one approach to alleviate the missing data issue would be to acquire additional sequencing lanes; however, this would increase the genotyping cost. The moderate SNP density (ca. 16 SNPs per Mbp) reported in the present study meant that there were few genes in LD with SNPs that could be targeted. An issue with reduced representation sequencing such as GBS is that less than 10% of the genome is sampled. Furthermore, during the GBS SNP calling workflow, reads are removed if they align to multiple regions of the genome and this is exacerbated in the current study due to the polyploidy of white clover. This reduces error but also typically results in fewer SNP markers retained compared to a diploid species such as perennial ryegrass (Faville *et al.*, 2018; Faville *et al.*, 2020b). Due to limitations of the current dataset, namely the low SNP density, linkage disequilibrium (LD) decay could not be determined for each population and an alternative approach was implemented. This involved focusing on identified intergenic outlier SNPs (ioSNPs) and calculating LD between these SNPs and all other SNPs within a 100 Kbp window (see section 6.3.5 for more discussion on LD).

6.3.3 Chapter 4 – Transcriptomic and proteomic analysis of water-soluble carbohydrate populations

A major limitation in the transcriptome and proteomic analyses in Chapter 4 was the low sample size for each population (only 5 – 6 individuals per population). This low sample

size was used due to the cost involved in transcriptomic and proteomic analyses. The SNP allelic variants detected from these data were therefore based on a small sample size comprising 10 individuals with a high WSC phenotype and 10 individuals with a low WSC phenotype. Additional studies investigating the genotypes and/or haplotypes for the 14 identified candidate genes from a larger sample size would help to strengthen the inferences that were made concerning candidate gene loci. While a phylogenetic framework was considered in analysing patterns of differential expression, profiles were not explicitly modelled as discussed in Voelckel *et al.* (2017). Such an approach might help to provide further evidence for candidate genes responding to selection.

6.3.4 Chapter 5 – Genomic analysis of naturalised soil moisture deficit clover populations

Compared to the previously discussed chapters, limitation and assumptions of a larger impact were observed in Chapter 5, in which white clover populations from contrasting soil moisture deficit (SMD) environments were investigated. Firstly, the environmental data utilised were from the 1970 – 2000 period and did not include environmental variation for the past 20 years because those data were not available from the same source. The majority of the populations had not been resown or oversown since the late 1990s so the environmental data did not describe the last ca. 18 years of the pastures' history. A question then is: how well does the environmental data represent the climatic conditions that the sites would have experienced over this more recent period? Important environmental variables including potential evapotranspiration, precipitation, and temperature were examined for 1970 – 2018 to determine if there were any major changes in environmental conditions between two time periods: 1970 – 2000 and 2000 – 2018 (**Figure 6.2** and **Figure 6.3**). Climate data were taken from four National Institute of Water and Atmospheric Research (NIWA) stations spread across the South Island/Te Waipounamu including: Nelson (-41.299, 173.226), Omarama (-43.7935, 171.7951), Christchurch (-44.526, 169.889) and Hokitika (-42.712,170.984) (NIWA, 2020). Permutation analysis of variance (ANOVA) tests were performed in R using the *aovp()* function from the "lmPerm" v 2.1.0 package (Wheeler & Torchiano, 2016) to determine if the means at each station were significantly different at $\alpha = 0.05$ for each environmental variable. These permutation ANOVAs were performed with 1e+7 permutations on each of the environmental variables: Penman (1948) potential evapotranspiration (mm; PET), total precipitation (mm), mean maximum air temperature (°C), mean air temperature (°C) and mean minimum air temperature (°C) with 'Station' (Hokitika, Nelson, Christchurch or Omarama) and 'Time Period' (1970 – 2000 or 2000 –

2018) as separate factors and an interaction term. These tests were run using the following code:

anova(aovp(Environmental_Variable~Station*Time_Period, data=x, maxIter=1000000, Ca=0.00000000000000001))

Where: Environmental_Variable is one of the five above-mentioned variables.

Following the ANOVAs, pairwise tests between the predicted means were performed using the *predictmeans()* function from the “predictmeans” v 1.0.1 package (Luo, Ganesh & Koolaard, 2018) to determine which means were significantly different using the following code:

predictmeans(aovp(Environmental_Variable~Station*Time_Period, data=x, maxIter=1000000, Ca=0.00000000000000001), Df=10, "Station:Time_Period", pairwise=TRUE)

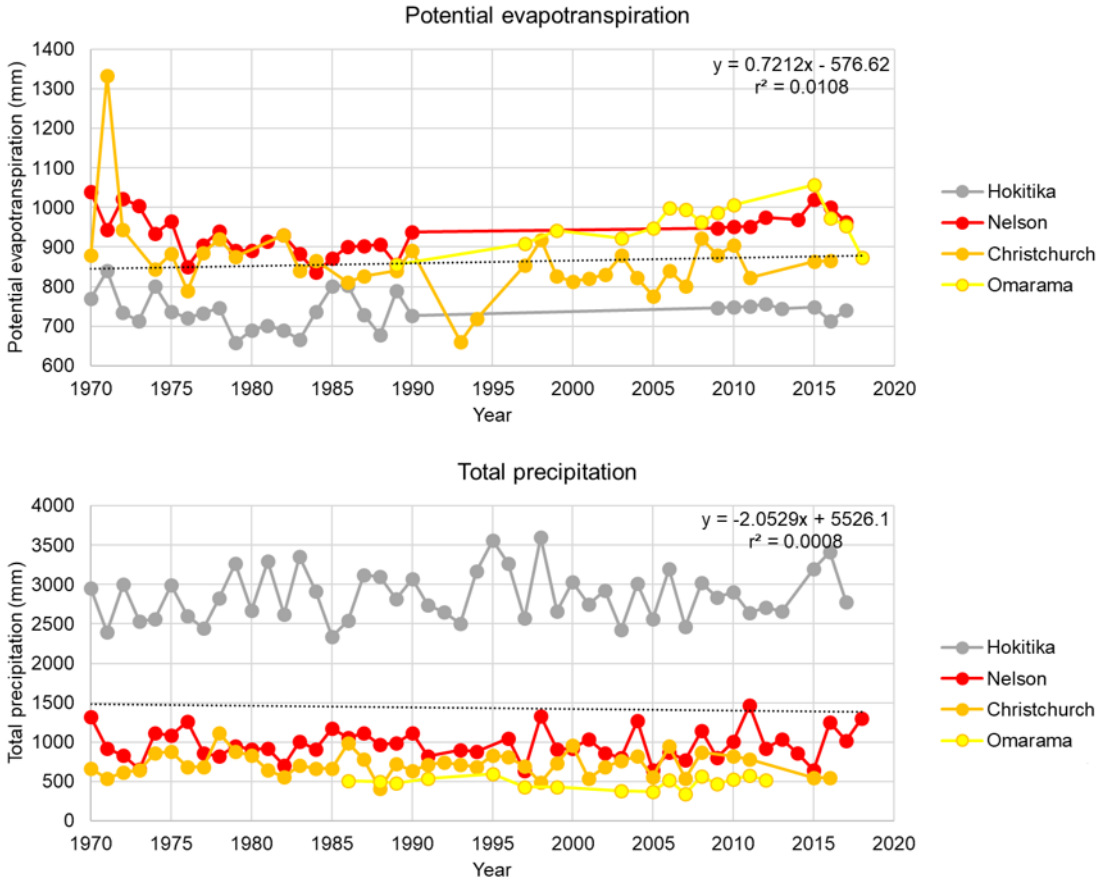


Figure 6.2 Precipitation variables including the annual mean of monthly Penman (1948) potential evapotranspiration (top) and of monthly total precipitation (bottom) from 1970 – 2018 for four locations in the South Island/Te Waipounamu of New Zealand/Aotearoa. Data collecting stations are located in Nelson (-41.299, 173.226), near Omarama (-43.7935, 171.7951), near Christchurch (-44.526, 169.889) and in Hokitika (-42.712,170.984). Line of best fit, coefficient of determination (r^2) and linear regression equation using all locations per variable are presented. Data were sourced from the New Zealand national climate database provided by National Institute of Water and Atmospheric Research (NIWA). *Note*: Omarama data are from late 1980s to late 2010s.

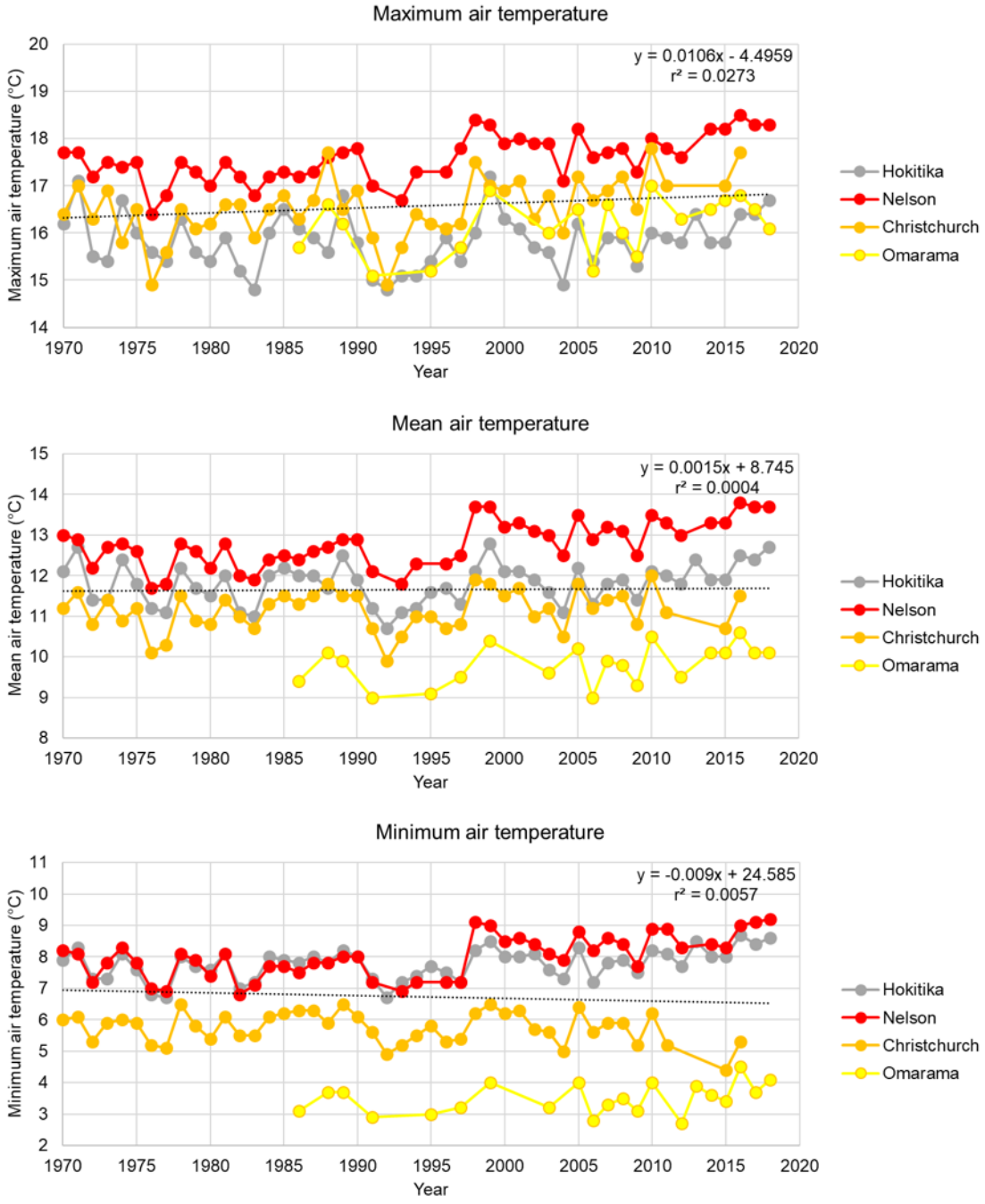


Figure 6.3 Temperature variables including the mean maximum temperature from daily maximums (top), mean air temperature calculated as 0.5 x (Max + Min) (middle) and mean minimum temperature from daily minimums (bottom) from 1970 – 2018 for four locations in the South Island/Te Waipounamu of New Zealand/Aotearoa. Stations are located in Nelson (-41.299, 173.226), near Omarama (-43.7935, 171.7951), near Christchurch (-44.526, 169.889) and in Hokitika (-42.712,170.984). Line of best fit, coefficient of determination (r^2) and linear regression equation using all locations per variable are presented. Data were sourced from the New Zealand national climate database provided by National Institute of Water and Atmospheric Research (NIWA). *Note*: Omarama data are from 1986 – 2018.

There were no significant differences in PET or precipitation for both the Time Period (p = 0.13 and 0.84, respectively) and interaction between the Station and Time Period (p = 0.11 and 1.00, respectively) (**Table 6.1**). However, there were significant differences ($p < 0.05$) between the two Time Periods for all three temperature variables (maximum temperature, mean temperature and minimum temperature) (**Table 6.1**). Pairwise tests reveal that only Nelson was significantly different between time periods for mean temperature (p = 0.0007), Nelson and Christchurch were significantly different between time periods for maximum temperature (p = 0.0077 and 0.0070, respectively) and Nelson and Hokitika were significantly different between time periods for minimum temperature (p = 0.0002 and 0.031, respectively) (**Table 6.2**). These temperature differences ranged from +0.36 to +0.84°C (**Table 6.2**).

Overall, therefore, precipitation-related variables used in the environmental association analyses in Chapter 5 appear to have accurately reflected the climatic conditions that the 17 pastures would have experienced across the full period of their existence. However, temperature related variables used in environmental association analyses may not have reflected the climatic conditions that the 17 pastures would have experienced over the last 18 years. Perhaps the WorldClim data was not the most appropriate for estimating temperature variables, but unless weather loggers are located at each site, there will always be some degree of error with interpretation of climate data.

Another potential caveat is that the soil data and site characteristics were recorded at the time of sample collection. These recent observations do not provide information on any variation in levels of minerals that might have occurred throughout the history of each site. However, this is likely to be of less concern, because soil characteristics were not important for differentiating “Dry” and “Wet” site populations based on principal component analysis (**Figure S5.4**, Appendix 4), and thus are unlikely to be confounders. In contrast, the SMD values were taken from van Ham *et al.* (2016) and were calculated from long-term SMD data obtained from NIWA (although specific years are not stated) and are likely to reasonably reflect climatic conditions that may have changed throughout the populations’ history. Finally, since no measures of reproductive fitness were made, there is uncertainty as to whether the individuals sampled from each site were locally adapted to each of the environmentally contrasting habitats. This is addressed in detail in section 5.5.3 (Chapter 5) and may be investigated using a common garden experiment.

Table 6.1 Permutation ANOVA based on 1e+7 permutations for the effect of Station (Hokitika, Nelson, Christchurch or Omarama) and Time period (1970 – 2000 to 2000 – 2018) and the interaction between Station and Time Period for five environmental variables.

Environmental variable	Response	DF	SS	MS	*p*-value (2 s.f.)	
Potential Evapotranspiration	Station	3	561701	187234	<2e-16	***
	Time Period	1	11280	11280	0.13	
	Station:Time Period	3	30132	10044	0.11	
	Residuals	99	474480	4793		
Total precipitation	Station	3	130845316	43615105	<2e-16	***
	Time Period	1	1988	1988	0.84	
	Station:Time Period	3	5447	1816	1.00	
	Residuals	145	7476838	51564		
Maximum temperature	Station	3	77.608	25.8694	<2e-16	***
	Time Period	1	5.217	5.2166	0.000016	***
	Station:Time Period	3	1.360	0.4533	0.19	
	Residuals	151	42.697	0.2828		
Mean temperature	Station	3	139.286	46.429	<2e-16	***
	Time Period	1	3.879	3.879	0.000063	***
	Station:Time Period	3	1.656	0.552	0.065	
	Residuals	151	34.006	0.225		
Minimum temperature	Station	3	373.92	124.639	<2e-16	***
	Time Period	1	2.99	2.987	0.00058	***
	Station:Time Period	3	5.29	1.765	0.00012	***
	Residuals	152	36.78	0.242		

Note: DF = degrees of freedom, SS = Sum of squares, MS = Mean square, s.f. = significant figures, *p*-value significance thresholds: * = $p < 0.05$, ** = $p < 0.01$ and *** = $p < 0.001$.

Table 6.2 Pairwise difference among two time periods for four environmental variables calculated by predicted means. Comparisons are comprised of the mean monthly (potential Evapotranspiration and total precipitation) and daily (maximum, mean and minimum temperature) data from each year within the time period (1970 – 2000 and 2000 – 2018). The difference between the years is presented in the "Diff." column.

Environmental variable	Location	Comparison	Diff.	*p*-value (2 s.f.)	
Potential Evapotranspiration	Hokitika	2000-2018 – 1970-2000	+7.15	0.81	
	Nelson	2000-2018 – 1970-2000	+51.82	0.10	
	Christchurch	2000-2018 – 1970-2000	-27.65	0.27	
	Omarama	2000-2018 – 1970-2000	+68.80	0.16	
Total precipitation	Hokitika	2000-2018 – 1970-2000	-16.31	0.82	
	Nelson	2000-2018 – 1970-2000	+8.35	0.90	
	Christchurch	2000-2018 – 1970-2000	+0.96	0.99	
	Omarama	2000-2018 – 1970-2000	-25.91	0.83	
Maximum temperature	Hokitika	2000-2018 – 1970-2000	+0.13	0.42	
	Nelson	2000-2018 – 1970-2000	+0.53	0.0077	**
	Christchurch	2000-2018 – 1970-2000	+0.58	0.0070	**
	Omarama	2000-2018 – 1970-2000	+0.37	0.17	
Mean temperature	Hokitika	2000-2018 – 1970-2000	+0.23	0.13	
	Nelson	2000-2018 – 1970-2000	+0.69	0.0007	***
	Christchurch	2000-2018 – 1970-2000	+0.20	0.23	
	Omarama	2000-2018 – 1970-2000	+0.28	0.24	
Minimum temperature	Hokitika	2000-2018 – 1970-2000	+0.36	0.031	*
	Nelson	2000-2018 – 1970-2000	+0.84	0.0002	***
	Christchurch	2000-2018 – 1970-2000	-0.17	0.32	
	Omarama	2000-2018 – 1970-2000	+0.19	0.43	

Note: Diff. = difference in predicted means, s.f. = significant figures, *p*-value significance thresholds: * = $p < 0.05$, ** = $p < 0.01$ and *** = $p < 0.001$.

6.3.5 Genomic methodologies implemented in Chapters 3 and 5

Methodologies used for identifying outlier SNPs in Chapter 3 included PCAdapt, BayeScan and KGD-F_{ST}. An alternative F_{ST} methodology for identifying outlier SNPs, OutFLANK, was implemented in Chapter 5. OutFLANK could not be utilised in Chapter 3 due to violation of the assumption of OutFLANK: that no pair of populations be more closely related than any other pair (Whitlock & Lotterhos, 2015). Because of the breeding structure, the two high WSC populations within a pool were more closely related than the two low WSC populations within a pool. When the data were run through OutFLANK, a χ^2 distribution could not be fitted because there were a large number of loci determined to be subject to balancing and diversifying selection, a situation which leads to a left skew and a long right tail for the distribution of *p*-values (variance is large and samples are not close to the mean). This was not the case for loci in the SMD GBS dataset and so OutFLANK could be implemented. BayeScan wasn't used in Chapter 5 and BayeScEnv was used instead. This was because BayeScEnv is extended from the

software BayeScan (Foll & Gaggiotti, 2008) to explicitly investigate local adaptation by incorporating environmental data (de Villemereuil & Gaggiotti, 2015).

The significance thresholds implemented for false discovery rate in Chapter 3 were more stringent than those in Chapter 5. The Bonferroni correction, often viewed as too stringent (Hirschhorn & Daly, 2005; Wang *et al.*, 2005), was used to minimise false positive loci identified by outlier detection methodologies in Chapter 3. A less stringent false discovery rate, such as the Benjamini-Hochberg or *q*-value (as implemented in Chapter 5), is probably as likely to minimise false positives while not reducing the true positives. Based on the functional annotations of the genes identified in the outlier detection methods from Chapter 3, few of them have any known direct role in WSC accumulation. This finding could be because the outlier detection methods did not detect biologically relevant genes associated with SNP loci. Alternatively, some of these genes are of interest but further studies are required before we understand their role in WSC accumulation.

Another observation was made that when phenotypic data were treated as qualitative (i.e. high vs low WSC, in outlier detection methodologies) rather than quantitative (continuous range of WSC values, in GWAS), fewer biologically relevant genes were identified. This could be due to the level of WSC in the assessed populations. The variation in WSC phenotypes was not accounted for in the outlier detection analyses. As observed in Chapter 4, the FNZLL-Low-End did not have a WSC phenotype as low as WNZLL-Low-End which may have impacted the ability to detect significant differences between the FNZLL-High-End and FNZLL-Low-End populations. The Widdup New Zealand small leaf (WNZSL) and FNZLL pools did not exhibit a large reduction in WSC compared to WNZLL, Widdup United States of America large leaf (WUSLL) and FNZSL (**Figure 2.4**, Chapter 2). This was not accounted for in Chapter 3 and may have impacted the outlier detection analyses to detect differences in the allele frequencies for the WNZSL and FNZLL pools. Hence, commonality between just the WNZLL, WUSLL and FNZSL for outlier loci could be investigated.

Linkage disequilibrium (LD) in Chapter 3 was estimated to decay to $r^2 = 0.25$ within 10 Kbp. For Chapter 5, a smaller distance was chosen ($r^2 = 0.25$ within 2 Kbp). This was because LD decay occurs over shorter distances in populations with large effective population sizes (*Ne*) (Hayes *et al.*, 2013). The expectation of r^2 is $\frac{1}{4N_eC+1}$ where *Ne* is the effective population size and *C* is the map distance in cM between loci (Sved, 1971). Hence, when *Ne* is large, r^2 is low. A conservative approach of a more rapid LD

decay was used in Chapter 5 because some degree of interpollination, and potentially migration, occurring in the naturalised populations was assumed. This was in comparison to only 20 – 30 parents placed in a crossing cage and pollinated with each other when developing the WSC populations (Widdup *et al.*, 2010). The theoretical basis of estimating LD assumes that the population is in Hardy-Weinberg equilibrium (HWE). Three of the assumptions of HWE were violated in the WSC populations (no selection, infinite population size and random mating). Selection had taken place on these populations, there were a relatively small number of ancestors and population structure was present due to non-random mating. Therefore, to more accurately estimate global LD, HWE filtering, a feature of KGD software (Dodds *et al.*, 2015) and easily implementable with GBS data, could have taken place in parallel to the outlier detection and GWAS analyses (**Figure 6.4**) as an alternative to using the SNPs filtered for the outlier detection and GWAS. This may have provided a more accurate estimate of global LD, however, until the scaffold assembly of white clover pseudomolecules are refined, accurate estimates of LD cannot be performed.

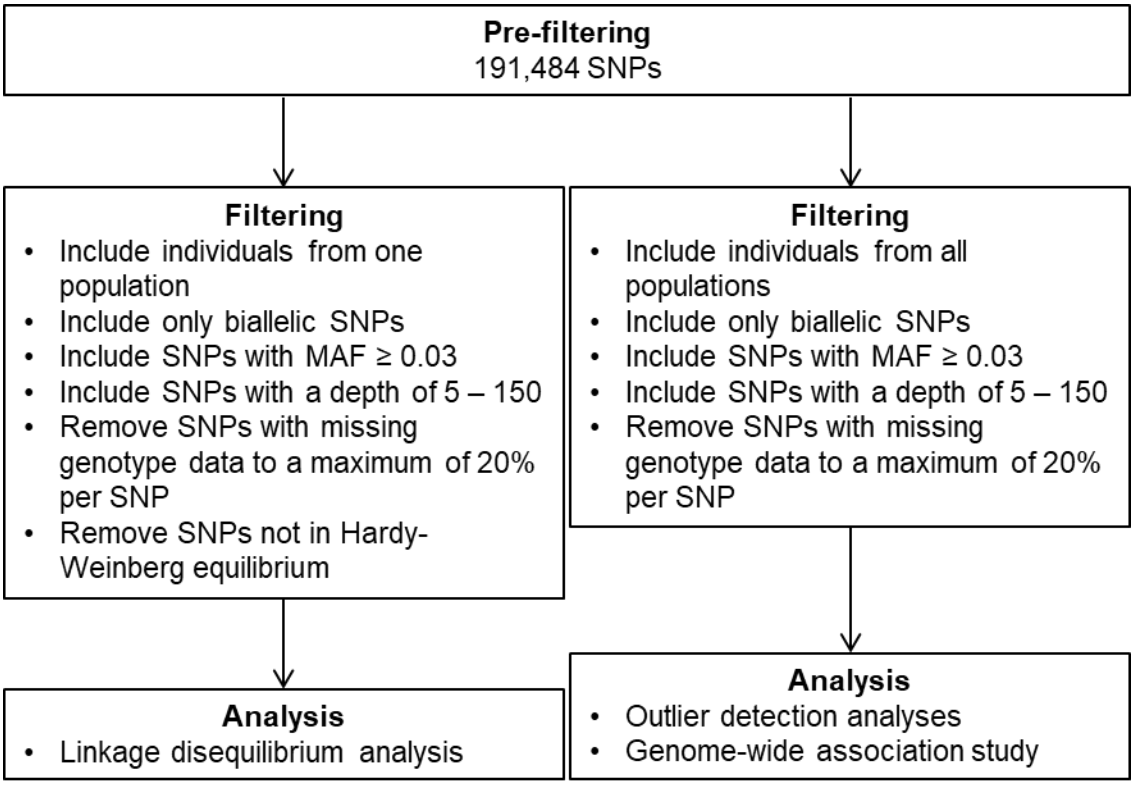


Figure 6.4 Workflow of SNP filtering to calculate linkage disequilibrium for one white clover population compared to original filtering methodology.
Note: SNPs = single nucleotide polymorphisms and MAF = minor allele frequency.

6.4 Potential impact to industry

Progress has been made in the current study to identify gene targets that impact levels of WSC in white clover. A large number of putative trait-associated SNPs were identified from outlier analyses; however, transcriptome and proteome analyses suggest that the genetic control of this trait is influenced by a small number of genes. Increasing WSC content in white clover leaves appears to be strongly affected by regulation of genes involved in the starch biosynthetic and degradation pathways. This is relevant to the forage breeding industry because it means that breeding for increased WSC might be made more efficient through leveraging specific SNP variants in carbohydrate metabolism genes such as *AMY* (chr8.jg3312.t1), *glgC* (chr5.jg7106.t1) and *WAXY* (chr3.jg7615.t1), for marker-assisted breeding for elevated foliar WSC.

These findings suggest that future work should seek to investigate allelic variants of both trans- and cis-acting carbohydrate metabolism genes, understand how broadly effective the gene variants are, i.e., are they effective in different genetic backgrounds, and their contribution to altering flux through the starch biosynthesis and degradation pathway. A further consideration will also be the potential role of epigenetics in influencing WSC phenotypes. This has not been investigated in the present work. The phenomenon of epigenetic regulation affects plant development and growth, and is of increasing interest to plant breeders (Gallusci *et al.*, 2017; Banta & Richards, 2018). Such further studies can be expected to create knowledge that will allow manipulation of WSC levels through targeted breeding programmes. These would be significant for reducing the environmental impact of the pastoral farming industry in New Zealand.

The genetic control of SMD tolerance in white clover requires further investigation before its potential impact on industry can be ascertained. The results from the current study suggest future investigation into carbohydrates metabolism and root system characteristic genes is warranted to elucidate the control of SMD locally adapted naturalised populations. Further studies are required to validate and specify gene targets for future breeding programmes.

REFERENCES

Abebe, T. D., Naz, A. A., & Léon, J. (2015). Landscape genomics reveal signatures of local adaptation in barley (*Hordeum vulgare* L.). *Frontiers in Plant Science, 6*(813).

Abernethy, G. A., & McManus, M. T. (1999). Tissue-specific Changes in the Pattern of Ubiquitin Conjugation of Leaf Proteins in *Festuca novae-zelandiae* in Response to a Water Deficit. *Journal of Plant Physiology, 154*(3), 404-407.

Acquaah, G. (2012). *Principles of Plant Genetics and Breeding* (2nd ed.). UK: Wiley-Blackwell.

Ahmad, M., & Uniyal, S. K. (2016). Is Recurving an Effective Strategy of *Trifolium repens* L. to Augment Reproduction? *Scientifica, 2016*, 4.

Albrechtsen, A., Nielsen, F. C., & Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular biology and evolution, 27*(11), 2534-2547.

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics, 12*(1), 246.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research, 19*(9), 1655-1664.

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-e0208614.

Alipour, H., Bihamta, M. R., Mohammadi, V., Peyghambari, S. A., Bai, G., & Zhang, G. (2017). Genotyping-by-Sequencing (GBS) Revealed Molecular Genetic Diversity of Iranian Wheat Landraces and Cultivars. *Frontiers in Plant Science, 8*, 1293-1293.

Alomar, D., Fuchslocher, R., Cuevas, J., Mardones, R., & Cuevas, E. (2009). Prediction of the composition of fresh pastures by near infrared reflectance or interactance-reflectance spectroscopy. *Chilean Journal of Agricultural Research, 69*(2), 198-206.

Amos, W., Driscoll, E., & Hoffman, J. I. (2011). Candidate genes versus genome-wide associations: which are better for detecting genetic susceptibility to infectious disease? *Proceedings of the Royal Society B: Biological Sciences, 278*(1709), 1183-1188.

Anderson, C. B., Franzmayr, B. K., Hong, S. W., Larking, A. C., van Stijn, T. C., Tan, R., . . . Griffiths, A. G. (2018). Protocol: a versatile, inexpensive, high-throughput plant genomic DNA extraction method suitable for genotyping-by-sequencing. *Plant Methods, 14*(1), 75.

Anderson, J. T., Panetta, A. M., & Mitchell-Olds, T. (2012). Evolutionary and Ecological Responses to Anthropogenic Climate Change. *Plant Physiology, 160*(4), 1728.

Anderson, J. T., Willis, J. H., & Mitchell-Olds, T. (2011). Evolutionary genetics of plant adaptation. *Trends in genetics : TIG, 27*(7), 258-266.

Andrew, R. L., Bernatchez, L., Bonin, A., Buerkle, C. A., Carstens, B. C., Emerson, B. C., . . . Rieseberg, L. H. (2013). A road map for molecular ecology. *Molecular Ecology, 22*(10), 2605-2626.

Annicchiarico, P., Barrett, B., Brummer, E. C., Julier, B., & Marshall, A. H. (2014). Achievements and Challenges in Improving Temperate Perennial Forage Legumes. *Critical Reviews in Plant Sciences, 34*(1-3), 327-380.

Annicchiarico, P., & Carelli, M. (2014). Origin of Ladino White Clover as Inferred from Patterns of Molecular and Morphophysiological Diversity. *Crop Science, 54*(6), 2696-2706.

Annicchiarico, P., & Piano, E. (1995). Variation within and among Ladino white clover ecotypes for agronomic traits. *Euphytica, 86*(2), 135-142.

Annicchiarico, P., & Piano, E. (2004). Indirect Selection for Root Development of White Clover and Implications for Drought Tolerance. *Journal of Agronomy and Crop Science, 190*(1), 28-34.

Arojju, S. K., Barth, S., Milbourne, D., Conaghan, P., Velmurugan, J., Hodkinson, T. R., & Byrne, S. L. (2016). Markers associated with heading and aftermath heading in perennial ryegrass full-sib families. *BMC Plant Biology, 16*(1), 160.

Ashby, R. L. (2019). *The development and implementation of genomic tools for the New Zealand Greenshell™ Mussel industry.* Doctor of Philosophy, University of Otago, Dunedin, New Zealand.

Ashraf, M., & Foolad, M. R. (2007). Roles of glycine betaine and proline in improving plant abiotic stress resistance. *Environmental and Experimental Botany, 59*(2), 206-216.

Atwood, S. S. (1940). Genetics of cross-incompatibility among self-incompatible plants of *Trifolium repens*. *Journal of the American Society of Agronomy, 32*, 955-968.

Australian Government. (2008). The biology of *Trifolium repens* L. (white clover).

Auzanneau, J., Huyghe, C., Julier, B., & Barre, P. (2007). Linkage disequilibrium in synthetic varieties of perennial ryegrass. *Theoretical and Applied Genetics, 115*(6), 837-847.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., . . . Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE, 3*(10), e3376.

Baker, C. G. J. (1997). *Industrial drying of foods*. US: Springer.

Baker, C. W., Givens, D. I., & Deaville, E. R. (1994). Prediction of organic matter digestibility in vivo of grass silage by near infrared reflectance spectroscopy: effect of calibration method, residual moisture and particle size. *Animal Feed Science and Technology, 50*(1), 17-26.

Ballicora, M. A., Iglesias, A. A., & Preiss, J. (2004). ADP-Glucose Pyrophosphorylase: A Regulatory Enzyme for Plant Starch Synthesis. *Photosynthesis Research, 79*(1), 1-24.

Ballizany, W. L., Hofmann, R. W., Jahufer, M. Z. Z., & Barrett, B. A. (2012). Multivariate associations of flavonoid and biomass accumulation in white clover (*Trifolium repens*) under drought. *Functional Plant Biology, 39*(2), 167-177.

Balloux, F., & Lugon-Moulin, N. (2002). The estimation of population differentiation with microsatellite markers. *Molecular Ecology, 11*(2), 155-165.

Banta, J. A., & Richards, C. L. (2018). Quantitative epigenetics and evolution. *Heredity, 121*(3), 210-224.

Barrett, B. A., Baird, I. J., & Woodfield, D. R. (2005). A QTL analysis of white clover seed production. *Crop Science, 45*(5), 1844-1850.

Barrett, B. A., Faville, M. J., Nichols, S. N., Simpson, W. R., Bryan, G. T., & Conner, A. J. (2015). Breaking through the feed barrier: options for improving forage genetics. *Animal Production Science, 55*(7), 883-892.

Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution, 23*(1), 38-44.

Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology, 13*(4), 969-980.

Beever, D. E. (1996). Meeting the protein requirements of ruminant livestock. *South African Journal of Animal Science, 26*, 20-26.

Bekele, W. A., Wight, C. P., Chao, S., Howarth, C. J., & Tinker, N. A. (2018). Haplotype-based genotyping-by-sequencing in oat genome research. *Plant biotechnology journal, 16*(8), 1452-1463.

Bell, G., & Gonzalez, A. (2009). Evolutionary rescue can prevent extinction following environmental change. *Ecology Letters, 12*(9), 942-948.

Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R., & Munoz, P. (2019). How can a high-quality genome assembly help plant breeders? *GigaScience, 8*(6).

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300.

Bennett, M. D., & Leitch, I. J. (2011). Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Annals of Botany, 107*(3), 467-590.

Berardo, N. (1997). Prediction of the chemical composition of white clover by near-infrared reflectance spectroscopy. *Grass and Forage Science, 52*, 27-32.

Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S., & Yockteng, R. (2017). Genetic diversity and association mapping in the Colombian Central Collection of *Solanum tuberosum* L. Andigenum group using SNPs markers. *PLOS ONE, 12*(3), e0173039.

Berthouly-Salazar, C., Thuillet, A.-C., Rhoné, B., Mariac, C., Ousseini, I. S., Couderc, M., . . . Vigouroux, Y. (2016). Genome scan reveals selection acting on genes linked to stress response in wild pearl millet. *Molecular Ecology, 25*(21), 5500-5512.

Biazzi, E., Nazzicari, N., Pecetti, L., Brummer, E. C., Palmonari, A., Tava, A., & Annicchiarico, P. (2017). Genome-Wide Association Mapping and Genomic Selection for Alfalfa (*Medicago sativa*) Forage Quality Traits. *PLOS ONE, 12*(1), e0169234.

Bilton, T., McEwan, J., Clarke, S., Brauning, R., van Stijn, T., Rowe, S., & Dodds, K. (2018). Linkage Disequilibrium Estimation in Low Coverage High-Throughput Sequencing Data. *Genetics, 209*(2), 389-400.

Bock, C. H., Poole, G. H., Parker, P. E., & Gottwald, T. R. (2010). Plant Disease Severity Estimated Visually, by Digital Photography and Image Analysis, and by Hyperspectral Imaging. *Critical Reviews in Plant Sciences, 29*(2), 59-107.

Bock, D. G., Kantar, M. B., Caseys, C., Matthey-Doret, R., & Rieseberg, L. H. (2018). Evolution of invasiveness by genetic accommodation. *Nature Ecology & Evolution, 2*(6), 991-999.

Bodner, G., Alsalem, M., Nakhforoosh, A., Arnold, T., & Leitner, D. (2017). RGB and Spectral Root Imaging for Plant Phenotyping and Physiological Research:

Experimental Setup and Imaging Protocols. *Journal of visualized experiments : JoVE*(126), 56251.

Bolens, L. (1981). Agronomes andalous du moyen-âge. Geneva: Librairie Droz.

Boller, B. C., & Nösberger, J. (1983). Effects of temperature and photoperiod on stolon characteristics, dry matter partitioning, and nonstructural carbohydrate concentration of two white clover ecotypes. *Crop Science, 23*(6), 1057-1062.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B, 26*, 211-252.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics, 23*(19), 2633-2635.

Branca, A., Paape, T. D., Zhou, P., Briskine, R., Farmer, A. D., Mudge, J., . . . Tiffin, P. (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proceedings of the National Academy of Sciences, 108*(42), E864.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., & Stephan, W. (1995). The Hitchhiking Effect on the Site Frequency Spectrum of DNA Polymorphisms. *Genetics, 140*(2), 783-796.

Brazauskas, G., Lenk, I., Pedersen, M. G., Studer, B., & Lübberstedt, T. (2011). Genetic variation, population structure, and linkage disequilibrium in European elite germplasm of perennial ryegrass. *Plant Science, 181*(4), 412-420.

Breseghello, F., & Sorrells, M. E. (2006). Association Analysis as a Strategy for Improvement of Quantitative Traits in Plants. *Crop Science, 46*(3), 1323-1330.

Brewbaker, J. L. (1955). V-leaf markings of white clover. *Journal of heredity, 46*(3), 115-123.

Brock, J. L., Albrecht, K. A., Tilbrook, J. C., & Hay, M. J. M. (2000). Morphology of white clover during development from seed to clonal populations in grazed pastures. *Journal of Agricultural Science, 135*(2), 103-111.

Brock, J. L., & Hay, M. J. M. (1996). A review of the role of grazing management on the growth and performance of white clover cultivars in lowland New Zealand pastures. *Special Publication-Agronomy Society of New Zealand*, 65-70.

Brock, J. L., & Hay, M. J. M. (2001). White clover performance in sown pastures: A biological/ecological perspective. *Proceedings of the New Zealand Grassland Association, 63*, 73-83.

Bryant, D., & Moulton, V. (2004). Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution, 21*(2), 255-265.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods, 12*(1), 59-60.

Burgess, D., Penton, A., Dunsmuir, P., & Dooner, H. (1997). Molecular cloning and characterization of ADP-glucose pyrophosphorylase cDNA clones isolated from pea cotyledons. *Plant Molecular Biology, 33*(3), 431-444.

Büttner, M. (2007). The monosaccharide transporter(-like) gene family in *Arabidopsis*. *FEBS Letters, 581*(12), 2318-2324.

Cahn, M. G., & Harper, J. L. (1976). The biology of the leaf mark polymorphism in *Trifolium repens* L. 1. Distribution of phenotypes at a local scale. *Heredity, 37*, 309-325.

Calenge, F., Saliba-Colombani, V., Mahieu, S., Loudet, O., Daniel-Vedele, F., & Krapp, A. (2006). Natural Variation for Carbohydrate Content in *Arabidopsis*. Interaction with Complex Traits Dissected by Quantitative Genetics. *Plant Physiology, 141*(4), 1630-1643.

Cañas, R. A., Yesbergenova-Cuny, Z., Simons, M., Chardon, F., Armengaud, P., Quilleré, I., . . . Hirel, B. (2017). Exploiting the Genetic Diversity of Maize Using a Combined Metabolomic, Enzyme Activity Profiling, and Metabolic Modeling Approach to Link Leaf Physiology to Kernel Yield. *The Plant Cell, 29*(5), 919.

Caradus, J. R. (1986). World checklist of white clover varieties. *New Zealand Journal of Experimental Agriculture, 14*(2), 119-164.

Caradus, J. R., Clifford, P. T. P., Chapman, D. F., Cousins, G. R., Williams, W. M., & Miller, J. E. (1997). Breeding and description of 'Grasslands Sustain', a medium-large-leaved white clover (*Trifolium repens* L.) cultivar. *New Zealand Journal of Agricultural Research, 40*(1), 1-7.

Caradus, J. R., Forde, M. B., Wewala, S., & Mackay, A. C. (1990). Description and classification of a white clover (*Trifolium repens* L.) germplasm collection from southwest Europe. *New Zealand Journal of Agricultural Research, 33*(3), 367-375.

Caradus, J. R., Hay, R. J. M., & Woodfield, D. R. (1996). Positioning of white clover cultivars in New Zealand. *Agronomy Society of New Zealand Special Publication No. 11 and Grassland Research and Practice Series No. 6* 45-49.

Caradus, J. R., McNabb, W., Woodfield, D. R., Waghorn, G. C., & Keogh, R. (1995). Improving quality characteristics of white clover. In J. G. Hampton & K. M. Pollock (Eds.), *Agronomy Society of New Zealand - Proceedings, Twenty-Fifth Annual Conference 1995/96* (Vol. 25, pp. 7-12). Lincoln Canterbury: Agronomy Soc New Zealand Inc.

Caradus, J. R., & Woodfield, D. R. (1997). Review: World checklist of white clover varieties II. *New Zealand Journal of Agricultural Research, 40*(2), 115-206.

Caradus, J. R., & Woodfield, D. R. (1998). Genetic control of adaptive root characteristics in white clover. *Plant and Soil, 200*(1), 63-69.

Caradus, J. R., Woodfield, D. R., & Stewart, A. V. (1996). *Overview and vision for white clover*. Christchurch: Agronomy Society of New Zealand.

Carciofi, M., Blennow, A., Jensen, S. L., Shaik, S. S., Henriksen, A., Buléon, A., . . . Hebelstrup, K. H. (2012). Concerted suppression of all starch branching enzyme genes in barley produces amylose-only starch granules. *BMC Plant Biology, 12*(1), 223.

Carnahan, H., Hill, H. D., Hanson, A., & Brown, K. (1955). Inheritance and frequencies of leaf markings in white clover. *Journal of heredity, 46*(3), 109-114.

Casler, M. D. (1988). Performance of orchardgrass, smooth bromegrass and ryegrass in binary mixtures with alfalfa. *Agronomy Journal, 80*, 509-514.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes|Genomes|Genetics, 1*(3), 171.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *American journal of human genetics, 19*(3 Pt 1), 233-257.

Caye, K., Jumentier, B., Lepeule, J., & François, O. (2019). LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular Biology and Evolution, 36*(4), 852-860.

Chapman, D. F., Parsons, A. J., & Schwinning, S. (1996). Management of clover in grazed pastures: expectations, limitations and opportunities. *Special Publication-Agronomy Society of New Zealand*, 55-64.

Chardon, F., Bedu, M., Calenge, F., Klemens, Patrick A. W., Spinner, L., Clement, G., . . . Krapp, A. (2013). Leaf Fructose Content Is Controlled by the Vacuolar Transporter SWEET17 in *Arabidopsis*. *Current Biology, 23*(8), 697-702.

Charlton, J. F. L., & Stewart, A. V. (1999). Pasture species and cultivars used in New Zealand - a list. *Proceedings of the New Zealand Grassland Association, 61*, 147-166.

Chen, H., & Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics, 12*(1), 35.

Chen, L.-Q., Hou, B.-H., Lalonde, S., Takanaga, H., Hartung, M. L., Qu, X.-Q., . . . Frommer, W. B. (2010). Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature, 468*(7323), 527-532.

Chen, L.-Q., Qu, X.-Q., Hou, B.-H., Sosso, D., Osorio, S., Fernie, A. R., & Frommer, W. B. (2012). Sucrose Efflux Mediated by SWEET Proteins as a Key Step for Phloem Transport. *Science, 335*(6065), 207.

Chen, T. H. H., & Murata, N. (2002). Enhancement of tolerance of abiotic stress by metabolic engineering of betaines and other compatible solutes. *Current Opinion in Plant Biology, 5*(3), 250-257.

Chessel, D., Dufour, A. B., & Thioulouse, J. (2004). The *ade4* Package - I: One-Table Methods. *R news, 4*(1), 5-10.

Chiou, T. J., & Bush, D. R. (1996). Molecular Cloning, Immunochemical Localization to the Vacuole, and Expression in Transgenic Yeast and Tobacco of a Putative Sugar Transporter from Sugar Beet. *Plant Physiology, 110*(2), 511-520.

Cirilli, M., Bassi, D., & Ciacciulli, A. (2016). Sugars in peach fruit: a breeding perspective. *Horticulture Research, 3*, 15067.

Clark, A., Mullan, B., & Porteous, A. (2011). Scenarios of Regional Drought under Climate Change. Wellington, New Zealand: NIWA.

Clark, D. A., Mathew, C., & Crush, J. R. (2001). More feed for New Zealand dairy systems. *Proceedings of the New Zealand Grassland Association, 63*, 283-288.

Cogan, N. O. I., Smith, K. F., Yamada, T., Francki, M. G., Vecchies, A. C., Jones, E. S., . . . Forster, J. W. (2005). QTL analysis and comparative genomics of herbage quality traits in perennial ryegrass (*Lolium perenne* L.). *Theoretical and Applied Genetics, 110*(2), 364-380.

Collins, R. (2002). *The effects of drought stress and winter stress on the persistence of white clover*. Paper presented at the Lowland Grasslands of Europe: Utilization and Development, FAO, Rome.

Collins, R. P., Helgadóttir, Á., Frankow-Lindberg, B. E., Skøt, L., Jones, C., & Skøt, K. P. (2012). Temporal changes in population genetic diversity and structure in red and white clover grown in three contrasting environments in northern Europe. *Annals of botany, 110*(6), 1341-1350.

Cong, B., Liu, J., & Tanksley, S. D. (2002). Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations. *Proceedings of the National Academy of Sciences, 99*(21), 13606.

Cope, A. L., O'Meara, B. C., & Gilchrist, M. A. (2020). Gene expression of functionally-related genes coevolves across fungal species: detecting coevolution of gene expression using phylogenetic comparative methods. *BMC genomics, 21*(1), 370.

Corander, J., Waldmann, P., & Sillanpää, M. J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics, 163*(1), 367-374.

Corson, D. C., Waghorn, G. C., Ulyatt, M. J., & Lee, J. (1999). NIRS: Forage analysis and livestock feeding. *Proceedings of the New Zealand Grassland Association, 61*, 127-132.

Cosgrove, G. P., Burke, J. L., Death, A. F., Hickey, M. J., Pacheco, D., & Lane, G. A. (2007). Ryegrasses with increased water soluble carbohydrate: evaluating the potential for grazing dairy cows in New Zealand. *Proceedings of the New Zealand Grassland Association, 69*, 179-185.

Cosgrove, G. P., Burke, J. L., Death, A. F., Lane, G. A., Fraser, K., & Pacheco, D. (2006). The effect of clover-rich diets on cows in mid lactation: production, behaviour and nutrient use. *Proceedings of the New Zealand Grassland Association, 68*, 267-273.

Cosgrove, G. P., Koolaard, J., Luo, D., Burke, J. L., & Pacheco, D. (2009). The composition of high sugar ryegrasses. *Proceedings of the New Zealand Grassland Association, 71*, 187-193.

Couchman, J. F. (1959). Storage of hay. I.—Effect of temperature on the 'Soluble' nitrogen, sugar and fat contents. *Journal of the Science of Food and Agriculture, 10*(10), 513-519.

Cousins, G., & Woodfield, D. R. (2006). *Effect of inbreeding on growth of white clover. Proceeding of the 13th Australasian Plant Breeding Conference.* Paper presented at the Breeding for Success: Diversity in Action, Christchurch, New Zealand.

Cox, B., Kislinger, T., & Emili, A. (2005). Integrating gene and protein expression data: pattern analysis and profile mining. *Methods, 35*(3), 303-314.

Crocker, W. (1938). Life-span of seeds. *Botanical Review, 4*, 235-274.

Crush, J. R., Care, D. A., Gourdin, A., & Woodfield, D. R. (2005). Root growth media effects on root morphology and architecture in white clover. *New Zealand Journal of Agricultural Research, 48*(2), 255-263.

Cvijović, I., Good, B. H., & Desai, M. M. (2018). The Effect of Strong Purifying Selection on Genetic Diversity. *Genetics, 209*(4), 1235.

Daday, H. (1958). Gene frequencies in wild populations of *Trifolium repens* L. III. World distribution. *Heredity, 12*(2), 169-184.

Daday, H. (1965). Gene frequencies in wild populations of *Trifolium repens* L. *Heredity, 20*(3), 355-365.

DairyNZ, & LIC. (2018). New Zealand dairy statistics 2017-2018. Hamilton, New Zealand.

Dalmannsdóttir, S., Helgadóttir, Á., & Gudleifsson, B. E. (2001). Fatty Acid and Sugar Content in White Clover in Relation to Frost Tolerance and Ice-encasement Tolerance. *Annals of Botany, 88*(4, Part 2), 753-759.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Group, G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158.

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., & Vigouroux, Y. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology, 22*(5), 1383-1399.

de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., & Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Molecular bioSystems, 5*(12), 1512-1526.

de Villemereuil, P., & Gaggiotti, O. E. (2015). A new F_{ST}-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution, 6*(11), 1248-1258.

Deinum, B., & Maassen, A. (1994). Effects of drying temperature on chemical composition and in vitro digestibility of forages. *Animal Feed Science and Technology, 46*, 75-86.

Deokar, A., Sagi, M., & Tar'an, B. (2019). Genome-wide SNP discovery for development of high-density genetic map and QTL mapping of ascochyta blight resistance in chickpea (*Cicer arietinum* L.). *Theoretical and Applied Genetics, 132*(6), 1861-1872.

Devlin, B., & Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics, 55*(4), 997-1004.

Dewhurst, R. J., Fisher, W. J., Tweed, J. K. S., & Wilkins, R. J. (2003). Comparison of grass and legume silages for milk production. 1. Production responses with different levels of concentrate. *Journal of Dairy Science, 86*(8), 2598-2611.

Dodds, K. G., McEwan, J. C., Brauning, R., Anderson, R. M., van Stijn, T. C., Kristjánsson, T., & Clarke, S. M. (2015). Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics, 16*(1), 1047.

Doebley, J., Stec, A., & Gustus, C. (1995). Teosinte Branched1 and the Origin of Maize: Evidence for Epistasis and the Evolution of Dominance. *Genetics, 141*(1), 333-346.

Doebley, J., Stec, A., & Hubbard, L. (1997). The evolution of apical dominance in maize. *Nature, 386*, 485.

Doi, K., Izawa, T., Fuse, T., Yamanouchi, U., Kubo, T., Shimatani, Z., . . . Yoshimura, A. (2004). Ehd1, a B-type response regulator in rice, confers short-day promotion of flowering and controls FT-like gene expression independently of Hd1. *Genes & Development, 18*(8), 926-936.

Donohue, K. (2014). The epigenetics of adaptation: Focusing on epigenetic stability as an evolving trait. *Evolution, 68*(3), 617-619.

Dracatos, P. M., Cogan, N. O. I., Sawbridge, T. I., Gendall, A. R., Smith, K. F., Spangenberg, G. C., & Forster, J. W. (2009). Molecular characterisation and genetic mapping of candidate genes for qualitative disease resistance in perennial ryegrass (*Lolium perenne* L.). *BMC Plant Biology, 9*(1), 62.

Dreccer, M. F., Barnes, L. R., & Meder, R. (2014). Quantitative dynamics of stem water soluble carbohydrates in wheat can be monitored in the field using hyperspectral reflectance. *Field Crops Research, 159*, 70-80.

Duchemin, W., Dupont, P.-Y., Campbell, M., Ganley, A., & Cox, M. (2015). HyLiTE: Accurate and flexible analysis of gene expression in hybrid and allopolyploid species. *BMC bioinformatics, 16*, 8.

Duforet-Frebourg, N., Bazin, E., & Blum, M. G. B. (2014). Genome Scans for Detecting Footprints of Local Adaptation Using a Bayesian Factor Model. *Molecular Biology and Evolution, 31*(9), 2483-2495.

Durgo, H., Klement, E., Hunyadi-Gulyas, E., Szucs, A., Kereszt, A., Medzihradszky, K. F., & Kondorosi, E. (2015). Identification of nodule-specific cysteine-rich plant peptides in endosymbiotic bacteria. *PROTEOMICS, 15*(13), 2291-2295.

Easton, H. S., Stewart, A. V., Lyons, T. B., Parris, M., & Charrier, S. (2009). Soluble carbohydrate content of ryegrass cultivars. *Proceedings of the New Zealand Grassland Association, 71*, 161-166.

Eckert, A. J., Bower, A. D., González-Martínez, S. C., Wegrzyn, J. L., Coop, G., & Neale, D. B. (2010a). Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology, 19*(17), 3789-3805.

Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez, S. C., & Neale, D. B. (2010b). Patterns of Population Structure and Environmental Associations to Aridity Across the Range of Loblolly Pine (*Pinus taeda* L., Pinaceae). *Genetics, 185*(3), 969.

Edwards, G. R., Parsons, A. J., Rasmussen, S., & Bryant, R. H. (2007). High sugar ryegrasses for livestock systems in New Zealand. *Proceedings of the New Zealand Grassland Association, 69*, 161-171.

El-Din El-Assal, S., Alonso-Blanco, C., Peeters, A. J. M., Raz, V., & Koornneef, M. (2001). A QTL for flowering time in *Arabidopsis* reveals a novel allele of CRY2. *Nature Genetics, 29*, 435.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution, 29*(1), 51-63.

Ellis, W., & Young, N. R. (1967). The characteristics of European, Mediterranean and other populations of white clover (*Trifolium repens* L.). *Euphytica, 16*(3), 330-340.

Ellison, N. W., Liston, A., Steiner, J. J., Williams, W. M., & Taylor, N. L. (2006). Molecular phylogenetics of the clover genus (*Trifolium* - Leguminosae). *Molecular Phylogenetics and Evolution, 39*(3), 688-705.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE, 6*(5).

Eltaher, S., Sallam, A., Belamkar, V., Emara, H. A., Nower, A. A., Salem, K. F. M., . . . Baenziger, P. S. (2018). Genetic Diversity and Population Structure of F(3:6) Nebraska Winter Wheat Genotypes Using Genotyping-By-Sequencing. *Frontiers in genetics, 9*, 76-76.

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome, 4*(3), 250-255.

Endler, A., Meyer, S., Schelbert, S., Schneider, T., Weschke, W., Peters, S. W., . . . Schmidt, U. G. (2006). Identification of a Vacuolar Sucrose Transporter in Barley and *Arabidopsis* Mesophyll Cells by a Tonoplast Proteomic Approach. *Plant Physiology, 141*(1), 196-207.

Erith, A. G. (1924). *White clover (Trifolium repens L.). A monograph.* London: Duckworth.

Ewing, G., Hermisson, J., Pfaffelhuber, P., & Rudolf, J. (2011). Selective sweeps for recessive alleles and for other modes of dominance. *Journal of Mathematical Biology, 63*(3), 399-431.

Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity, 103*, 285.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics, 131*(2), 479-491.

Fang, C., Luo, J., & Wang, S. (2019). The Diversity of Nutritional Metabolites: Origin, Dissection, and Application in Crop Breeding. *Frontiers in Plant Science, 10*(1028).

Faville, M. J., Cao, M., Schmidt, J., Ryan, D. L., Ganesh, S., Jahufer, M. Z. Z., . . . Barrett, B. A. (2020a). Divergent Genomic Selection for Herbage Accumulation and Days-To-Heading in Perennial Ryegrass. *Agronomy, 10*(3), 340.

Faville, M. J., Ganesh, S., Cao, M., Jahufer, M. Z. Z., Bilton, T. P., Easton, H. S., . . . Barrett, B. A. (2018). Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theoretical and Applied Genetics, 131*(3), 703-720.

Faville, M. J., Griffiths, A. G., Baten, A., Cao, M., Ashby, R. L., Ghamkhar, K., . . . Webber, Z. (2020b). Genomic assessment of white clover and perennial ryegrass genetic resources. *Journal of New Zealand Grasslands, 82*, 27-34.

Faville, M. J., Griffiths, A. G., Jahufer, M. Z. Z., & Barrett, B. A. (2012). Progress towards marker-assisted selection in forages. *Proceedings of the New Zealand Grassland Association, 74*, 189-194.

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics, 155*(3), 1405.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology, 37*(12), 4302-4315.

Ficklin, S. P., Dunwoodie, L. J., Poehlman, W. L., Watson, C., Roche, K. E., & Feltus, F. A. (2017). Discovering Condition-Specific Gene Co-Expression Patterns Using Gaussian Mixture Models: A Cancer Case Study. *Scientific Reports, 7*(1), 8617.

Fijarczyk, A., & Babik, W. (2015). Detecting balancing selection in genomes: limits and prospects. *Molecular Ecology, 24*(14), 3529-3545.

Filippou, P., Bouchagier, P., Skotti, E., & Fotopoulos, V. (2014). Proline and reactive oxygen/nitrogen species metabolism is involved in the tolerant response of the invasive plant species Ailanthus altissima to drought and salinity. *Environmental and Experimental Botany, 97*, 1-10.

Finch, S., & Percival, D. (2017). White Clovers. Retrieved March 2017, from http://www.specseed.co.nz/portfolio-item/clover/

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S. (2003). Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology, 54*(1), 357-374.

Flint-Garcia, S. A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., . . . Buckler, E. S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal, 44*(6), 1054-1064.

Foll, M., & Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics, 180*(2), 977-993.

Frame, D. J., Rosier, S. M., Noy, I., Harrington, L. J., Carey-Smith, T., Sparrow, S. N., . . . Dean, S. M. (2020). Climate change attribution and the economic costs of extreme weather events: a study on damages from extreme rainfall and drought. *Climatic Change*.

Frame, J. (2003). *Trifolium repens* L. Retrieved March 2017, from http://www.fao.org/ag/AGP/AGPC/doc/Gbase/data/pf000350.htm

François, O., & Durand, E. (2010). Spatially explicit Bayesian clustering models in population genetics. *Molecular Ecology Resources, 10*(5), 773-784.

François, O., Martins, H., Caye, K., & Schoville, Sean D. (2016). Controlling false discoveries in genome scans for selection. *Molecular Ecology, 25*(2), 454-469.

Frary, A., Nesbitt, T. C., Frary, A., Grandillo, S., Knaap, E. v. d., Cong, B., . . . Tanksley, S. D. (2000). *fw2.2*: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size. *Science, 289*(5476), 85.

Frayling, T. M. (2014). Genome-wide association studies: the good, the bad and the ugly. *Clinical medicine (London, England), 14*(4), 428-431.

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution, 6*(8), 925-929.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics, 196*(4), 973.

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution, 30*(7), 1687-1699.

Fridman, E., Carrari, F., Liu, Y.-S., Fernie, A. R., & Zamir, D. (2004). Zooming In on a Quantitative Trait for Tomato Yield Using Interspecific Introgressions. *Science, 305*(5691), 1786.

Fridman, E., Pleban, T., & Zamir, D. (2000). A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proceedings of the National Academy of Sciences of the United States of America, 97*(9), 4718-4723.

Fu, F.-F., & Xue, H.-W. (2010). Coexpression Analysis Identifies Rice Starch Regulator1, a Rice AP2/EREBP Family Transcription Factor, as a Novel Rice Starch Biosynthesis Regulator. *Plant Physiology, 154*(2), 927.

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics, 147*, 915-925.

Futschik, A., & Schlötterer, C. (2010). The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics, 186*(1), 207.

Gaedeke, N., Klein, M., Kolukisaoglu, U., Forestier, C., Müller, A., Ansorge, M., . . . Martinoia, E. (2001). The *Arabidopsis thaliana* ABC transporter AtMRP5 controls root development and stomata movement. *The EMBO journal, 20*(8), 1875-1887.

Gallusci, P., Dai, Z., Génard, M., Gauffretau, A., Leblanc-Fournier, N., Richard-Molard, C., . . . Brunel-Muguet, S. (2017). Epigenetics for Plant Improvement: Current Knowledge and Modeling Avenues. *Trends in Plant Science, 22*(7), 610-623.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., . . . Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology, 22*(14), 3766-3779.

George, J., Dobrowolski, M. P., Jong, E. v. Z. d., Cogan, N. O. I., Smith, K. F., & Forster, J. W. (2006). Assessment of genetic diversity in cultivars of white clover (*Trifolium repens* L.) detected by SSR polymorphisms. *Genome, 49*, 919-930.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., . . . Weissman, J. S. (2003). Global analysis of protein expression in yeast. *Nature, 425*(6959), 737-741.

Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., . . . Lusis, A. J. (2011). Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *PLOS Genetics, 7*(6), e1001393.

Gibon, Y., Usadel, B., Blaesing, O. E., Kamlage, B., Hoehne, M., Trethewey, R., & Stitt, M. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in *Arabidopsis* rosettes. *Genome Biology, 7*(8), 1-23.

Gibson, P. B., & Cullen, N. J. (2015). Synoptic and sub-synoptic circulation effects on wind resource variability – A case study from a coastal terrain setting in New Zealand. *Renewable Energy, 78*, 253-263.

Gibson, P. B., & Hollowell, E. A. (1966). *White clover*. Washington, D.C.: United States Department of Agriculture.

Gibson, S. I. (2005). Control of plant development and gene expression by sugar signaling. *Current Opinion in Plant Biology, 8*(1), 93-102.

Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., . . . Aebersold, R. (2012). Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics, 11*(6), O111.016717.

Gillett, J. M., & Taylor, N. L. (2001). *The World of Clovers*. Ames, Iowa, USA: Iowa State University Press.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLoS ONE, 9*(2), e90346.

Gödde, M., & Conrad, R. (2000). Influence of soil properties on the turnover of nitric oxide and nitrous oxide by nitrification and denitrification at constant temperature and moisture. *Biology and Fertility of Soils, 32*(2), 120-128.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., . . . Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research, 36*(10), 3420-3435.

Goudet, J., & Jombart, T. (2015). hierfstat: Estimation and Tests of Hierarchical F-Statistics. R package version 0.04-22. Retrieved from https://CRAN.R-project.org/package=hierfstat

Grant, P. R., & Grant, B. R. (2002). Unpredictable Evolution in a 30-Year Study of Darwin's Finches. *Science, 296*(5568), 707.

Grant, V. (1981). *Plant speciation*. New York, USA: Columbia University Press.

Greenhill, W. L., Couchman, J. F., & De Freitas, J. (1961). Storage of hay. III.—effect of temperature and moisture on loss of dry matter and changes in composition. *Journal of the Science of Food and Agriculture, 12*(4), 293-297.

Griffiths, A. G., Barrett, B. A., Simon, D., Khan, A. K., Bickerstaff, P., Anderson, C. B., . . . Jones, C. S. (2013). An integrated genetic linkage map for white clover (*Trifolium repens* L.) with alignment to *Medicago*. *BMC Genomics, 14*, 388.

Griffiths, A. G., Moraga, R., Tausen, M., Gupta, V., Bilton, T. P., Campbell, M. A., . . . Andersen, S. U. (2019). Breaking Free: The Genomics of Allopolyploidy-Facilitated Niche Expansion in White Clover. *The Plant Cell, 31*(7), 1466-1487.

Günther, T., & Coop, G. (2013). Robust Identification of Local Adaptation from Allele Frequencies. *Genetics, 195*(1), 205.

Guo, X., Cericola, F., Fè, D., Pedersen, M. G., Lenk, I., Jensen, C. S., . . . Janss, L. L. (2018). Genomic Prediction in Tetraploid Ryegrass Using Allele Frequencies Based on Genotyping by Sequencing. *Frontiers in Plant Science, 9*(1165).

Gustine, D. L., & Huff, D. R. (1999). Genetic Variation within among White Clover Populations from Managed Permanent Patures of the Northeastern USA. *Crop Science, 39*, 524-530.

Gustine, D. L., & Sanderson, M. A. (2001). Quantifying Spatial and Temporal Genotypic Changes in White Clover Populations by RAPD Technology. *Crop Science, 41*(1), 143-148.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*, 149-161.

Hamrick, J. L., & Godt, M. J. W. (1996). Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 351*(1345), 1291-1298.

Hansen, K. D., Langmead, B., & Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology, 13*(10), R83.

Harberd, D. J. (1963). Observations on natural clones of *Trifolium repens* L. *New Phytologist, 62*(2), 198-204.

Harr, B., Kauer, M., & Schlötterer, C. (2002). Hitchhiking mapping: A population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences, 99*(20), 12949-12954.

Harris, S. L., Auldist, M. J., Clark, D. A., & Jansen, E. B. L. (1998). Effects of white clover content in the diet on herbage intake, milk production and milk composition of New Zealand dairy cows housed indoors. *Journal of Dairy Research, 65*(03), 389-400.

Harris, S. L., Clark, D. A., Auldist, M. J., Waugh, C. D., & Laboyrie, P. G. (1997). *Optimum white clover content for dairy pastures.* Paper presented at the Proceedings of the Conference New Zealand Grassland Association.

Harris, S. L., Clark, D. A., Waugh, C. D., & Clarkson, F. H. (1996). Nitrogen fertiliser effects on white clover in dairy pastures. *Agronomy Society of New Zealand Special Publication No. 11 and Grassland Research and Practice Series No. 6*, 119-124.

Hartl, D. L., & Clark, A. G. (2007). *Principles of Population Genetics.* (4th ed.). Sutherland: Sinauer.

Hayes, B. J., Cogan, N. O. I., Pembleton, L. W., Goddard, M. E., Wang, J., Spangenberg, G. C., & Forster, J. W. (2013). Prospects for genomic selection in forage plant species. *Plant Breeding, 132*(2), 133-143.

Hermisson, J. (2009). Who believes in whole-genome scans for selection? *Heredity, 103*(4), 283-284.

Hermisson, J., & Pennings, P. S. (2005). Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics, 169*(4), 2335-2352.

Hernández, H. G., Tse, M. Y., Pang, S. C., Arboleda, H., & Forero, D. A. (2013). Optimizing methodologies for PCR-based DNA methylation analysis. *BioTechniques, 55*(4), 181-197.

Heslot, N., Rutkoski, J., Poland, J., Jannink, J.-L., & Sorrells, M. E. (2013). Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PloS one, 8*(9), e74612-e74612.

Hetta, M., Mussadiq, Z., Wallsten, J., Halling, M., Swensson, C., & Geladi, P. (2017). Prediction of nutritive values, morphology and agronomic characteristics in forage maize using two applications of NIRS spectrometry. *Acta Agriculturae Scandinavica, Section B — Soil & Plant Science, 67*(4), 326-333.

Higgs, R. J., Cosgrove, G. P., Burke, J. L., Lane, G. A., Pacheco, D., Fraser, K., . . . Ford, J. L. (2010). *Effect of white clover containing either high or low concentrations of water-soluble carbohydrate on metabolic indicators of protein degradation in the rumen of dairy cows* (Vol. 70): New Zealand Society of Animal Production.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet, 38*(6), 226-231.

Hintze, J. L. (2008). *Power analysis and sample size system (PASS) for windows User's Guide I*. Kaysville, Utah, USA: NCSS.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics, 6*(2), 95-108.

Hofmann, R. W., & Jahufer, M. Z. Z. (2011). Tradeoff between Biomass and Flavonoid Accumulation in White Clover Reflects Contrasting Plant Strategies. *PLOS ONE, 6*(4), e18949.

Hofmann, R. W., Swinny, E. E., Bloor, S. J., Markham, K. R., Ryan, K. G., Campbell, B. D., . . . Fountain, D. W. (2000). Responses of Nine *Trifolium repens* L. Populations to Ultraviolet-B Radiation: Differential Flavonol Glycoside Accumulation and Biomass Production. *Annals of Botany, 86*(3), 527-537.

Hoglund, J. H., & Brock, J. L. (1978). Regulation of nitrogen fixation in a grazed pasture. *New Zealand Journal of Agricultural Research, 21*(1), 73-82.

Holderegger, R., Buehler, D., Gugerli, F., & Manel, S. (2010). Landscape genetics of plants. *Trends in Plant Science, 15*(12), 675-683.

Huang, J., Wei, M., Zhang, G., Tan, F., Fang, W., & Huang, J. (2010). Breeding of high yield and sugar sugarcane variety Guitang 23 and its cultivation techniques. *Guangxi Agricultural Sciences, 41*(9), 916-919.

Huang, W., Ratkowsky, D. A., Hui, C., Wang, P., Su, J., & Shi, P. (2019a). Leaf Fresh Weight Versus Dry Weight: Which is Better for Describing the Scaling

Relationship between Leaf Biomass and Leaf Area for Broad-Leaved Plants? *Forests, 10*(3), 256.

Huang, W., Su, X., Ratkowsky, D. A., Niklas, K. J., Gielis, J., & Shi, P. (2019b). The scaling relationships of leaf biomass vs. leaf surface area of 12 bamboo species. *Global Ecology and Conservation, 20*, e00793.

Hudson, R. R., Kreitman, M., & Aguadé, M. (1987). A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics, 116*(1), 153.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., . . . Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research, 47*(D1), D309-D314.

Humphreys, M. O. (1989). Water-soluble carbohydrates in perennial ryegrass breeding. III. Relationships with herbage production, digestibility and crude protein content. *Grass and Forage Science, 44*(4), 423-430.

Inostroza, L., Bhakta, M., Acuña, H., Vásquez, C., Ibáñez, J., Tapia, G., . . . Munoz, P. (2018). Understanding the Complexity of Cold Tolerance in White Clover using Temperature Gradient Locations and a GWAS Approach. *The Plant Genome, 11*(3).

Inostroza, L., Lobos, I., Acuna, H., Vasquez, C., Tapia, G., & Monzon, G. (2017). NIR-Prediction of water-soluble carbohydrate in white clover and its genetic relationship with cold tolerance. *Chilean Journal of Agricultural Research, 77*(3), 218-225.

Intergovernmental Panel on Climate, C. (2014). *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Jafari, A., Connolly, V., Frolich, A., & Walsh, E. J. (2003). A note on estimation of quality parameters in perennial ryegrass by near infrared reflectance spectroscopy. *Irish Journal of Agricultural and Food Research, 42*(2), 293-299.

Jahufer, M. Z. Z., Barrett, B. A., Griffiths, A. G., & Woodfield, D. R. (2003). DNA fingerprinting and genetic relationships among white clover cultivars. *Proceedings of the New Zealand Grassland Association, 65*, 163-169.

Jahufer, M. Z. Z., Cooper, M., Bray, R. A., & Ayres, J. F. (1999). Evaluation of white clover (*Trifolium repens* L.) populations for summer moisture stress adaptation in Australia. *Australian Journal of Agricultural Research, 50*(4), 561-574.

Jahufer, M. Z. Z., & Luo, D. (2018). DeltaGen: a comprehensive decision support tool for plant breeders. *Crop Science, 58*, 1118-1131.

Jahufer, M. Z. Z., Rogers, H., & Rogers, M. J. (2001). *White clover*. Department of Natural Resources and Environment, Victoria.

Jelinek, J., & Madzo, J. (2016). DREAM: A Simple Method for DNA Methylation Profiling by High-throughput Sequencing. In S. Li & H. Zhang (Eds.), *Chronic Myeloid Leukemia: Methods and Protocols* (pp. 111-127). New York, NY: Springer New York.

Jermyn, M. A. (1956). A New Method for determining Ketohexoses in the Presence of Aldohexoses. *Nature, 177*(4497), 38-39.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research, 36*(suppl_2), W5-W9.

Jolly, L., Pompeo, F., van Heijenoort, J., Fassy, F., & Mengin-Lecreulx, D. (2000). Autophosphorylation of Phosphoglucosamine Mutase from *Escherichia coli*. *Journal of Bacteriology, 182*(5), 1280.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics, 24*(11), 1403-1405.

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

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England), 30*(9), 1236-1240.

Junaid, A., Kumar, H., Rao, A. R., Patil, A. N., Singh, N. K., & Gaikwad, K. (2018). Unravelling the epigenomic interactions between parental inbreds resulting in an altered hybrid methylome in pigeonpea. *DNA Research, 25*(4), 361-373.

Kagan, I. A., Anderson, M. L., Kramer, K. J., Seman, D. H., Lawrence, L. M., & Smith, S. R. (2020). Seasonal and Diurnal Variation in Water-Soluble Carbohydrate Concentrations of Repeatedly Defoliated Red and White Clovers in Central Kentucky. *J Equine Vet Sci, 84*, 102858.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35*, 401-417.

Kamber, G., McDonald, C., & Price, G. (2013). Drying out: Investigating the economic effects of drought in New Zealand (Vol. Reserve Bank of New Zealand Analytical Note Series, pp. 1-31).

Kanehisa, M., & Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein Science, 29*(1), 28-35.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., . . . Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics, 42*(4), 348-354.

Kang, Y., Sakiroglu, M., Krom, N., Stanton-Geddes, J., Wang, M., Lee, Y.-C., . . . Udvardi, M. (2015). Genome-wide association of drought-related and biomass traits with HapMap SNPs in *Medicago truncatula*. *Plant, Cell & Environment, 38*(10), 1997-2011.

Karsten, H. D., & MacAdam, J. W. (2001). Effect of drought on growth, carbohydrates, and soil water use by perennial ryegrass, tall fescue, and white clover. *Crop Science, 41*(1), 156-166.

Kassambara, A. (2019). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2.3. Retrieved from https://CRAN.R-project.org/package=ggpubr

Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters, 7*(12), 1225-1241.

Kerepesi, I., & Galiba, G. (2000). Osmotic and Salt Stress-Induced Alteration in Soluble Carbohydrate Content in Wheat Seedlings. *Crop Science, 40*(2), 482-487.

Khanlou, K. M., Vandepitte, K., Asl, L. K., & Bockstaele, E. V. (2011). Towards an optimal sampling strategy for assessing genetic variation within and among white clover (*Trifolium repens* L.) cultivars using AFLP. *Genetics and Molecular Biology, 34*, 252-258.

Kim, S. J., & Kim, W. T. (2013). Suppression of *Arabidopsis* RING E3 ubiquitin ligase AtATL78 increases tolerance to cold stress and decreases tolerance to drought stress. *FEBS Letters, 587*(16), 2584-2590.

Kim, T.-H., Lee, B.-R., Jung, W.-J., Kim, K.-Y., Avice, J.-C., & Ourry, A. (2004). De novo protein synthesis in relation to ammonia and proline accumulation in water stressed white clover. *Functional Plant Biology, 31*(8), 847-855.

Kim, Y., & Nielsen, R. (2004). Linkage Disequilibrium as a Signature of Selective Sweeps. *Genetics, 167*(3), 1513.

Kim, Y., & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics, 160*(2), 765-777.

Kingston-Smith, A. H., & Theodorou, M. K. (2000). Tansley Review No. 118. Post-Ingestion Metabolism of Fresh Forage. *The New Phytologist, 148*(1), 37-55.

Kiyosue, T., Abe, H., Yamaguchi-Shinozaki, K., & Shinozaki, K. (1998). ERD6, a cDNA clone for an early dehydration-induced gene of *Arabidopsis*, encodes a putative sugar transporter1. *Biochimica et Biophysica Acta (BBA) - Biomembranes, 1370*(2), 187-191.

Kiyosue, T., Yoshiba, Y., Yamaguchi-Shinozaki, K., & Shinozaki, K. (1996). A nuclear gene encoding mitochondrial proline dehydrogenase, an enzyme involved in proline metabolism, is upregulated by proline but downregulated by dehydration in *Arabidopsis*. *The Plant cell, 8*(8), 1323-1335.

Kjærgaard, T. (2003). A plant that changed the world: the rise and fall of clover 1000-2000. (Theme Issue: The native, naturalized and exotic - plants and animals in human history). *Landscape Research, 28*(1), 41-49.

Klein, M., Perfus-Barbeoch, L., Frelet, A., Gaedeke, N., Reinhardt, D., Mueller-Roeber, B., . . . Forestier, C. (2003). The plant multidrug resistance ABC transporter AtMRP5 is involved in guard cell hormonal signalling and water use. *The Plant Journal, 33*(1), 119-129.

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.

Kojima, S., Takahashi, Y., Kobayashi, Y., Monna, L., Sasaki, T., Araki, T., & Yano, M. (2002). Hd3a, a rice ortholog of the *Arabidopsis* FT gene, promotes transition to flowering downstream of Hd1 under short-day conditions. *Plant Cell Physiology, 43*(10), 1096-1105.

Kolbe, A., Tiessen, A., Schluepmann, H., Paul, M., Ulrich, S., & Geigenberger, P. (2005). Trehalose 6-phosphate regulates starch synthesis via posttranslational redox activation of ADP-glucose pyrophosphorylase. *Proceedings of the National Academy of Sciences of the United States of America, 102*(31), 11118.

Kölliker, R., Jones, E. S., Jahufer, M. Z. Z., & Forster, J. W. (2001). Bulked AFLP analysis for the assessment of genetic diversity in white clover (*Trifolium repens* L.). *Euphytica, 121*(3), 305-315.

Kooyers, N. J., & Olsen, K. M. (2012). Rapid evolution of an adaptive cyanogenesis cline in introduced North American white clover (*Trifolium repens* L.). *Molecular Ecology, 21*(10), 2455-2468.

Kooyers, N. J., & Olsen, K. M. (2013). Searching for the bull/'s eye: agents and targets of selection vary among geographically disparate cyanogenesis clines in white clover (*Trifolium repens* L.). *Heredity, 111*(6), 495-504.

Küchenmeister, K., Kuechenmeister, F., Kayser, M., Wrage-Mönnig, N., & Isselstein, J. (2013). Influence of drought stress on nutritive value of perennial forage legumes. *International Journal of Plant Production, 7*(4), 693-710.

Lai, Y.-T., Yeung, C. K. L., Omland, K. E., Pang, E.-L., Hao, Y., Liao, B.-Y., . . . Li, S.-H. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences, 116*(6), 2152.

Lam, D., Luu, P.-L., Song, J. Z., Qu, W., Risbridger, G. P., Lawrence, M. G., . . . Stirzaker, C. (2020). Comprehensive evaluation of targeted multiplex bisulphite PCR sequencing for validation of DNA methylation biomarker panels. *Clinical Epigenetics, 12*(1), 90.

Lambert, M. G., Clark, D. A., & Litherland, A. J. (2004). Advances in pasture management for animal productivity and health. *New Zealand Veterinary Journal, 52*(6), 311-319.

Lancashire, J. A., Ralston, M. P., & Scott, D. J. (1985). Contamination of white clover seed crops by buried seeds *Producing Herbage Seeds, 2*, 61-65.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics, 9*(1), 559.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods, 9*(4), 357-359.

Larned, S. T., Scarsbrook, M. R., Snelder, T. H., Norton, N. J., & Biggs, B. J. F. (2004). Water quality in low-elevation streams and rivers of New Zealand: Recent state and trends in contrasting land-cover classes. *New Zealand Journal of Marine and Freshwater Research, 38*(2), 347-366.

Larned, S. T., Snelder, T., Unwin, M. J., & McBride, G. B. (2016). Water quality in New Zealand rivers: current state and trends. *New Zealand Journal of Marine and Freshwater Research, 50*(3), 389-417.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software; Vol 1, Issue 1 (2008)*.

Le, T. N., & McQueen-Mason, S. J. (2006). Desiccation-tolerant plants in dry environments. *Reviews in Environmental Science and Bio/Technology, 5*(2), 269.

Ledgard, S. F., & Steele, K. W. (1992). Biological nitrogen fixation in mixed legume/grass pastures. *Plant and Soil, 141*(1), 137-153.

Lee, B.-R., Jin, Y.-L., Jung, W.-J., Avice, J.-C., Morvan-Bertrand, A., Ourry, A., . . . Kim, T.-H. (2008). Water-deficit accumulates sugars by starch degradation—not by de novo synthesis—in white clover leaves (*Trifolium repens*). *Physiologia Plantarum, 134*(3), 403-411.

Lee, B.-R., Jin, Y. L., Avice, J.-C., Cliquet, J.-B., Ourry, A., & Kim, T.-H. (2009). Increased proline loading to phloem and its effects on nitrogen uptake and assimilation in water-stressed white clover (*Trifolium repens*). *New Phytologist, 182*(3), 654-663.

Lee, M. R. F., Jones, E. L., Moorby, J. M., Humphreys, M. O., Theodorou, M. K., & Scollan, N. D. (2001). Production responses from lambs grazed on *Lolium*

perenne selected for an elevated water-soluble carbohydrate concentration. *Animal Research, 50*(6), 441-449.

Lee, M. V., Topper, S. E., Hubler, S. L., Hose, J., Wenger, C. D., Coon, J. J., & Gasch, A. P. (2011). A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular systems biology, 7*, 514-514.

Lertrat, K., & Pulam, T. (2007). Breeding for increased sweetness in sweet corn. *International Journal of Plant Breeding, 1*(1), 27-30.

Lewis, G., Schrire, B., Mackinder, B., & Lock, M. (2005). *Legumes of the World*. Surrey, UK: Royal Botanic Gardens, Kew.

Lewis, J. (1973). Longevity of crop and weed seeds: Survival after 20 years in soil. *Weed Research, 13*(2), 179-191.

Li, F., Lei, H., Zhao, X., Tian, R., & Li, T. (2012). Characterization of Three Sorbitol Transporter Genes in Micropropagated Apple Plants Grown under Drought Stress. *Plant Molecular Biology Reporter, 30*(1), 123-130.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760.

Li, J., Wu, L., Foster, R., & Ruan, Y.-L. (2017a). Molecular regulation of sucrose catabolism and sugar transport for development, defence and phloem function. *Journal of Integrative Plant Biology, 59*(5), 322-335.

Li, S., Bashline, L., Zheng, Y., Xin, X., Huang, S., Kong, Z., . . . Gu, Y. (2016). Cellulose synthase complexes act in a concerted fashion to synthesize highly aggregated cellulose in secondary cell walls of plants. *Proceedings of the National Academy of Sciences, 113*(40), 11348.

Li, X., Han, Y., Wei, Y., Acharya, A., Farmer, A. D., Ho, J., . . . Brummer, E. C. (2014). Development of an Alfalfa SNP Array and Its Use to Evaluate Patterns of Population Structure and Linkage Disequilibrium. *PLOS ONE, 9*(1), e84329.

Li, Y., Zhang, X.-X., Mao, R.-L., Yang, J., Miao, C.-Y., Li, Z., & Qiu, Y.-X. (2017b). Ten Years of Landscape Genomics: Challenges and Opportunities. *Frontiers in Plant Science, 8*, 2136.

Li, Z., Shi, P., & Peng, Y. (2013). Improved drought tolerance through drought preconditioning associated with changes in antioxidant enzyme activities, gene expression and osmoregulatory solutes accumulation in white clover (*Trifolium repens* L.). *Plant OMICS, 6*(6), 481-489.

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

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., . . . Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature, 462*(7271), 315-322.

Lister, S. J., & Dhanoa, M. S. (1998). Comparison of calibration models for the prediction of forage quality traits using near infrared spectroscopy. *Journal of Agricultural Science, 131*(2), 237-243.

Liu, J., Van Eck, J., Cong, B., & Tanksley, S. D. (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proceedings of the National Academy of Sciences of the United States of America, 99*(20), 13302-13306.

Liu, Y., Beyer, A., & Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell, 165*(3), 535-550.

Livingston, D. P., 3rd, Hincha, D. K., & Heyer, A. G. (2009). Fructan and its relationship to abiotic stress tolerance in plants. *Cellular and molecular life sciences : CMLS, 66*(13), 2007-2023.

Long, A. D., & Langley, C. H. (1999). The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits. *Genome Research, 9*(8), 720-731.

López-González, C., Juárez-Colunga, S., Morales-Elías, N. C., & Tiessen, A. (2019). Exploring regulatory networks in plants: transcription factors of starch metabolism. *PeerJ, 7*, e6841-e6841.

Lorenz, A. J. (2013). Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment. *G3: Genes|Genomes|Genetics, 3*(3), 481.

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular ecology, 23*(9), 2178-2192.

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology, 24*(5), 1031-1046.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550.

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., . . . Costich, D. E. (2013). Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLoS Genetics, 9*(1), e1003215.

Ludlow, M. M., & Muchow, R. C. (1990). A Critical Evaluation of Traits for Improving Crop Yields in Water-Limited Environments. In N. C. Brady (Ed.), *Advances in Agronomy* (Vol. 43, pp. 107-153): Academic Press.

Luo, D., Ganesh, S., & Koolaard, J. (2018). predictmeans: Calculate Predicted Means for Linear Models. R package version 1.0.1.

Luo, J., Sun, X. Z., Pacheco, D., Ledgard, S. F., Lindsey, S. B., Hoogendoorn, C. J., . . . Watkins, N. L. (2015). Nitrous oxide emission factors for urine and dung from sheep fed either fresh forage rape (*Brassica napus* L.) or fresh perennial ryegrass (*Lolium perenne* L.). *Animal, 9*(3), 534-543.

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources, 17*(1), 67-77.

Ma, Y., Coyne, C. J., Grusak, M. A., Mazourek, M., Cheng, P., Main, D., & McGee, R. J. (2017). Genome-wide SNP identification, linkage map construction and QTL mapping for seed mineral concentrations and contents in pea (*Pisum sativum* L.). *BMC Plant Biology, 17*(1), 43.

Macdonald, S. J., & Long, A. D. (2004). A Potential Regulatory Polymorphism Upstream of *hairy* Is Not Associated With Bristle Number Variation in Wild-Caught *Drosophila*. *Genetics, 167*(4), 2127.

Maier, T., Güell, M., & Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Letters, 583*(24), 3966-3973.

Malinowski, D. P., Belesky, D. P., & Fedders, J. (1998). Photosynthesis of White Clover (*Trifolium repens* L.) Germplasms with Contrasting Leaf Size. *Photosynthetica, 35*(3), 419-427.

Marshall, A. H., Michaelson-Yeates, T., & Williams, I. H. (1999). How Busy Are Bees - modelling the pollination of clover *IGER Innovations 3* (pp. 17-21). Aberystwyth, UK: Aberystwyth University.

Marshall, A. H., Williams, T. A., Abberton, M. T., Michaelson-Yeates, T. P. T., Olyott, P., & Powell, H. G. (2004). Forage quality of white clover (*Trifolium repens* L.) × Caucasian clover (*T. ambiguum* M. Bieb.) hybrids and their grass companion when grown over three harvest years. *Grass and Forage Science, 59*(1), 91-99.

Marten, G. C., Brink, G. E., Buxton, D. R., Halgerson, J. L., & Hornstein, J. S. (1984). Near infrared reflectance spectroscopy analysis of forage quality in four legume species. *Crop Science, 24*, 1179-1182.

Mascher, M., Wu, S., Amand, P. S., Stein, N., & Poland, J. (2013). Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley. *PLoS ONE, 8*(10), e76925.

Mathew, L. S., Seidel, M. A., George, B., Mathew, S., Spannagl, M., Haberer, G., . . . Malek, J. A. (2015). A genome-wide survey of date palm cultivars supports two major subpopulations in *Phoenix dactylifera*. *G3: Genes, Genomes, Genetics, 5*(7), 1429-1438.

McCouch, S. (2004). Diversifying Selection in Plant Breeding. *PLOS Biology, 2*(10), e347.

McDowell, R. W., Larned, S. T., & Houlbrooke, D. J. (2009). Nitrogen and phosphorus in New Zealand streams and rivers: Control and impact of eutrophication and the influence of land management. *New Zealand Journal of Marine and Freshwater Research, 43*(4), 985-995.

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics, 157*(4), 1819-1829.

Michell, P. J. (1973). Relations between fibre and water soluble carbohydrate contents of pasture species and their digestibility and voluntary intake by sheep. *Australian Journal of Experimental Agriculture and Animal Husbandry, 13*, 165-170.

Mickelbart, M. V., Hasegawa, P. M., & Bailey-Serres, J. (2015). Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. *Nature Reviews Genetics, 16*(4), 237-251.

Miller, L. A., Moorby, J. M., Davies, D. R., Humphreys, M. O., Scollan, N. D., MacRae, J. C., & Theodorou, M. K. (2001). Increased concentration of water-soluble carbohydrate in perennial ryegrass (*Lolium perenne* L.): milk production from late-lactation dairy cows. . *Grass and Forage Science, 56*, 383-394.

Miller, L. A., Theodorou, M. K., MacRae, J. C., Evans, R. T., Humphreys, M. O., Scollan, N. D., & Moorby, J. M. (2000, 11-13 September 2000). *Efficiency of nitrogen use by dairy cows offered perennial ryegrass with high water soluble carbohydrate concentrations.* Paper presented at the Proceedings of the 6th Research Conference of the British Grassland Society, Aberdeen.

Minitab LLC. (2017). Minitab 18 Statistical Software (Version 18.1). State College, PA. Retrieved from www.minitab.com

Miryeganeh, M., & Saze, H. (2020). Epigenetic inheritance and plant evolution. *Population Ecology, 62*(1), 17-27.

Miskimen, K. L. S., Chan, E. R., & Haines, J. L. (2017). Assay for Transposase-Accessible Chromatin Using Sequencing (ATAC-seq) Data Analysis. *Current Protocols in Human Genetics, 92*(1), 20.24.21-20.24.13.

Misra, A., McKnight, T. D., & Mandadi, K. K. (2018). Bromodomain proteins GTE9 and GTE11 are essential for specific BT2-mediated sugar and ABA responses in *Arabidopsis thaliana*. *Plant Molecular Biology, 96*(4), 393-402.

Montgomery, E. G. (1911). *Correlation studies in corn.* 24th Annual Report, Agricultural Experimental Station, Nebraska, MO, USA.

Moorby, J. M., Evans, R. T., Scollan, N. D., Macraet, J. C., & Theodorou, M. K. (2006). Increased concentration of water-soluble carbohydrate in perennial ryegrass (*Lolium perenne* L.). Evaluation in dairy cows in early lactation. *Grass and Forage Science, 61*(1), 52-59.

Moose, S. P., & Mumm, R. H. (2008). Molecular Plant Breeding as the Foundation for 21st Century Crop Improvement. *Plant Physiology, 147*(3), 969.

Moradi, M. H., Nejati-Javaremi, A., Moradi-Shahrbabak, M., Dodds, K. G., & McEwan, J. C. (2012). Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *BMC Genetics, 13*(1), 10.

Moran, J. (2005). How the rumen works. *Tropical dairy farming : feeding management for small holder dairy farmers in the humid tropics* (pp. 312): Landlinks Press.

Morell, M. K., Bloom, M., Knowles, V., & Preiss, J. (1987). Subunit Structure of Spinach Leaf ADPglucose Pyrophosphorylase. *Plant physiology, 85*(1), 182-187.

Mouchel, C. F., Briggs, G. C., & Hardtke, C. S. (2004). Natural genetic variation in *Arabidopsis* identifies BREVIS RADIX, a novel regulator of cell proliferation and elongation in the root. *Genes & Development, 18*(6), 700-714.

Müller-Röber, B., Nast, G., & Willmitzer, L. (1995). Isolation and expression analysis of cDNA clones encoding a small and a large subunit of ADP-glucose pyrophosphorylase from sugar beet. *Plant Molecular Biology, 27*(1), 191-197.

Muller-Rober, B. T., Kossmann, J., Hannah, L. C., Willmitzer, L., & Sonnewald, U. (1990). One of two different ADP-glucose pyrophosphorylase genes from potato responds strongly to elevated levels of sucrose. *Mol Gen Genet, 224*(1), 136-146.

Nagy, R., Grob, H., Weder, B., Green, P., Klein, M., Frelet-Barrand, A., . . . Martinoia, E. (2009). The *Arabidopsis* ATP-binding cassette protein AtMRP5/AtABCC5 is a high affinity inositol hexakisphosphate transporter involved in guard cell signaling and phytate storage. *The Journal of biological chemistry, 284*(48), 33614-33622.

Nakamura, M., Toyota, M., Tasaka, M., & Morita, M. T. (2011). An *Arabidopsis* E3 Ligase, SHOOT GRAVITROPISM9, Modulates the Interaction between Statoliths and F-Actin in Gravity Sensing. *The Plant Cell, 23*(5), 1830.

Nakata, P. A., Greene, T. W., Anderson, J. M., Smith-White, B. J., Okita, T. W., & Preiss, J. (1991). Comparison of the primary sequences of two potato tuber ADP-glucose pyrophosphorylase subunits. *Plant Molecular Biology, 17*(5), 1089-1093.

Narum, S. R., & Hess, J. E. (2011). Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources, 11*(s1), 184-194.

Newstrom-Lloyd, L. E. (2013). Pollination in New Zealand. In D. J. R. (Ed.), *Ecosystem services in New Zealand – conditions and trends*. Manaaki Whenua Press, Lincoln, New Zealand.

Nie, Z., Tremblay, G. F., Bélanger, G., Berthiaume, R., Castonguay, Y., Bertrand, A., . . . Han, J. (2009). Near-infrared reflectance spectroscopy prediction of neutral detergent-soluble carbohydrates in timothy and alfalfa. *Journal of Dairy Science, 92*(4), 1702-1711.

Nielsen, R., Hubisz, M. J., & Clark, A. G. (2004). Reconstituting the Frequency Spectrum of Ascertained Single-Nucleotide Polymorphism Data. *Genetics, 168*(4), 2373.

Niklas, K. J., Cobb, E. D., Niinemets, Ü., Reich, P. B., Sellin, A., Shipley, B., & Wright, I. J. (2007). "Diminishing returns" in the scaling of functional leaf traits across and within species groups. *Proceedings of the National Academy of Sciences, 104*(21), 8891.

NIWA. (2020). The National Climate Database. Retrieved 06/05/2020, from https://cliflo.niwa.co.nz/

Nocek, J. E., & Russell, J. B. (1988). Protein and energy as an integrated system. Relationship of ruminal protein and carbohydrate availability to microbial synthesis and milk protein. *Journal of Dairy Science, 70*, 2070-2107.

Nurminsky, D. I. (2001). Genes in sweeping competition. *Cellular and Molecular Life Sciences, 58*(1), 125-134.

Nybom, H. (2004). Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology, 13*(5), 1143-1155.

NZIER. (2016). How valuable is that plant species? Application of a method for enumerating the contribution of selected plant species to New Zealand's GDP. (pp. 212). Wellington, New Zealand.

NZIER. (2019). The importance of crop protection products for the New Zealand economy (pp. 31). Wellington, New Zealand.

Oleksyk, T. K., Smith, M. W., & O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1537), 185-205.

Olsen, K. M., Sutherland, B. L., & Small, L. L. (2007). Molecular evolution of the Li/li chemical defence polymorphism in white clover (*Trifolium repens* L.). *Molecular Ecology, 16*(19), 4180-4193.

Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation, 15*(12).

Pacheco, D., & Waghorn, G. C. (2008). Dietary nitrogen – definitions, digestion, excretion and consequences of excess for grazing ruminants. *Proceedings of the New Zealand Grassland Association, 70*, 107-116.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics, 26*, 419-420.

Parida, A. K., Dagaonkar, V. S., Phalak, M. S., & Aurangabadkar, L. P. (2008). Differential responses of the enzymes involved in proline biosynthesis and degradation in drought tolerant and sensitive cotton genotypes during drought stress and recovery. *Acta Physiologiae Plantarum, 30*(5), 619-627.

Patel, M., Milla-Lewis, S., Zhang, W., Templeton, K., Reynolds, W. C., Richardson, K., . . . Sathish, P. (2015). Overexpression of ubiquitin-like LpHUB1 gene confers

drought tolerance in perennial ryegrass. *Plant Biotechnology Journal, 13*(5), 689-699.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature, 335*, 721.

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLOS Genetics, 2*(12), e190.

Pauls, S. U., Nowak, C., Bálint, M., & Pfenninger, M. (2013). The impact of global climate change on genetic diversity within populations and species. *Molecular Ecology, 22*(4), 925-946.

Pelletier, S., Tremblay, G. F., Bertrand, A., Bélanger, G., Castonguay, Y., & Michaud, R. (2010). Drying procedures affect non-structural carbohydrates and other nutritive value attributes in forage samples. *Animal Feed Science and Technology, 157*(3), 139-150.

Pembleton, L. W., Drayton, M. C., Bain, M., Baillie, R. C., Inch, C., Spangenberg, G. C., . . . Cogan, N. O. I. (2016). Targeted genotyping-by-sequencing permits cost-effective identification and discrimination of pasture grass species and cultivars. *Theoretical and Applied Genetics, 129*(5), 991-1005.

Penman, H. L. (1948). Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 193*(1032), 120-145.

Penning, P. D., Parsons, A. J., Orr, R. J., Harvey, A., & Yarrow, N. H. (1995). Dietary preference of heifers for grass or clover, with and without romensin slow-release anti-bloat boluses. *Animal Science, 60*, 550.

Pérez-Figueroa, A., García-Pereira, M. J., Saura, M., Rolán-Alvarez, E., & Caballero, A. (2010). Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology, 23*(10), 2267-2276.

Pertl-Obermeyer, H., Trentmann, O., Duscha, K., Neuhaus, H. E., & Schulze, W. X. (2016). Quantitation of Vacuolar Sugar Transporter Abundance Changes Using QconCAT Synthtetic Peptides. *Frontiers in Plant Science, 7*(411).

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE, 7*(5), e37135.

Phan, H. T. T., Rybak, K., Bertazzoni, S., Furuki, E., Dinglasan, E., Hickey, L. T., . . . Tan, K.-C. (2018). Novel sources of resistance to Septoria nodorum blotch in the Vavilov wheat collection identified by genome-wide association studies. *Theoretical and Applied Genetics, 131*(6), 1223-1238.

Piaskowski, J. L., Brown, D., & Campbell, K. G. (2016). Near-Infrared Calibration of Soluble Stem Carbohydrates for Predicting Drought Tolerance in Spring Wheat. *Agronomy Journal, 108*, 285-293.

Pilon-Smits, E. A. H., Terry, N., Sears, T., & van Dun, K. (1999). Enhanced drought resistance in fructan-producing sugar beet. *Plant Physiology and Biochemistry, 37*(4), 313-317.

Pittman, J. J., Arnall, D. B., Interrante, S. M., Wang, N., Raun, W. R., & Butler, T. J. (2016). Bermudagrass, Wheat, and Tall Fescue Crude Protein Forage Estimation using Mobile-Platform, Active-Spectral and Canopy-Height Data. *Crop Science, 56*(2), 870-881.

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., . . . Jannink, J.-L. (2012a). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome, 5*(3), 103-113.

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012b). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLOS ONE, 7*(2), e32253.

Ponnala, L., Wang, Y., Sun, Q., & van Wijk, K. J. (2014). Correlation of mRNA and protein abundance in the developing maize leaf. *The Plant Journal, 78*(3), 424-440.

Portela, A., & Esteller, M. (2010). Epigenetic modifications and human disease. *Nature Biotechnology, 28*(10), 1057-1068.

Prakash, G., Kumar, S., & Mikawlrawng, K. (2014). Metabolic Pathways Engineering: an emerging approach in abiotic stress tolerance in plants. *Journal of Pharmacognosy and Phytochemistry, 3*, 104-107.

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

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics, 155*(2), 945-959.

Przeworski, M. (2002). The Signature of Positive Selection at Randomly Chosen Loci. *Genetics, 160*(3), 1179.

Pullanagari, R., Kereszturi, G., Yule, I., & Irwin, M. (2015). *Determination of pasture quality using airborne hyperspectral imaging*. Paper presented at the SPIE Remote Sensing, Toulouse, France.

Pullanagari, R. R., Kereszturi, G., & Yule, I. (2018). Integrating Airborne Hyperspectral, Topographic, and Soil Data for Estimating Pasture Quality Using Recursive Feature Elimination with Random Forest Regression. *Remote Sensing, 10*(7), 1117.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics, 81*(3), 559-575.

Purugganan, M. D., Boyles, A. L., & Suddith, J. I. (2000). Variation and Selection at the *CAULIFLOWER* Floral Homeotic Gene Accompanying the Evolution of Domesticated *Brassica oleracea*. *Genetics, 155*(2), 855.

Pyke, N. B., Rolston, M. P., & Woodfield, D. R. (2004). National and export trends in herbage seed production. *Proceedings of the New Zealand Grassland Association, 66*, 95-102.

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.1). R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.

Rago, R., Mitchen, J., & Wilding, G. (1990). DNA fluorometric assay in 96-well tissue culture plates using Hoechst 33258 after cell lysis by freezing in distilled water. *Analytical Biochemistry, 191*(1), 31-34.

Raquin, A. L., Brabant, P., RhonÉ, B., Balfourier, F., Leroy, P., & Goldringer, I. (2008). Soft selective sweep near a gene that increases plant height in wheat. *Molecular Ecology, 17*(3), 741-756.

Rasmussen, J., Eriksen, J., Jensen, E. S., Esbensen, K. H., & Høgh-Jensen, H. (2007). In situ carbon and nitrogen dynamics in ryegrass–clover mixtures: Transfers, deposition and leaching. *Soil Biology and Biochemistry, 39*(3), 804-815.

Rasmussen, S., Parsons, A. J., Xue, H., & Newman, J. A. (2009). High sugar grasses–harnessing the benefits of new cultivars through growth management. *Proceedings of the New Zealand Grassland Association, 71*, 167-175.

Reguera, M., Peleg, Z., & Blumwald, E. (2012). Targeting metabolic pathways for genetic engineering abiotic stress-tolerance in crops. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1819*(2), 186-194.

Reinert, S., Osthoff, A., Léon, J., & Naz, A. A. (2019). Population Genetics Revealed a New Locus That Underwent Positive Selection in Barley. *International journal of molecular sciences, 20*(1), 202.

Reisinger, A., Mullan, B., Manning, M., Wratt, D., & Nottage, R. (2010). Global and local climate change scenarios to support adaptation in New Zealand. In R. A. C. Nottage, D. S. Wratt, J. F. Bornman & K. Jones (Eds.), *Climate change adaptation in New Zealand: future scenarios and some sectoral perspectives* (pp. 26-43). Wellington: New Zealand Climate Change Centre.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology, 24*(17), 4348-4370.

Rellstab, C., Zoller, S., Tedder, A., Gugerli, F., & Fischer, M. C. (2013). Validation of SNP Allele Frequencies Determined by Pooled Next-Generation Sequencing in Natural Populations of a Non-Model Plant Species. *PLoS ONE, 8*(11), e80422.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., . . . Buckler, E. S. t. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America, 98*(20), 11479-11484.

Resende, R. M. S., Casler, M. D., & de Resende, M. D. V. (2014). Genomic Selection in Forage Breeding: Accuracy and Methods. *Crop Science, 54*(1), 143-156.

Rey, O., Eizaguirre, C., Angers, B., Baltazar-Soares, M., Sagonas, K., Prunier, J. G., & Blanchet, S. (2020). Linking epigenetics and biological conservation: Towards a conservation epigenetics perspective. *Functional Ecology, 34*(2), 414-427.

Rivas, A., Singh, R., Horne, D., Roygard, J., Matthews, A., & Hedley, M. J. (2017). Denitrification potential in the subsurface environment in the Manawatu River catchment, New Zealand: Indications from oxidation-reduction conditions, hydrogeological factors, and implications for nutrient management. *Journal of Environmental Management, 197*, 476-489.

Rizhsky, L., Liang, H., Shuman, J., Shulaev, V., Davletova, S., & Mittler, R. (2004). When Defense Pathways Collide. The Response of *Arabidopsis* to a Combination of Drought and Heat Stress. *Plant Physiology, 134*(4), 1683-1696.

Roberts, H. A. (1981). Seed banks in soils. *Advances in Applied Biology, 6*, 1-55.

Rohlfs, R. V., & Nielsen, R. (2015). Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. *Systematic Biology, 64*(5), 695-708.

Rontein, D., Basset, G., & Hanson, A. D. (2002). Metabolic Engineering of Osmoprotectant Accumulation in Plants. *Metabolic Engineering, 4*(1), 49-56.

Rosa, M., Prado, C., Podazza, G., Interdonato, R., González, J. A., Hilal, M., & Prado, F. E. (2009). Soluble sugars—Metabolism, sensing and abiotic stress: A complex network in the life of plants. *Plant Signaling & Behavior, 4*(5), 388-393.

Rose, C. J., Chapman, J. R., Marshall, S. D. G., Lee, S. F., Batterham, P., Ross, H. A., & Newcomb, R. D. (2011). Selective sweeps at the organophosphorus insecticide resistance locus, Rop-1, have affected variation across and beyond the α-esterase gene cluster in the Australian sheep blowfly, *Lucilia cuprina*. *Molecular Biology and Evolution, 28*(6), 1835-1846.

Ross-Ibarra, J., Morrell, P. L., & Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences, 104*(suppl 1), 8641-8648.

Rosyara, U. R., De Jong, W. S., Douches, D. S., & Endelman, J. B. (2016). Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome, 9*(2).

Royston, P. (1995). Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 44*(4), 547-551.

Ruckle, M., Bernasconi, L., Kölliker, R., Zeeman, S., & Studer, B. (2018). Genetic diversity of diurnal carbohydrate accumulation in white clover (*Trifolium repens* L.). *Agronomy, 8*(4), 47.

Ruckle, M. E., Meier, M. A., Frey, L., Eicke, S., Kolliker, R., Zeeman, S. C., & Studer, B. (2017). Diurnal leaf starch content: An orphan trait in forage legumes. *Agronomy-Basel, 7*(1), 15.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., . . . Lander, E. S. (2006). Positive Natural Selection in the Human Lineage. *Science, 312*(5780), 1614-1620.

Saccheri, I. J., Rousset, F., Watts, P. C., Brakefield, P. M., & Cook, L. M. (2008). Selection and gene flow on a diminishing cline of melanic peppered moths. *Proceedings of the National Academy of Sciences of the United States of America, 105*(42), 16212-16217.

Sackton, T. B., & Clark, N. (2019). Convergent evolution in the genomics era: new insights and directions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 374*(1777), 20190102-20190102.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*(4), 406-425.

Sakai, T., Mochizuki, S., Haga, K., Uehara, Y., Suzuki, A., Harada, A., . . . Okada, K. (2012). The wavy growth 3 E3 ligase family controls the gravitropic response in *Arabidopsis* roots. *The Plant journal : for cell and molecular biology, 70*(2), 303-314.

Sakiroglu, M., & Brummer, E. C. (2017). Identification of loci controlling forage yield and nutritive value in diploid alfalfa using GBS-GWAS. *Theoretical and Applied Genetics, 130*(2), 261-268.

Sambe, M. A. N., He, X., Tu, Q., & Guo, Z. (2015). A cold-induced myo-inositol transporter-like gene confers tolerance to multiple abiotic stresses in transgenic tobacco plants. *Physiologia Plantarum, 153*(3), 355-364.

Sax, K. (1923). The Association of Size Differences with Seed-Coat Pattern and Pigmentation in *Phaseolus vulgaris*. *Genetics, 8*(6), 552-560.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., . . . Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nat Meth, 9*(7), 676-682.

Schlenke, T. A., & Begun, D. J. (2004). Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proceedings of the National Academy of Sciences of the United States of America, 101*(6), 1626-1631.

Schneider, S., Beyhl, D., Hedrich, R., & Sauer, N. (2008). Functional and Physiological Characterization of *Arabidopsis INOSITOL TRANSPORTER1*, a Novel Tonoplast-Localized Transporter for *myo*-Inositol. *The Plant Cell, 20*(4), 1073.

Schoville, S. D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., & Manel, S. (2012). Adaptive Genetic Variation on the Landscape: Methods and Cases. *Annual Review of Ecology, Evolution, and Systematics, 43*(1), 23-43.

Schulze, W. X., Reinders, A., Ward, J., Lalonde, S., & Frommer, W. B. (2003). Interactions between co-expressed *Arabidopsis* sucrose transporters in the split-ubiquitin system. *BMC biochemistry, 4*, 3-3.

Schulze, W. X., Schneider, T., Starck, S., Martinoia, E., & Trentmann, O. (2012). Cold acclimation induces changes in *Arabidopsis* tonoplast protein abundance and activity and alters phosphorylation of tonoplast monosaccharide transporters. *The Plant Journal, 69*(3), 529-541.

Selbie, D. R., Buckthought, L. E., & Shepherd, M. A. (2015). Chapter Four - The Challenge of the Urine Patch for Managing Nitrogen in Grazed Pasture Systems. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 129, pp. 229-292): Academic Press.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution, 8*(8), 907-917.

Shetty, N., Gislum, R., Jensen, A. M. D., & Boelt, B. (2012). Development of NIR calibration models to assess year-to-year variation in total non-structural carbohydrates in grasses using PLSR. *Chemometrics and Intelligent Laboratory Systems, 111*(1), 34-38.

Shorten, P. R., Leath, S. R., Schmidt, J., & Ghamkhar, K. (2019). Predicting the quality of ryegrass using hyperspectral imaging. *Plant Methods, 15*(1), 63.

Singh, M., Kumar, J., Singh, S., Singh, V. P., & Prasad, S. M. (2015). Roles of osmoprotectants in improving salinity and drought tolerance in plants: a review. *Reviews in Environmental Science and Bio/Technology, 14*(3), 407-426.

Smith, C., Karunaratne, S., Badenhorst, P., Cogan, N., Spangenberg, G., & Smith, K. (2020). Machine Learning Algorithms to Predict Forage Nutritive Value of In Situ Perennial Ryegrass Plants Using Hyperspectral Canopy Reflectance Data. *Remote Sensing, 12*(6).

Smith, D. (1973). Influence of drying and storage conditions on nonstructural carbohydrate analysis of herbage tissue—a review. *Grass and Forage Science, 28*(3), 129-134.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research, 23*(1), 23-35.

Soltis, D. E., Soltis, P. S., & Tate, J. A. (2003). Advances in the study of polyploidy since Plant speciation. *New Phytologist, 161*(1), 173-191.

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., & Neale, D. B. (2013). Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics & Genomes, 9*(4), 901-911.

Statistics New Zealand. (2012). Dairy industry 'mooooving' forward Retrieved February, 2017, from http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/yearbook/environment/agriculture/dairy.aspx

Stitt, M., & Zeeman, S. C. (2012). Starch turnover: pathways, regulation and role in growth. *Current Opinion in Plant Biology, 15*(3), 282-292.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America, 100*(16), 9440-9445.

Storfer, A., Patton, A., & Fraik, A. K. (2018). Navigating the Interface Between Landscape Genetics and Landscape Genomics. *Frontiers in Genetics, 9*(68).

Sturman, A., & Wanner, H. (2001). A Comparative Review of the Weather and Climate of the Southern Alps of New Zealand and the European Alps. *Mountain Research and Development, 21*(4), 359-369.

Suckling, F. E. T. (1950). The passage of white clover sleeds through the body of sheep and the effect on germination capacity. *Proceedings of the New Zealand Grassland Association, 12*, 108-121.

Suckling, F. E. T., & Charlton, J. F. L. (1978). A review of the significance of buried legume seeds with particular reference to New Zealand agriculture. *New Zealand Journal of Experimental Agriculture, 6*(3), 211-215.

Sul, J. H., & Eskin, E. (2013). Mixed models can correct for population structure for genomic regions under selection. *Nature Reviews Genetics, 14*(4), 300-300.

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

Sun, C., Palmqvist, S., Olsson, H., Borén, M., Ahlandsberg, S., & Jansson, C. (2003). A Novel WRKY Transcription Factor, SUSIBA2, Participates in Sugar Signaling in Barley by Binding to the Sugar-Responsive Elements of the *iso1* Promoter. *The Plant Cell, 15*(9), 2076.

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., . . . Zheng, H. (2013). SLAF-seq: An Efficient Method of Large-Scale De Novo SNP Discovery and Genotyping Using High-Throughput Sequencing. *PLOS ONE, 8*(3), e58700.

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology, 2*(2), 125-141.

Swarts, K., Li, H., Romero Navarro, J. A., An, D., Romay, M. C., Hearne, S., . . . Bradbury, P. J. (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *The Plant Genome, 7*(3).

Syed, F., Grunenwald, H., & Caruccio, N. (2009). Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition. *Nature Methods, 6*(11), i-ii.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., . . . Mering, C. V. (2019). STRING v11: protein-protein association networks with

increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res, 47*(D1), D607-d613.

Tahir, F., Hassani, A., Kouadria, M., & Rezzoug, W. (2019). Study of Morpho-Physiological and Biochemical Behavior of Cultivated Legume (*Lens culinaris Medik* Ssp *culinaris*) in Dry Area of Algeria. *Ukrainian Journal of Ecology, 9*(4), 535-541.

Taiz, L., & Zeiger, E. (2010a). Chapter 2: Genome Organization and Gene Expression. *Plant Physiology* (5th ed., pp. 35-63). Sunderland, MA, USA: Sinauer Associates.

Taiz, L., & Zeiger, E. (2010b). Chapter 6: Solute Transport. *Plant Physiology* (5th ed., pp. 131-159). Sunderland, MA, USA: Sinauer Associates.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics, 123*(3), 585.

Takahashi, Y., Shomura, A., Sasaki, T., & Yano, M. (2001). Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the α subunit of protein kinase CK2. *Proceedings of the National Academy of Sciences of the United States of America, 98*(14), 7922-7927.

Tan, J., Wang, C., Xiang, B. I. N., Han, R., & Guo, Z. (2013). Hydrogen peroxide and nitric oxide mediated cold- and dehydration-induced myo-inositol phosphate synthase that confers multiple resistances to abiotic stresses. *Plant, Cell & Environment, 36*(2), 288-299.

Tas, B. M., Taweel, H. Z., Smit, H. J., Elgersma, A., Dijkstra, J., & Tamminga, S. (2005). Effects of Perennial Ryegrass Cultivars on Intake, Digestibility, and Milk Yield in Dairy Cows. *Journal of Dairy Science, 88*(9), 3240-3248.

Tas, B. M., Taweel, H. Z., Smit, H. J., Elgersma, A., Dijkstra, J., & Tamminga, S. (2006). Effects of Perennial Ryegrass Cultivars on Milk Yield and Nitrogen Utilization in Grazing Dairy Cows. *Journal of Dairy Science, 89*(9), 3494-3500.

Taweel, H. Z., Tas, B. M., Smit, H. J., Elgersma, A., Dijkstra, J., & Tamminga, S. (2005). Effects of feeding perennial ryegrass with an elevated concentration of water-soluble carbohydrates on intake, rumen function and performance of dairy cows. *Animal Feed Science and Technology, 121*(3), 243-256.

Taweel, H. Z., Tas, B. M., Smit, H. J., Elgersma, A., Dijkstra, J., & Tamminga, S. (2006). Grazing behaviour, intake, rumen function and milk production of dairy cows offered Lolium perenne containing different levels of water-soluble carbohydrates. *Livestock Science, 102*(1), 33-41.

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., & Gaut, B. S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proceedings of the National Academy of Sciences of the United States of America, 98*(16), 9161-9166.

Tenaillon, M. I., U'Ren, J., Tenaillon, O., & Gaut, B. S. (2004). Selection Versus Demography: A Multilocus Investigation of the Domestication Process in Maize. *Molecular Biology and Evolution, 21*(7), 1214-1225.

Teshima, K. M., Coop, G., & Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Research, 16*(6), 702-712.

Thomas, R. G. (1987a). The structure of the mature plant. In M. J. Baker & W. M. Williams (Eds.). Wallingford: CAB International.

Thomas, R. G. (1987b). Vegetative growth and development. In M. J. Baker & W. M. Williams (Eds.). Wallingford: CAB International.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., . . . Su, Z. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic acids research, 45*(W1), W122-W129.

Tiessen, A., Hendriks, J. H. M., Stitt, M., Branscheid, A., Gibon, Y., Farré, E. M., & Geigenberger, P. (2002). Starch Synthesis in Potato Tubers Is Regulated by Post-Translational Redox Modification of ADP-Glucose Pyrophosphorylase. *The Plant Cell, 14*(9), 2191.

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

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ, 1*, e203.

Torkamaneh, D., Laroche, J., & Belzile, F. (2016). Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. *PLoS ONE, 11*(8), e0161333.

Trabucco, A., & Zomer, R. (2019). Global Aridity Index and Potential Evapotranspiration (ET0) Climate Database v2. In figshare (Ed.).

Trucchi, E., Mazzarella, A. B., Gilfillan, G. D., Lorenzo, M. T., Schönswetter, P., & Paun, O. (2016). BsRADseq: screening DNA methylation in natural populations of non-model species. *Molecular ecology, 25*(8), 1697-1713.

Turkington, R. O. Y., & Burdon, J. J. (1983). The Biology of Canadian Weeds: 57. *Trifolium repens* L. *Canadian Journal of Plant Science, 63*(1), 243-266.

Turner, L. B. (1990). The Extent and Pattern of Osmotic Adjustment in White Clover (*Trifolium repens* L.) During the Development of Water Stress. *Annals of Botany, 66*(6), 721-727.

Turner, L. B., Cairns, A. J., Armstead, I. P., Ashton, J., Skøt, K., Whittaker, D., & Humphreys, M. O. (2006). Dissecting the regulation of fructan metabolism in perennial ryegrass (*Lolium perenne*) with quantitative trait locus mapping. *New Phytologist, 169*(1), 45-58.

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

Uga, Y., Sugimoto, K., Ogawa, S., Rane, J., Ishitani, M., Hara, N., . . . Yano, M. (2013). Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions. *Nature Genetics, 45*(9), 1097-1102.

Ulyatt, M. J. (1997). Can protein utilisation from pasture be improved? *Proceedings of the New Zealand Society of Animal Production, 57*, 4-8.

Ulyatt, M. J., Lancashire, J. A., & Jones, W. T. (1977). The nutritive value of legumes. *Proceedings of the New Zealand Grassland Association, 38*, 107-118.

Umina, P. A., Weeks, A. R., Kearney, M. R., McKechnie, S. W., & Hoffmann, A. A. (2005). A rapid shift in a classic clinal pattern in *Drosophila* reflecting climate change. *Science, 308*(5722), 691-693.

Unamba, C. I. N., Nag, A., & Sharma, R. K. (2015). Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. *Frontiers in Plant Science, 6*, 1074.

van Den Bosch, J., Black, I. K., Cousins, G. R., & Woodfield, D. R. (1993). Enhanced drought tolerance in white clover. *Proceedings of the New Zealand Grassland Association, 55*, 97-101.

van Gurp, T. P., Wagemaker, N. C. A. M., Wouters, B., Vergeer, P., Ouborg, J. N. J., & Verhoeven, K. J. F. (2016). epiGBS: reference-free reduced representation bisulfite sequencing. *Nature Methods, 13*(4), 322-324.

van Ham, R., O'Callaghan, M., Geurts, R., Ridgway, H. J., Ballard, R., Noble, A., . . . Wakelin, S. A. (2016). Soil moisture deficit selects for desiccation tolerant *Rhizobium leguminosarum* bv. *trifolii*. *Applied Soil Ecology, 108*, 371-380.

Van Harsselaar, J. K., Lorenz, J., Senning, M., Sonnewald, U., & Sonnewald, S. (2017). Genome-wide analysis of starch metabolism genes in potato (*Solanum tuberosum* L.). *BMC Genomics, 18*(1), 37.

Vaseva, I., Akiscan, Y., Demirevska, K., Anders, I., & Feller, U. (2011). Drought stress tolerance of red and white clover-comparative analysis of some chaperonins and dehydrins. *Scientia Horticulturae, 130*(3), 653-659.

Vaseva, I. I., Anders, I., & Feller, U. (2014). Identification and expression of different dehydrin subclasses involved in the drought response of *Trifolium repens*. *Journal of Plant Physiology, 171*(3-4), 213-224.

Vavilov, N. I. (1992). *Origin and geography of cultivated plants*. Cambridge: Cambridge University Press.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Vigouroux, Y., McMullen, M., Hittinger, C. T., Houchins, K., Schulz, L., Kresovich, S., . . . Doebley, J. (2002a). Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proceedings of the National Academy of Sciences, 99*(15), 9650.

Vigouroux, Y., McMullen, M., Hittinger, C. T., Houchins, K., Schulz, L., Kresovich, S., . . . Doebley, J. (2002b). Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proceedings of the National Academy of Sciences, 99*(15), 9650-9655.

Villand, P., Olsen, O.-A., & Kleczkowski, L. A. (1993). Molecular characterization of multiple cDNA clones for ADP-glucose pyrophosphorylase from *Arabidopsis thaliana*. *Plant Molecular Biology, 23*(6), 1279-1284.

Voelckel, C., Gruenheit, N., & Lockhart, P. (2017). Evolutionary Transcriptomics and Proteomics: Insight into Plant Adaptation. *Trends in Plant Science, 22*(6), 462-471.

Voelckel, C., Mirzaei, M., Reichelt, M., Luo, Z., Pascovici, D., Heenan, P. B., . . . Lockhart, P. J. (2010). Transcript and protein profiling identify candidate gene sets of potential adaptive significance in New Zealand *Pachycladon*. *BMC Evolutionary Biology, 10*(1), 151.

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics, 13*(4), 227-232.

VSN International. (2015). Genstat for Windows 18th Edition.

Wang, J., Drayton, M. C., George, J., Cogan, N. O. I., Baillie, R. C., Hand, M. L., . . . Smith, K. F. (2010). Identification of genetic factors influencing salt stress

tolerance in white clover (*Trifolium repens* L.) by QTL analysis. *Theoretical and Applied Genetics, 120*(3), 607-619.

Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods, 9*(8), 808-810.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics, 6*(2), 109-118.

Weber, H., Heim, U., Borisjuk, L., & Wobus, U. (1995). Cell-type specific, coordinate expression of two ADP-glucose pyrophosphorylase genes in relation to starch biosynthesis during seed development of Vicia faba L. *Planta, 195*(3), 352-361.

Weber, H., Heim, U., Golombek, S., Borisjuk, L., Manteuffel, R., & Wobus, U. (1998). Expression of a yeast-derived invertase in developing cotyledons of *Vicia narbonensis* alters the carbohydrate state and affects storage functions. *Plant J, 16*(2), 163-172.

Wedderburn, M. E., Adam, K. D., Greaves, L. A., & Carter, J. L. (1996). Effect of oversown ryegrass (*Lolium perenne*) and white clover (*Trifolium repens*) on the genetic structure of New Zealand hill pastures. *New Zealand Journal of Agricultural Research, 39*(1), 41-52.

Wei, T., & Simko, V. (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Retrieved from https://github.com/taiyun/corrplot

Wei, X., Liu, F., Chen, C., Ma, F., & Li, M. (2014). The *Malus domestica* sugar transporter gene family: identifications based on genome and expression profiling related to the accumulation of fruit sugars. *Frontiers in Plant Science, 5*(569).

Weigel, D., & Nordborg, M. (2005). Natural Variation in *Arabidopsis*. How Do We Find the Causal Genes? *Plant Physiology, 138*(2), 567.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution, 38*(6), 1358-1370.

Werner, J. D., Borevitz, J. O., Warthmann, N., Trainer, G. T., Ecker, J. R., Chory, J., & Weigel, D. (2005). Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. *Proceedings of the National Academy of Sciences of the United States of America, 102*(7), 2460-2465.

Westbrooks, F. E., & Tesar, M. B. (1955). Tap root survival of ladino clover. *Agronomy Journal, 47*(9), 403-410.

Wheeler, B., & Torchiano, M. (2016). lmPerm: Permutation Tests for Linear Models. R package version 2.1.0.

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of F_{ST}. *The American Naturalist, 186*, S24-S36.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software, 40*, i01.

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

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage

crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics, 18*(1), 586.

Widdup, K. H., & Barrett, B. A. (2011). Achieving persistence and productivity in white clover. *Pasture Persistence – Grassland Research and Practice, Series 15*, 173-180.

Widdup, K. H., Ford, J. L., Barrett, B. A., & Woodfield, D. R. (2010). Development of white clover populations with higher concentrations of water soluble carbohydrate. *Proceedings of the New Zealand Grassland Association, 72*, 277-282.

Williams, W. M. (1983). Chapter 25 White Clover. In G. S. Wratt & H. C. Smith (Eds.), *Plant Breeding in New Zealand* (pp. 221-228). Wellington, NZ: Butterworths/DSIR.

Williams, W. M. (1987a). Adaptive variation. In M. J. Baker & W. M. Williams (Eds.), *White Clover* (pp. 299-321). Wallingford, UK: CAB International.

Williams, W. M. (1987b). Genetics and breeding. In M. J. Baker & W. M. Williams (Eds.), *White Clover*. Wallingford, UK: CABI Publishing.

Williams, W. M., Easton, H. S., & Jones, C. S. (2007). Future options and targets for pasture plant breeding in New Zealand. *New Zealand Journal of Agricultural Research, 50*, 223-248.

Williams, W. M., Ellison, N. W., Ansari, H. A., Verry, I. M., & Hussain, S. W. (2012). Experimental evidence for the ancestry of allotetraploid *Trifolium repens* and creation of synthetic forms with value for plant breeding. *BMC Plant Biology, 12*(55).

Woodfield, D. R., & Caradus, J. R. (1996). Factors affecting white clover persistence in New Zealand pastures. *Proceedings of the New Zealand Grassland Association, 58*, 229-235.

Woodfield, D. R., & Clark, D. A. (2009). Do forage legumes have a role in modern dairy farming systems? *Irish Journal of Agricultural and Food Research, 48*(2), 137-147.

Woodfield, D. R., Clifford, P. T. P., Baird, I. J., & Cousins, G. R. (Eds.). (1995). *Gene flow and estimated isolation requirements for transgenic white clover.* University of California, Oakland, USA.

Woodfield, D. R., Clifford, P. T. P., baird, I. J., Cousins, G. R., Miller, J. E., Widdup, K. H., & Caradus, J. R. (2003). Grasslands Tribute: a multi-purpose white clover for Australasia. *Proceedings of the New Zealand Grassland Association, 65*, 157-162.

Woodfield, D. R., Clifford, P. T. P., Cousins, G. R., Ford, J. L., Baird, I. J., Miller, J. E., . . . Caradus, J. R. (2001). Grasslands Kopu II and Crusader: new generation white clovers. *Proceedings of the New Zealand Grassland Association, 63*, 103-108.

Wormit, A., Trentmann, O., Feifer, I., Lohr, C., Tjaden, J., Meyer, S., . . . Neuhaus, H. E. (2006). Molecular Identification and Physiological Characterization of a Novel Monosaccharide Transporter from *Arabidopsis* Involved in Vacuolar Sugar Transport. *The Plant Cell, 18*(12), 3476-3490.

Wright, S. (1939). The distribution of self-sterility alleles in populations. *Genetics, 24*(4), 538.

Wright, S. (1951). The genetical structure of populations. *Annals of eugenics, 15*(4), 323-354.

Wright, S. (1978). *Evolution and the genetics of populations*. Vol. 4. Variability within and among natural populations. Chicago: University of Chicago Press.

Wright, S. J., Cui Zhou, D., Kuhle, A., & Olsen, K. M. (2017). Continent-Wide Climatic Variation Drives Local Adaptation in North American White Clover. *Journal of Heredity, 109*(1), 78-89.

Yamada, K., Osakabe, Y., Mizoi, J., Nakashima, K., Fujita, Y., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2010). Functional analysis of an *Arabidopsis thaliana* abiotic stress-inducible facilitated diffusion transporter for monosaccharides. *J Biol Chem, 285*(2), 1138-1146.

Yamada, T., & Kawaguchi, T. (1972). Dissemination of pasture plants by livestock. 2. Recovery, viability and emergence of some pasture plant seeds passed through the digestive tract of the dairy cow. *Journal of Japanese Society of Grassland Science, 18*(1), 8-15.

Yamazaki, M., Shimada, T., Takahashi, H., Tamura, K., Kondo, M., Nishimura, M., & Hara-Nishimura, I. (2008). *Arabidopsis* VPS35, a Retromer Component, is Required for Vacuolar Protein Sorting and Involved in Plant Growth and Leaf Senescence. *Plant and Cell Physiology, 49*(2), 142-156.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., . . . Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics, 42*(7), 565-569.

Yang, Z., Nie, G., Pan, L., Zhang, Y., Huang, L., Ma, X., & Zhang, X. (2017). Development and validation of near-infrared spectroscopy for the prediction of forage quality parameters in *Lolium multiflorum*. *PeerJ, 5*, e3867-e3867.

Yano, M., Katayose, Y., Ashikari, M., Yamanouchi, U., Monna, L., Fuse, T., . . . Sasaki, T. (2000). *Hd1*, a Major Photoperiod Sensitivity Quantitative Trait Locus in Rice, Is Closely Related to the *Arabidopsis* Flowering Time Gene *CONSTANS*. *The Plant Cell, 12*(12), 2473.

Yu, J., & Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology, 17*(2), 155-160.

Yu, X., Bai, G., Liu, S., Luo, N., Wang, Y., Richmond, D. S., . . . Jiang, Y. (2013). Association of candidate genes with drought tolerance traits in diverse perennial ryegrass accessions. *J Exp Bot, 64*(6), 1537-1551.

Yu, X., Pijut, P. M., Byrne, S., Asp, T., Bai, G., & Jiang, Y. (2015). Candidate gene association mapping for winter survival and spring regrowth in perennial ryegrass. *Plant Science, 235*, 37-45.

Yuan, Y., Zhang, Q., Zeng, S., Gu, L., Si, W., Zhang, X., . . . Wang, L. (2017). Selective sweep with significant positive selection serves as the driving force for the differentiation of japonica and indica rice cultivars. *BMC Genomics, 18*(1), 307.

Zeven, A. C. (1991). Four hundred years of cultivation of Dutch white clover landraces. *Euphytica, 54*(1), 93-99.

Zhang, M., Kimatu, J. N., Xu, K., & Liu, B. (2010a). DNA cytosine methylation in plant development. *Journal of Genetics and Genomics, 37*(1), 1-12.

Zhang, T., Yu, L.-X., Zheng, P., Li, Y., Rivera, M., Main, D., & Greene, S. L. (2015). Identification of Loci Associated with Drought Resistance Traits in Heterozygous

Autotetraploid Alfalfa (*Medicago sativa* L.) Using Genome-Wide Association Studies with Genotyping by Sequencing. *PLOS ONE, 10*(9), e0138931.

Zhang, Y. (2016). On The Use of P-Values in Genome Wide Disease Association Mapping. *Journal of Biometrics & Biostatistics, 7*(3), 1-2.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., . . . Buckler, E. S. (2010b). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics, 42*(4), 355-360.

Zhao, X., Bushman, B. S., Zhang, X., Robbins, M. D., Larson, S. R., Robins, J. G., & Thomas, A. (2017). Association of candidate genes with heading date in a diverse Dactylis glomerata population. *Plant Science, 265*, 146-153.

Zhou, Z. J., Morel, J., Parsons, D., Kucheryavskiy, S. V., & Gustaysson, A. M. (2019). Estimation of yield and quality of legume and grass mixtures using partial least squares and support vector machine analysis of spectral data. *Computers and Electronics in Agriculture, 162*, 246-253.

Zhu, A., Ibrahim, J. G., & Love, M. I. (2018). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics, 35*(12), 2084-2092.

Zhuo, C., Wang, T., Lu, S., Zhao, Y., Li, X., & Guo, Z. (2013). A cold responsive galactinol synthase gene from *Medicago falcata* (MfGolS1) is induced by myo-inositol and confers multiple tolerances to abiotic stresses. *Physiologia Plantarum, 149*(1), 67-78.

APPENDIX 1

Chapter 2 Supplementary Material

SUPPLEMENTARY TABLES

Table S2.1 Latin square design generated in GenStat (v 18). First replicate block is from column 1 to 10, second replicate block is from column 11 – 20, and third replicate block is from column 21 – 30, blocks are separated by vertical dashed lines. High = high water-soluble carbohydrate (WSC), and Low = low WSC. Colour codes correspond to generations: Parent = Parent generation (yellow), Mid = Middle generation (blue), and End = End generation (green). Table continued on next page.

	Columns														
Rows	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	FNZSL-High-Mid	WNZSL-Parent	WNZSL-High-End	WNZSL-Low-End	WNZSL-High-Mid	WNZLL-High-Mid	FNZLL-Low-End	WNZLL-Parent	WNZLL-Parent	FNZSL-Parent	WUSLL-Low-Mid	FNZSL-Low-End	FNZLL-Parent	WUSLL-Low-End	WNZSL-Parent
2	FNZLL-Low-End	FNZLL-High-Mid	WNZLL-Low-Mid	FNZSL-Parent	FNZLL-Low-Mid	WNZSL-High-Mid	FNZLL-High-End	WNZSL-High-End	WUSLL-Low-Mid	WNZLL-Low-End	FNZSL-Low-End	WNZLL-Parent	WNZLL-High-End	WUSLL-High-End	FNZSL-Parent
3	WNZSL-Parent	WUSLL-High-Mid	FNZLL-Parent	FNZSL-Parent	FNZLL-Parent	WUSLL-Parent	FNZLL-High-Mid	WUSLL-Parent	WNZSL-High-Mid	WNZSL-Low-Mid	FNZLL-Low-Mid	WNZLL-High-Mid	FNZLL-Low-End	WUSLL-Low-Mid	WNZSL-High-End
4	FNZSL-Parent	WNZSL-Low-Mid	WNZLL-High-Mid	WUSLL-High-Mid	FNZSL-High-Mid	FNZLL-High-End	WNZLL-Low-End	FNZLL-Low-Mid	WUSLL-Parent	WUSLL-Low-End	FNZLL-Parent	WNZLL-High-End	WUSLL-Parent	WNZSL-High-End	FNZSL-Low-End
5	FNZLL-Parent	FNZLL-Low-End	FNZSL-Parent	FNZSL-High-End	WUSLL-Low-Mid	WNZLL-Parent	WNZLL-High-End	WNZSL-Parent	WUSLL-Low-End	WUSLL-Parent	WUSLL-High-End	WUSLL-High-Mid	WNZLL-Low-Mid	FNZSL-Low-Mid	WNZSL-Low-End
6	WNZLL-Low-End	WNZSL-Parent	WNZSL-High-Mid	WUSLL-Low-End	FNZLL-Low-End	FNZSL-High-Mid	WNZSL-Low-End	WNZLL-High-Mid	FNZLL-Parent	WUSLL-High-End	WNZLL-High-End	WUSLL-Parent	FNZLL-Parent	WNZLL-Low-Mid	WNZLL-Parent
7	WNZSL-Low-End	FNZSL-Parent	FNZLL-Low-Mid	WUSLL-High-End	FNZLL-High-End	FNZLL-Low-End	FNZSL-Parent	WNZSL-High-Mid	WNZLL-High-End	WUSLL-High-Mid	WUSLL-Parent	FNZLL-Parent	FNZSL-High-End	WNZLL-Parent	WUSLL-Low-Mid
8	WUSLL-Low-Mid	FNZLL-Low-Mid	FNZSL-Low-Mid	WUSLL-Parent	WNZSL-Low-Mid	FNZSL-Parent	FNZSL-Low-End	FNZLL-High-Mid	FNZSL-Parent	FNZLL-Parent	WNZLL-Low-End	WNZSL-Low-End	WUSLL-High-End	WUSLL-Parent	FNZLL-High-End
9	FNZSL-High-End	FNZSL-Parent	FNZSL-Low-End	WNZSL-Parent	WUSLL-Parent	WNZLL-High-End	WUSLL-Parent	WUSLL-Low-Mid	WNZLL-Parent	FNZLL-High-Mid	WNZSL-High-End	WNZLL-Low-Mid	WNZLL-High-Mid	WNZSL-Parent	WUSLL-High-End
10	WNZLL-High-End	FNZLL-High-End	WNZSL-Low-Mid	WUSLL-Parent	FNZSL-Low-End	WUSLL-Low-Mid	WUSLL-Parent	FNZSL-Parent	WUSLL-High-End	FNZLL-Parent	WUSLL-High-Mid	WUSLL-Low-End	WNZLL-Parent	WNZSL-Parent	FNZSL-Parent
11	WUSLL-High-Mid	WNZLL-Parent	FNZLL-Low-End	WNZLL-Low-Mid	WNZSL-Low-End	WNZLL-Low-End	WUSLL-Low-End	FNZSL-High-Mid	FNZLL-Parent	WNZLL-Parent	FNZSL-High-End	WUSLL-Parent	FNZLL-High-Mid	FNZLL-Low-Mid	FNZLL-Parent
12	WNZSL-Low-Mid	WNZLL-Parent	WUSLL-Parent	WNZLL-Parent	WNZSL-Parent	FNZSL-Low-Mid	WNZSL-Parent	FNZSL-High-End	FNZSL-High-Mid	WUSLL-Low-Mid	FNZLL-Low-End	FNZLL-High-End	WNZLL-Low-End	FNZLL-Parent	WNZLL-High-Mid
13	WUSLL-Low-End	WUSLL-Low-Mid	FNZLL-High-End	WNZLL-Parent	FNZSL-Parent	WNZSL-Low-End	WUSLL-High-End	FNZLL-Low-End	FNZSL-High-End	WNZSL-High-End	WUSLL-Parent	FNZLL-Parent	FNZSL-Low-Mid	WNZLL-High-Mid	WNZLL-High-End
14	WNZSL-Parent	WNZSL-High-End	FNZLL-Parent	WUSLL-Low-Mid	FNZLL-High-Mid	WNZSL-Parent	FNZSL-Parent	WUSLL-Parent	FNZLL-Low-End	FNZSL-Low-End	FNZLL-High-End	FNZSL-High-Mid	WNZSL-Low-End	WNZLL-High-End	WNZSL-High-Mid
15	WNZSL-High-End	FNZLL-Parent	WNZSL-Low-End	FNZLL-Low-Mid	WUSLL-Low-End	WUSLL-High-Mid	WNZLL-Low-Mid	WNZLL-Low-End	FNZLL-High-Mid	WNZLL-High-Mid	FNZSL-Low-Mid	WNZSL-Parent	FNZSL-Parent	FNZLL-High-End	FNZLL-Parent
16	WNZLL-Parent	WNZSL-High-Mid	FNZLL-High-Mid	WNZLL-High-End	FNZSL-Parent	WNZSL-Parent	WUSLL-Low-Mid	WNZSL-Parent	WNZSL-Low-End	WUSLL-Parent	FNZSL-Parent	WNZLL-Low-End	WUSLL-Low-End	FNZSL-High-End	FNZLL-Low-End
17	FNZSL-Parent	WNZLL-Low-Mid	FNZSL-High-End	FNZSL-Low-End	FNZSL-Low-Mid	FNZLL-High-Mid	WNZSL-Low-Mid	FNZLL-Parent	FNZLL-High-End	WNZLL-Parent	FNZSL-High-Mid	FNZLL-Low-End	FNZSL-Parent	WUSLL-Parent	FNZLL-Low-Mid
18	WUSLL-High-End	FNZSL-Low-End	FNZSL-High-Mid	WNZSL-High-End	WNZLL-Low-End	FNZSL-Parent	WUSLL-High-Mid	FNZLL-High-End	WUSLL-Parent	WNZLL-Low-Mid	FNZLL-Parent	FNZSL-High-End	WNZSL-Parent	WNZSL-High-Mid	WUSLL-Parent
19	FNZSL-Low-Mid	WUSLL-High-End	WUSLL-Parent	WNZSL-Parent	WUSLL-Parent	FNZSL-High-End	WNZSL-Parent	WNZLL-High-End	WNZLL-High-Mid	FNZSL-Parent	WNZSL-High-Mid	FNZLL-Low-Mid	FNZSL-High-Mid	WNZLL-Parent	WNZLL-Parent
20	FNZSL-Low-End	WNZLL-High-Mid	WNZSL-Parent	FNZLL-Parent	WNZSL-Parent	WNZSL-Low-Mid	WNZLL-Parent	FNZSL-Low-Mid	WNZLL-Low-End	WNZLL-High-End	WNZSL-Low-End	FNZSL-Parent	WUSLL-High-Mid	FNZLL-Parent	FNZSL-High-Mid
21	WUSLL-Parent	WNZLL-Low-End	WNZLL-Parent	FNZLL-High-Mid	FNZLL-Parent	WUSLL-Parent	FNZLL-Parent	FNZSL-Low-End	WNZSL-High-End	FNZSL-Low-Mid	WNZLL-Low-Mid	WNZLL-Parent	WNZSL-High-Mid	FNZSL-Parent	WUSLL-High-Mid
22	FNZLL-Low-Mid	FNZSL-High-End	WUSLL-High-End	FNZSL-High-Mid	WNZLL-Parent	WNZLL-Low-Mid	WNZLL-High-Mid	WUSLL-Low-End	WNZSL-Low-Mid	FNZLL-Low-End	WNZSL-Parent	FNZSL-Parent	FNZSL-Low-End	WNZLL-Low-End	FNZSL-Low-Mid
23	WNZLL-Low-Mid	WNZLL-High-End	FNZSL-Parent	WNZLL-High-Mid	WUSLL-High-End	WUSLL-Low-End	WNZLL-Parent	WNZSL-Low-End	FNZSL-Low-Mid	WNZSL-High-Mid	WNZSL-Parent	FNZLL-High-Mid	WNZSL-Low-Mid	FNZSL-High-Mid	FNZSL-High-End
24	WNZLL-High-Mid	WUSLL-Parent	WUSLL-High-Mid	FNZLL-Low-End	WNZSL-High-End	WNZLL-Parent	WNZSL-High-Mid	WUSLL-High-End	WNZSL-Parent	FNZLL-High-End	FNZSL-Parent	WNZSL-Low-Mid	WNZLL-Parent	WNZSL-Low-End	WNZSL-Parent
25	FNZLL-Parent	WNZSL-Low-End	WUSLL-Low-Mid	FNZSL-Low-Mid	WNZLL-High-End	FNZLL-Parent	FNZSL-High-End	WNZLL-Parent	WNZLL-Low-Mid	WNZSL-Parent	WNZLL-Parent	WNZSL-High-End	FNZLL-Low-Mid	WNZSL-Low-Mid	WUSLL-Low-End
26	WNZLL-Parent	WUSLL-Parent	WNZLL-Low-End	WNZSL-High-Mid	WUSLL-High-Mid	WUSLL-High-End	WNZSL-High-End	FNZSL-Parent	WNZSL-Parent	FNZLL-Low-Mid	FNZLL-High-Mid	FNZSL-Low-Mid	WNZSL-Parent	FNZLL-Low-End	WUSLL-Parent
27	FNZLL-High-End	FNZSL-Low-Mid	WNZLL-Parent	WNZLL-Low-End	WNZLL-High-Mid	FNZLL-Low-Mid	FNZSL-High-Mid	WNZLL-Low-Mid	FNZSL-Low-End	WNZSL-Low-End	WNZLL-Parent	WUSLL-Low-Mid	WUSLL-Parent	WUSLL-High-Mid	WNZSL-Low-Mid
28	FNZLL-High-Mid	WUSLL-Low-End	WNZLL-High-End	WNZSL-Low-Mid	FNZSL-High-End	FNZLL-Parent	FNZSL-Low-Mid	FNZLL-Parent	FNZLL-Low-Mid	WNZSL-Parent	WNZLL-High-Mid	WNZSL-High-Mid	FNZLL-High-End	FNZSL-Low-End	WNZLL-Low-Mid
29	WUSLL-Parent	FNZSL-High-Mid	WNZSL-Parent	FNZLL-Parent	WNZLL-Parent	FNZSL-Low-End	FNZLL-Parent	WNZSL-Low-Mid	WUSLL-High-Mid	FNZSL-High-End	WUSLL-Low-End	WUSLL-High-End	WNZSL-High-End	FNZLL-High-Mid	WNZLL-Low-End
30	WNZSL-High-Mid	FNZLL-Parent	WUSLL-Low-End	FNZLL-High-End	WNZLL-Low-Mid	WNZSL-High-End	FNZLL-Low-Mid	WUSLL-High-Mid	FNZSL-Parent	FNZSL-High-Mid	WNZSL-Low-Mid	WNZSL-Parent	WUSLL-Low-Mid	FNZSL-Parent	FNZLL-High-Mid

Table S2.1 (continued)

Rows	Columns 16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	FNZLL-High-End	FNZSL-High-End	FNZLL-Low-Mid	FNZSL-Parent	WNZLL-High-End	FNZLL-High-Mid	WUSLL-High-End	WNZSL-Low-Mid	FNZSL-Low-Mid	WUSLL-Parent	WUSLL-Parent	WNZLL-Low-End	WNZLL-Low-Mid	FNZLL-Parent	WUSLL-High-Mid
2	FNZSL-High-Mid	WUSLL-Parent	WNZLL-High-Mid	WNZSL-Low-Mid	WUSLL-Parent	FNZSL-Low-Mid	WUSLL-High-Mid	WNZSL-Parent	WNZSL-Parent	FNZLL-Parent	FNZLL-Parent	WNZSL-Low-End	WNZLL-Parent	FNZSL-High-End	WUSLL-Low-End
3	FNZSL-Low-Mid	FNZSL-Parent	FNZSL-High-End	WNZLL-Low-Mid	FNZLL-High-End	WUSLL-Low-End	FNZSL-Low-End	WNZLL-Parent	WUSLL-High-End	FNZSL-High-Mid	WNZLL-Low-End	WNZSL-Parent	WNZLL-High-End	WNZSL-Low-End	WNZLL-Parent
4	WNZSL-Low-End	FNZLL-High-Mid	FNZLL-Low-End	WNZLL-Parent	FNZLL-Parent	WNZSL-Parent	WNZLL-Low-Mid	WUSLL-Low-Mid	FNZSL-Parent	FNZSL-High-End	FNZSL-Low-Mid	WUSLL-High-End	WNZSL-High-Mid	WNZSL-Parent	WNZLL-Parent
5	WUSLL-Parent	WNZLL-High-Mid	FNZSL-Low-End	FNZSL-Parent	WNZLL-Parent	FNZLL-High-End	WNZSL-Parent	WNZLL-Low-End	FNZSL-High-Mid	WNZSL-High-End	WNZSL-High-Mid	FNZLL-Parent	WNZSL-Low-Mid	FNZLL-Low-Mid	FNZLL-High-Mid
6	FNZSL-Parent	FNZSL-Low-Mid	FNZLL-High-End	WUSLL-Low-Mid	FNZSL-High-End	FNZSL-Parent	WNZLL-Parent	FNZSL-Low-End	WNZSL-Low-Mid	WUSLL-Parent	WNZSL-Parent	WUSLL-High-Mid	FNZLL-Low-Mid	FNZLL-High-Mid	WNZSL-High-End
7	WNZLL-Low-End	WNZSL-Parent	FNZSL-High-Mid	FNZSL-Low-End	WUSLL-Parent	WNZSL-Low-Mid	WNZSL-High-End	WNZLL-Parent	WNZSL-Parent	FNZLL-Parent	FNZLL-High-Mid	WUSLL-Low-End	WNZLL-High-Mid	FNZSL-Low-Mid	WNZLL-Low-Mid
8	WNZLL-Parent	WNZSL-High-End	WNZSL-Parent	FNZSL-High-Mid	WUSLL-High-Mid	WNZLL-High-Mid	FNZLL-Parent	FNZLL-Low-End	WNZSL-High-Mid	WUSLL-Low-End	WNZLL-Low-Mid	WNZLL-High-End	WNZSL-Parent	WNZLL-Parent	FNZSL-High-End
9	FNZLL-Parent	FNZLL-Low-End	FNZLL-Parent	WUSLL-High-Mid	WNZSL-High-Mid	WNZLL-Low-End	FNZSL-Parent	WUSLL-Low-End	WNZSL-Low-End	FNZLL-Low-Mid	FNZLL-High-End	FNZSL-Low-Mid	WNZLL-Parent	FNZSL-High-Mid	WNZSL-Low-Mid
10	FNZLL-Parent	WNZSL-High-Mid	WNZLL-Parent	WNZLL-Low-End	WNZSL-High-End	FNZSL-High-Mid	FNZLL-High-Mid	WNZSL-Low-End	FNZLL-Low-End	WNZLL-Low-Mid	FNZLL-Low-Mid	FNZSL-High-End	WNZSL-Parent	WNZLL-High-Mid	FNZSL-Low-Mid
11	WUSLL-High-End	WNZSL-Low-Mid	FNZSL-Parent	WNZLL-High-End	FNZSL-Low-Mid	WUSLL-Low-Mid	WNZLL-High-Mid	WUSLL-Parent	FNZSL-Low-End	WNZSL-Parent	WNZSL-Parent	WNZSL-High-End	FNZLL-High-End	FNZSL-Parent	WNZSL-High-Mid
12	FNZSL-Parent	WUSLL-Low-End	FNZLL-High-Mid	WNZSL-High-Mid	WNZSL-Low-End	WNZSL-High-End	WNZLL-High-End	FNZLL-Low-Mid	WNZLL-Low-Mid	FNZSL-Parent	WUSLL-High-End	FNZSL-Low-End	FNZLL-Parent	WUSLL-High-Mid	WUSLL-Parent
13	WUSLL-High-Mid	WNZSL-Parent	WNZLL-Low-End	WUSLL-Parent	WNZSL-Parent	FNZSL-Low-End	WNZSL-High-Mid	FNZLL-Parent	WNZLL-Parent	FNZLL-High-Mid	FNZSL-Parent	WNZLL-Low-Mid	FNZSL-High-Mid	WNZSL-Low-Mid	FNZLL-Low-Mid
14	WNZSL-Low-Mid	WUSLL-High-End	FNZSL-Low-Mid	FNZLL-Low-Mid	FNZSL-Parent	WNZLL-Low-Mid	WUSLL-Parent	WNZLL-High-Mid	WNZLL-Parent	WNZLL-Low-End	WUSLL-High-Mid	WNZLL-Parent	FNZSL-High-End	WUSLL-Low-End	FNZLL-Parent
15	WNZLL-Parent	FNZSL-Low-End	WUSLL-High-End	FNZSL-High-End	WNZSL-Low-Mid	WNZLL-High-End	FNZSL-High-Mid	WUSLL-Parent	WUSLL-Parent	WNZSL-Parent	WNZLL-Parent	WNZSL-High-Mid	FNZSL-Parent	WUSLL-Low-Mid	FNZLL-Low-End
16	FNZSL-Low-End	WNZLL-Parent	WNZSL-Low-Mid	FNZLL-High-End	WUSLL-High-End	FNZLL-Low-Mid	WUSLL-Parent	FNZSL-High-Mid	WNZLL-High-Mid	WUSLL-High-Mid	WNZSL-High-End	FNZLL-Parent	FNZSL-Low-Mid	WNZLL-Low-Mid	FNZLL-Parent
17	WNZSL-Parent	WUSLL-High-Mid	WNZSL-Parent	WNZLL-High-Mid	WNZLL-Low-End	WNZLL-Parent	FNZLL-Parent	WNZSL-High-Mid	WNZSL-High-End	WNZSL-Low-End	WUSLL-Low-End	WUSLL-Low-Mid	WUSLL-Parent	WUSLL-High-End	WNZLL-High-End
18	WUSLL-Low-End	FNZSL-Parent	WNZSL-Low-End	FNZLL-Parent	FNZLL-High-Mid	WNZLL-Parent	FNZLL-Low-Mid	WNZLL-High-End	WUSLL-Low-Mid	FNZSL-Low-Mid	WNZSL-Low-Mid	WNZLL-Parent	FNZLL-Low-End	WNZSL-Parent	WNZLL-High-Mid
19	FNZLL-High-Mid	WNZSL-Low-End	FNZLL-Parent	WNZSL-High-End	FNZLL-Low-End	WUSLL-High-Mid	WUSLL-Low-Mid	WNZLL-Low-Mid	WUSLL-Low-End	FNZLL-High-End	FNZSL-Parent	WNZSL-Low-Mid	FNZLL-Parent	WNZLL-Low-End	FNZSL-Low-End
20	WUSLL-Low-Mid	WNZLL-Low-Mid	FNZSL-Parent	FNZLL-Low-End	WUSLL-Low-End	WNZSL-High-Mid	FNZSL-High-End	FNZLL-High-End	FNZLL-Low-Mid	WUSLL-High-End	WNZLL-Parent	WUSLL-Parent	FNZLL-High-Mid	WNZSL-High-End	WUSLL-Parent
21	FNZSL-High-End	FNZLL-High-End	WNZLL-High-End	WUSLL-Low-End	FNZLL-Low-Mid	WNZSL-Low-End	WNZSL-Low-Mid	WUSLL-High-End	FNZSL-Parent	WNZLL-High-Mid	FNZSL-High-Mid	WNZSL-Parent	WUSLL-Low-Mid	FNZLL-Low-End	WNZSL-Parent
22	WNZSL-High-Mid	FNZLL-Parent	WNZSL-High-End	WNZSL-Parent	WNZLL-Parent	WUSLL-Parent	WNZSL-Low-End	FNZLL-High-Mid	FNZLL-Parent	WUSLL-Low-Mid	WNZLL-High-End	FNZLL-High-End	WUSLL-High-Mid	WUSLL-Parent	FNZSL-Parent
23	WNZSL-High-End	WNZLL-Parent	WUSLL-High-Mid	WUSLL-Parent	WNZSL-Parent	WUSLL-Parent	FNZLL-Low-End	FNZLL-Parent	FNZLL-Parent	FNZSL-Parent	WUSLL-Low-Mid	FNZLL-Low-Mid	WNZLL-Low-End	FNZSL-Low-End	FNZLL-High-End
24	FNZLL-Low-Mid	WNZLL-High-End	WNZLL-Low-Mid	FNZLL-High-Mid	WUSLL-Low-Mid	FNZLL-Parent	FNZSL-Parent	FNZSL-Low-Mid	FNZSL-High-End	FNZSL-Low-End	WUSLL-Parent	FNZSL-High-Mid	WUSLL-Low-End	FNZLL-Parent	WNZLL-Low-End
25	WUSLL-Parent	FNZSL-High-Mid	WUSLL-Parent	WUSLL-High-End	WNZLL-High-Mid	FNZSL-Parent	WNZSL-Parent	WUSLL-High-Mid	WNZLL-Low-End	WNZSL-High-Mid	FNZLL-Low-End	FNZLL-High-Mid	FNZSL-Low-End	FNZLL-High-End	FNZSL-Parent
26	WNZLL-Low-Mid	WUSLL-Low-Mid	WUSLL-Low-End	FNZLL-Parent	FNZSL-Parent	FNZLL-Parent	FNZLL-High-End	FNZSL-High-End	WNZLL-High-End	WNZSL-Low-Mid	FNZSL-Low-End	WNZLL-High-Mid	WNZSL-Low-End	WNZLL-Parent	FNZSL-High-Mid
27	FNZLL-Low-End	FNZLL-Parent	WNZSL-High-Mid	WNZSL-Parent	FNZLL-Parent	WNZSL-Parent	WUSLL-Low-End	FNZSL-Parent	FNZLL-High-Mid	WNZLL-High-End	FNZSL-High-End	FNZSL-Parent	WNZSL-High-End	WUSLL-Parent	WUSLL-High-End
28	WNZSL-Parent	WNZLL-Low-End	WUSLL-Parent	WNZLL-Parent	FNZSL-High-Mid	WUSLL-High-End	WNZLL-Parent	WNZSL-High-End	WUSLL-High-Mid	FNZLL-Low-End	WNZSL-Low-End	FNZSL-Parent	WUSLL-Parent	FNZSL-Parent	WUSLL-Low-Mid
29	WNZLL-High-End	FNZLL-Low-Mid	WUSLL-Low-Mid	WNZSL-Low-End	WNZLL-Low-Mid	FNZLL-Low-End	FNZSL-Low-Mid	FNZSL-Parent	FNZLL-High-End	WNZLL-Parent	WNZLL-High-Mid	WUSLL-Parent	FNZSL-Parent	WNZSL-High-Mid	WNZSL-Parent
30	WNZLL-High-Mid	WUSLL-Parent	WNZLL-Parent	FNZSL-Low-Mid	FNZSL-Low-End	FNZSL-High-End	WNZLL-Low-End	WNZSL-Parent	WUSLL-Parent	WNZLL-Parent	FNZLL-Parent	FNZLL-Low-End	WUSLL-High-End	WNZLL-High-End	WNZSL-Low-End

Table S2.2 Regression information of four standard curves used for the estimation of anthrone (ANTH)-determined high molecular weight (ANTH-HMW) and low molecular weight (ANTH-LMW) water-soluble carbohydrate fractions, using inulin and sucrose as standards, respectively. Standard curves were used to determine HMW-ANTH and LMW-ANTH concentrations in white clover leaves from a subset of samples across the five pools.

Standards	Regression equation	Coefficient of determination (r^2)	*p*-value	*n*
LMW day 1	y = 192.55x - 33.817	0.962	0.001*	3-6
HMW day 1	y = 126.58x - 14.256	0.987	<0.001*	6
LMW day 2	y = 191.77x - 25.826	0.962	0.003*	15
HMW day 2	y = 122.29x - 14.269	0.988	0.001*	15

Note: LMW = low molecular weight, HMW = high molecular weight, and *n* = number of replicates used for 0 – 100 µg mL^{-1} standards.
* denotes $p < 0.05$ significance threshold.

Table S2.3 Mean, maximum (Max) and minimum (Min) concentration of anthrone-determined low molecular weight (LMW) and high molecular weight (HMW) water-soluble carbohydrate (WSC) levels in white clover leaves using 32 samples per pool.

WSC size fraction	Pool	*n*	Min (g kg^{-1})	Mean (g kg^{-1})	Max (g kg^{-1})
LMW	WNZLL	32	34.1	63.0	96.0
	WNZSL	32	26.1	50.7	96.6
	WUSLL	32	14.3	48.7	84.6
	FNZLL	32	18.8	50.8	96.6
	FNZSL	32	8.8	48.0	101.7
	Mean	**160**	**8.8**	**52.2**	**101.7**
HMW	WNZLL	32	0.0	2.6	6.2
	WNZSL	32	0.0	2.3	6.1
	WUSLL	32	0.0	2.9	6.7
	FNZLL	32	0.0	2.9	6.4
	FNZSL	32	0.0	2.7	5.9
	Mean	**160**	**0.0**	**2.7**	**6.7**

Note: W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL = large leaf, SL = small leaf, and *n* = number of samples.

Table S2.4 Population fitted values for water-soluble carbohydrate (WSC) levels after spatial and treatment effects are taken into account. Standard error and confidence intervals are presented.

Pool	Population	*n*	Mean	SE	lower CI	upper CI
WNZLL	Low-End	20	92.0	6.55	71.0	112.9
	Low-Mid	20	123.3	6.55	102.4	144.2
	Parent	40	134.2	4.76	118.1	150.2
	High-Mid	20	170.5	6.55	149.6	191.4
	High-End	20	191.6	6.55	170.7	212.5
WNZSL	Low-End	20	119.1	6.55	98.2	140
	Low-Mid	20	119.7	6.55	98.8	140.6
	Parent	40	119.2	4.77	103.2	135.3
	High-Mid	20	167.9	6.55	147	188.8
	High-End	20	188.9	6.55	167.9	209.8
WUSLL	Low-End	20	93.4	6.55	72.4	114.3
	Low-Mid	20	104.3	6.55	83.4	125.2
	Parent	40	138.5	4.77	122.4	154.6
	High-Mid	20	151.9	6.55	131	172.8
	High-End	20	162.1	6.55	141.2	183
FNZLL	Low-End	20	129.8	6.55	108.8	150.7
	Low-Mid	20	120.3	6.55	99.4	141.3
	Parent	40	143.2	4.77	127.1	159.3
	High-Mid	20	168.4	6.55	147.4	189.3
	High-End	20	184.6	6.55	163.6	205.5
FNZSL	Low-End	20	74.8	6.55	53.8	95.7
	Low-Mid	20	83.2	6.55	62.3	104.1
	Parent	40	101.9	4.76	85.8	117.9
	High-Mid	20	159.6	6.55	138.7	180.6
	High-End	20	173.5	6.55	152.6	194.4

Note: *n* = number of individuals per population used in analysis, SE = standard error, CI = 95% confidence interval of the population mean, W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL = large leaf, SL = small leaf, Low = low WSC, High = high WSC, Parent = Parent generation, Mid = Middle generation, and End = End generation. Mean WSC is in g kg^{-1} DM.

Table S2.5 Means of leaf and petiole measurements for a subset of white clover plants grouped into leaf size classes. Mean values and analysis of variance (ANOVA) tests for differences in means were calculated from log-transformed data. Back transformed means are reported, and the difference in means is calculated by subtracting the back transformed SL mean from the back transformed LL mean.

Leaf size class	*n*	Lw (cm)	Ll (cm)	LA (cm^2)	Pl (cm)
LL	105	2.3	2.3	3.7	6.8
SL	60	1.9	1.9	2.5	5.4
Difference in means		0.4**	0.4**	1.1**	1.4**

Note: *n* = number of samples, Lw = leaf width, Ll = leaf length, LA = leaf area, Pl = petiole length, LL = large leaf, and SL = small leaf.
** indicates *p*-value from ANOVA < 0.001 at α = 0.05.

Table S2.6 Variance of population means for a subset of 11 white clover populations calculated from 15 samples for each population for each measurement. Calculations are based on logged data and variance is reported to 3 significant figures.

Population	Lw	Ll	LA	Pl
WNZLL-Parent	0.00232	0.00237	0.00548	0.0224
WUSLL- Parent	0.00752	0.0220	0.0539	0.0184
WNZSL- Parent	0.00255	0.00428	0.0118	0.0165
FNZLL- Parent	0.00301	0.000949	0.00531	0.0176
FNZSL- Parent	0.00652	0.0101	0.0290	0.00933
WUSLL-Low-Mid	0.00693	0.00640	0.0257	0.0524
WUSLL-High-Mid	0.00115	0.00249	0.00550	0.0150
WNZSL-Low-End	0.00321	0.00908	0.0195	0.0377
FNZSL-High-End	0.00949	0.0109	0.0362	0.0152
FNZLL-Low-End	0.00944	0.0125	0.0423	0.0159
FNZLL-High-End	0.00597	0.0153	0.0392	0.0233
Mean	0.00528	0.00876	0.0249	0.0222

Note: Lw = leaf width, Ll = leaf length, LA = leaf area, Pl = petiole length, W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL = large leaf, SL = small leaf, Parent = Parent generation, Mid = Middle generation, End = End generation, Low = low water-soluble carbohydrate (WSC), and High = high WSC.

Table S2.7 Variance of individual means for a subset of 11 white clover populations calculated from five replicates from each individual for each measurement. Calculations are based on logged data and variance is reported to a maximum of 3 significant figures.

Population	Individual	Lw	Ll	LA	Pl
WNZLL-Parent	1	0.00204	0.00149	0.00597	0.0338
WNZLL- Parent	2	0.00298	0.00219	0.00856	0.00989
WNZLL- Parent	3	0.00109	0.00116	0.00442	0.00341
WUSLL- Parent	1	0.00117	0.00147	0.00486	0.00219
WUSLL- Parent	2	0.00175	0.00632	0.0141	0.00577
WUSLL- Parent	3	0.000801	0.00464	0.00777	0.00211
WNZSL- Parent	1	0.00146	0.00245	0.00721	0.00421
WNZSL- Parent	2	0.00180	0.00298	0.00730	0.00582
WNZSL- Parent	3	0.000578	0.00376	0.00715	0.00950
FNZLL- Parent	1	0.00104	0.000932	0.00240	0.00874
FNZLL- Parent	2	0.00222	0.000717	0.00374	0.0121
FNZLL- Parent	3	0.00143	0.000534	0.000835	0.0171
FNZSL- Parent	1	0.00196	0.00567	0.0132	0.0147
FNZSL- Parent	2	0.000364	0.000327	0.00128	0.00706
FNZSL- Parent	3	0.00750	0.00788	0.0303	0.00597
WUSLL-Low-Mid	1	0.0000898	0.000514	0.000779	0.00326
WUSLL- Low-Mid	2	0.000450	0.000735	0.00102	0.00322
WUSLL- Low-Mid	3	0.00172	0.000808	0.00411	0.00421
WUSLL-High-Mid	1	0.00118	0.000484	0.00282	0.0101
WUSLL- High-Mid	2	0.000940	0.00116	0.00339	0.000192
WUSLL- High-Mid	3	0.00148	0.00190	0.00646	0.0172
WNZSL-Low-End	1	0.00157	0.000312	0.00291	0.00548
WNZSL-Low-End	2	0.000583	0.00403	0.00661	0.0130
WNZSL-Low-End	3	0.00364	0.0115	0.0259	0.00409
FNZSL-High-End	1	0.00401	0.00240	0.0125	0.00554
FNZSL- High-End	2	0.00889	0.00308	0.0212	0.00487
FNZSL- High-End	3	0.00255	0.00205	0.00599	0.00527
FNZLL-Low-End	1	0.00112	0.00136	0.00410	0.00393
FNZLL- Low-End	2	0.00536	0.00242	0.0140	0.00284
FNZLL- Low-End	3	0.00132	0.00176	0.00535	0.0138
FNZLL-High-End	1	0.00189	0.000614	0.00350	0.0229
FNZLL- High-End	2	0.000883	0.00112	0.00326	0.00710
FNZLL- High-End	3	0.00433	0.00823	0.0242	0.00145
Mean		0.00213	0.00264	0.00810	0.0082

Note: Lw = leaf width, Ll = leaf length, LA = leaf area, Pl = petiole length, W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL = large leaf, SL = small leaf, Parent = Parent generation, Mid = Middle generation, End = End generation, Low = low water-soluble carbohydrate (WSC), and High = high WSC.

Table S2.8 Population fitted values for leaf area after spatial and treatment effects are taken into account. Standard error and confidence intervals are presented.

Pool	Population	n	Mean	SE	lower CI	upper CI
WNZLL	Low-End	15	12.32	1.02	9.2	15.9
	Low-Mid	15	13.75	1.07	10.4	17.5
	Parent	30	13.15	0.82	10.3	16.4
	High-Mid	15	15.89	1.15	12.3	19.9
	High-End	15	16.07	1.16	12.5	20.1
WNZSL	Low-End	15	11.50	0.98	8.5	15.0
	Low-Mid	15	12.75	1.03	9.6	16.4
	Parent	30	10.94	0.75	8.4	13.9
	High-Mid	15	16.83	1.19	13.1	21.0
	High-End	15	16.42	1.17	12.8	20.5
WUSLL	Low-End	13	12.66	1.09	9.4	16.5
	Low-Mid	14	15.42	1.17	11.8	19.5
	Parent	30	17.07	0.94	13.8	20.7
	High-Mid	15	18.21	1.24	14.4	22.5
	High-End	15	16.20	1.17	12.6	20.3
FNZLL	Low-End	15	13.21	1.05	10.0	16.9
	Low-Mid	15	15.09	1.13	11.6	19.0
	Parent	30	18.58	0.98	15.2	22.3
	High-Mid	15	19.30	1.27	15.3	23.7
	High-End	15	19.77	1.29	15.8	24.3
FNZSL	Low-End	15	9.77	0.91	7.0	13.0
	Low-Mid	15	11.27	0.97	8.3	14.7
	Parent	30	11.27	0.76	8.6	14.3
	High-Mid	15	10.52	0.94	7.7	13.9
	High-End	15	9.50	0.89	6.8	12.7

Note: *n* = number of individuals per population used in analysis, SE = standard error, CI = 95% confidence interval of the population mean, W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL = large leaf, SL = small leaf, Low = low water-soluble carbohydrate (WSC), High = high WSC, Parent = Parent generation, Mid = Middle generation, and End = End generation. Mean leaf area is in cm^2.

Table S2.9 Shapiro-Wilk Normality Test *p*-values and optimal transformation determined by Box-Cox transformation analysis for both water-soluble carbohydrate and leaf area traits in each pool and combined pool datasets.

	WSC	LA					
Pool	***p*-value**	***p*-value**	**λ**	**log-Likelihood**	**Optimal Transformation**	**log_{10}(LA) *p*-value**	**√LA *p*-value**
WNZLL	0.33†	0.29†					
WNZSL	0.38†	0.00066	0 (0.18)	-59.73	log_{10}	0.22†	
WUSLL	0.23†	0.49†					
FNZLL	0.24†	0.13†					
FNZSL	0.21†	0.00064	0 (-0.02)	-54.70	log_{10}	0.49†	
Pools combined	0.24†	1.82e-05	0.5 (0.38)	-538.89	square root		0.52†

Note: WSC = water-soluble carbohydrate, LA = leaf area, and λ = best lambda value to nearest half based on the maximum log-Likelihood (95% confidence interval) with exact λ value corresponding to log-Likelihood presented in parentheses. log_{10}(LA) *p*-value and square root LA *p*-value determined by Shapiro-Wilk Normality Test. W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL = large leaf, and SL = small leaf.

† indicates *p*-value from Shapiro-Wilk Normality Test is above 0.05 threshold and data follow a normal distribution.

Table S2.10 Linear regression results for water-soluble carbohydrate and leaf area interaction for each population in each pool.

Pool	Model	Adj r^2	F-value	*p*-value	Random effect	Coefficient	SE	t-value	*p*-value		Equation
WNZLL	WSC ~ 0 + Population * LA	0.97	215.3	<2.2e-16	Low-End	15.38	6.45	2.39	0.021	*	WSC = 15.38 - 0.5 x LA
					Low-Mid	7.65	3.96	1.93	0.059		WSC = 7.65 + 0.34 x LA
					Parent	16.12	2.46	6.56	2.98e-08	***	WSC = 16.12 - 0.19 x LA
					High-Mid	12.16	4.26	2.85	0.0063	**	WSC = 12.16 + 0.26 x LA
					High-End	21.81	4.24	5.15	4.49e-06	***	WSC = 21.81 - 0.105 x LA
					LA	-0.19	0.17	-1.13	0.27		
					Low-End : LA	-0.31	0.55	-0.56	0.58		
					Low-Mid : LA	0.53	0.35	1.51	0.14		
					High-Mid : LA	0.45	0.29	1.54	0.13		
					High-End : LA	0.085	0.31	0.27	0.79		
WNZSL	WSC ~ 0 + Population * LA	0.97	213.2	<2.2e-16	Low-End	9.35	2.80	3.34	0.0016	**	WSC = 9.35 + 0.206 x LA
					Low-Mid	4.51	2.83	1.59	0.12		WSC = 4.51 + 0.6 x LA
					Parent	10.23	1.67	6.13	1.36e-07	***	WSC = 10.23 + 0.16 x LA
					High-Mid	12.57	2.21	5.68	6.94e-07	***	WSC = 12.57 + 0.196 x LA
					High-End	18.17	4.86	3.74	0.00047	***	WSC = 18.17 + 0.35 x LA
					LA	0.16	0.13	1.18	0.24		
					Low-End : LA	0.046	0.26	0.18	0.86		
					Low-Mid : LA	0.44	0.26	1.73	0.089		
					High-Mid : LA	0.036	0.17	0.20	0.84		
					High-End : LA	0.19	0.31	0.59	0.56		
WUSLL	WSC ~ 0 + Population * LA	0.95	100.2	<2.2e-16	Low-End	2.52	3.30	0.76	0.45		WSC = 2.52 + 0.56 x LA
					Low-Mid	6.01	4.03	1.49	0.14		WSC = 6.01 + 0.29 x LA
					Parent	6.84	2.25	3.04	0.0039	**	WSC = 6.84 + 0.37 x LA
					High-Mid	14.35	3.81	3.76	0.00047	***	WSC = 14.35 + 0.03 x LA
					High-End	10.04	3.39	2.96	0.0048	**	WSC = 10.04 + 0.32 x LA
					LA	0.37	0.12	3.03	0.004	**	
					Low-End : LA	0.19	0.26	0.76	0.45		
					Low-Mid : LA	-0.079	0.26	-0.30	0.77		
					High-Mid : LA	-0.34	0.22	-1.50	0.14		
					High-End : LA	-0.051	0.22	-0.23	0.82		

Table S2.10 (continued)

Pool	Model	Adj r^2	F-value	*p*-value	Random effect	Coefficient	SE	t-value	*p*-value†		Equation
FNZLL	WSC ~ 0 + Population * LA	0.97	180	<2.2e-16	Low-End	7.11	3.62	1.97	0.055		WSC = 7.11 + 1.03 x LA
					Low-Mid	9.98	3.37	2.96	0.0047	**	WSC = 9.98 + 1.06 x LA
					Parent	18.69	3.25	5.75	5.28e-07	***	WSC = 18.69 - 0.23 x LA
					High-Mid	13.42	4.38	3.07	0.0035	**	WSC = 13.42 + 1.11 x LA
					High-End	13.58	2.75	4.93	9.44e-06	***	WSC = 13.58 + 0.67 x LA
					LA	-0.23	0.16	-1.38	0.17		
					Low-End : LA	0.67	0.31	2.14	0.037	*	
					Low-Mid : LA	0.36	0.26	1.39	0.17		
					High-Mid : LA	0.39	0.28	1.37	0.18		
					High-End : LA	0.44	0.21	2.12	0.04	*	
FNZSL	WSC ~ 0 + Population * LA	0.94	95.8	<2.2e-16	Low-End	6.68	3.24	2.06	0.045	*	WSC = 6.68 + 0.06 x LA
					Low-Mid	8.99	6.84	1.32	0.19		WSC = 8.99 - 0.02 x LA
					Parent	7.17	1.86	3.85	0.00034	***	WSC = 7.17 + 0.23 x LA
					High-Mid	11.54	4.56	2.53	0.015	*	WSC = 11.54 + 0.37 x LA
					High-End	21.11	3.97	5.32	2.48e-06	***	WSC = 21.11 - 0.2 x LA
					LA	0.23	0.14	1.71	0.093		
					Low-End : LA	-0.17	0.35	-0.48	0.63		
					Low-Mid : LA	-0.25	0.56	-0.45	0.65		
					High-Mid : LA	0.14	0.42	0.34	0.74		
					High-End : LA	-0.43	0.40	-1.087	0.28		

Note: Model = linear model used for regression, WSC = water-soluble carbohydrate, LA = leaf area, Adj r^2 = adjusted coefficient of determination, and SE = standard error. The green shading in the "*Coefficient*" column shows the intercept for each population (respective populations are in the "*Random effect*" column), and the blue shading shows the slope of the line for each Parent population in each pool. To get the slope of the line for each population, the parental slope is added to each of the population slopes (unshaded and below dashed line) within each pool. The "*Equation*" column gives the linear regression equation for each population with the intercept and combined slope values presented. The t-value is a measure of how many standard deviations the coefficient estimate is far away from zero. If t-statistic values are relatively far away from zero and are large relative to the standard error, it indicates that a relationship exists.

† = *p*-values corresponding to each coefficient and is the probability of observing any value equal or larger than the t-value.

Significance codes: *** ≤ 0.0001, ** = 0.001 – 0.01, * = 0.01 – 0.05, no symbol ≥ 0.05 at α = 0.05

Table S2.11 Mean key environmental variables, including temperature and hours of sunlight, for October and November when selections were originally made (2001 – 2004) and the current experiment (2017). Means for each variable for each month are presented as well as the mean for October and November combined to give mean values for each year.

			Environmental variable			
Year	**Month**	***n***	**Tdry**	**Tmax**	**Tmin**	**Sun**
	Oct	31	14.1	18.5	10.3	3.3
2001	Nov	30	15.1	19.2	10.6	4.6
	Oct and Nov	61	14.6	18.8	10.4	3.9
	Oct	31	10.4	14.3	5.8	5.2
2002	Nov	30	11.6	15.9	8.1	3.9
	Oct and Nov	61	11.0	15.1	6.9	4.5
	Oct	17	11.5	16.2	6.5	5.3
2003	Nov	30	13.2	17.1	9.2	4.5
	Oct and Nov	47	12.6	16.8	8.2	4.8
	Oct	31	12.2	16.3	9.1	3.0
2004	Nov	30	15.1	19.1	10.5	6.5
	Oct and Nov	61	13.6	17.7	9.8	4.7
	Oct	31	13.1	17.1	10.1	3.9
2017	Nov	30	15.3	20.6	10.4	7.2
	Oct and Nov	61	14.2	18.8	10.3	5.6

Note: Oct = October, Nov = November, *n* = number of days, Tdry = spot value of dry bulb temperature (air temperature) at 9am in degrees Celsius (°C), Tmax = Maximum temperature over 24 hours in °C, Tmin = Minimum temperature over 24 hours in °C, and Sun = hours of sunshine in 24 hours. Data were sourced from the New Zealand/Aotearoa national climate database provided by National Institute of Water and Atmospheric Research (NIWA).

Table S2.12 Pairwise differences among five years for three environmental variables calculated by Tukey's honest significant differences test. Comparisons are comprised of the mean from daily data from October and November for each year (n = 47 – 61). The difference between years is presented in the "*Difference*" column. Adjusted *p*-values for important year comparisons are underlined and in bold.

Tmax			Tmin			Tdry		
Comparison	**Difference**	***p*-value**	**Comparison**	**Difference**	***p*-value**	**Comparison**	**Difference**	***p*-value**
2002 – 2001	-3.7	<0.001	2002 – 2001	-3.5	<0.001	2002 – 2001	-3.6	<0.001
2003 – 2001	-2.1	0.0011	2003 – 2001	-2.2	0.0056	2003 – 2001	-2.0	<0.001
2004 – 2001	-1.1	0.14	2004 – 2001	-0.7	0.80	2004 – 2001	-1.0	0.13
2017 – 2001	-0.02	1.00	2017 – 2001	-0.2	1.00	2017 – 2001	-0.4	0.85
2003 – 2002	1.7	0.014	2003 – 2002	1.3	0.24	2003 – 2002	1.6	0.0039
2004 – 2002	2.6	<0.001	2004 – 2002	2.8	<0.001	2004 – 2002	2.6	<0.001
2017 – 2002	3.7	<0.001	2017 – 2002	3.3	<0.001	2017 – 2002	3.2	<0.001
2004 – 2003	0.9	0.40	2004 – 2003	1.5	0.11	2004 – 2003	1.0	0.20
2017 – 2003	2.0	0.0012	2017 – 2003	2.0	0.014	2017 – 2003	1.6	0.006
2017 – 2004	1.1	**0.15**	2017 – 2004	0.5	**0.93**	2017 – 2004	0.6	**0.65**

Note: Tmax = maximum temperature over 24 hours in °C, Tmin = minimum temperature over 24 hours in °C, and Tdry = spot value of dry bulb temperature (air temperature) at 9am in degrees Celsius (°C).

Table S2.13 Number of individuals infected by downy mildew per population prior to leaf area and water-soluble carbohydrate determination in November 2017. Counts of number of individuals infected were determined by a presence/absence score for each of the 900 individuals in the white clover plot.

Population	Number of individuals infected	*n*	% infected
WNZLL-Low-End	1	30	3.3
WNZLL-Low-Mid	1	30	3.3
WNZLL-Parent	0	60	0.0
WNZLL-High-Mid	0	30	0.0
WNZLL-High-End	2	30	6.7
WNZSL-Low-End	1	30	3.3
WNZSL-Low-Mid	1	30	3.3
WNZSL-Parent	4	60	6.7
WNZSL-High-Mid	9	30	30
WNZSL-High-End	2	30	6.7
WUSLL-Low-End	15	30	50
WUSLL-Low-Mid	6	30	20
WUSLL-Parent	5	60	8.3
WUSLL-High-Mid	3	30	10
WUSLL-High-End	15	30	50
FNZLL-Low-End	0	30	0.0
FNZLL-Low-Mid	0	30	0.0
FNZLL-Parent	2	60	3.3
FNZLL-High-Mid	0	30	0.0
FNZLL-High-End	1	30	3.3
FNZSL-Low-End	0	30	0.0
FNZSL-Low-Mid	0	30	0.0
FNZSL-Parent	2	60	3.3
FNZSL-High-Mid	0	30	0.0
FNZSL-High-End	0	30	0.0

Note: W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL = large leaf, SL = small leaf, Low = low water-soluble carbohydrate (WSC), High = high WSC, Parent = Parent generation, Mid = Middle generation, End = End generation, *n* = number of individuals in the white clover plot for each population, and % infected = percentage of individuals infected by downy mildew per population.

SUPPLEMENTARY FIGURES

Figure S2.1 The white clover trial block at AgResearch Grasslands Research Centre, Palmerston North, New Zealand/Aotearoa in early October 2017.

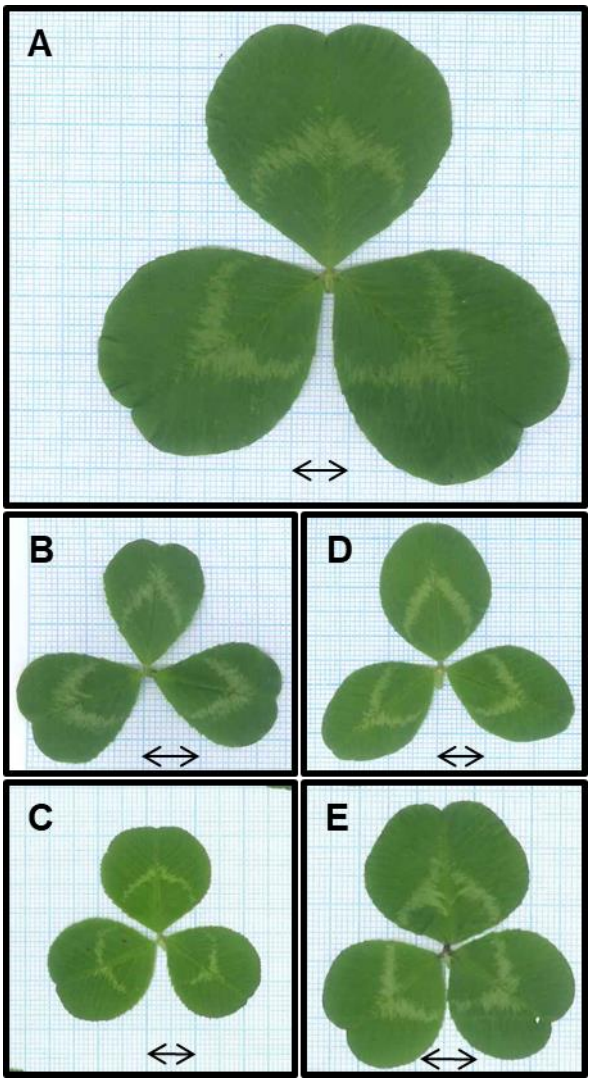


Figure S2.2 Leaf shape and size variation examples of one trifoliate leaf from five white clover populations. **A**) FNZLL-High-Mid. **B**) WNZSL-Low-End. **C**) WUSLL-Low-Mid. **D**) WNZSL-High-End. **E**) WUSLL-High-Mid.

Note: W = Widdup, F = Ford, NZ = New Zealand/Aotearoa, US = United States of America, LL = large leaf, SL = small leaf, Low = low water-soluble carbohydrate (WSC), High = high WSC, Mid = Middle generation, and End = End generation. Black arrows indicate 1 cm for each photo.

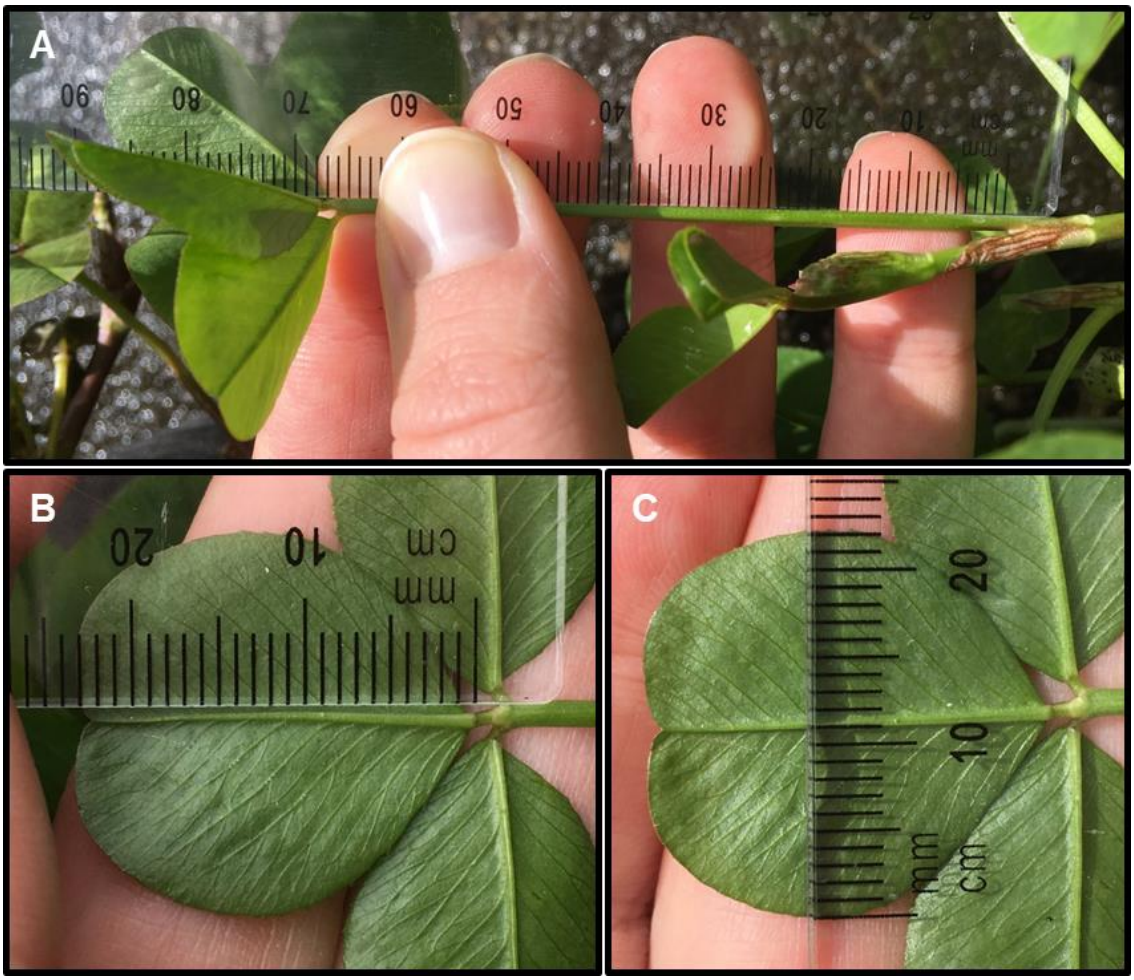


Figure S2.3 Locations of where leaf measurements were taken. **A**) Petiole length of first opened leaf of a stolon. **B**) Middle leaflet length. **C**) Middle leaflet width at widest point.

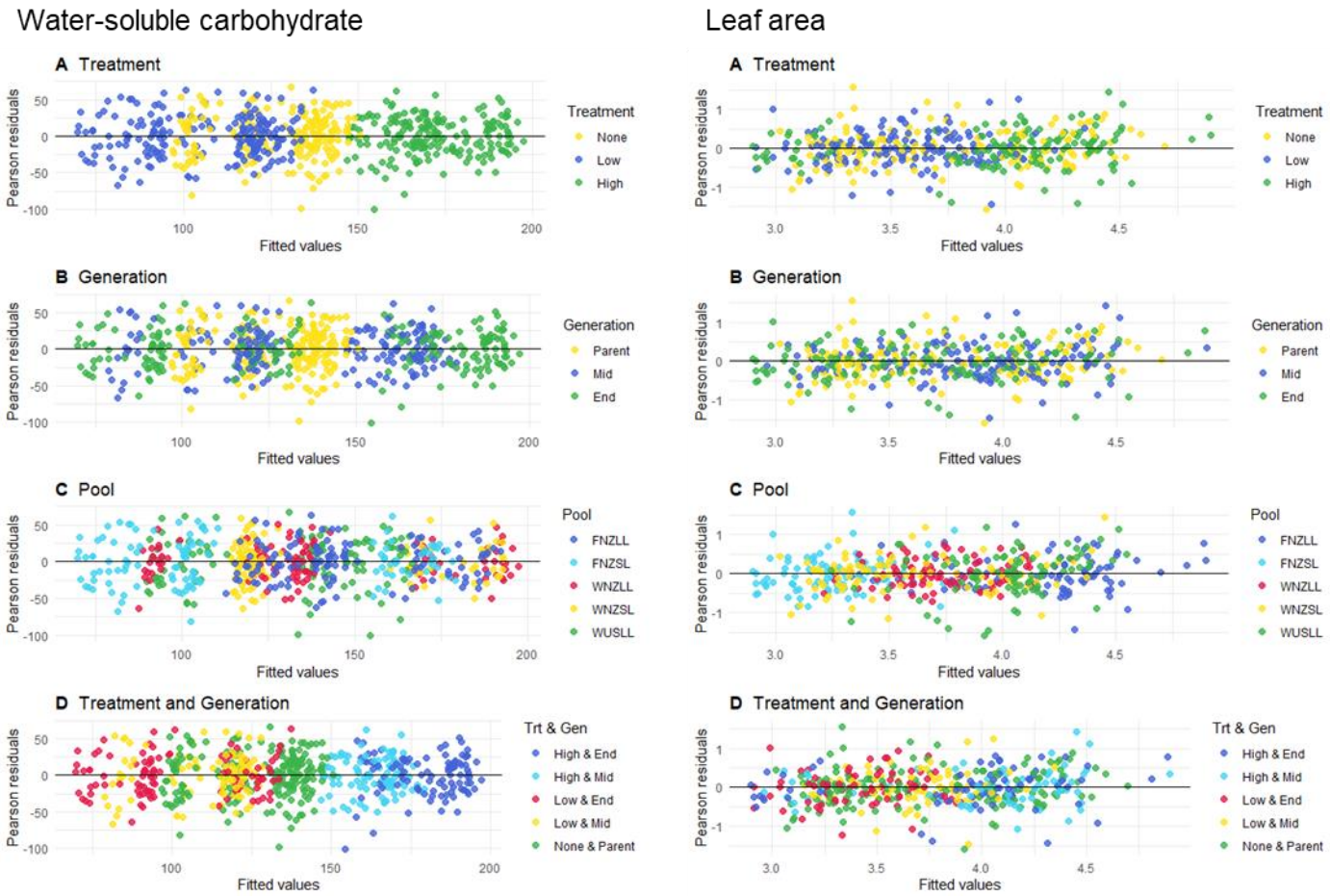


Figure S2.4 Residual plots for water-soluble carbohydrate (WSC) values by treatments (left) and square root leaf area values by treatment (right). Individuals were categorised into: **A**) Treatment = None (i.e., Parent), Low and High, **B**) Generation = Parent, Mid and End, **C**) Pool = FNZLL, FNZSL, WNZLL, WNZSL and WUSLL, **D**) Treatment and Generation combination = combination of treatment and generation. *Note*: Low = low WSC, High = high WSC, Trt = treatment, and Gen = generation.

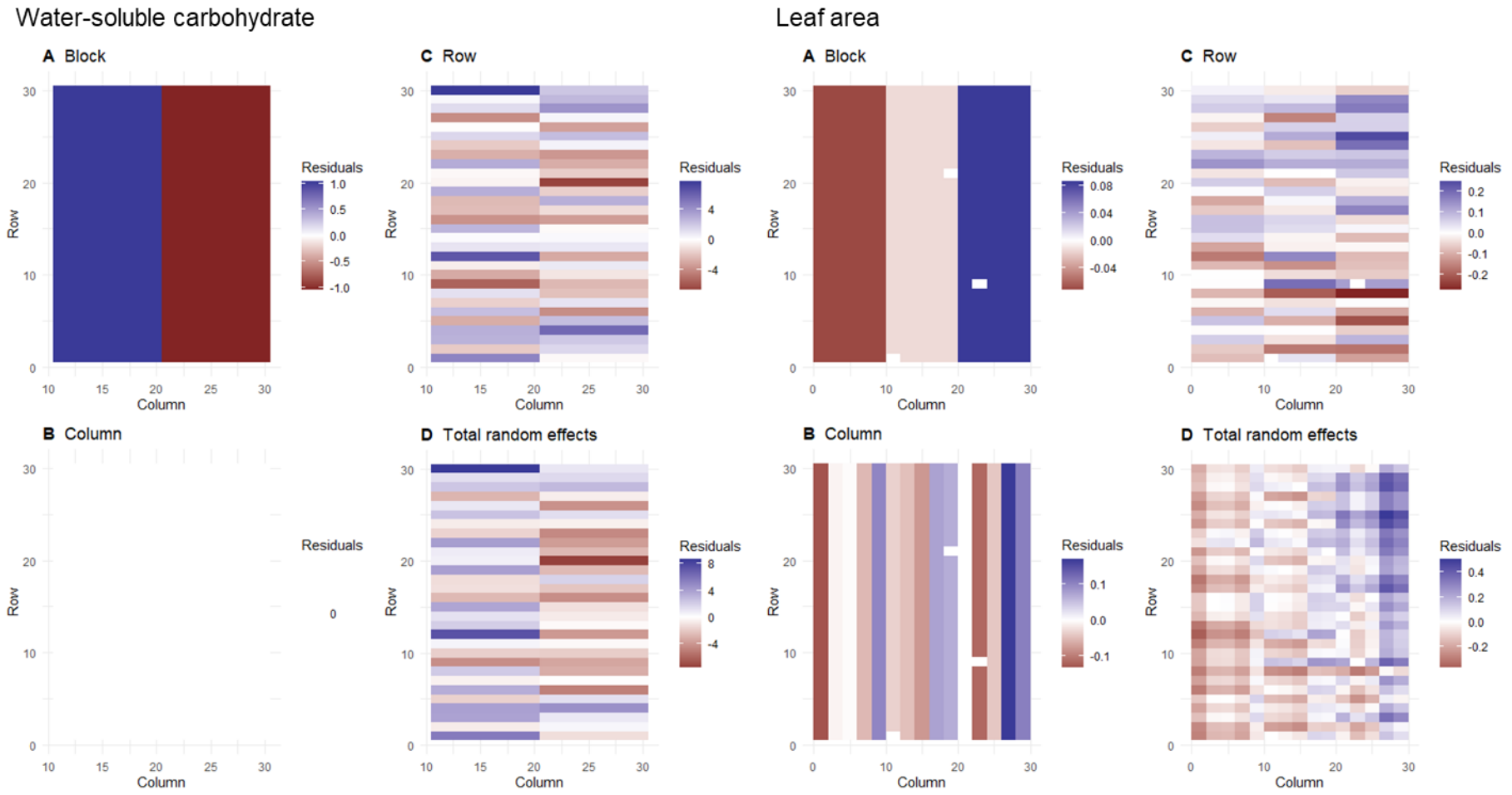


Figure S2.5 Variance explained by random effects for 600 individuals assessed for water-soluble carbohydrate (left) and 447 individuals assessed for leaf area (right). **A**) Variance explained by block, **B**) column variance effects nested within each block, **C**) row variance effects nested within each block, **D**) block, column and row random effects combined. Colour scale represents positive Pearson residuals (dark blue) to negative residuals (dark red).

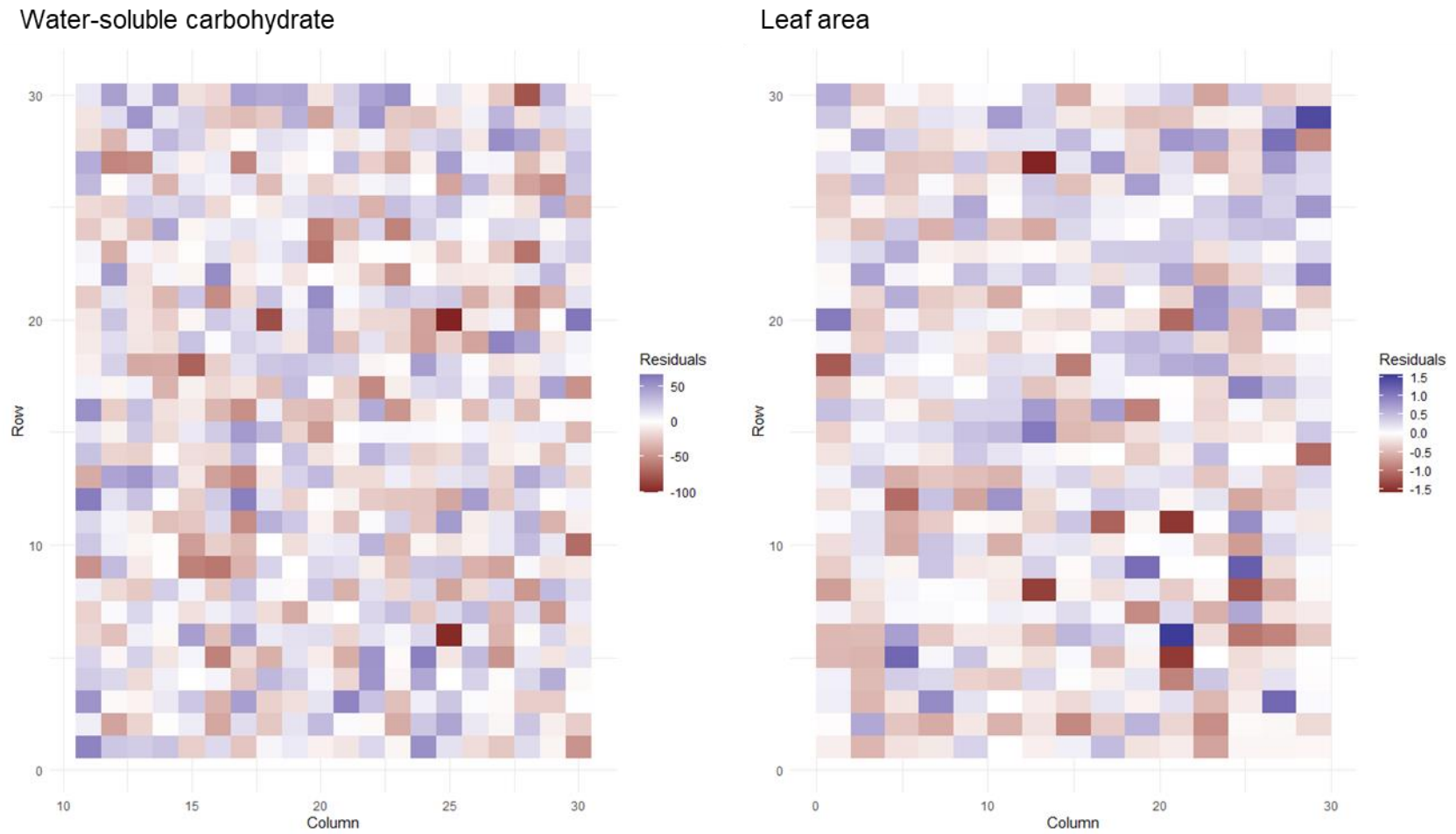


Figure S2.6 Spatial residual plot for individual water-soluble carbohydrate (left) and individual leaf area (right) after accounting for spatial and treatment effects. Colour scale represents positive Pearson residuals (dark blue) to negative residuals (dark red).

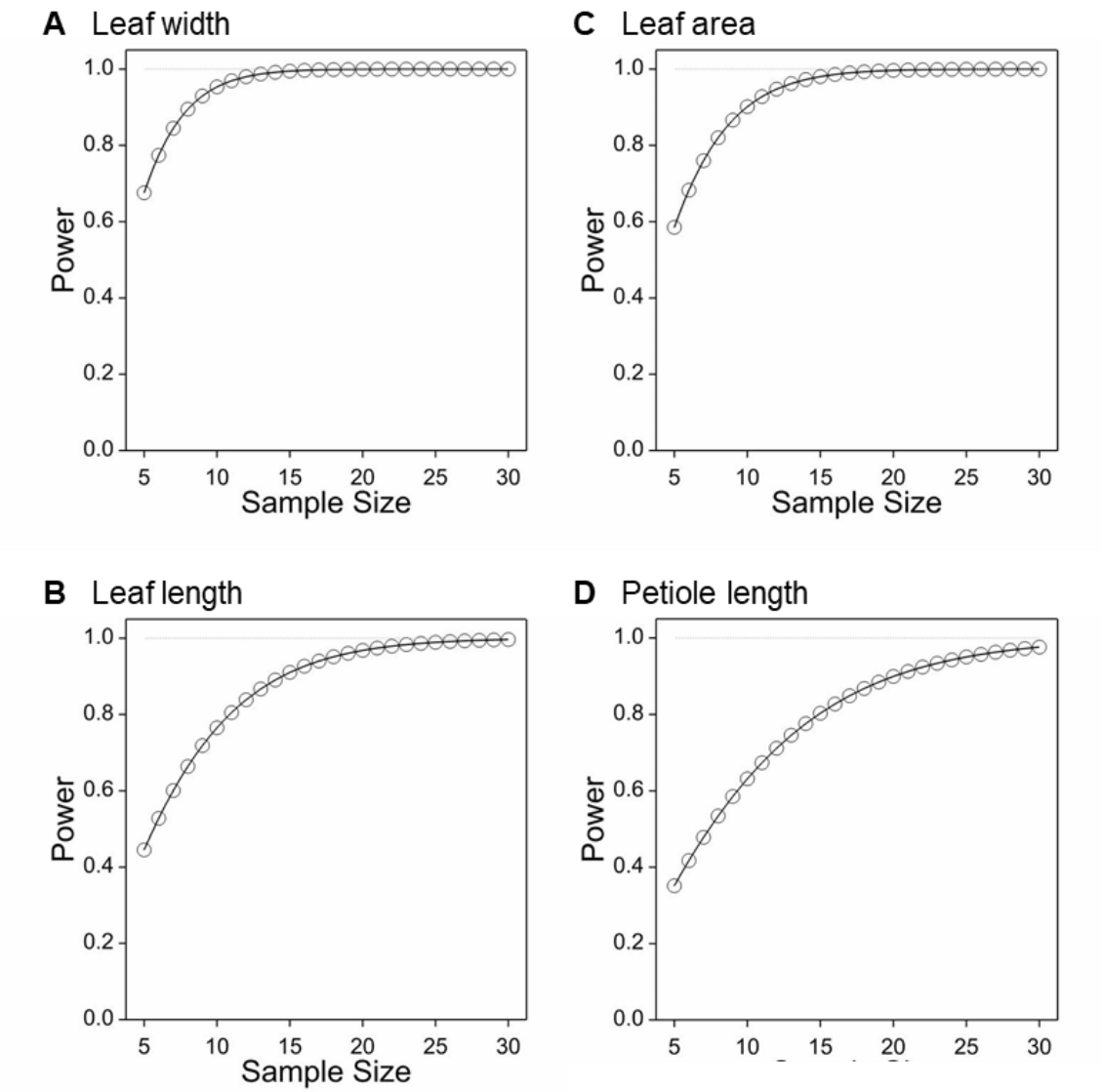


Figure S2.7 Power plots of **A**) leaf width, **B**) leaf length, **C**) leaf area and **D**) petiole length for number of individuals required for each population. Calculations are based on log-transformed data.

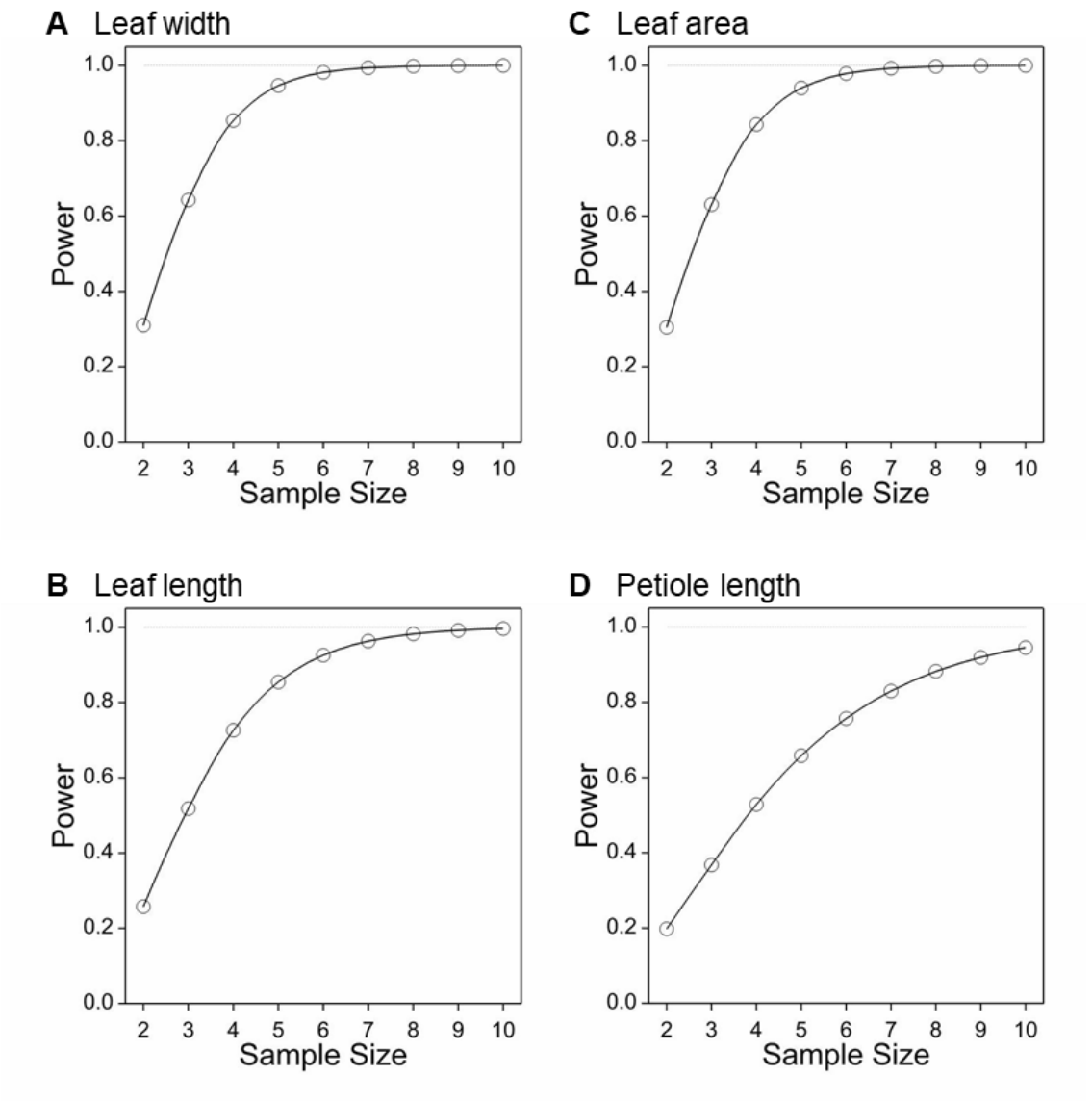


Figure S2.8 Power plots of **A**) leaf width, **B**) leaf length, **C**) leaf area and **D**) petiole length for number of leaves required per individual. Calculations are based on log-transformed data.

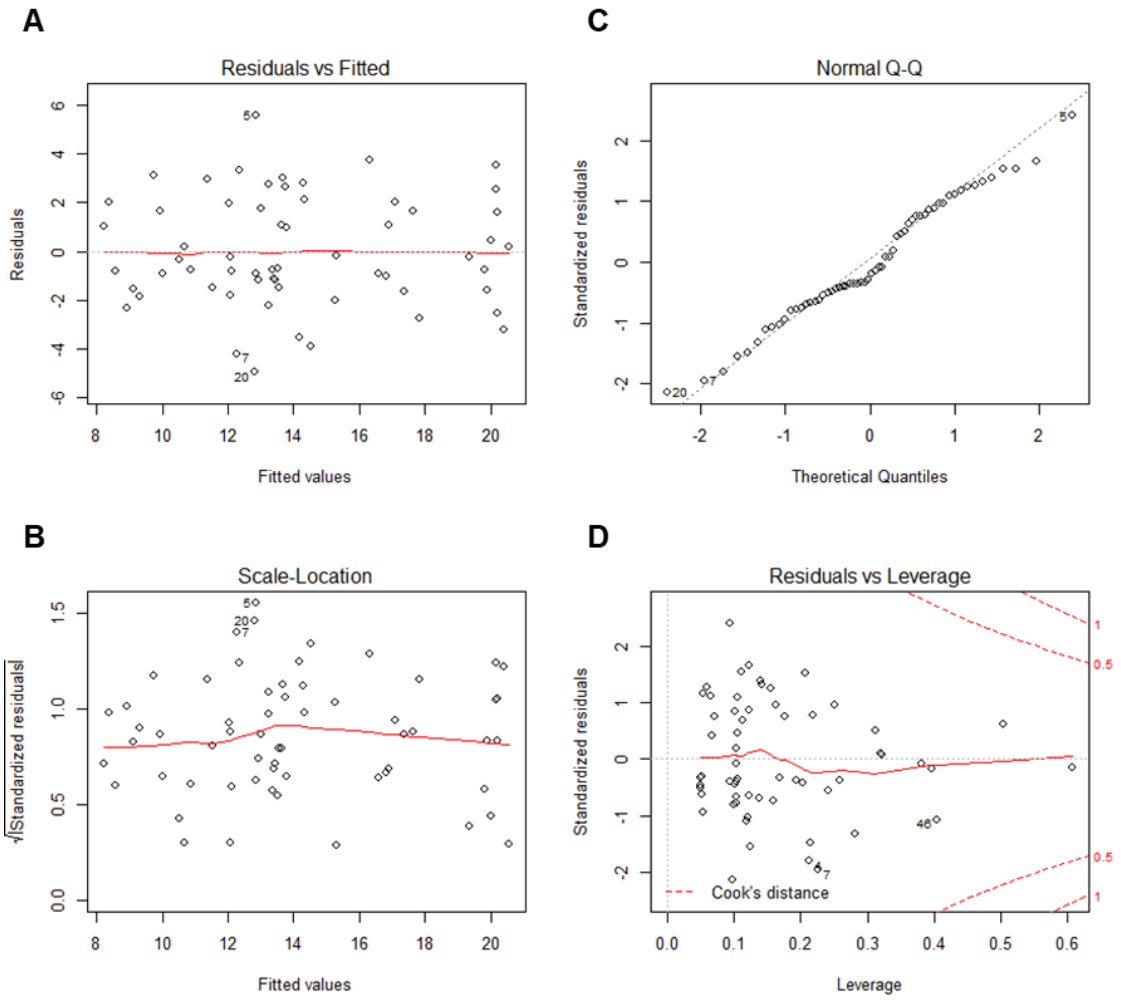


Figure S2.9 Example residual plots of WNZLL pool from linear regression analysis. **A**) the residual errors versus their fitted values. **B**) scale location plot showing the square root of the standardized residuals versus the fitted values. **C**) Q-Q plot of standardised residuals versus theoretical quantiles. **D**) leverage of each residual point. Contour lines for Cook's distance area superimposed on the plot.

APPENDIX 2

Chapter 3 Supplementary Material

SUPPLEMENTARY METHODS

Number of K_{PC} detection for PCAdapt analysis

To determine the number of principal components (K_{PC}) that separate the high water-soluble carbohydrate (WSC) from the low WSC populations, the Parent populations were removed from each pool, then each of the five pools were analysed in PCAdapt separately. Scree plots were produced (**Figure S3.11**, Appendix 2) and visually assessed for the optimal K_{PC} value as proposed by Cattell (1966), where components in the portion of the steep curve, before the first point that starts the flat line curve are retained. The scree plots PCAdapt produces have the proportion of explained variance (PEV) accounted for on the *y*-axis, not the eigenvalues like regular scree tests. Therefore, to be able to determine K_{PC} by the Kaiser-Guttman criterion, the PEV must be converted into the eigenvalues for the components of interest. PCAdapt gives the singular values for each component, which is the K_{PC} ordered square root of the proportion of variance explained by each component. Therefore, the PEV is the singular value squared (**Equation S3.1**).

$$PEV = SV^2 \qquad \textbf{Equation S3.1}$$

Where: *PEV* is the proportion of explained variance, and *SV* is the singular value.

The PEV is then converted into the eigenvalues for the component of interest (**Equation S3.2**).

$$E = PEV \times T \qquad \textbf{Equation S3.2}$$

Where: *E* is the eigenvalue for the component of interest; *PEV* is the proportion of variance accounted for; and *T* is the total eigenvalues of the correlation matrix. The total eigenvalues of the correlation matrix equals the number of principal component values on the scree plot.

Finally, the eigenvalues for the first three principal components (PCs) for each pool are calculated, and the components with eigenvalues greater than 1 are retained for each pool. The PCs with eigenvalues less than 1 explain less of the total variance than a single variable does on average, therefore by choosing K_{PC} with eigenvalues

greater than 1 we maintain the PCs that express more of the variability than each of the original variables.

Quantile-Quantile (Q-Q) plots were constructed using PCAdapt to confirm the presence of outliers (data not presented). Q-Q plots show the observed association *p*-value for all single nucleotide polymorphisms (SNPs) on the *y*-axis and the expected uniform distribution of *p*-values under the null hypothesis of no association on the *x*-axis, thus checking the distribution of the *p*-values of the test-statistic. Most of the *p*-values follow the expected distribution, however SNPs in strong association deviate from the expected as the smallest *p*-values are smaller than expected, therefore confirming outlier SNPs are present. Once the K_{PC} value is determined, PCAdapt creates a vector of z-scores that is used to measure what extent a SNP is related to the K_{PC}. The Mahalanobis distance is used as the test statistic for detecting outliers and is based on the z-scores retrieved when regressing SNPs with K_{PC}.

Cut-off threshold for outlier single nucleotide polymorphism detection in PCAdapt

The Bonferroni correction (Bonferroni, 1936) was used to control for false-discovery of outlier SNPs. The Bonferroni correction is often viewed as too conservative and it leads to an increase in false-negatives (Type II error i.e., not identifying outliers when they are there). However, when identifying differentiating SNPs, greater certainty around definite outlier SNPs is more important than not detecting a few differentiating SNPs due to the type II error. Therefore, the cut-off threshold for outlier detection was determined using the Bonferroni correction and investigated at two α thresholds (**Equation S3.3**).

$$threshold = \frac{\alpha}{N} \qquad \textbf{Equation S3.3}$$

Where: α is the probability of rejection the null hypothesis when the null hypothesis is true, and N is the number of SNPs tested.

References

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research, 1*(2), 245-276.

SUPPLEMENTARY TABLES

Table S3.1 Genome position and gene annotation for SNPs with large -$\log_{10}$(*p*-values) identified from the genome-wide association study (GWAS).

Chromosome	SNP position (bp)	SNP ID	-$\log_{10}$(*p*-value) for WSC	-$\log_{10}$(*p*-value) for SSS	Genomic region	Gene model ID and Gene annotation (Ann)	Potential function of gene and codon change
1 (Tr_{To} – 1)	2,338,028	1_2338028	3.45	4.4	exon	chr1.jg302.t1 Ann: VPS35B Vacuolar protein sorting-associated protein 35B (*Arabidopsis thaliana*)	Protein storage and vacuole biogenesis. Retarding the senescence of leaves. ATC to GTC changes isoleucine to valine.
5 (Tr_{To} – 5)	47,903,593	5_47903593	4.07	3.05	promoter	855 bp upstream from start codon of chr5.jg7106.t1 Ann: *glgC* Glucose-1-phosphate adenylyltransferase small subunit 1, chloroplastic (*Vicia faba*)	Starch biosynthesis and glycan biosynthesis.
9 (Tr_{Tp} – 1)	23,070,656	9_23070656	3.02	3.24	exon	chr9.jg3440.t1 Ann: UPL6 E3 ubiquitin-protein ligase UPL6 (*A. thaliana*)	Protein post-translational modifications. Response to water deficit and cold stress. CAA to GAA changes glutamine to glutamic acid.

Note: A total of 605 individuals were used for the GWAS with a mean of 25 individuals per population (*n* = 24). A total of 122 individuals were used from the WNZLL pool, 83 from WNZSL, 136 from WUSLL, 127 from FNZLL and 137 from FNZSL.
bp = base pairs, ID = identifier, WSC = water-soluble carbohydrate, SSS = Soluble sugars and starch, Tr_{To} = white clover *Trifolium occidentale*-derived subgenome, Tr_{Tp} = white clover *T. pallescens*-derived subgenome.

Table S3.2 Proportion of explained variance (PEV) of scree plots for multiple principal component values and their corresponding eigenvalues for each pool. Individuals per pool only included high and low water-soluble carbohydrate populations and not the Parent population.

Pool	*n*	K_{PC}	K_{PC} ordered square root of PEV	PEV (3 d.p.)	Eigenvalue
WNZLL	188	1	0.336	0.113	**2.26***
		2	0.172	0.030	0.59
		3	0.168	0.029	0.56
WNZSL	186	1	0.303	0.092	**1.84***
		2	0.176	0.031	0.62
		3	0.176	0.031	0.62
WUSLL	188	1	0.320	0.102	**2.05***
		2	0.209	0.044	0.87
		3	0.182	0.033	0.66
FNZLL	184	1	0.349	0.122	**2.44***
		2	0.194	0.038	0.75
		3	0.183	0.033	0.67
FNZSL	181	1	0.347	0.120	**2.41***
		2	0.217	0.047	0.94
		3	0.183	0.033	0.67

Note: *n* = number of individuals, K_{PC} = number of principal components, d.p. = decimal points.

* Eigenvalue greater than 1.

Table S3.3 Bonferroni cut-off threshold for outlier SNP detection at two alpha levels for PCAdapt analysis. The number of outlier SNPs detected for each pool and the number of outlier SNPs in common between more than two pools are presented. Outlier SNPs are based on the first principal component. -$\log_{10}$(*p*-value) thresholds for each alpha within each pool are displayed in **Figure S3.13**, Appendix 2.

		α = 0.05				**α = 0.01**			
Pool	**Number of SNPs**	***p*-value**	**-$\log_{10}$(*p*-value)**	**Outlier SNPs**	**Common SNPs**	***p*-value**	**-$\log_{10}$(*p*-value)**	**Outlier SNPs**	**Common SNPs**
WNZLL	11,061	4.52e-06	5.34	165	13	9.04e-07	6.04	117	10
WNZSL	11,479	4.36e-06	5.36	94	14	8.71e-07	6.06	69	13
WUSLL	11,171	4.48e-06	5.35	119	18	8.95e-07	6.05	81	9
FNZLL	10,979	4.55e-06	5.34	203	18	9.11e-07	6.04	145	11
FNZSL	10,976	4.56e-06	5.34	99	10	9.11e-07	6.04	59	7
Total	14,743	3.39e-06	5.47	643	36	6.78e-07	6.17	446	25
Mean	11,133.2	4.49e-06	5.34	136	14.6	8.98e-07	6.05	94.2	10

Note: Number of SNPs = Number of SNPs used for the analysis, *p*-value = Bonferroni corrected *p*-value cut-off with corresponding -$\log_{10}$(*p*-value), Outlier SNPs = Number of outlier SNPs detected, Common SNPs = Number of outlier SNPs in common between more than two pools.

Table S3.4 Number of outlier SNPs detected per pool based on F_{ST} analyses. Outlier SNPs are separated by pool with total and mean also presented. **A**) Outlier SNPs detected at two *q*-value alpha thresholds from BayeScan in each pool with "common SNPs" referring to outlier SNPs present in more than two pools. **B**) Outlier SNPs detected from KGD-F_{ST} analysis. The KGD-F_{ST} "Outlier SNPs" column refers to the total number of SNPs with F_{ST} values greater than 0.3 and the "common SNPs" column refers to outlier SNPs that are present in more than two pools at an F_{ST} higher than 0.3.

	A) BayeScan	**α = 0.05**		**α = 0.01**		**B) KGD-F_{ST}**	**F_{ST} > 0.3**	
Pool	**Number of SNPs**	**Outlier SNPs**	**Common SNPs**	**Outlier SNPs**	**Common SNPs**	**Number of SNPs**	**Outlier SNPs**	**Common SNPs**
WNZLL	14,598	88	15	59	11	14,743	373	128
WNZSL	14,620	68	8	45	5	14,743	203	76
WUSLL	14,623	90	12	59	5	14,743	207	90
FNZLL	14,627	58	13	37	4	14,743	387	121
FNZSL	14,620	53	7	32	4	14,743	301	97
Total	14,743	329	27	217	14	14,743	1,188	229
Mean	14,617.6	71.4	11	46.4	5.8	14,743	294.2	102.4

Note: Number of SNPs = Number of SNPs used for the analysis, Outlier SNPs = Number of outlier SNPs detected, Common SNPs = Number of outlier SNPs in common between more than two pools.

Table S3.5 Linkage disequilibrium (LD) measured as the squared correlation of allele counts (r^2) for four intergenic outlier SNPs within a 100 Kbp window.

Outlier SNP	Pool†	Population	Nearby SNP (bp)	Distance between SNPs (bp)	r^2
2_6673787	FNZLL†	Parent	6663484	-10,303	**0.25**
			6673760	-27	0.05
			6673787	0	**1.00**
		High-Mid	6663484	-10,303	0.11
	WUSLL†	Parent	6673760	-27	0.06
		High-Mid	6663465	-10,303	0.01
			6673760	-27	**0.31**
			6673787	0	**1.00**
		High-End	6663465	-10,303	**1.00**
			6673760	-27	**1.00**
	WNZSL	High-Mid	6673760	-27	0.06
		High-End	6673760	-27	**0.37**
6_31429353	WUSLL†	Low-End	31429364	11	**1.00**
			31429365	12	**1.00**
		Low-Mid	31429364	11	**1.00**
			31429365	12	**1.00**
		Parent	31429276	-77	0.03
			31429278	-75	0.03
			31429334	-19	0.01
			31429364	11	**0.55**
			31429365	12	**0.72**
		High-Mid	31429276	-77	0.18
			31429278	-75	0.13
			31429334	-19	**0.27**
			31429364	11	0.14
			31429365	12	**0.63**
		High-End	31429276	-77	0.10
			31429278	-75	0.04
			31429334	-19	0.12
			31429364	11	0.06
			31429365	12	**1.00**
	FNZLL†	Low-Mid	31429276	-77	0.002
			31429278	-75	0.002
			31429334	-19	**0.54**
			31429364	11	0.008
			31429365	12	**1.00**
		Parent	31429276	-77	0.09
			31429278	-75	0.09
			31429364	11	0.19
			31429365	12	0.19
		High-Mid	31429276	-77	0.12
			31429278	-75	0.12
			31429364	11	**1.00**
			31429365	12	**1.00**

Table S3.5 (Continued)

Outlier SNP	Pool†	Population	Nearby SNP (bp)	Distance between SNPs (bp)	r^2
6_31429365	WUSLL†	Low-End	31429353	-12	**1.00**
			31429364	-1	**1.00**
		Low-Mid	31429353	-12	**1.00**
			31429364	-1	**1.00**
		Parent	31429276	-89	0.04
			31429278	-87	0.04
			31429334	-31	0.03
			31429353	-12	**0.72**
			31429364	-1	**0.74**
		High-Mid	31429276	-89	0.18
			31429278	-87	0.12
			31429334	-31	**0.49**
			31429353	-12	**0.63**
			31429364	-1	0.18
		High-End	31429276	-89	0.10
			31429278	-87	0.04
			31429334	-31	0.12
			31429353	-12	**1.00**
			31429364	-1	0.06
	FNZLL†	Low-Mid	31429276	-89	0.002
			31429278	-87	0.002
			31429334	-31	**0.54**
			31429353	-12	**1.00**
			31429364	-1	0.008
		Parent	31429353	-12	0.19
			31429364	-1	**1.00**
		High-Mid	31429276	-89	0.12
			31429278	-87	0.12
			31429353	-12	**1.00**
			31429364	-1	**1.00**
13_4850703	FNZLL	Parent	13_4850696	-7	0.16
			13_4850721	18	**0.35**
	WNZSL	Low-Mid	13_4850721	18	**1.00**

Note: High = high water-soluble carbohydrate (WSC), Low = low WSC, Parent = Parent generation, Mid = Middle generation, End = End generation, bp = base pairs. Solid lines show separation between SNPs and dashed lines show separation between pools for each SNP. r^2 values above the 0.25 threshold are indicated in bold, with † indicating the SNP was an outlier within the pool.

SUPPLEMENTARY FIGURES

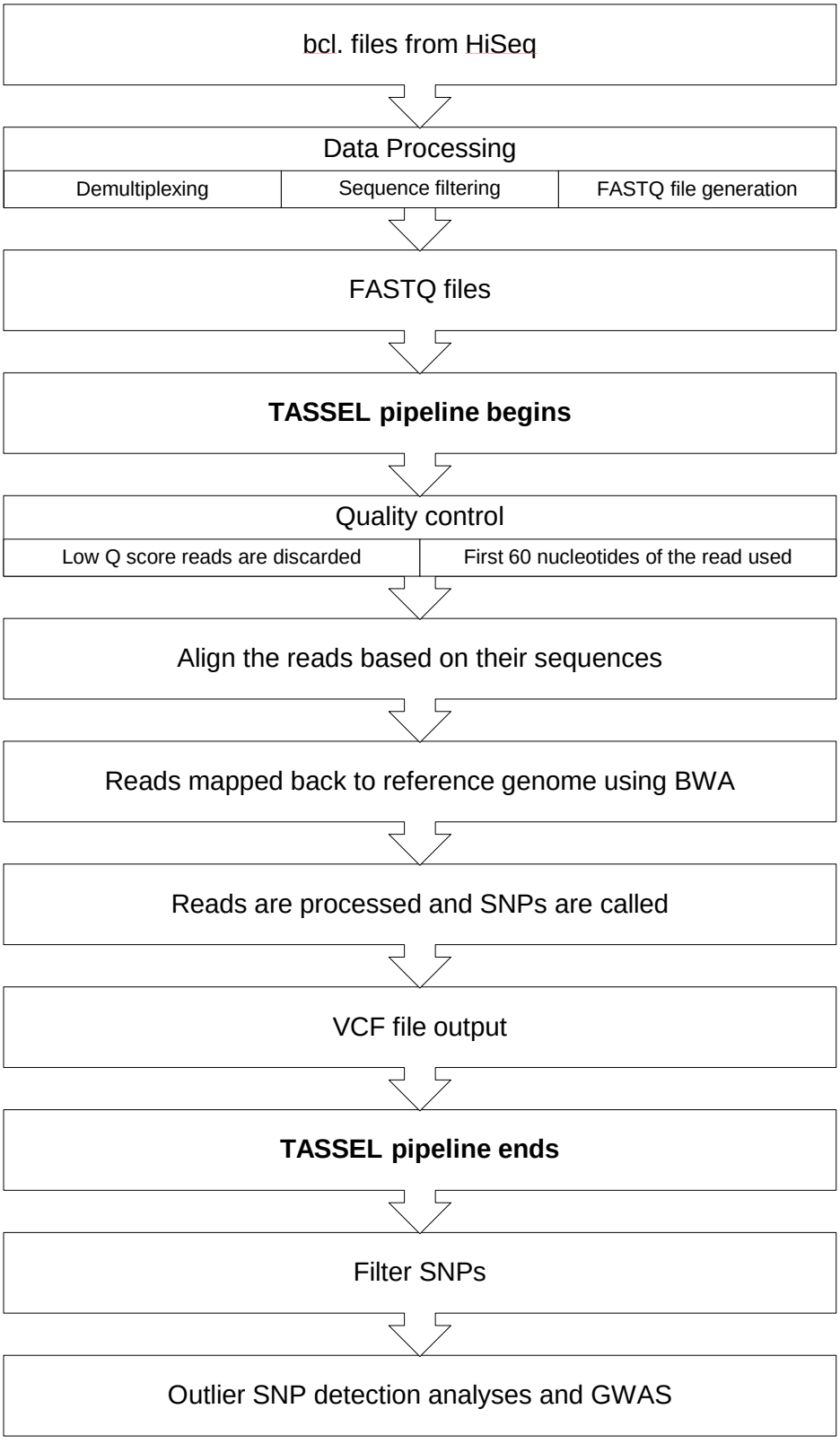


Figure S3.1 Bioinformatics workflow summary using TASSEL v 5.0 in this study.
Note: bcl = raw files containing base calls and quality scores, BWA = Burrows-Wheeler Alignment, GWAS = genome-wide association study, Q = quality, SNPs = single nucleotide polymorphisms, and VCF = variant call format.

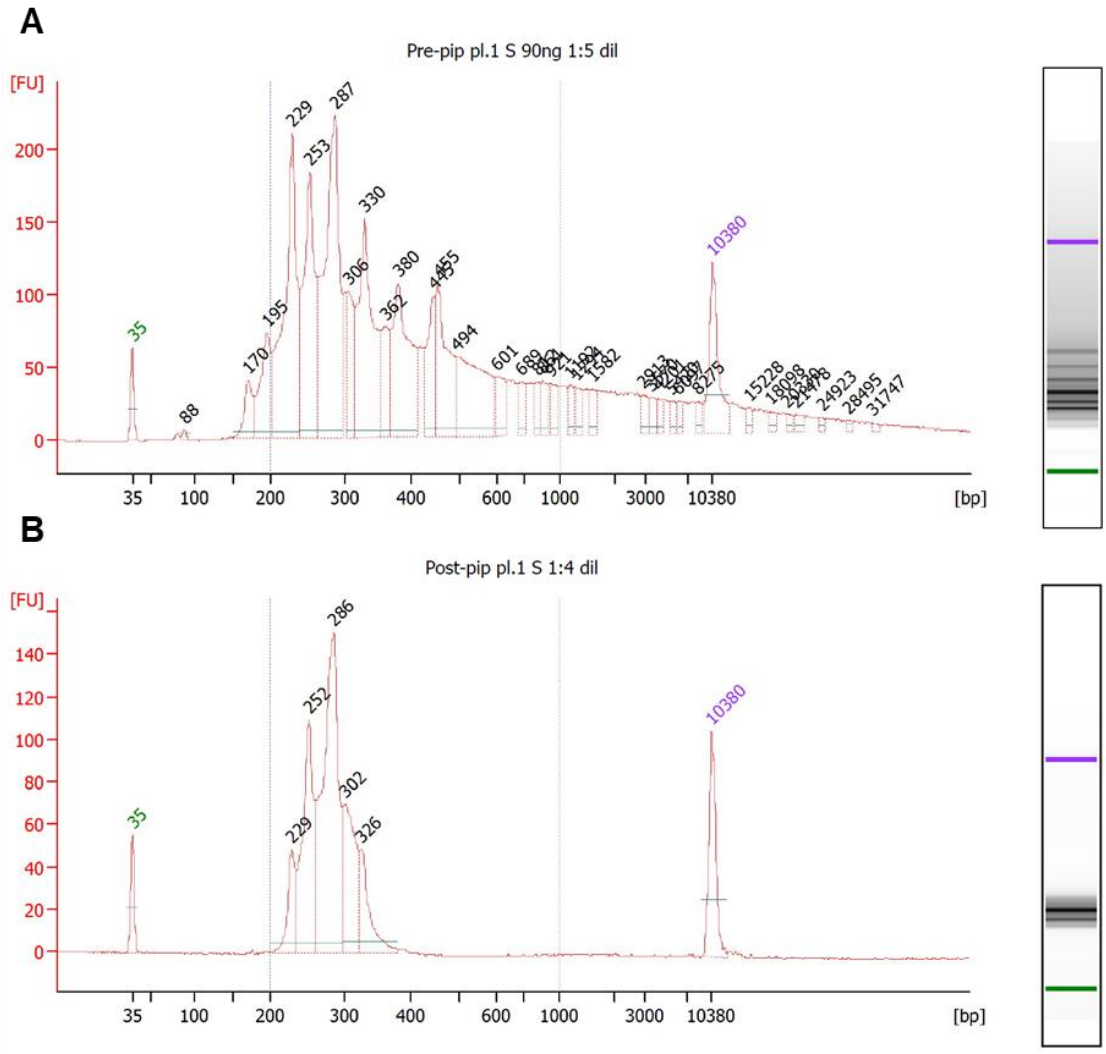


Figure S3.2 Electropherograms of an example GBS library generated using *PstI* and *MspI* restriction enzymes in a double digest. Fluorescence intensity of fragments (in fluorescence units, FU) is presented on the *y*-axis and fragment lengths (in base pairs, bp) is on the *x*-axis. **A**) pre-size selection library. *Note*: adapter dimer peak at 88 bp. **B**) post-size selection library. Size selection was for fragments within the range 193 – 300 bp. *Note*: size standard peaks at 35 bp and 10,380 bp in green and blue respectively.

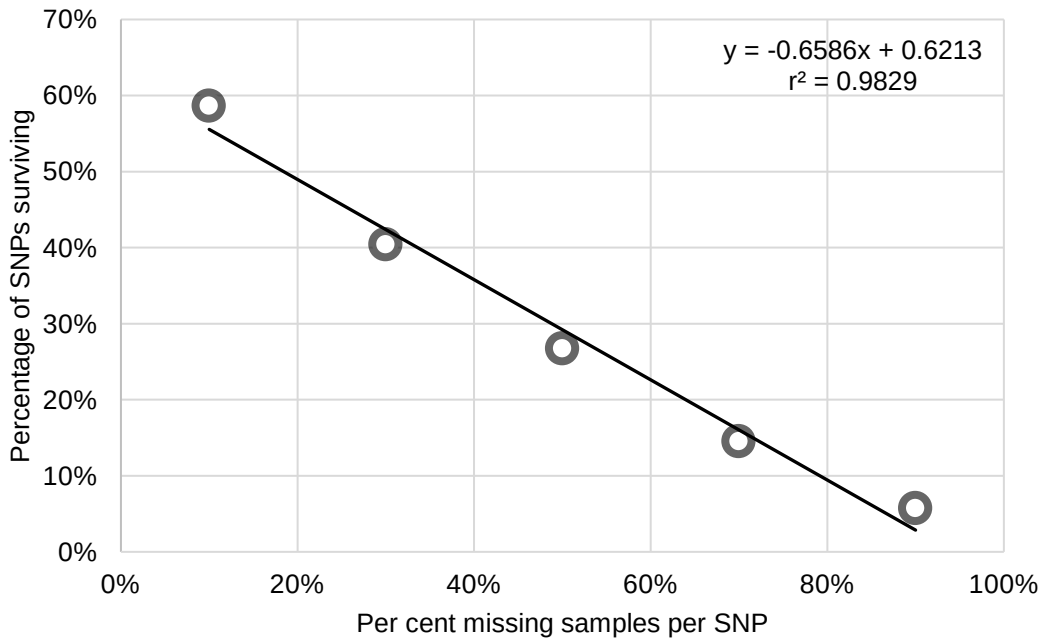


Figure S3.3 Percentage of SNPs surviving at missing samples per SNP threshold, from the mean samples from two pools (WNZLL and FNZSL, *n* = 463). SNPs were filtered on depth (5 to 150) and no multiallelic SNPs, prior to missingness threshold investigation. Line of best fit and coefficient of determination (square of the Pearson correlation, r^2) are presented.

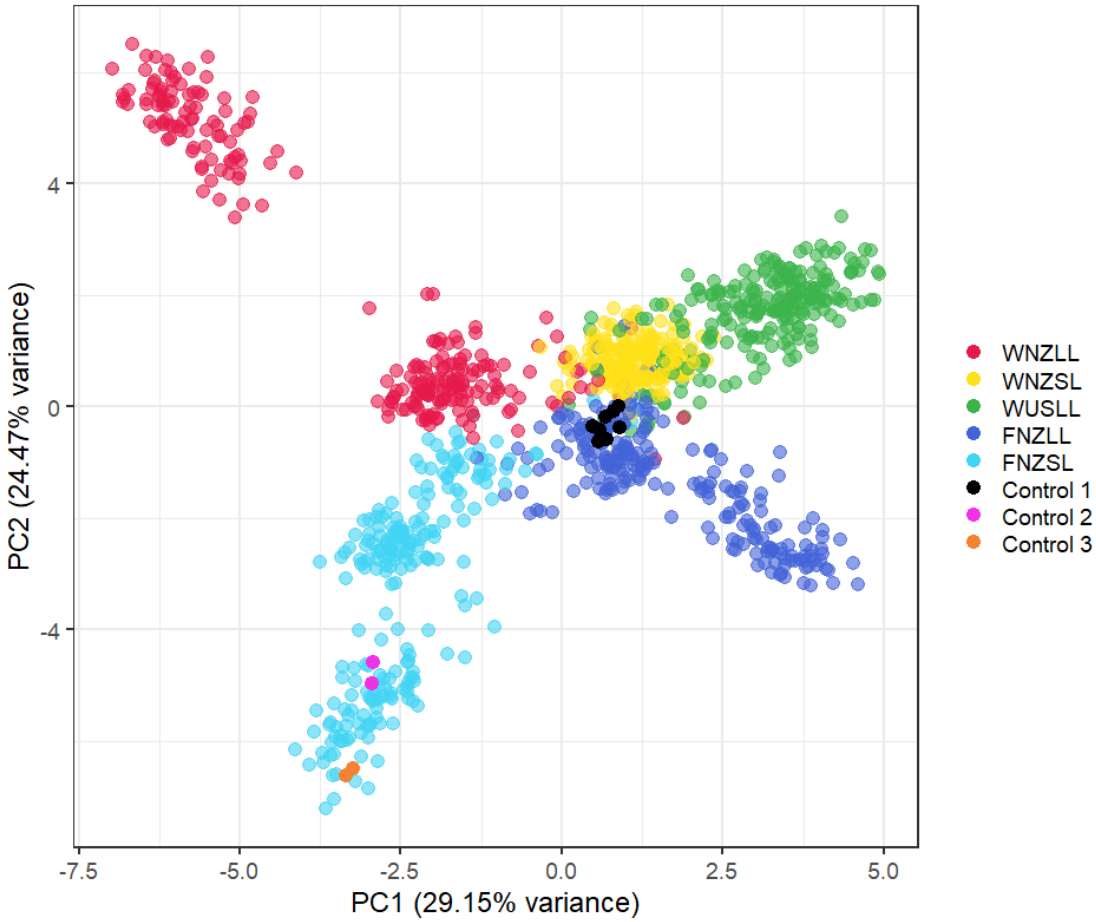


Figure S3.4 Principal component analysis plot for white clover individuals (*n* = 1,128) from five pools (WNZLL, WNZSL, WUSLL, FNZLL and FNZSL). Each point represents an individual, where the colour of the point corresponds to their pool. Thirteen repeats of a control DNA sample (one per 96-plex GBS library) are represented by black dots near the centre of the plot. Control 2, represented by two pink dots, is a DNA sample from FNZSL-Low-Mid repeated in the same GBS library (library 10). Control 3, represented by two orange dots, is a DNA sample from FNZSL-Low-End repeated in the same GBS library (library 11).

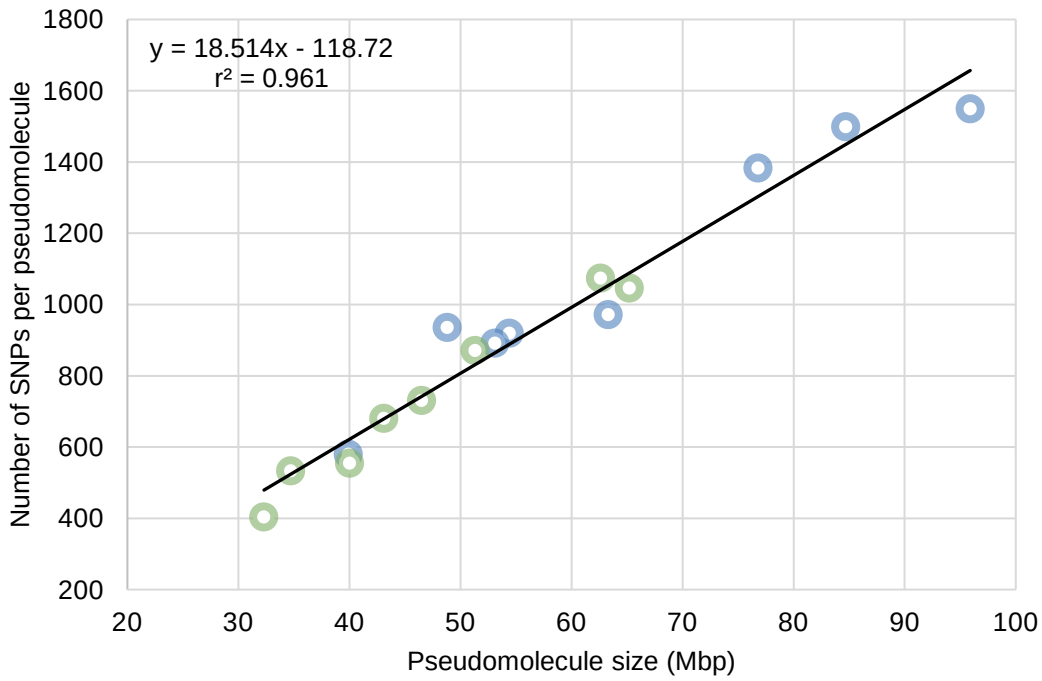


Figure S3.5 Relationship between the number of filtered SNPs per pseudomolecule and pseudomolecule size (Mbp) using samples from all pools (*n* = 1,113). SNPs were filtered to include only biallelic SNPs, a minimum and maximum read depth range of 5 to 150, maximum missing data to 20% per SNP and including SNPs with a minor allele frequency of ≥ 0.03. Data points in blue represent the Tr_{To} pseudomolecule and the Tr_{Tp} pseudomolecule are in green. Line of best fit and coefficient of determination (square of the Pearson correlation, r^2) are presented.

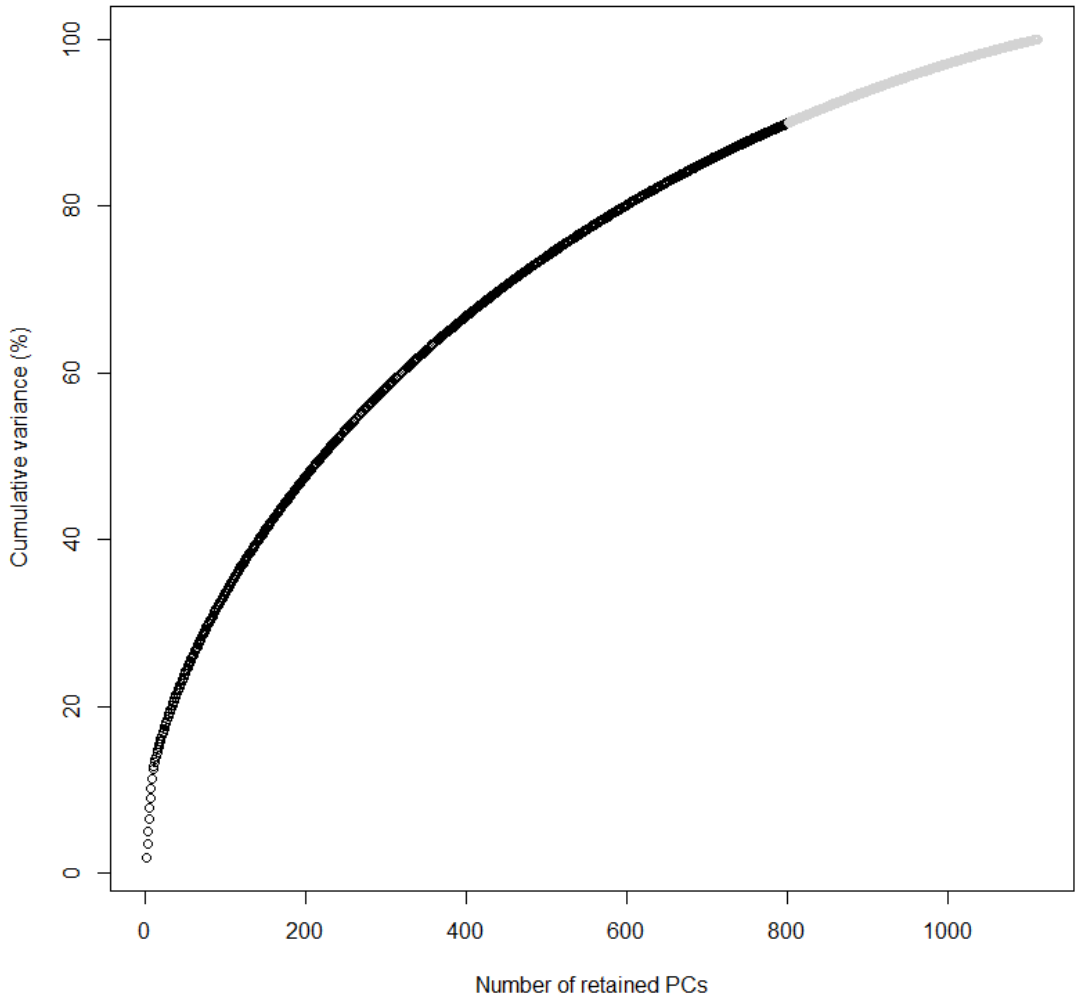


Figure S3.6 Cumulative variance explained by the principal component analysis (PCA) relative to the number of principal components (PCs) retained prior to the *K*-means analysis. Black circles represent the PCs retained (800) and grey circles represent the PCs not retained.

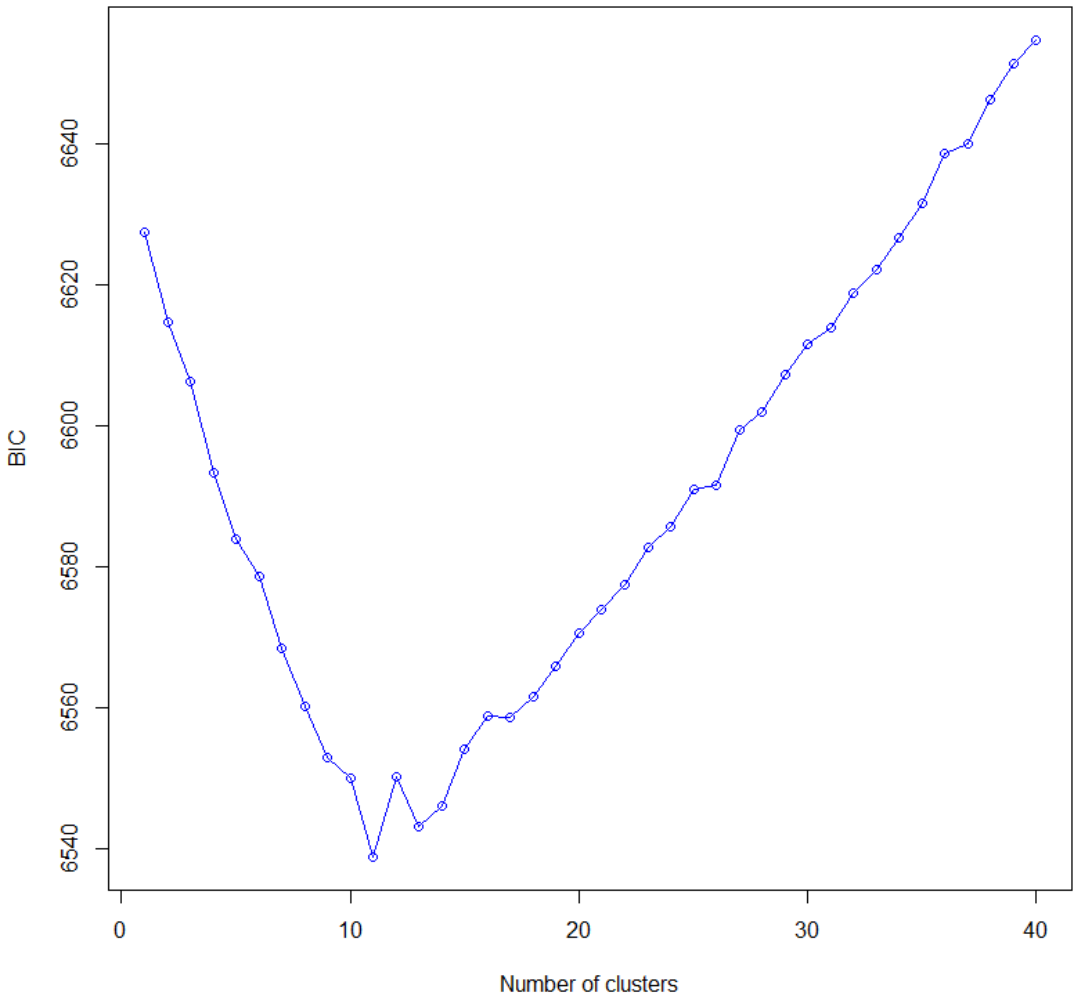


Figure S3.7 Selection of the optimal number of clusters for discriminant analysis of principal components (DAPC) using *K*-means algorithm and the lowest Bayesian information criterion (BIC). The graph shows a clear decrease of BIC until *K* = 11 clusters as the most likely value of *K*, after which BIC increases.

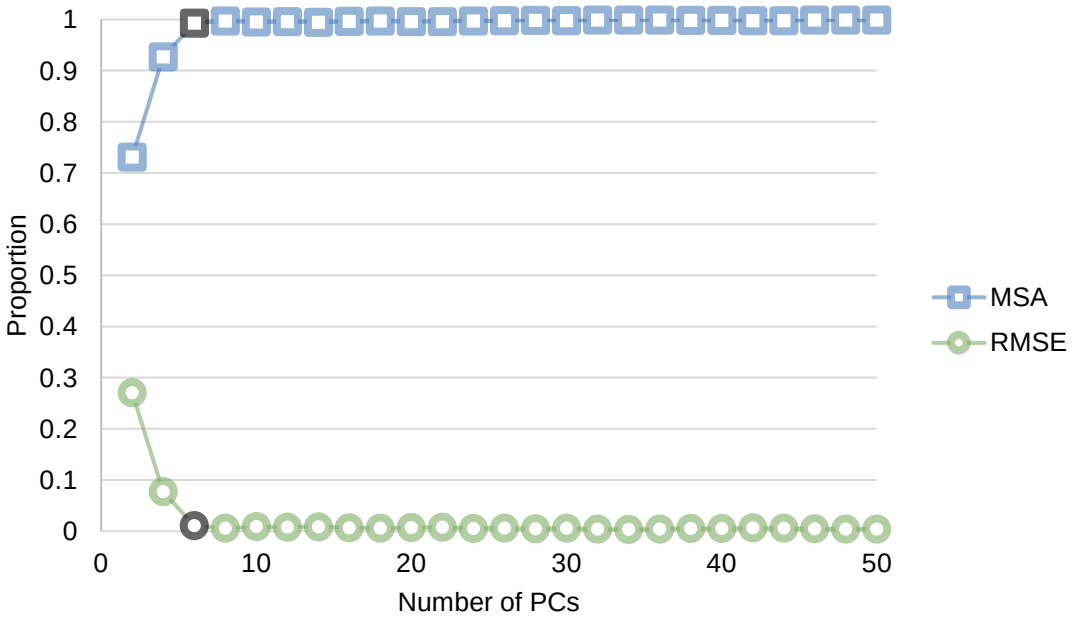


Figure S3.8 Cross-validation results from discriminant analysis of principal components (DAPC) for *K* = 11. Proportion of successful assignment of the validation set (10% of the data) is presented on the *y*-axis and the number of principal components (PCs) is presented on the *x*-axis (every second PC is represented). Mean values from 100 replicate runs for each PC are presented for mean successful assignment (MSA) as blue squares, and root mean square error (RMSE) as green circles. RMSE and MSA values plateau at six PCs as indicated by the black points.

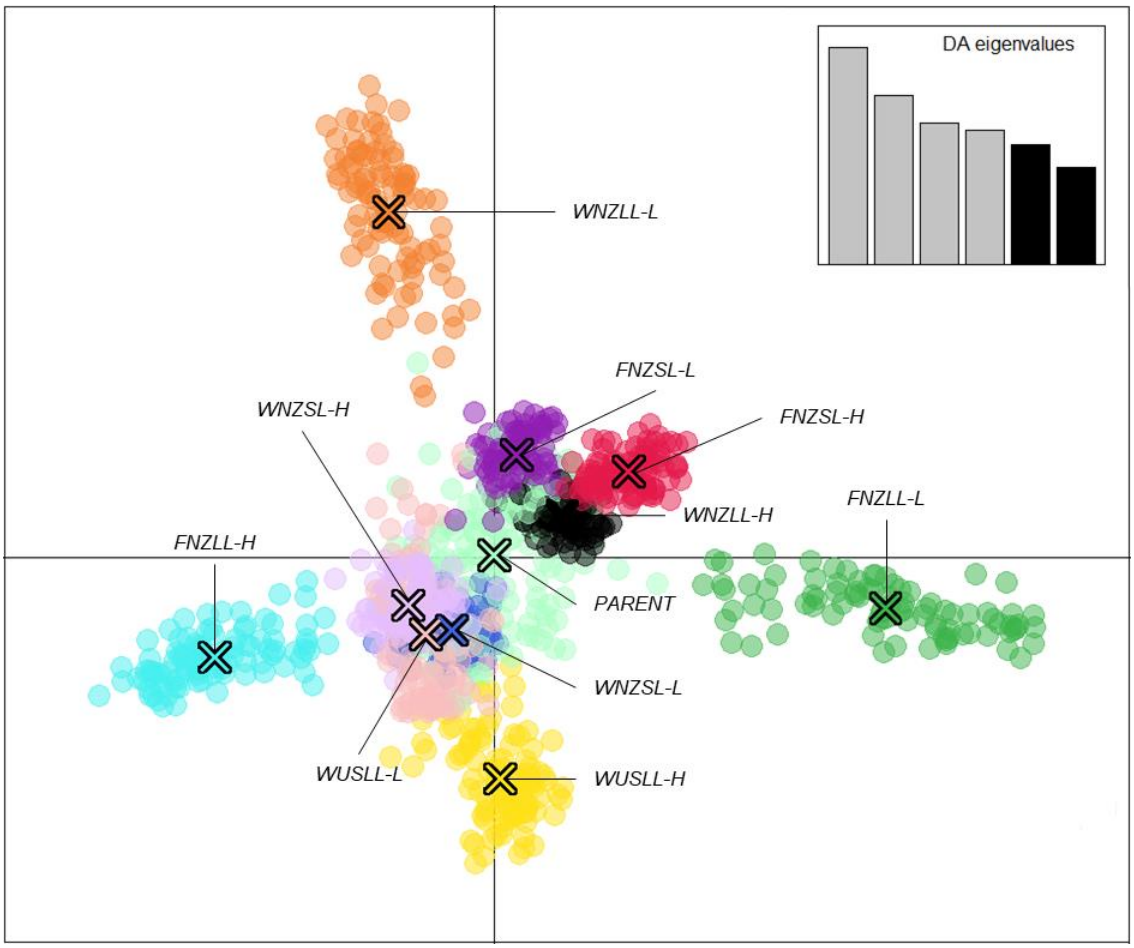


Figure S3.9 Discriminant analysis of principal component (DAPC) scatter plot of 1,113 individuals using 14,743 SNPs based on 11 assigned genetic clusters. Six PCs and six discriminant functions (DF) were retained for analyses to describe the relationship between the genetic clusters. The scatterplot shows the fifth and sixth DFs from the DAPC analysis with the scree plot of eigenvalues of the linear discriminant analysis (LDA) shown in the inset. Populations are labelled and colour coded by *K* = 11 as determined from the *K*-means clustering algorithm. Each dot represents a single individual and the centre of each cluster, as determined by a minimum spanning tree based on the squared distances between populations, is indicated by a cross.

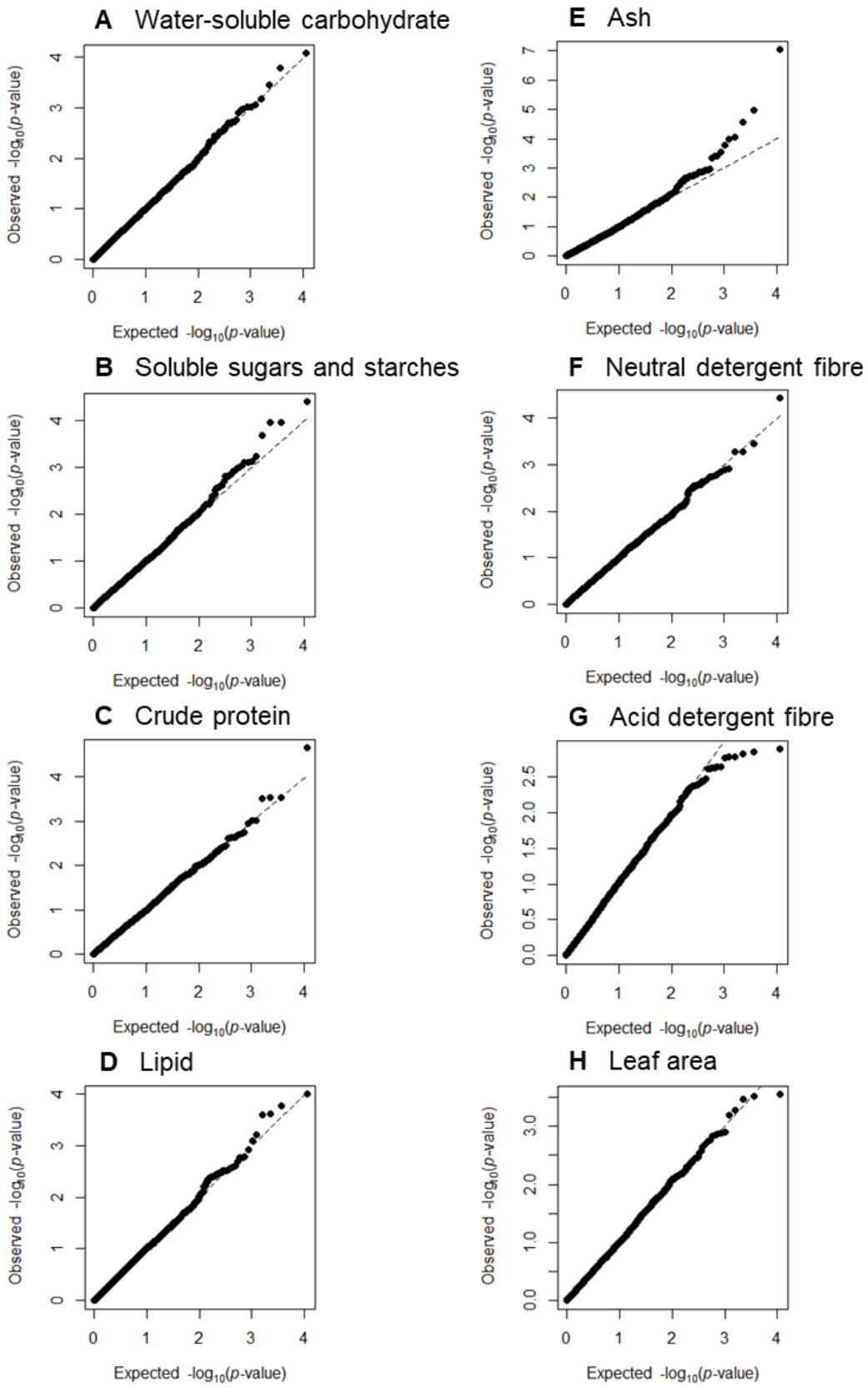


Figure S3.10 Quantile-Quantile (Q-Q) plots of expected *p*-values on *x*-axis and observed *p*-values on *y*-axis for each SNP in eight phenotypic traits investigated in the genome-wide association study (GWAS) (**Figure 3.4**). Most *p*-values are similar to the expected diagonal in the Q-Q plots, which indicates the appropriateness of the GWAS model.

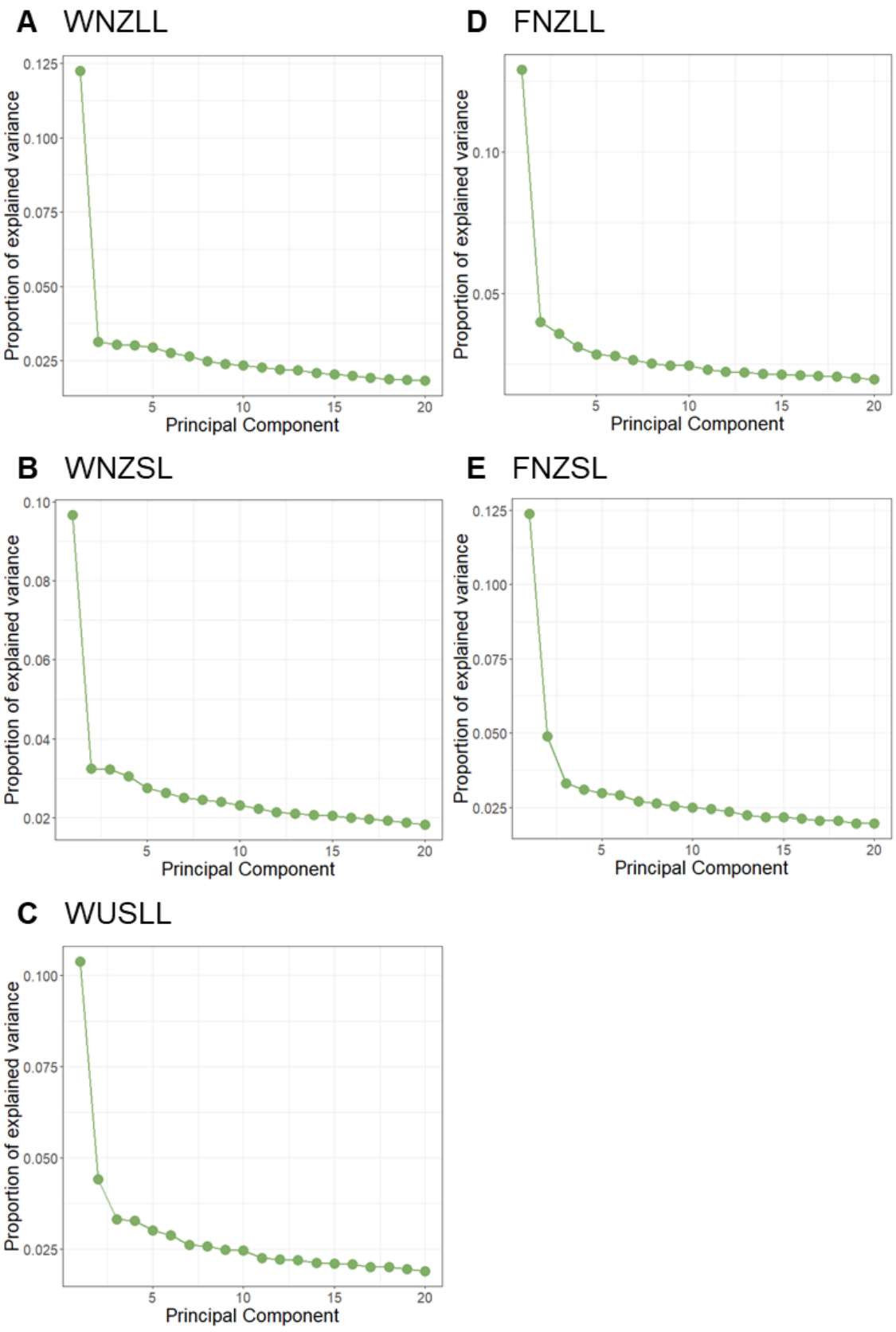


Figure S3.11 Scree plots for each pool determined in PCAdapt analysis. Proportion of explained variance is displayed on the *y*-axis with K_{PC} values (number of principal components) from 1 to 20 displayed on the *x*-axis.

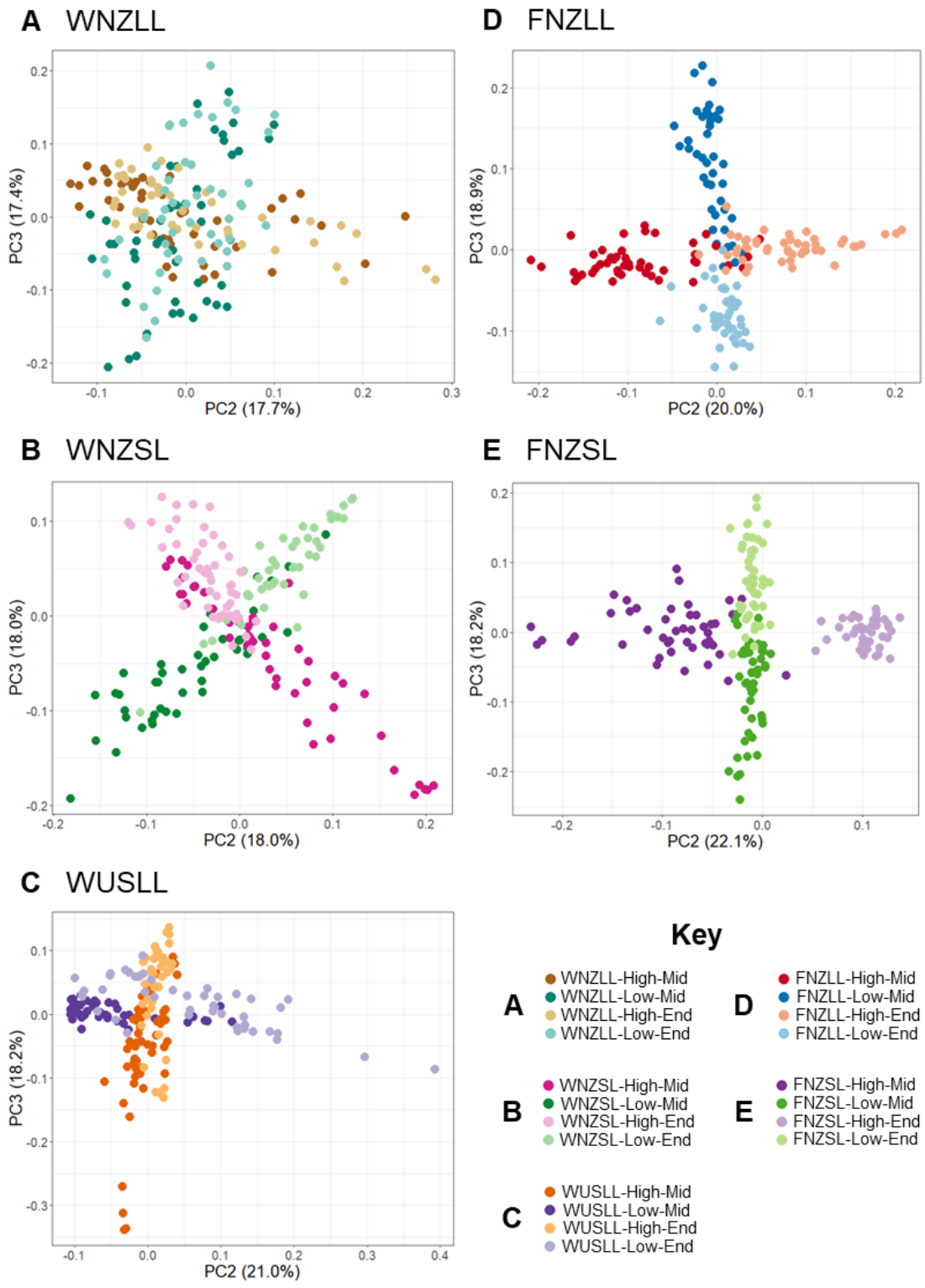


Figure S3.12 Score plots from PCAdapt analysis for each pool with principal component (PC) 2 and PC 3 displayed. Each dot represents a single individual and the colour corresponds to individuals from the same population. Each pool has four populations as the Parent populations was excluded from the analysis. Population information is displayed in the key in the bottom right corner.

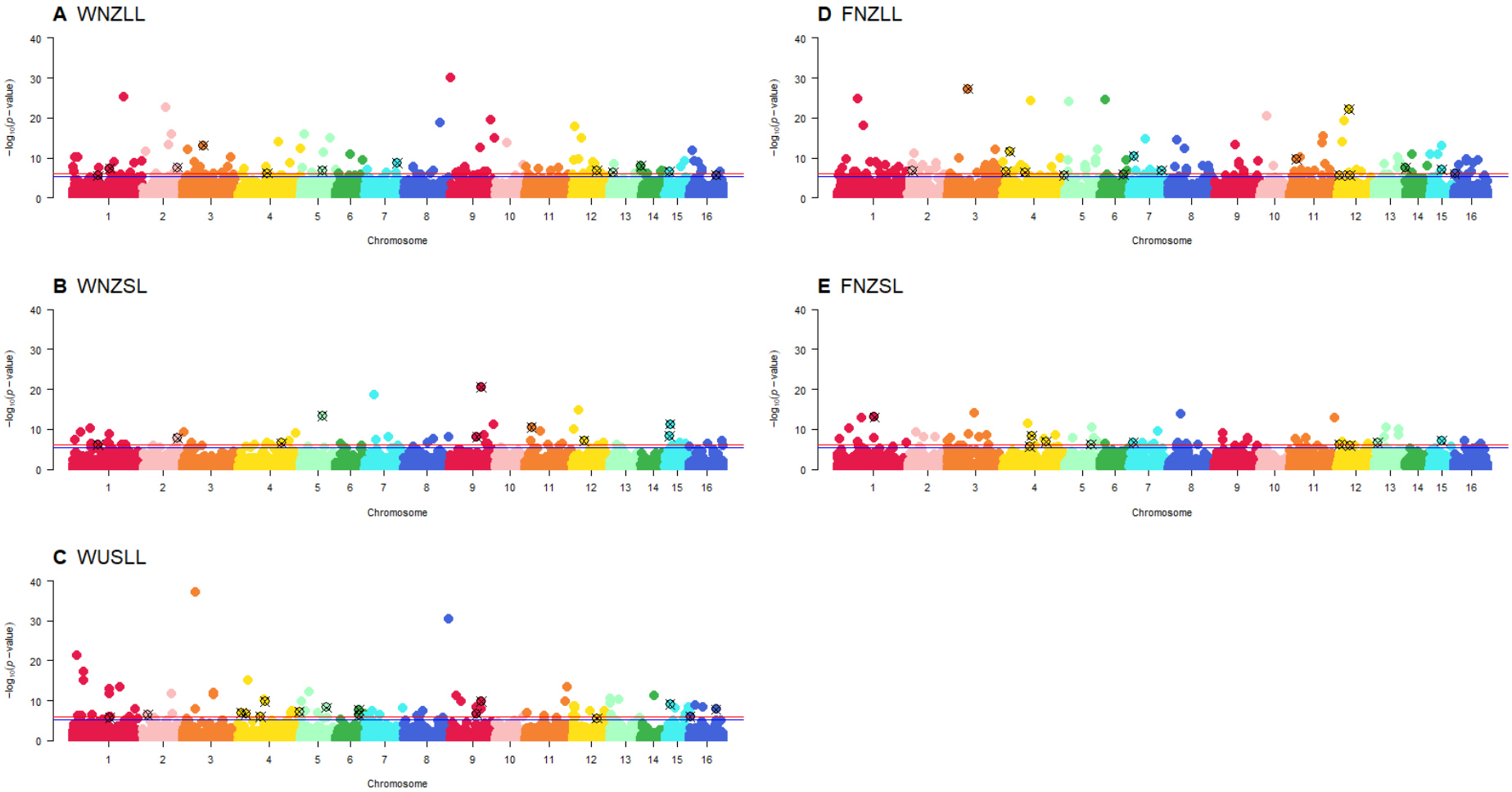


Figure S3.13 Manhattan plots representing the outlier SNPs from PCAdapt analysis comparing the high and low water-soluble carbohydrate populations within each pool. **A**) Outlier SNPs for WNZLL pool, *n* = 188, SNPs = 11,061. **B**) Outlier SNPs for WNZSL pool, *n* = 186, SNPs = 11,479. **C**) Outlier SNPs for WUSLL pool, *n* = 195, SNPs = 11,171. **D**) Outlier SNPs for FNZLL pool, *n* = 182, SNPs = 10,979. **E**) Outlier SNPs for FNZSL pool, *n* = 184, SNPs = 10,976. -$\log_{10}$(*p*-values) are plotted against physical map position of SNPs with subgenomes of corresponding

chromosomes (i.e., pseudomolecules) similarly coloured (Tr_{To} 1 – 8 and Tr_{Tp} 9 – 16). Significant loci lie above the FDR thresholds as denoted by the red (α = 0.01) and blue (α = 0.05) solid lines (See **Table S3.3**, Appendix 2 for threshold values). SNPs highlighted by black symbols are present in more than two pools with *p*-values above the FDR thresholds.

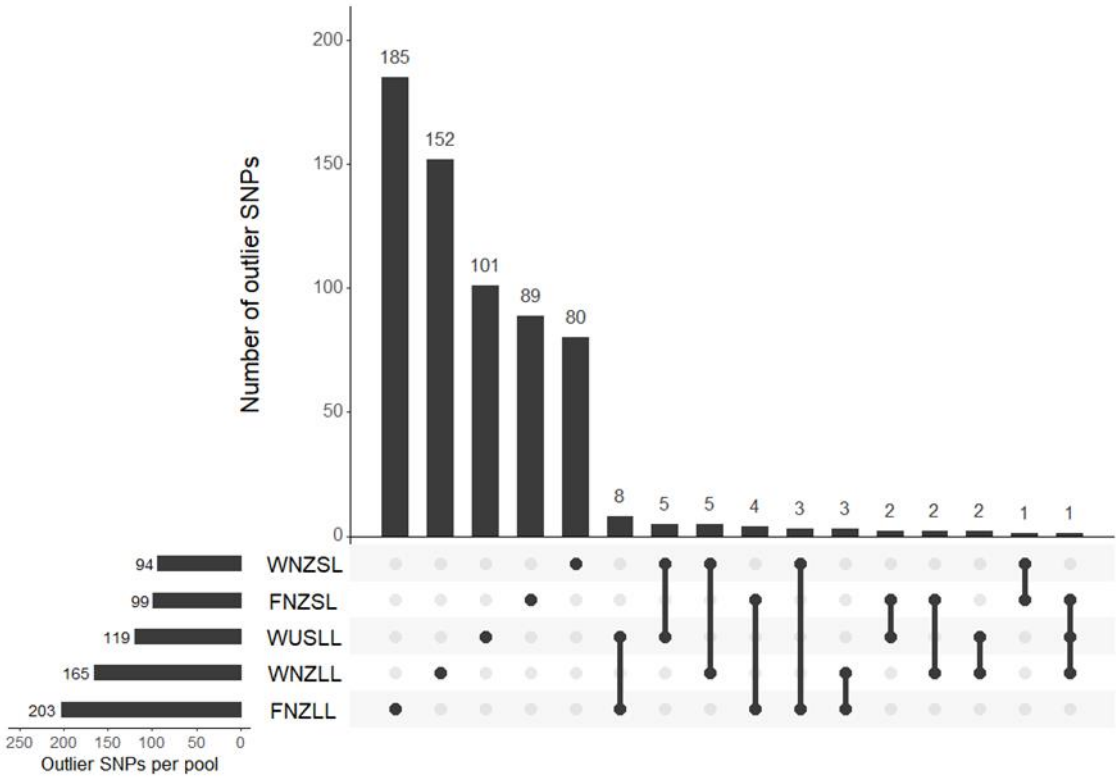


Figure S3.14 Upset plot of outlier SNPs per pool detected by the PCAdapt analysis. The total number of outliers detected in each pool is displayed by the horizontal bars on the left of the plot. The vertical bars represent the number of shared outliers. Within the matrix, the first five dark circles indicate unique outliers detected only in that pool. When dark circles are connected by vertical bars it indicates outliers are in common between multiple pools. The plot is sorted by the number of outliers detected and uniqueness of the pool.

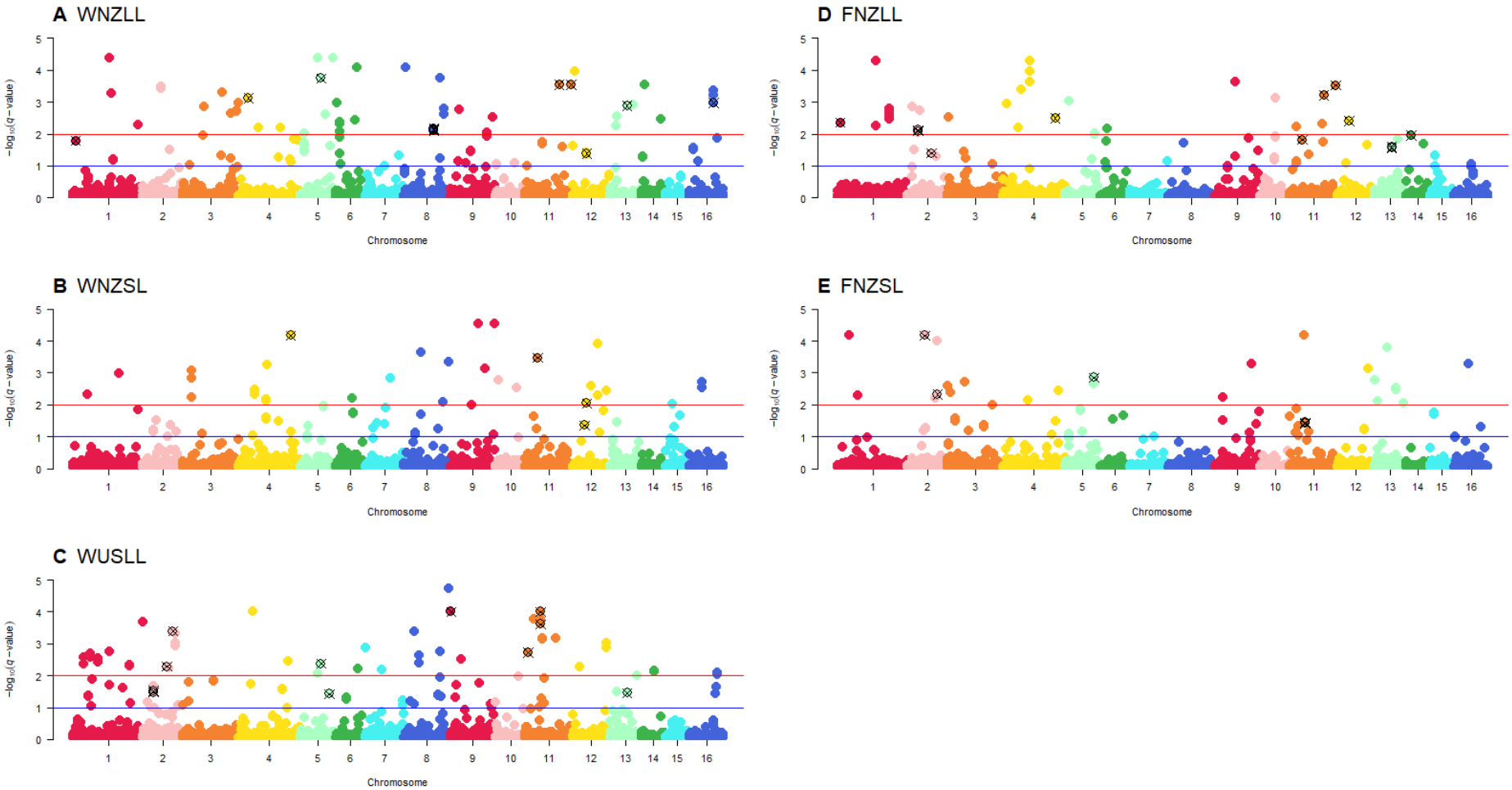


Figure S3.15 Manhattan plots representing the outlier SNPs from BayeScan analysis comparing the high and low water-soluble carbohydrate populations within each pool. **A**) Outlier SNPs for WNZLL pool, *n* = 188, SNPs = 14,598. **B**) Outlier SNPs for WNZSL pool, *n* = 186, SNPs = 14,620. **C**) Outlier SNPs for WUSLL pool, *n* = 195, SNPs = 14,623. **D**) Outlier SNPs for FNZLL pool, *n* = 182, SNPs = 14,626. **E**) Outlier SNPs for FNZSL pool, *n* = 184, SNPs = 14,619. -$\log_{10}$(*q*-values) are plotted against physical map position of SNPs with subgenomes of corresponding

chromosomes (i.e., pseudomolecules) similarly coloured (Tr_{To} 1 – 8 and Tr_{Tp} 9 – 16). Significant loci lie above the FDR thresholds as denoted by the red (α = 0.01) and blue (α = 0.05) solid lines. SNPs highlighted by black symbols are present in more than two pools with *q*-values above the FDR thresholds.

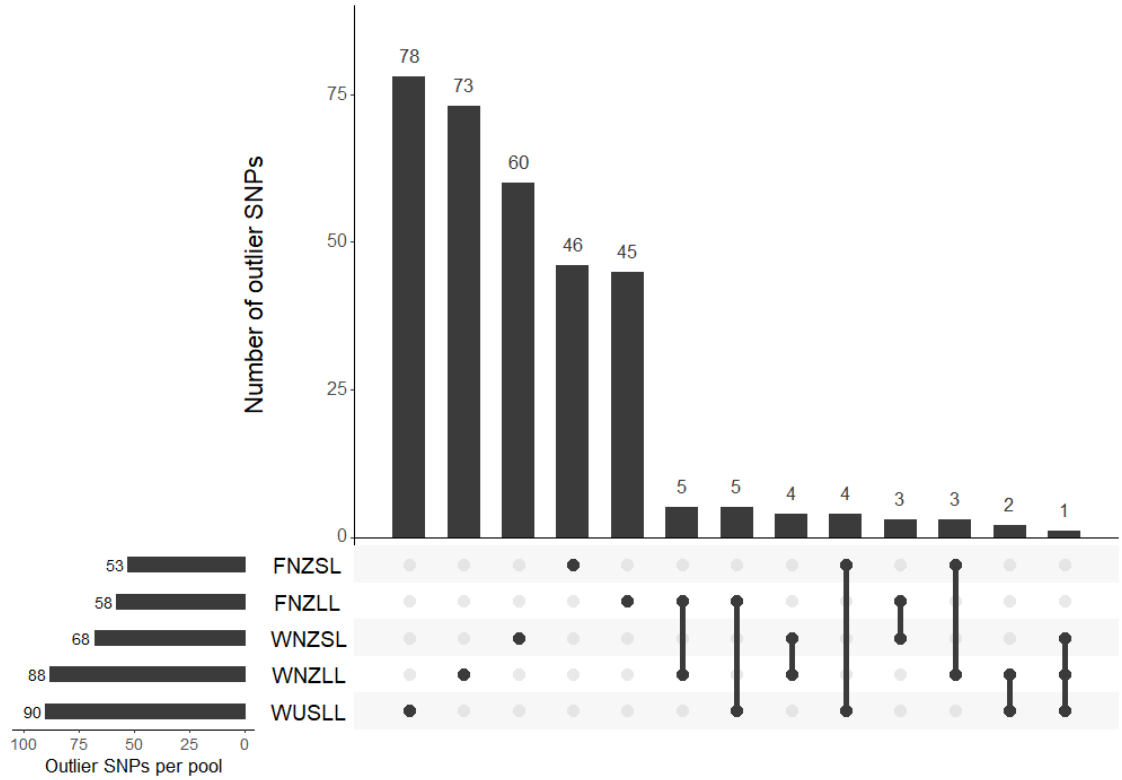


Figure S3.16 Upset plot of outlier SNPs per pool detected by the BayeScan analysis. The total number of outliers detected in each pool is displayed by the horizontal bars on the left of the plot. The vertical bars represent the number of shared outliers. Within the matrix, the first five dark circles indicate unique outliers detected only in that pool. When dark circles are connected by vertical bars it indicates outliers are in common between multiple pools. The plot is sorted by the number of outliers detected and uniqueness of the pool.

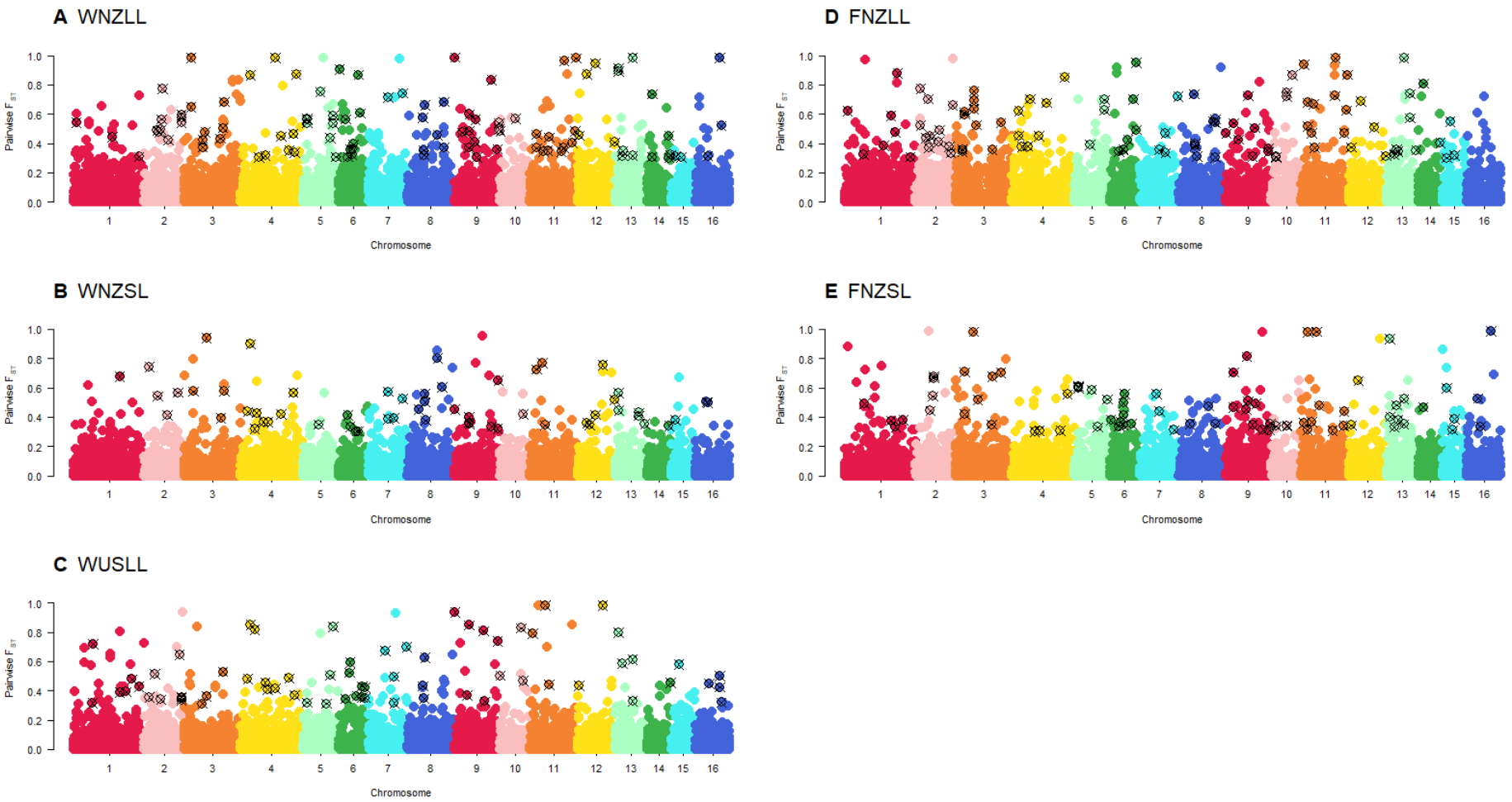


Figure S3.17 Manhattan plots representing the outlier SNPs from KGD-F_{ST} analysis comparing the high and low water-soluble carbohydrate populations within each pool using 14,743 SNP markers. **A**) Outlier SNPs for WNZLL pool, *n* = 188. **B**) Outlier SNPs for WNZSL pool, *n* = 186. **C**) Outlier SNPs for WUSLL pool, *n* = 195. **D**) Outlier SNPs for FNZLL pool, *n* = 182. **E**) Outlier SNPs for FNZSL pool, *n* = 184. Pairwise F_{ST} values

are plotted against physical map position of SNPs with subgenomes of corresponding pseudomolecules (i.e., chromosomes) similarly coloured (Tr_{To} 1 – 8 and Tr_{Tp} 9 – 16). The SNPs with black symbols are present in two or more pools at an F_{ST} greater than 0.3.

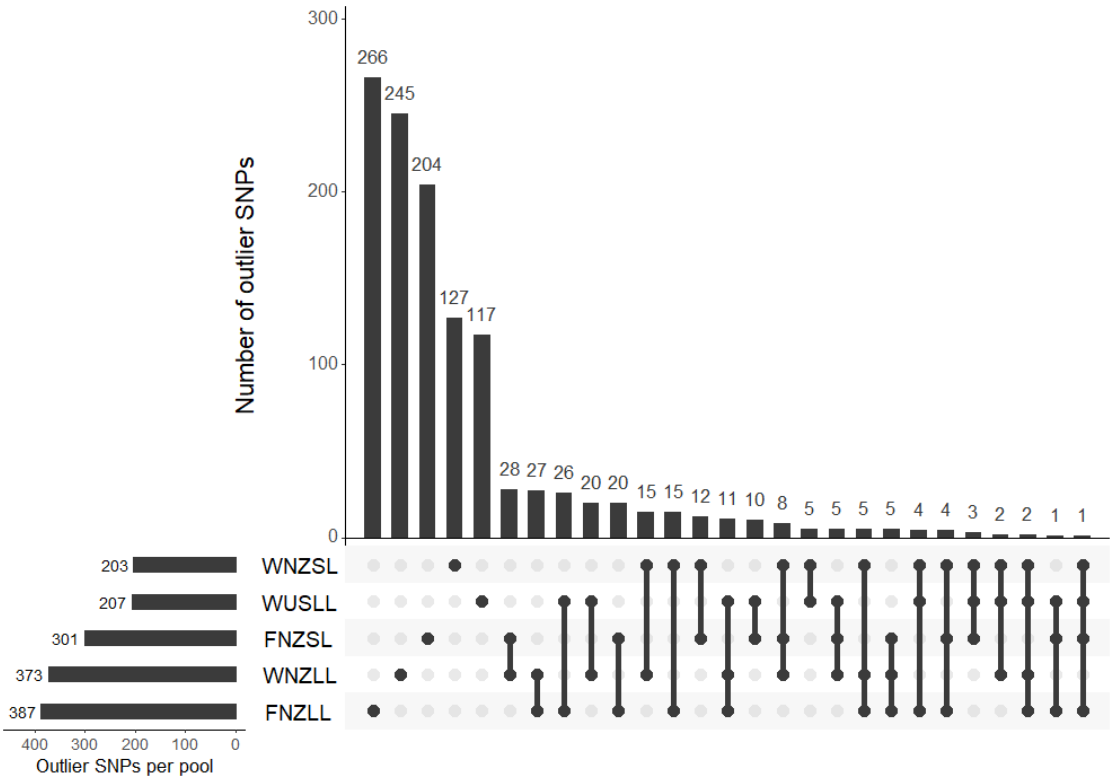


Figure S3.18 Upset plot of SNPs with F_{ST} values greater than 0.3 per pool detected by the KGD-F_{ST} analysis. The total number of outliers detected in each pool is displayed by the horizontal bars on the left of the plot. The vertical bars represent the number of shared outliers. Within the matrix, the first five dark circles indicate unique outliers detected only in that pool. When dark circles are connected by vertical bars it indicates outliers are in common between multiple pools. The plot is sorted by the number of outliers detected and uniqueness of the pool.

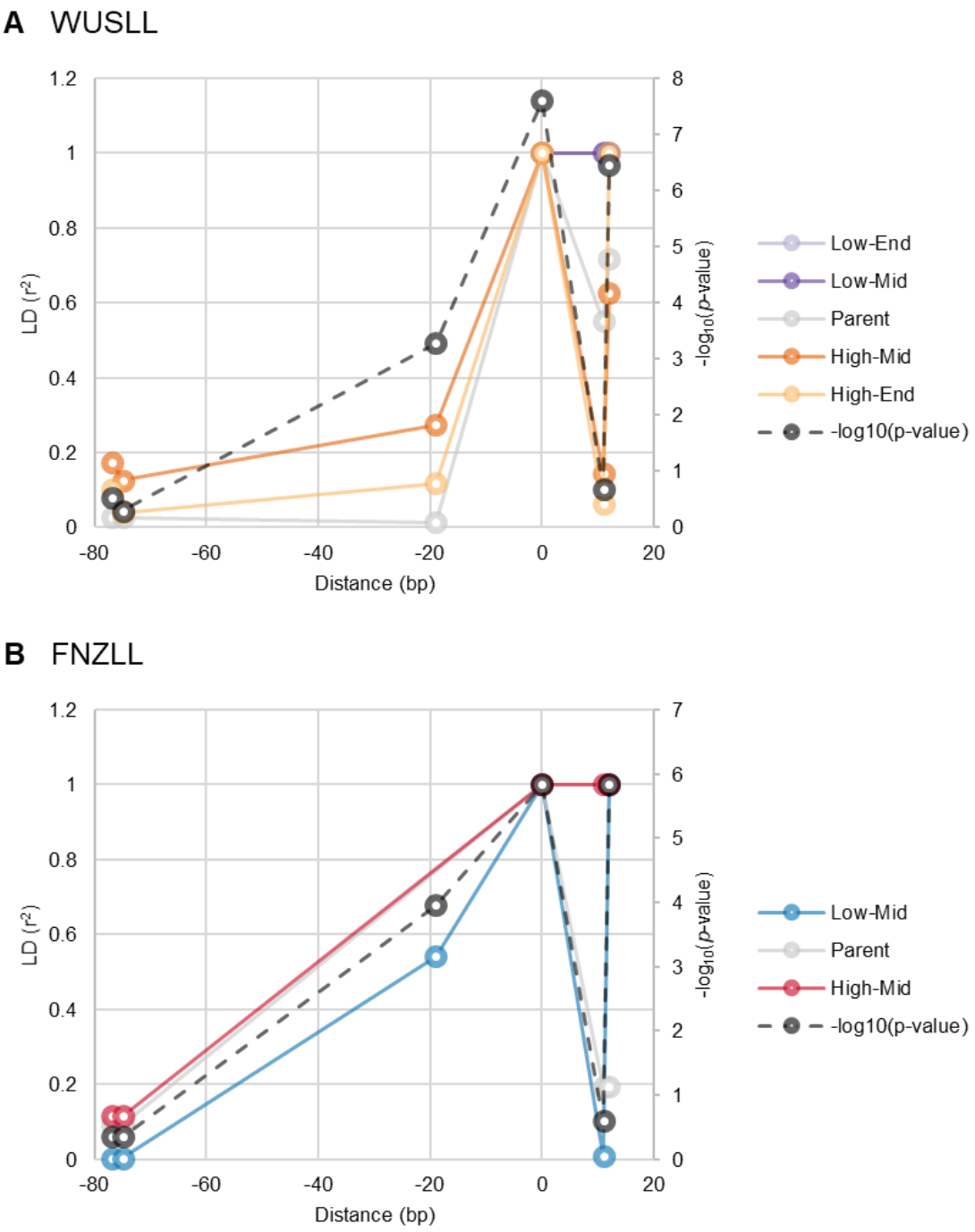


Figure S3.19 Linkage disequilibrium (LD) and -$\log_{10}$(*p*-value) in a 100 base pair (bp) window for 6_31429353 in the WUSLL pool (**A**) and FNZLL pool (**B**). r^2 values for populations are joined by solid lines with primary axis on the left-hand side. -$\log_{10}$(*p*-values) for the pool are joined by a dotted line with secondary axis on the right-hand side. Distance from the SNP (at zero) is the horizontal axis.

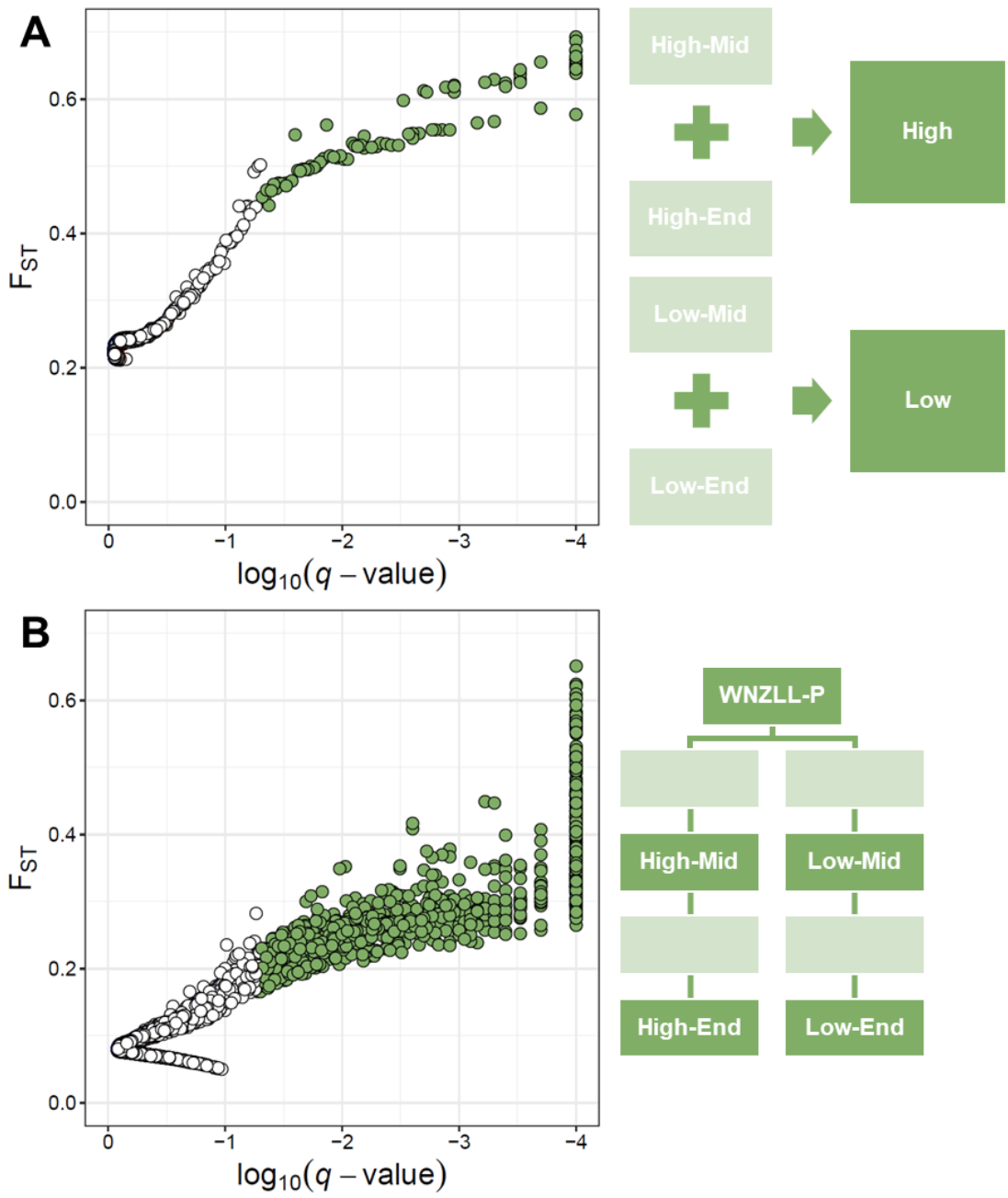


Figure S3.20 BayeScan test for selection on 14,743 SNPs in the WNZLL pool. **A**) Outlier analysis using the *High* vs *Low* comparison, comprising of high water-soluble carbohydrate (WSC) populations grouped together and low WSC populations grouped together and excluding the Parent population (see corresponding diagram to the right of **A**). This was the structure used in the outlier SNP detection analysis. **B**) Comparative BayeScan analysis using all five populations in the WNZLL pool (see corresponding diagram to the right of **B**).
Note: White circles represent SNPs under neutral selection (either balancing or purifying) and green circles represent SNPs under divergent selection at a false discovery rate of 0.05. Populations in dark green boxes in the diagrams were used in the analysis. *q*-values of 0 were changed to 0.0001 before log conversion, leading to a high number of SNPs sitting at -4 on the *x*-axis.

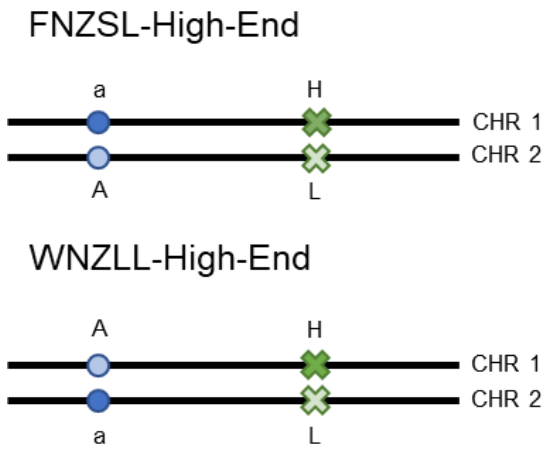


Figure S3.21 Phase state diagram for SNP on pseudomolecule 16 at 32,428,574 bp for FNZSL and WNZLL pools. Allele 'a' for SNP 16_32428574 is in coupling with the gene for high water-soluble carbohydrate (WSC) and allele 'A' is in repulsion for the FNZSL-High-End population. Allele 'A' is in coupling with the gene for high WSC and allele 'a' is in repulsion for the WNZLL-High-End population.
Note: circles represent SNP position and crosses represent gene location. H = high WSC, L = low WSC, CHR 1 = homologous chromosome 1, and CHR 2 = homologous chromosome 2.

A

SNP 2	SNP 1	
	A	**a**
B	**AB**	**aB**
	6	**0**
b	**Ab**	**ab**
	14	**0**

B

Haplotype frequency

AB	**6 ÷ 20**
	0.3
aB	**0 ÷ 0**
	0
Ab	**14 ÷ 20**
	0.7
ab	**0 ÷ 20**
	0

Allele frequency

A	**20 ÷ 20**
	1
a	**0 ÷ 20**
	0
B	**6 ÷ 20**
	0.3
b	**14 ÷ 20**
	0.7

C

$$r = \left(\frac{(AB \times ab) - (aB \times Ab)}{A \times a \times B \times b}\right)^{\frac{1}{2}}$$

$$r = \left(\frac{(0.3 \times 0) - (0 \times 0.7)}{1 \times 0 \times 0.3 \times 0.7}\right)^{\frac{1}{2}}$$

$$r = \left(\frac{(0)-(0)}{0}\right)^{\frac{1}{2}}$$

$$r^2 = (NaN)^2$$

$$r^2 = NaN$$

Figure S3.22 Example of linkage disequilibrium (LD) calculation between two loci using 20 individuals. **A**) number of individuals with each haplotype. **B**) Haplotype and allele frequencies. **C**) Equations to calculate r^2 using example haplotype and allele frequencies.

APPENDIX 3

Chapter 4 Supplementary Material

SUPPLEMENTARY METHODS

Protein extraction and sequential window acquisition of all theoretical fragment ion spectra

Sample preparation

The following procedure was followed to extract proteins from leaf tissue. Leaf tissue was ground to a fine powder in liquid Nitrogen. Fifty mg of the leaf powder was then resuspended in 10% trichloroacetic acid in acetone, 0.07% β-mercaptoethanol, followed by incubation at -20°C for 1 h. The extract was centrifuged for 30 minutes at 16,000 g, and the pellet was collected and washed with 1.5 ml of 100% acetone followed by centrifugation for 30 minutes at 16,000 g. The acetone washing step was repeated three times for the complete removal of pigments, lipids, and other lipophilic molecules. The colourless resulting pellet was dried in a vacuum centrifuge and resuspended in 2% SDS in 8 M urea, 100 mM Tris–HCl (pH 8.8), followed by reduction and alkylation with 10 mM DTT and 20 mM iodoacetamide. Methanol/Chloroform precipitation was performed, and the resulting protein pellet was reconstituted in 8 M urea, 100 mM Tris-HCl (pH 8.8) in water. Protein concentrations were determined using BCA assay (Thermo Scientific, USA) as per the manufacturer's instruction. Sample proteolysis with Lys-C (100:1 protein to enzyme ratio) was performed at 28°C overnight followed by digestion with trypsin (100:1 protein to enzyme ratio) at 37°C for 6 h. The pH was adjusted to approximately 3 using a final concentration of approximately 1% TFA, and each sample desalted using a Stage tips containing Styrene Divinyl Benzene (Empore SDB-RPS 47 mm extraction disk, SUPLCO). Briefly, stage tips were self-packed into pipette tips, peptides were bound to the stage-tip, washed with 0.2% TFA and finally eluted with 80% acetonitrile: 5% ammonium hydroxide. Peptides were dried by vacuum centrifuge and then reconstituted in 200 mM HEPES pH 8.8 followed by peptides concentration determination using the Pierce quantitative colorimetric peptide assay (Thermo Scientific, USA). Five µL for each sample was diluted with 5 µL of loading buffer (0.1% formic acid) to be used for sequential window acquisition of all theoretical fragment ion spectra (SWATH) analysis. SWATH were acquired in random order with one blank run after each sample.

High pH RP-HPLC

For ion library generation through high pH (HpH) fractionation, 20 µg of each sample were pooled and fractionated by high pH RP-HPLC. The pooled sample was first cleaned up by C18 stage tip then dried and resuspended in mobile phase buffer A (5

mM ammonium hydroxide solution (pH 10.5)). The composition of buffer B was 5 mM ammonia solution with 90% Acetonitrile (pH 10.5). After sample loading and washing with 3% buffer B for 10 minutes at a flow rate of 300 µL/min, the buffer B concentration was increased from 3% to 30% over 55 minutes and then to 70% between 55 to 65 minutes and to 90% between 65-70mins. The eluent was collected every 2 minutes at the beginning of the gradient and at one-minute intervals for the rest of the gradient.

2D-IDA

Following HpH-RP-HPLC separation, 17 fractions were concatenated (0 – 85 min), dried and resuspended in 20 µL of loading buffer. 10 µL per fraction was taken for 2D IDA analysis.

Data dependent acquisition (IDA)

Sample (10 µL) was injected onto a reverse-phase peptide trap for pre-concentration and desalted with loading buffer, at 10 µL min^{-1} for 3 minutes. The peptide trap was then switched into line with the analytical column. Peptides were eluted from the column using a linear solvent gradient from mobile phase A: mobile phase B (95:5) to mobile phase A: mobile phase B (65:35) at 5 µL min^{-1} over a 120 minute period. After peptide elution, the column was cleaned with 95% buffer B for 6 minutes and then equilibrated with 95% buffer A for 5 minutes before next sample injection. The reverse phase nanoLC eluent was subject to positive ion nanoflow electrospray analysis in an information dependant acquisition mode (IDA). In the IDA mode a TOF-MS survey scan was acquired (m/z 350-1500, 0.25 second) with the 20 most intense multiply charged ions (2+ – 5+; exceeding 200 counts per second) in the survey scan sequentially subjected to MS/MS analysis. MS/MS spectra were accumulated for 100 milliseconds in the mass range m/z 100 – 1800 with rolling collision energy.

Data independent acquisition (SWATH)

Sample (10 µL) was injected onto a reverse-phase peptide trap for pre-concentration and desalted with loading buffer, at 10 µL min^{-1} for 3 minutes. The peptide trap was then switched into line with the analytical column. Peptides were eluted from the column using a linear solvent gradient from mobile phase A: mobile phase B (95:5) to mobile phase A: mobile phase B (65:35) at 5 µL min^{-1} over a 120 minute period. After peptide elution, the column was cleaned with 95% buffer B for 6 minutes and then equilibrated with 95% buffer A for 5 minutes before next sample injection. The reverse phase nanoLC eluent was subject to positive ion nanoflow electrospray analysis in a data independent

acquisition (SWATH). In SWATH mode, first a TOFMS survey scan was acquired (m/z 350 – 1500, 50 msec) then the 100 predefined m/z ranges were sequentially subjected to MS/MS analysis. MS/MS spectra were accumulated for 30 milliseconds in the mass range m/z 350-1500 with rolling collision energy optimised for lowed m/z in m/z window +10%.

Additional information includes: Trap column: 1 cm x 300 µm, Prontosil 120 C18H, 5 µm (Dr. Maisch GmbH); Analytical column: 15 cm x 300 µm, 3C18-CL-120, 3 µm particles - 120 Å pores (Eksigent); Loading buffer: 0.1% formic acid; Mobile phase A: 0.1% formic acid; Mobile phase B: 99.9% acetonitrile, 0.1% formic acid; NanoLC system: Eksigent Ultra nanoLC system (Eksigent) and Mass spectrometer: Triple TOF 6600 (Sciex).

Data processing: Database searches for IDA data

The data files generated by 2D-IDA-MS analysis were searched with ProteinPilot (v5.0) (Sciex) using the Paragon™ algorithm in thorough mode. Carbamidomethylation of Cys residues was selected as a fixed modification. An Unused Score cut-off was set to 1.3 (95% confidence for identification), and global protein FDR of 1%. A FASTA file containing a list of 68,557 *T. repens* proteins was used as a reference for protein identification (Griffiths *et al.*, 2019).

Data processing: SWATH extraction and quantitation

The ion library was constructed from 2D-IDAs and contained 9,854 proteins. Ion library and SWATH data files were imported into PeakView v2.2 (developed by AB SCIEX Corporation, Marsh drive Foster City, California USA). Protein peak area information in SWATH data were extracted using PeakView with the following parameters:

Top 6 most intense fragments of each peptide were extracted from the SWATH data sets (75 ppm mass tolerance, 5 minute retention time window). Shared and modified peptides were excluded. After data processing, peptides (max 100 peptides per protein) with confidence ≥ 99% and FDR ≤ 1% (based on chromatographic feature after fragment extraction) were used for quantitation.

Variant calling using STAR and GATK

Variant calling from RNA-Seq data was performed using the RNA-Seq short variant discovery workflow as a baseline, avaliable at https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels- but not all the

steps were applied. First, reads were mapped using Spliced Transcripts Alignment to a Reference (STAR) with the following settings:

STAR --runThreadN 8 --genomeDir {dir} --readFilesCommand zcat --outFilterMultimapNmax 1 --outSAMtype BAM SortedByCoordinate --outSAMunmapped Within KeepPairs --outFileNamePrefix {of} --readFilesIn {r1} {r2}

The most important setting being –outFilterMultimapNmax 1 because it only allows the reads to map once to the genome, to avoid the read mapping to both subgenomes.

Then ReadGroups are added to the reads where each sample is its own group. This was required to have on the mapped reads in order for Genome Analysis Toolkit (GATK) to accept the reads. Duplicates were then marked. This step isn't necessary since there should be no duplicates when each read is only mapped once, but it was used it because the Split'N'Trim required the output to be marked (duplicate or not). A "homebrew" script was then applied to fix the chromosome lengths in the BAM files, since STAR adds +1 to the length of each chromosome, which produces errors in GATK. Split'N'Trim was applied to prepare the reads for HaplotypeCaller in GATK, where it reformats alignments that span introns. The Base Quality was also reformatted ("recalibrated") since STAR and GATK have different definitions of quality. Finally, HaplotyperCaller was run on the reads to obtain the variants. Filtering was then applied using SelectVariants.

Permutation analysis of variance and correlation tests

Permutation analysis of variance (ANOVA) tests were performed in R using the *aovp()* function from the "*lmPerm*" v 2.1.0 package. These permutation ANOVAs were performed with 1e+7 permutations on each of the 49 transcripts and 13 proteins with pool (W or F) and selection (H or L) as separate factors and an interaction term. These tests were run in a loop using the following code:

for(i in 3:dim(x)[2]) {print(colnames(x)[i]); print(anova(aovp(x[,i]~x$Pool*x$Selection, Ca=0.000001, maxIter=10000000)))}

Where: x is a table of the log expression changes for the 49 transcripts and 13 proteins.

The correlations were also performed in R using the function *cor.test()* from the "*stats*" package. Six correlations were performed in total, three for each dataset. The three correlations included: the $\log_2$ fold change of the two pairwise comparisons, the

log_2 expression means for both of the high water-soluble carbohydrate (WSC) populations (WH and FH) and the log_2 expression means for the low WSC populations (WL and FL). Data points, Pearson correlation coefficient (r), 95% confidence interval for r and *p*-values were plotted using *ggscatter()* from the package "*ggpubr*" v 0.2.3 (Kassambara, 2019).

Genome-wide association study

A mixed linear model implemented in the R package "*rrBLUP*" (Endelman, 2011) was used to perform an association analysis on the 20 individuals from the transcriptomic dataset. Markers for this analysis were filtered to retain single nucleotide polymorphism (SNPs) with a minimum depth of 5, a minor allele frequency above 0.03, and 50% missing data before the *A.mat()* function was used to impute missing values using the EM algorithm (Poland *et al.*, 2012). This resulted in 1,025,071 SNPs used in the analysis. Population structure and family relatedness was accounted for with a kinship matrix calculated by *rrBLUP* from the transcriptomic data. To account for multiple testing, a Bonferroni correction was applied and markers passing the threshold at an α of 0.05 were considered statistically significant (Bonferroni, 1936). Manhattan and Q-Q plots were created and gene model IDs were identified for highly significant SNPs.

References

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome, 4*(3), 250-255.

Griffiths, A. G., Moraga, R., Tausen, M., Gupta, V., Bilton, T. P., Campbell, M. A., . . . Andersen, S. U. (2019). Breaking Free: The Genomics of Allopolyploidy-Facilitated Niche Expansion in White Clover. *The Plant Cell, 31*(7), 1466-1487.

Kassambara, A. (2019). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2.3. Retrieved from https://CRAN.R-project.org/package=ggpubr

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., . . . Jannink, J.-L. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome, 5*(3), 103-113.

SUPPLEMENTARY TABLES

Table S4.1 Samples used in transcriptomic and proteomic studies and their corresponding population.

Analysis	Population			
	WH	WL	FH	FL
Transcriptomic	36-20	37-16	73-18	72-12
	36-24	37-28	73-28	72-28
Proteomic	36-16	37-14	73-14	72-16
	36-22	37-20	73-22	72-22
	36-26	37-30	73-24	72-26
Transcriptomic and Proteomic	36-12	37-12	73-12	72-14
	36-14	37-18	73-16	72-24
	36-18	37-26	73-20	72-30

Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

Table S4.2 Extracted RNA quality and quantity results from Nanodrop spectrophotometer, Qubit and LabChip.

Population	Sample name	260/280	260/230	RIN	ug μL^{-1}
WH	36-12	2.06	1.99	8.5	78.8
WH	36-14	2.17	2.84	8.4	128
WH	36-18	2.18	2.32	8.6	127
WH	36-20	2.18	2.52	8.6	114
WH	36-24	2.17	2.45	7.7	173
WL	37-12	2.17	2.19	9.1	447
WL	37-16	2.17	2.02	7.8	119
WL	37-18	2.18	2.54	8.9	268
WL	37-26	2.16	2.40	8.7	358
WL	37-28	2.18	2.37	9.1	390
FH	73-12	2.19	2.31	8.4	341
FH	73-16	2.19	2.29	8.2	161
FH	73-18	2.21	2.30	8.2	433
FH	73-20	2.13	2.18	8.8	297
FH	73-28	2.20	2.24	8.9	132
FL	72-12	2.20	2.39	8.9	288
FL	72-14	2.19	2.41	7.7	497
FL	72-24	2.17	2.45	7.7	337
FL	72-28	2.18	2.07	7.9	206
FL	72-30	2.22	2.29	8.0	161

Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

Table S4.3 RNA quality and sequencing data output of each sample.

Sample	Pop	RIN	RNA con (ng uL^{-1})	260/280	260/230	Raw reads	Clean reads	Raw bases (G)	Clean bases (G)	Effective rate (%)	Error rate (%)	Q20(%)	Q30(%)	GC content (%)
36-12	WH	8.3	34	2.83	1.21	21,948,851	21,638,332	6.6	6.5	98.59	0.03	97.83	93.35	41.41
36_14	WH	8.4	60	2.5	1.77	22,880,962	22,623,851	6.9	6.8	98.88	0.03	97.42	92.64	41.01
36_18	WH	8.4	184	2.09	2.04	24,876,246	24,514,516	7.5	7.4	98.55	0.03	97.82	93.36	41.54
36_20	WH	8.3	110	2.12	1.72	28,809,430	28,385,921	8.6	8.5	98.53	0.03	97.45	92.69	41.47
36_24	WH	8.4	156	2.00	2.29	24,817,364	24,465,693	7.4	7.3	98.58	0.03	97.4	92.56	41.52
37_12	WL	8.8	280	2.06	1.94	22,806,427	22,215,568	6.8	6.7	97.41	0.03	97.67	93.18	41.6
37_16	WL	8.5	110	1.90	1.83	28,298,123	27,706,578	8.5	8.3	97.91	0.03	97.56	92.94	41.66
37_18	WL	8.7	216	2.16	2.20	27,361,385	26,799,374	8.2	8.0	97.95	0.03	97.58	92.98	41.8
37_26	WL	8.8	320	2.13	1.88	21,372,452	20,951,629	6.4	6.3	98.03	0.03	97.55	92.91	42.49
37_28	WL	8.7	204	2.32	2.32	23,055,321	22,638,072	6.9	6.8	98.19	0.03	97.48	92.78	41.8
73_12	FH	8.6	276	2.03	1.87	22,059,863	21,472,752	6.6	6.4	97.34	0.03	97.58	92.94	41.11
73_16	FH	8.4	88	2.00	1.52	20,249,670	19,772,439	6.1	5.9	97.64	0.03	97.69	93.22	41.3
73_18	FH	8.8	282	2.31	1.91	21,335,811	20,959,767	6.4	6.3	98.24	0.03	97.47	92.69	41.62
73_20	FH	8.6	188	2.41	2.00	20,461,374	20,096,871	6.1	6.0	98.22	0.03	97.56	92.95	41.69
73_28	FH	6.9	12	2.00	0.18	23,816,867	23,273,673	7.1	7.0	97.72	0.03	97.41	92.59	41.5
72_12	FL	8.7	146	2.09	1.74	20,194,683	19,820,154	6.1	5.9	98.15	0.03	97.62	93.05	40.99
72_14	FL	8.4	336	2.15	2.15	27,199,801	26,670,064	8.2	8.0	98.05	0.03	97.62	93.06	41.52
72_24	FL	8.1	226	2.06	2.20	28,283,303	27,709,902	8.5	8.3	97.97	0.03	97.61	93.03	41.86
72_28	FL	7.5	114	2.17	1.50	28,517,251	27,931,288	8.6	8.4	97.95	0.03	97.65	93.1	41.11
72_30	FL	7.9	104	2.26	1.53	25,374,383	24,857,368	7.6	7.5	97.96	0.03	97.58	92.97	41.51

Note: Sample = sample name, Pop = population, RIN = RNA integrity number, RNA con = RNA concentration, 260/280 = ratio of absorbance at 260 nm and 280 nm, 260/230 = ratio of absorbance at 260 nm and 230 nm, Raw reads = total amount of reads of raw data (equals the amount of read1 and read2), Clean reads = total amount of reads of clean data (the amount of read1 and read 2), Raw bases = (Raw reads) * (sequence length), calculating in G, Clean bases = (Clean reads) * (sequence length), calculating in G, Effective rate = (Clean reads/Raw reads) * 100, Error rate = base error rate, Q20 and Q30 = (Base count of Phred value > 20 or 30) / (Total base count), GC content = (G & C base count) / (Total base count), FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

Table S4.4 Log_2 fold change and associated *p*-values (unadjusted and Benjamini-Hochberg adjusted) for 26 of the 151 transcripts identified as related to carbohydrate metabolism in the WH-WL pairwise comparison (PWC). Information is provided for WH-WL and FH-FL PWCs only.

Gene model ID	Gene	Function	WH-WL			FH-FL		
			LFC	*p*-value	Adj *p*-value	LFC	*p*-value	Adj *p*-value
jg30894.t1	***ARL2***	ADP-ribosylation factor-like protein 2	-0.44	0.0015**	0.026*	-0.16	0.26	0.66
jg28239.t1	***BAM***	Chain A, Beta-amylase	1.73	<0.001***	0.017*	0.24	0.033*	0.25
jg28239.t2	***BAM***	beta-amylase	1.86	<0.001***	0.0097**	0.28	0.021*	0.20
jg33907.t1	***BFRUCT***	Beta-fructofuranosidase	1.13	0.0042**	0.049*	1.39	0.0033**	0.065†
jg47865.t1	***BGAL***	beta-galactosidase 3	-0.24	<0.001***	<0.001***	-0.13	0.0086**	0.12
jg71018.t1	***BGAL***	Beta-glucosidase	-1.11	0.0012**	0.022*	-0.38	0.033*	0.25
jg71018.t2	***BGAL***	Beta-glucosidase	-1.60	<0.001***	<0.001***	-0.50	0.021*	0.19
jg42591.t1	***Bgluc***	glucan endo-1,3-beta-glucosidase-like	-1.43	0.0036**	0.044*	-0.20	0.0067**	0.10
jg18640.t1	***CESA4***	cellulose synthase A catalytic subunit 4 [UDP-forming]	-0.34	0.0017**	0.027*	-0.12	0.058†	0.33
jg9527.t1	***CHIB1***	acidic endochitinase	2.07	<0.001***	0.0099**	0.12	0.15	0.52
jg35402.t1	***CHR12***	probable ATP-dependent DNA helicase CHR12 isoform X1	0.52	0.0017**	0.027*	0.54	<0.001***	0.025*
jg59275.t1	***GCST***	aminomethyltransferase, mitochondrial	0.89	0.0020**	0.030*	0.11	0.24	0.63
jg17067.t1	***glgC***	glucose-1-phosphate adenylyltransferase large subunit 1	2.86	<0.001***	0.0013**	1.26	<0.001***	0.018*
jg5855.t1	***glgC***	glucose-1-phosphate adenylyltransferase large subunit 1	1.04	0.0038**	0.046*	1.02	0.0018**	0.044*
jg15808.t1	***IPK1***	inositol-pentakisphosphate 2-kinase	-0.64	0.0031**	0.041*	-0.32	0.061†	0.34
jg14031.t2	***LI***	cyanogenic beta-glucosidase	-1.15	0.0022**	0.033*	-0.87	0.0052**	0.086†
jg1287.t1	***MYH11***	myosin-11 isoform X1	-0.78	<0.001***	<0.001***	-0.78	0.0038**	0.070†
jg54907.t1	***PGL2***	probable 6-phosphogluconolactonase 2	1.35	<0.001***	0.014*	0.19	0.19	0.58
jg68180.t1	***PME***	pectin methylesterase	-2.40	<0.001***	0.0082**	-0.16	0.048*	0.31
jg53847.t1	***PPE8B***	pectinesterase/pectinesterase inhibitor PPE8B	1.67	<0.001***	<0.001***	1.09	0.0019**	0.046*

Table S4.4 (continued)

Gene model ID	Gene	Function	WH-WL			FH-FL		
			LFC	*p*-value	Adj *p*-value	LFC	*p*-value	Adj *p*-value
jg59017.t1	***rham***	probable rhamnogalacturonate lyase B isoform X1	-1.07	0.0023**	0.033*	-1.25	<0.001***	<0.001***
jg68993.t1	***UXS***	UDP-glucuronic acid decarboxylase 2-like	-1.61	<0.001***	0.0023**	-0.12	0.36	0.74
jg19853.t1	***VIP2***	inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase VIP2 isoform X2	-0.86	0.0011**	0.020*	-0.38	0.038*	0.27
jg55201.t1	***VIP2***	inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase VIP2 isoform X1	-0.97	<0.001***	0.0069**	-0.65	0.012*	0.14
jg55238.t1	***VIP2***	inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase VIP2 isoform X1	-0.90	0.0010**	0.019*	-0.41	0.035*	0.26
jg67692.t1	***WAXY***	granule-bound starch synthase 1, chloroplastic/amyloplastic-like	1.50	<0.001***	0.0065**	0.26	0.033*	0.25

Note: *p*-value significance thresholds: † = p < 0.1, * = p < 0.05, ** p < 0.01 and *** p < 0.001. Colours correspond to $\log_2$ fold change (LFC) value where: dark green = LFC > +1.5, medium green = LFC +1.5 – +1, light green = LFC +1 – 0, light blue = LFC 0 – -1, medium blue = LFC -1 – -1.5, dark blue = LFC < -1.5. FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

Table S4.5 Log_2 fold change and associated *p*-values (unadjusted and Benjamini-Hochberg adjusted) for 13 of the 42 proteins identified as related to carbohydrate metabolism in the WH-WL pairwise comparison (PWC). Information is provided for WH-WL and FH-FL PWCs only.

			WH-WL			FH-FL		
Gene model ID	**Gene**	**Function**	**LFC**	***p*-value**	**Adj *p*-value**	**LFC**	***p*-value**	**Adj *p*-value**
chr16.jg2307.t1	***AMY***	alpha-amylase 3, chloroplastic	1.67	0.0087**	0.78	1.61	0.18	0.63
chr8.jg3312.t1	***AMY***	alpha-amylase 3, chloroplastic	1.68	0.015*	0.87	1.57	0.18	0.63
chr1.jg12949.t1	***BAM***	beta-amylase	2.76	0.0043**	0.61	2.44	0.093†	0.52
chr12.jg1584.t2	***BAM***	beta-amylase	4.08	0.015*	0.87	2.57	0.093†	0.52
chr12.jg4565.t1	***BAM***	Chain A, Beta-amylase	3.45	0.026*	0.91	2.81	0.065†	0.46
chr8.jg2580.t1	***BGAL***	beta-galactosidase 1	-4.74	0.041*	0.92	-1.70	0.39	0.80
chr4.jg10686.t1	***DBR***	2-alkenal reductase (NADP(+)-dependent)	-1.52	0.041*	0.92	-1.55	0.24	0.69
chr16.jg6075.t1	***glgA***	probable starch synthase 4, chloroplastic/amyloplastic isoform X1	1.30	0.015*	0.87	1.31	0.065†	0.46
chr9.jg6601.t1	***glgA***	soluble starch synthase 1, chloroplastic/amyloplastic	1.38	0.041*	0.92	1.24	0.13	0.58
chr12.jg7613.t1	***glgB***	1,4-alpha-glucan-branching enzyme 1, chloroplastic/amyloplastic isoform X1	1.60	0.041*	0.92	1.24	0.31	0.74
chr5.jg7106.t1	***glgC***	glucose-1-phosphate adenylyltransferase small subunit 2, chloroplastic	1.51	0.015*	0.87	1.01	0.59	0.88
chr1.jg10570.t1	***GWD2***	alpha-glucan water dikinase, chloroplastic	1.38	0.015*	0.87	1.65	0.093†	0.52
chr3.jg8461.t1	***ISA3***	isoamylase 3, chloroplastic isoform X1	1.53	0.041*	0.92	1.60	0.13	0.58

Note: *p*-value significance thresholds: † = p < 0.1, * = p < 0.05 and ** p < 0.01. Colours correspond to log_2 fold change (LFC) value where: dark green = LFC > +1.5, medium green = LFC +1.5 – +1, light green = LFC +1 – 0, light blue = LFC 0 – -1, medium blue = LFC -1 – -1.5, dark blue = LFC < -1.5. FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

Table S4.6 Permutation ANOVA based on 1e+7 permutations for the effect of pool (W – WNZLL; F – FNZLL) and selection (H – high WSC; L – low WSC) and the interaction between pool and selection for the expression of 49 genes. A total of 20 individuals were used for each of the 49 genes. Genes are alphabetically ordered and genes coloured in light green represent the 26 candidate genes under selection as identified in the correlation analysis.

Gene model ID and Gene	Response	df	SS	MS	*p*-value (2 s.f.)
jg60509.t1 ***ACP1***	Pool	1	1.692	1.692	0.41
	Sel	1	10.241	10.241	0.031*
	Pool:Sel	1	4.970	4.970	0.14
	Residuals	16	33.299	2.081	
jg30894.t1 ***ARL2***	Pool	1	0.029	0.029	0.60
	Sel	1	0.697	0.697	0.0090**
	Pool:Sel	1	0.079	0.079	0.39
	Residuals	16	1.474	0.092	
jg28239.t1 ***BAM***	Pool	1	0.000	0.000	1
	Sel	1	12.484	12.484	0.020*
	Pool:Sel	1	1.570	1.570	0.36
	Residuals	16	28.942	1.809	
jg28239.t2 ***BAM***	Pool	1	0.010	0.010	0.94
	Sel	1	13.529	13.529	0.015*
	Pool:Sel	1	1.580	1.580	0.35
	Residuals	16	27.652	1.728	
jg33907.t1 ***BFRUCT***	Pool	1	0.203	0.203	0.70
	Sel	1	13.468	13.468	0.0072**
	Pool:Sel	1	0.709	0.709	0.47
	Residuals	16	21.007	1.313	
jg19321.t1 ***BGAL***	Pool	1	0.698	0.698	0.58
	Sel	1	18.968	18.969	0.012*
	Pool:Sel	1	8.454	8.454	0.071
	Residuals	16	35.745	2.234	
jg19398.t1 ***BGAL***	Pool	1	0.008	0.008	0.93
	Sel	1	16.833	16.833	0.0014**
	Pool:Sel	1	7.917	7.917	0.015*
	Residuals	16	16.555	1.035	
jg45994.t1 ***BGAL***	Pool	1	1.237	1.237	0.38
	Sel	1	10.642	10.642	0.018*
	Pool:Sel	1	1.887	1.887	0.27
	Residuals	16	23.953	1.497	
jg47865.t1 ***BGAL***	Pool	1	6.347	6.347	0.17
	Sel	1	70.462	70.462	0.00043***
	Pool:Sel	1	0.596	0.596	0.66
	Residuals	16	49.428	3.089	
jg71018.t1 ***BGAL***	Pool	1	0.106	0.106	0.69
	Sel	1	9.681	9.681	0.0016**
	Pool:Sel	1	0.082	0.082	0.73
	Residuals	16	10.603	0.663	

Table S4.6 (continued)

Gene model ID and Gene	Response	df	SS	MS	*p*-value (2 s.f.)
jg71018.t2 ***BGAL***	Pool	1	0.029	0.029	0.81
	Sel	1	12.097	12.097	0.00018***
	Pool:Sel	1	0.032	0.032	0.80
	Residuals	16	8.055	0.503	
jg42591.t1 ***Bgluc***	Pool	1	1.336	1.336	0.39
	Sel	1	21.570	21.570	0.0010**
	Pool:Sel	1	0.114	0.114	0.80
	Residuals	16	26.487	1.655	
jg45631.t1 ***Bgluc***	Pool	1	5.907	5.907	0.28
	Sel	1	30.846	30.846	0.024*
	Pool:Sel	1	0.059	0.059	0.91
	Residuals	16	77.272	4.830	
jg18640.t1 ***CESA4***	Pool	1	2.854	2.854	0.41
	Sel	1	27.251	27.251	0.024*
	Pool:Sel	1	0.697	0.697	0.68
	Residuals	16	66.908	4.182	
jg9527.t1 ***CHIB1***	Pool	1	1.058	1.058	0.48
	Sel	1	10.183	10.183	0.040*
	Pool:Sel	1	1.568	1.568	0.39
	Residuals	16	32.391	2.025	
jg35402.t1 ***CHR12***	Pool	1	0.130	0.130	0.30
	Sel	1	1.697	1.697	0.0018**
	Pool:Sel	1	0.220	0.220	0.18
	Residuals	16	1.839	0.115	
jg37163.t1 ***COBL4***	Pool	1	0.360	0.360	0.74
	Sel	1	16.278	16.278	0.037*
	Pool:Sel	1	3.226	3.226	0.32
	Residuals	16	49.848	3.116	
jg3576.t1 ***CSLH1***	Pool	1	0.359	0.359	0.32
	Sel	1	2.485	2.485	0.015*
	Pool:Sel	1	0.202	0.202	0.46
	Residuals	16	5.403	0.338	
jg55076.t1 ***DCAF1***	Pool	1	3.299	3.299	0.19
	Sel	1	34.800	34.800	0.0050**
	Pool:Sel	1	3.299	3.299	0.19
	Residuals	16	41.077	2.567	
jg62181.t1 ***FAM135B***	Pool	1	0.066	0.066	0.59
	Sel	1	1.185	1.185	0.037*
	Pool:Sel	1	0.271	0.271	0.28
	Residuals	16	3.532	0.221	
jg22077.t1 ***FRK***	Pool	1	0.077	0.077	0.18
	Sel	1	3.067	3.067	<2.2e-16***
	Pool:Sel	1	1.184	1.184	7.44e-05***
	Residuals	16	0.635	0.040	

Table S4.6 (continued)

Gene model ID and Gene	Response	df	SS	MS	*p*-value (2 s.f.)
jg42427.t1 ***GALM***	Pool	1	1.917	1.917	0.26
	Sel	1	8.839	8.839	0.027*
	Pool:Sel	1	5.003	5.003	0.083
	Residuals	16	23.193	1.450	
jg59275.t1 ***GCST***	Pool	1	7.807	7.807	0.015*
	Sel	1	6.219	6.219	0.028*
	Pool:Sel	1	0.466	0.466	0.54
	Residuals	16	17.857	1.116	
jg17067.t1 ***glgC***	Pool	1	0.385	0.385	0.60
	Sel	1	17.165	17.165	0.0022**
	Pool:Sel	1	1.749	1.749	0.27
	Residuals	16	21.079	1.317	
jg5855.t1 ***glgC***	Pool	1	0.003	0.003	0.94
	Sel	1	7.136	7.136	0.0015**
	Pool:Sel	1	0.157	0.157	0.58
	Residuals	16	7.729	0.483	
jg15808.t1 ***IPK1***	Pool	1	0.078	0.078	0.47
	Sel	1	2.061	2.061	0.0028**
	Pool:Sel	1	0.058	0.058	0.53
	Residuals	16	2.384	0.149	
jg15602.t1 ***IRE1***	Pool	1	3.509	3.509	0.042*
	Sel	1	15.604	15.604	0.00058***
	Pool:Sel	1	3.023	3.023	0.056
	Residuals	16	11.219	0.701	
jg14031.t2 ***LI***	Pool	1	0.001	0.001	0.97
	Sel	1	11.061	11.061	0.00049***
	Pool:Sel	1	0.011	0.011	0.89
	Residuals	16	9.445	0.590	
jg32500.t1 ***MAC12.5***	Pool	1	0.223	0.223	0.37
	Sel	1	1.742	1.742	0.021*
	Pool:Sel	1	0.388	0.388	0.24
	Residuals	16	4.253	0.266	
jg61904.t1 ***MAC12.5***	Pool	1	0.161	0.161	0.53
	Sel	1	2.942	2.942	0.0077**
	Pool:Sel	1	1.854	1.854	0.033*
	Residuals	16	5.805	0.363	
jg36439.t1 ***Mdh***	Pool	1	1.372	1.373	0.59
	Sel	1	27.295	27.295	0.033*
	Pool:Sel	1	22.123	22.123	0.050
	Residuals	16	76.924	4.808	
jg1287.t1 ***MYH11***	Pool	1	0.050	0.050	0.63
	Sel	1	4.000	4.000	0.00026***
	Pool:Sel	1	0.387	0.387	0.19
	Residuals	16	3.328	0.208	

Table S4.6 (continued)

Gene model ID and Gene	Response	df	SS	MS	*p*-value (2 s.f.)
jg866.t1 ***PGL1***	Pool	1	0.744	0.744	0.77
	Sel	1	58.013	58.013	0.023*
	Pool:Sel	1	23.242	23.242	0.12
	Residuals	16	140.995	8.812	
jg869.t1 ***PGL1***	Pool	1	0.468	0.468	0.79
	Sel	1	43.604	43.604	0.022*
	Pool:Sel	1	21.151	21.151	0.09
	Residuals	16	103.928	6.496	
jg54907.t1 ***PGL2***	Pool	1	1.898	1.898	0.063
	Sel	1	4.067	4.067	0.0070**
	Pool:Sel	1	2.242	2.242	0.043*
	Residuals	16	7.865	0.492	
jg68180.t1 ***PME***	Pool	1	0.006	0.007	0.97
	Sel	1	18.485	18.485	0.033*
	Pool:Sel	1	1.815	1.815	0.47
	Residuals	16	53.336	3.334	
jg21861.t1 ***PME22***	Pool	1	0.827	0.827	0.70
	Sel	1	31.184	31.184	0.036*
	Pool:Sel	1	28.419	28.419	0.043*
	Residuals	16	91.959	5.748	
jg53847.t1 ***PPE8B***	Pool	1	2.072	2.072	0.035*
	Sel	1	12.094	12.094	7.61e-05***
	Pool:Sel	1	0.776	0.776	0.17
	Residuals	16	6.081	0.380	
jg59017.t1 ***rham***	Pool	1	1.176	1.176	0.12
	Sel	1	9.612	9.612	0.00032***
	Pool:Sel	1	0.168	0.168	0.54
	Residuals	16	6.861	0.429	
jg51134.t1 ***RRT1***	Pool	1	0.048	0.048	0.69
	Sel	1	2.705	2.705	0.011*
	Pool:Sel	1	1.599	1.599	0.038*
	Residuals	16	4.870	0.304	
jg46513.t1 ***TRA2***	Pool	1	0.018	0.018	0.81
	Sel	1	1.358	1.358	0.019*
	Pool:Sel	1	0.678	0.678	0.10
	Residuals	16	3.846	0.240	
jg68993.t1 ***UXS***	Pool	1	0.168	0.168	0.54
	Sel	1	6.678	6.678	0.0011**
	Pool:Sel	1	1.364	1.364	0.096
	Residuals	16	6.952	0.435	
jg19853.t1 ***VIP2***	Pool	1	0.226	0.226	0.53
	Sel	1	5.839	5.839	0.0029**
	Pool:Sel	1	0.159	0.159	0.59
	Residuals	16	8.354	0.522	

Table S4.6 (continued)

Gene model ID and Gene	Response	df	SS	MS	*p*-value (2 s.f.)
jg55201.t1 ***VIP2***	Pool	1	0.122	0.122	0.61
	Sel	1	6.371	6.371	0.0027**
	Pool:Sel	1	0.128	0.128	0.60
	Residuals	16	7.408	0.463	
jg55238.t1 ***VIP2***	Pool	1	0.380	0.380	0.40
	Sel	1	5.918	5.918	0.0030**
	Pool:Sel	1	0.079	0.079	0.70
	Residuals	16	8.025	0.502	
jg67692.t1 ***WAXY***	Pool	1	0.453	0.453	0.55
	Sel	1	8.259	8.259	0.020*
	Pool:Sel	1	1.077	1.077	0.36
	Residuals	16	19.463	1.216	
jg62183.t1 ***XK2***	Pool	1	0.358	0.358	0.16
	Sel	1	2.663	2.663	0.00091***
	Pool:Sel	1	0.554	0.554	0.083
	Residuals	16	2.583	0.161	
jg70400.t1 ***XTH1***	Pool	1	4.009	4.010	0.26
	Sel	1	23.162	23.162	0.015*
	Pool:Sel	1	7.161	7.161	0.14
	Residuals	16	47.521	2.970	
jg12044.t1 ***XYL1***	Pool	1	5.066	5.066	0.12
	Sel	1	18.483	18.483	0.0073**
	Pool:Sel	1	4.567	4.567	0.14
	Residuals	16	30.231	1.889	

Note: Sel = selection, df = degrees of freedom, SS = Sum of squares, MS = Mean square, s.f. = significant figures, *p*-value significance thresholds: * = $p < 0.05$, ** = $p < 0.01$ and *** = $p < 0.001$.

Table S4.7 Permutation ANOVA based on 1e+7 permutations for the effect of pool (W – WNZLL; F – FNZLL) and selection (h – high WSC; L – low WSC) and the interaction between pool and selection for the expression of 13 proteins. A total of 24 individuals were used for each of the 13 proteins. Proteins are alphabetically ordered based on gene name and proteins coloured in light green represent the 13 candidate proteins under selection as identified in the correlation analysis.

Gene model ID and Gene	Response	df	SS	MS	*p*-value (2 s.f.)
chr16.jg2307.t1 ***AMY***	Pool	1	0.390	0.390	0.26
	Sel	1	3.071	3.071	0.0051**
	Pool:Sel	1	0.004	0.004	1.00
	Residuals	20	5.917	0.296	
chr8.jg3312.t1 ***AMY***	Pool	1	0.310	0.310	0.34
	Sel	1	2.947	2.947	0.0085**
	Pool:Sel	1	0.014	0.013	0.84
	Residuals	20	6.619	0.331	
chr1.jg12949.t1 ***BAM***	Pool	1	1.927	1.927	0.17
	Sel	1	11.363	11.363	0.0024**
	Pool:Sel	1	0.049	0.049	0.83
	Residuals	20	19.296	0.965	
chr12.jg1584.t2 ***BAM***	Pool	1	1.492	1.492	0.31
	Sel	1	17.253	17.253	0.0016**
	Pool:Sel	1	0.664	0.664	0.50
	Residuals	20	27.741	1.387	
chr12.jg4565.t1 ***BAM***	Pool	1	1.781	1.781	0.25
	Sel	1	16.118	16.118	0.0020**
	Pool:Sel	1	0.130	0.130	0.75
	Residuals	20	25.625	1.281	
chr8.jg2580.t1 ***BGAL***	Pool	1	3.072	3.072	0.32
	Sel	1	13.572	13.572	0.046*
	Pool:Sel	1	3.294	3.294	0.31
	Residuals	20	60.135	3.007	
chr4.jg10686.t1 ***DBR***	Pool	1	0.581	0.581	0.15
	Sel	1	2.303	2.303	0.0081**
	Pool:Sel	1	0.002	0.002	1.00
	Residuals	20	5.149	0.257	
chr16.jg6075.t1 ***glgA***	Pool	1	0.296	0.296	0.066
	Sel	1	0.873	0.873	0.0038**
	Pool:Sel	1	0.000	0.000	1.00
	Residuals	20	1.557	0.078	
chr9.jg6601.t1 ***glgA***	Pool	1	0.003	0.003	0.87
	Sel	1	0.907	0.907	0.018*
	Pool:Sel	1	0.034	0.034	0.62
	Residuals	20	2.717	0.136	
chr12.jg7613.t1 ***glgB***	Pool	1	0.023	0.022	0.76
	Sel	1	1.468	1.468	0.023*
	Pool:Sel	1	0.203	0.203	0.37
	Residuals	20	4.872	0.244	

Table S4.7 (continued)

Gene model ID and Gene	Response	df	SS	MS	*p*-value (2 s.f.)
chr5.jg7106.t1 ***glgC***	Pool	1	0.005	0.005	0.82
	Sel	1	0.569	0.569	0.033*
	Pool:Sel	1	0.495	0.495	0.045*
	Residuals	20	2.141	0.107	
chr1.jg10570.t1 ***GWD2***	Pool	1	0.011	0.011	0.85
	Sel	1	2.116	2.116	0.0073**
	Pool:Sel	1	0.103	0.103	0.55
	Residuals	20	5.240	0.262	
chr3.jg8461.t1 ***ISA3***	Pool	1	0.012	0.011	0.85
	Sel	1	2.511	2.511	0.012*
	Pool:Sel	1	0.006	0.006	1.00
	Residuals	20	6.308	0.315	

Note: Sel = selection, df = degrees of freedom, SS = Sum of squares, MS = Mean square, s.f. = significant figures, *p*-value significance thresholds: * = $p < 0.05$ and ** = $p < 0.01$.

Table S4.8 KEGG pathways for two pairwise comparisons (PWCs) of differentially expressed transcripts (DET) and differentially expressed proteins (DAPs) ordered for metabolism, genetic information processing, environmental information processing, cellular processes and organismal systems. Values represent hits for each PWC. Maps were limited to only include maps with more than five hits.

		Transcriptomics		Proteomics	
Pathway	**Map**	**WH-WL**	**FH-FL**	**WH-WL**	**FH-FL**
1. Metabolism					
1.1 Global and overview maps	Metabolic pathway	147	76	59	127
	Biosynthesis of secondary metabolites	92	52	28	74
	Microbial metabolism in diverse environments	20	13	13	37
	Carbon metabolism	15	12	8	27
	Biosynthesis of amino acids	14	10		16
1.2 Carbohydrate metabolism	Glycolysis / Gluconeogenesis	6	6		10
	Starch and sucrose metabolism	11	6	10	8
	Amino sugar and nucleotide sugar metabolism	9			9
	Pyruvate metabolism		6		8
	Glyoxylate and dicarboxylate metabolism	6			11
	Citrate cycle (TCA cycle)				10
	Inositol phosphate metabolism	9			
1.3 Energy metabolism	Oxidative phosphorylation				7
	Carbon fixation in photosynthetic organisms				8
1.4 Lipid metabolism	Glycerophospholipid metabolism	8	8		
1.5 Nucleotide metabolism	Purine metabolism	11			
1.6 Amino acid metabolism	Cysteine and methionine metabolism	11			9
	Valine, leucine and isoleucine degradation				6
	Tyrosine metabolism				6
	Arginine and proline metabolism				6
	Glycine, serine and threonine metabolism	6			
1.7 Metabolism of other amino acids	Glutathione metabolism				7
1.8 Metabolism of cofactors and vitamins	Porphyrin and chlorophyll metabolism	9			
	Folate biosynthesis	6			
1.9 Biosynthesis of other secondary metabolites	Phenylpropanoid biosynthesis	8	9		

Table S4.8 (continued)

Pathway	Map	Transcriptomics		Proteomics	
		WH-WL	FH-FL	WH-WL	FH-FL
2. Genetic Information Processing					
2.1 Transcription	Spliceosome	15			6
2.2 Translation	Ribosome	39	9		
	Ribosome biogenesis in eukaryotes	26			
	Aminoacyl-tRNA biosynthesis	6			6
	RNA transport	17			7
	mRNA surveillance pathway	8			
2.3 Folding, sorting and degradation	Protein processing in endoplasmic reticulum	15	6		16
	RNA degradation	7			7
	Ubiquitin mediated proteolysis	15	11		
2.4 Replication and repair	Homologous recombination	7			
3. Environmental Information Processing					
3.1 Signal transduction	MAPK signalling pathway – plant	14	6		
	Plant hormone signal transduction	13			
	Phosphatidylinositol signalling system	7			
4. Cellular Processes					
4.1 Transport and catabolism	Endocytosis	12	7		
	Peroxisome	6			
4.2 Cell growth and death	Cell cycle	14	7		
	Oocyte meiosis	10	7		
	Cellular senescence	9			
5. Organismal Systems					
5.1 Environmental adaptation	Plant-pathogen interaction	14	9		6

Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

SUPPLEMENTARY FIGURES

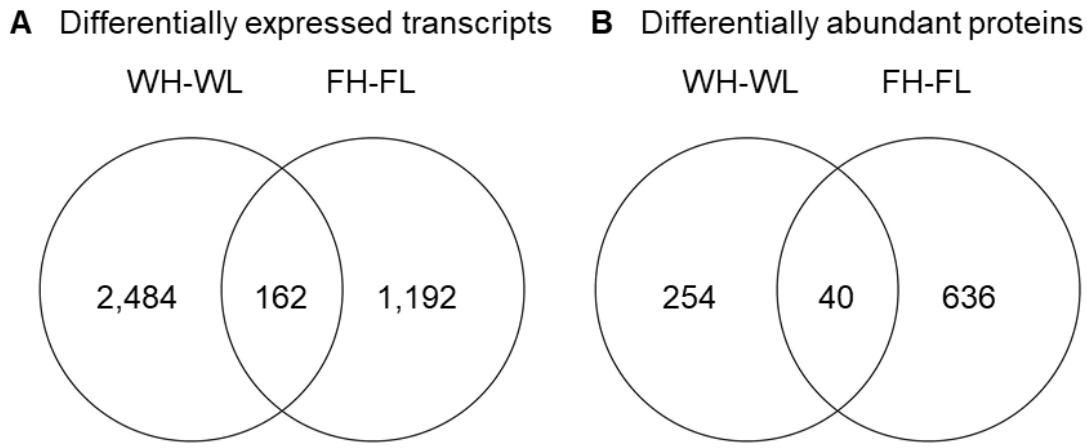


Figure S4.1 Venn diagram overlap between differentially expressed transcripts (**A**) and differentially abundant proteins (**B**) for the WH-WL and FH-FL pairwise comparisons. *Note:* FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

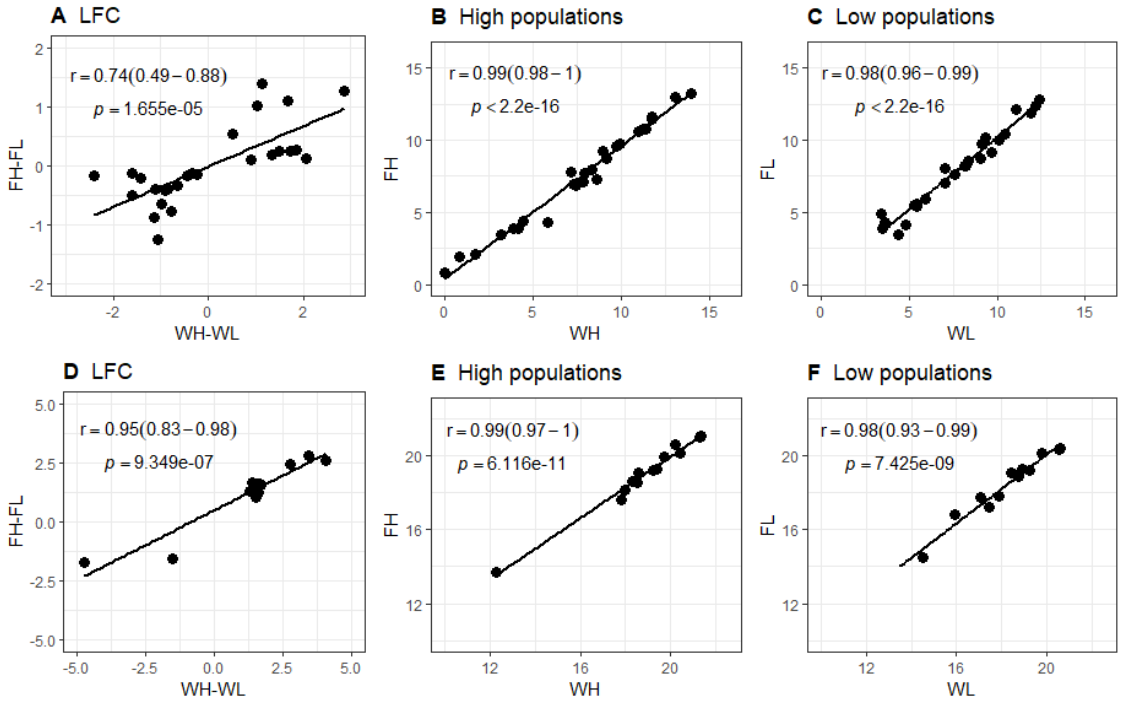


Figure S4.2 Correlation scatterplots for log_2 fold changes and mean log_2 expression values for 26 transcripts (**A** – **C**) and 13 proteins (**D** – **F**). **A** and **D** show the log_2 fold change (LFC) relationship between the WH-WL pairwise comparison (PWC) and FH-FL PWC. **B** and **E** show the mean log_2 expression values for the two high populations. **C** and **F** show the mean log_2 expression values for the two low populations. Pearson correlation coefficient (r) values, 95% confidence interval for the correlation coefficient (in parentheses) and *p*-values for lines of best fit are shown. *Note*: Plots A, D, E and F have *x*- and *y*-axes that do not begin at zero. FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

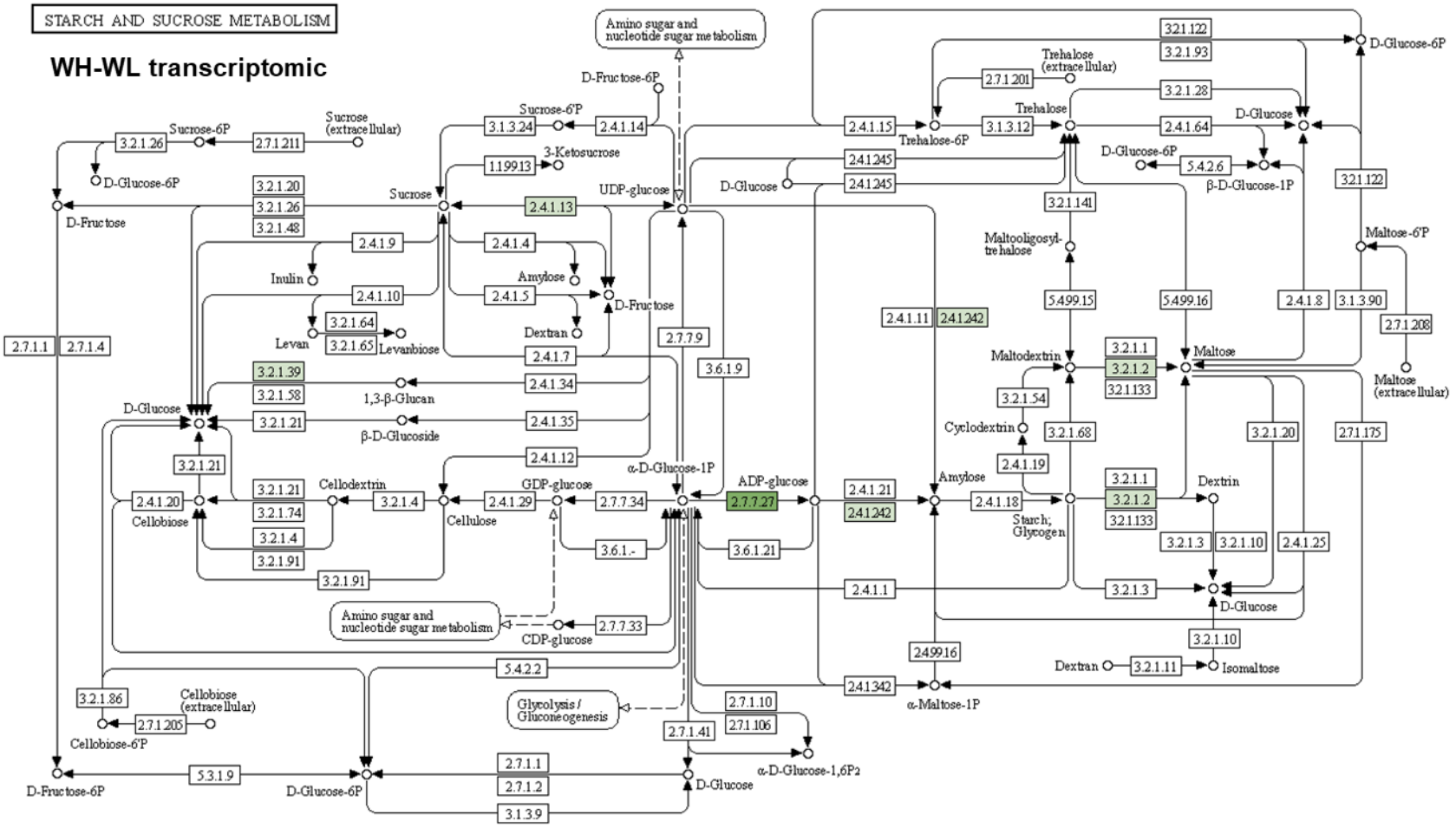


Figure S4.3 Starch and sucrose metabolism map for WH-WL based on transcriptomic data. Map only displays genes that are up-regulated and are colour coded based on the $\log_2$ fold change (LFC), where a LFC between 0 and 2 is coloured light green and a LFC above 2 is coloured dark green. *Note*: WH = WNZLL-High-End, WL = WNZLL-Low-End; where: NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

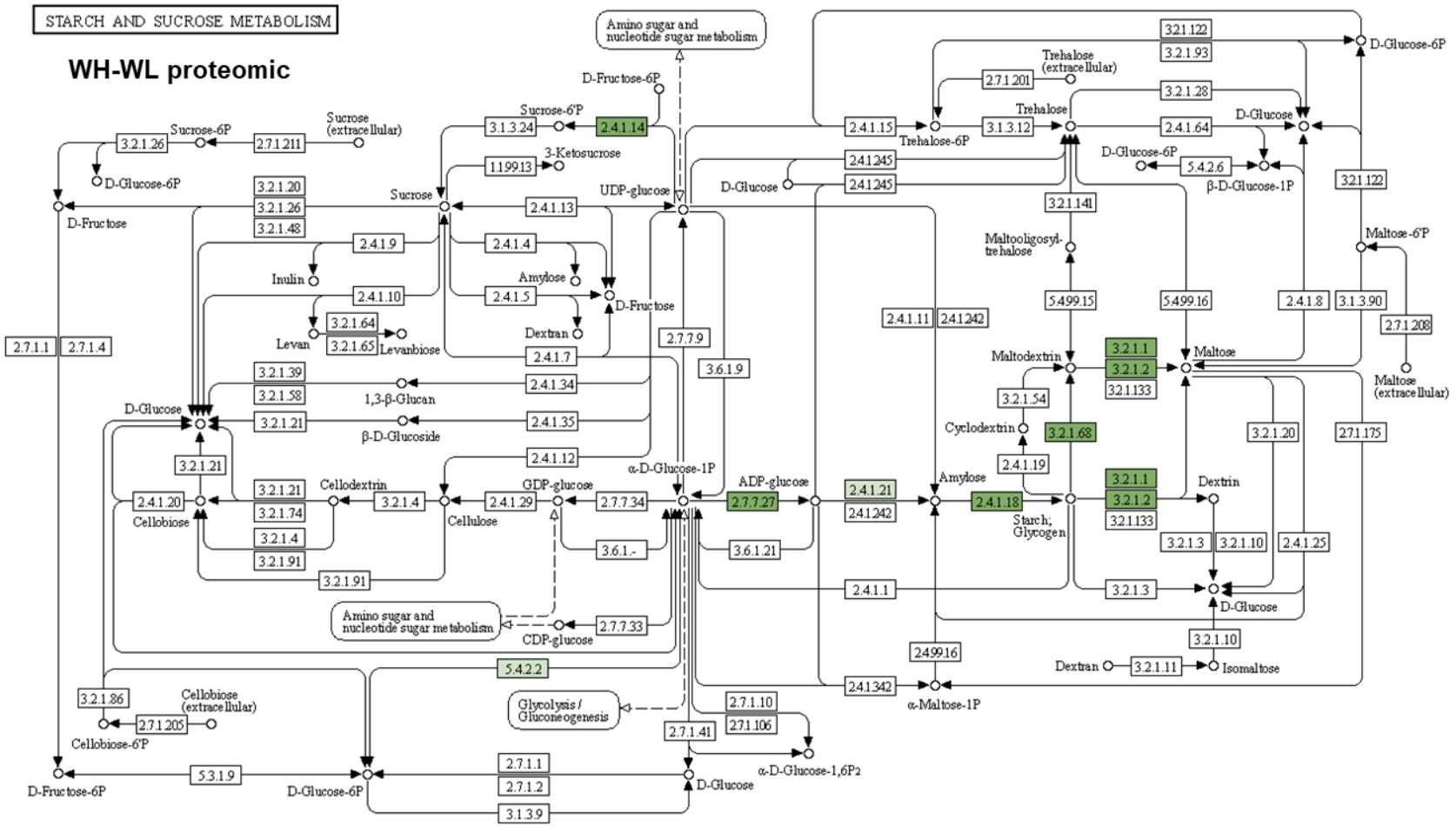


Figure S4.4 Starch and sucrose metabolism map for WH-WL based on proteomic data. Map only displays genes that are up-regulated and are colour coded based on the $\log_2$ fold change (LFC), where a LFC between 0 and 1.5 is coloured light green and a LFC above 1.5 is is coloured dark green. *Note:* WH = WNZLL-High-End, WL = WNZLL-Low-End; where: NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

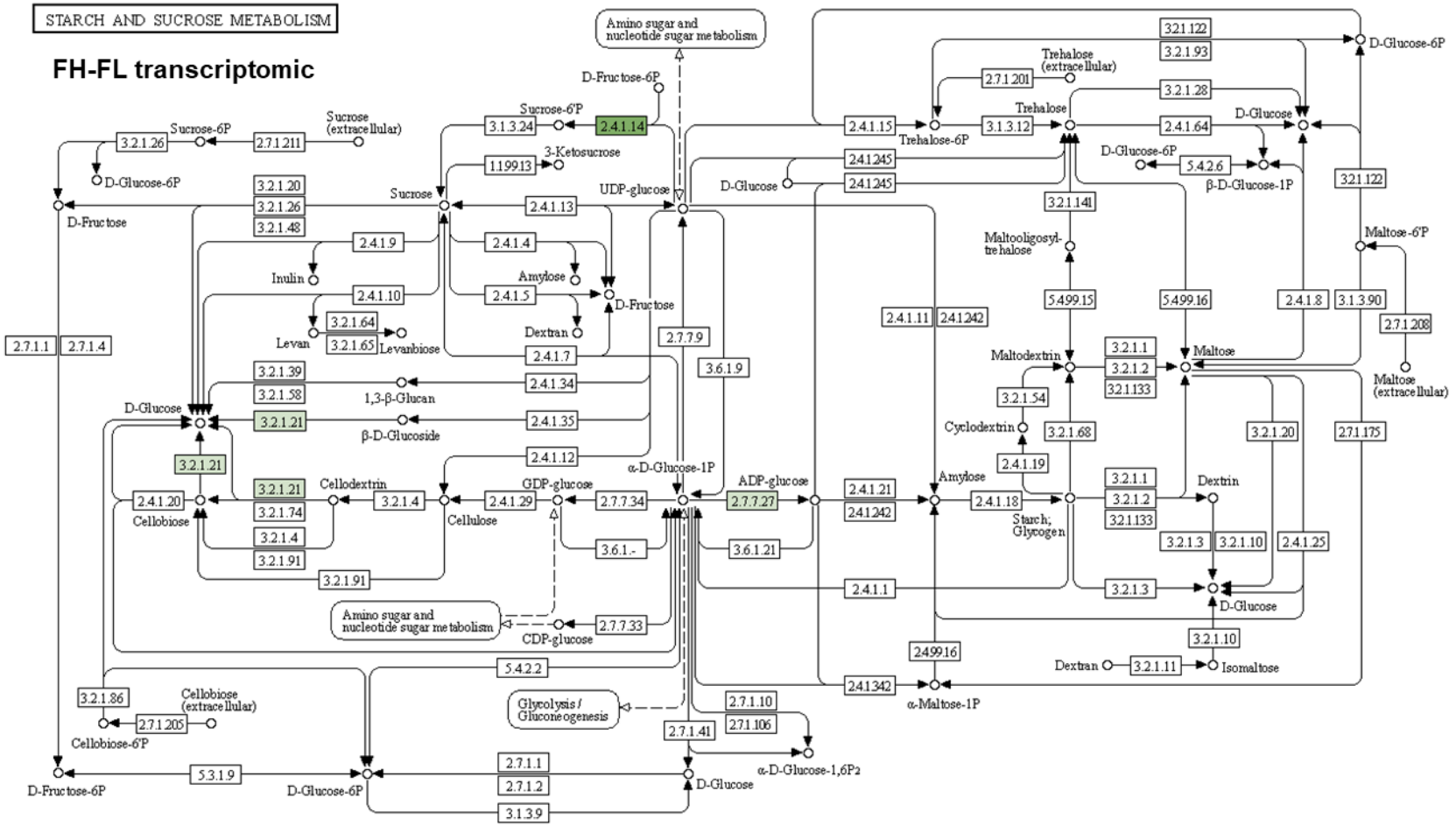


Figure S4.5 Starch sucrose metabolism map for FH-FL based on transcriptomic data. Map only displays genes that are up-regulated and are colour coded based on the log_2 fold change (LFC), where a LFC between 0 and 2 is coloured light green and a LFC above 2 is coloured dark green. *Note:* FH = FNZLL-High-End, FL = FNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation.

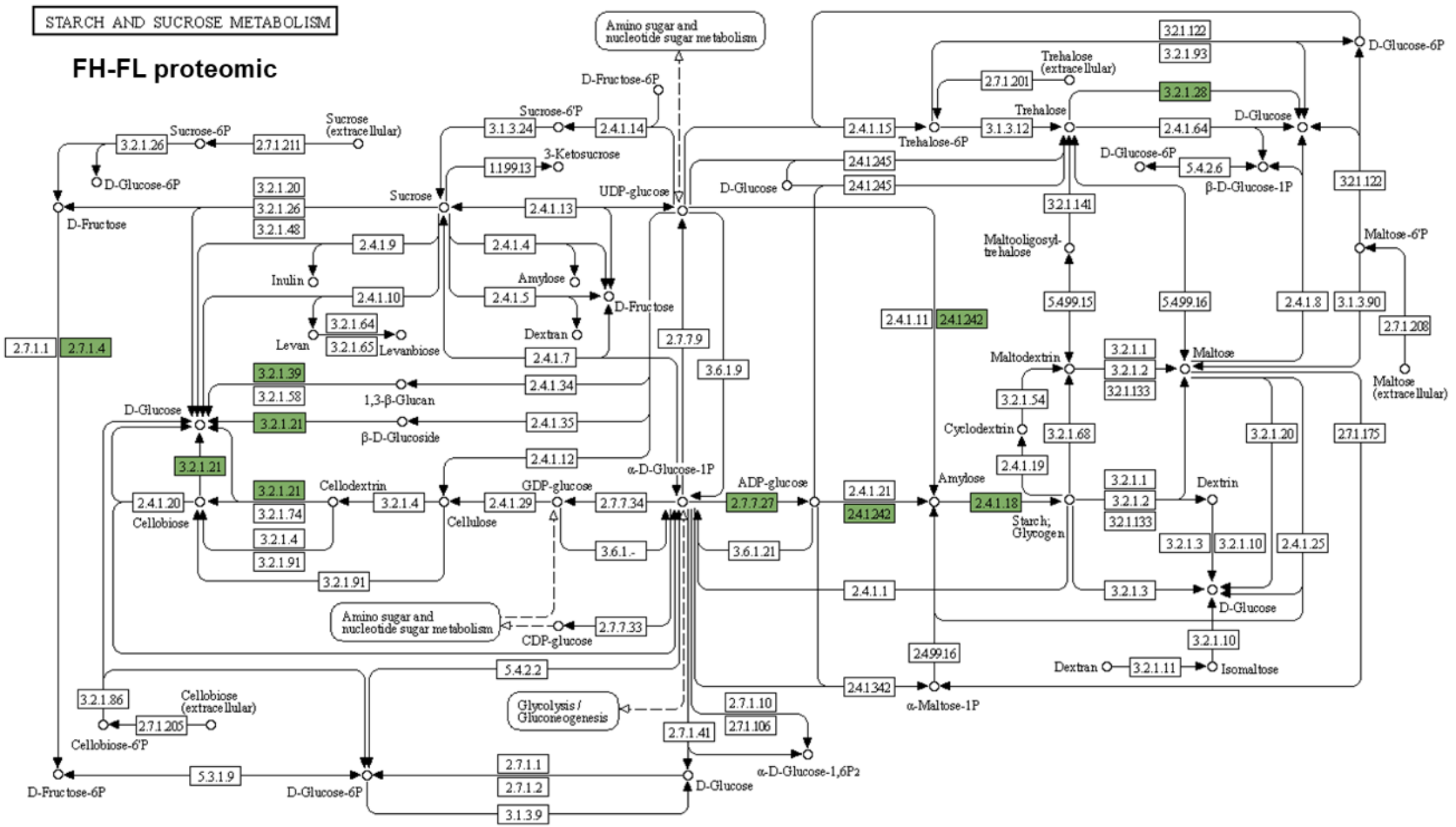
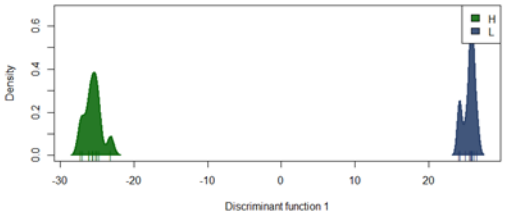
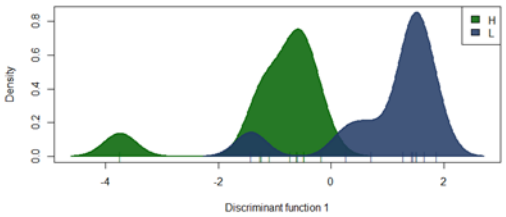


Figure S4.6 Starch sucrose metabolism map for FH-FL based on proteomic data. Map only displays genes that are up-regulated and are colour coded based on the $\log_2$ fold change (LFC), where a LFC between 0 and 1.5 is coloured light green and a LFC above 1.5 is coloured dark green. *Note:* FH = FNZLL-High-End, FL = FNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation.

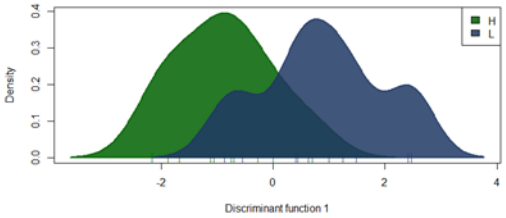
A *AMY* chr8.jg3312.t1 (94)



B *AMY* chr16.jg2307.t1 (124)

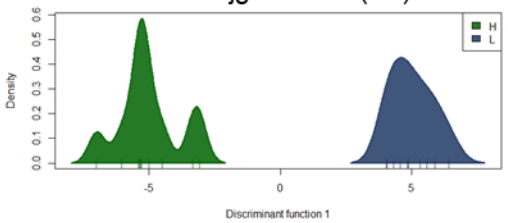


C *BAM* chr1.jg12949.t1 (86)

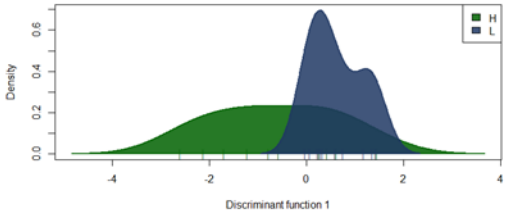


	Pop	WH	WH	WH	WH	WH	FH	FH	FH	FH	FH	WL	WL	WL	WL	WL	FL	FL	FL	FL	FL
	Ind	36-12	36-14	36-18	36-20	36-24	73-12	73-16	73-18	73-20	73-28	37-12	37-16	37-18	37-26	37-28	72-12	72-14	72-24	72-28	72-30
Chr	Pos																				
8	24305522	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	1	0	0	0
16	16396421	2	2	2	2	2	0	1	2	1	1	0	0	0	0	2	0	0	?	0	0
16	16396431	2	2	2	2	2	0	1	2	1	1	0	0	0	0	2	0	1	?	0	0
1	89197128	0	?	?	0	0	0	0	2	2	0	0	0	0	0	0	?	0	0	0	0
1	89197136	0	?	?	0	0	2	1	2	2	2	0	1	1	1	0	?	0	0	0	0
1	89197186	0	?	?	0	0	0	0	2	2	0	0	0	0	0	0	?	0	0	0	0
1	89197469	0	0	0	0	0	0	0	0	?	?	?	0	?	0	1	?	1	2	0	2
1	89197479	0	0	0	0	0	0	0	0	?	?	?	1	?	0	1	?	1	2	0	2
1	89197518	0	0	0	0	0	0	0	0	?	?	1	0	?	0	1	?	1	2	0	2
1	89197858	1	0	1	1	2	2	2	2	2	2	0	0	1	1	0	0	1	1	2	1

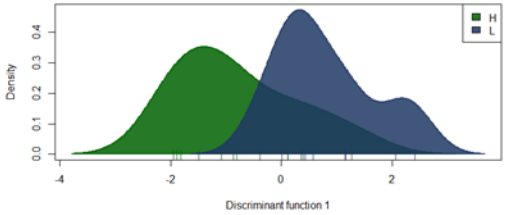
D *BAM* chr12.jg1584.t2 (63)



E *BAM* chr12.jg4565.t1 (33)



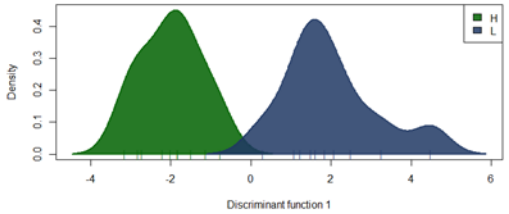
F *glgA* chr9.jg6601.t1 (121)



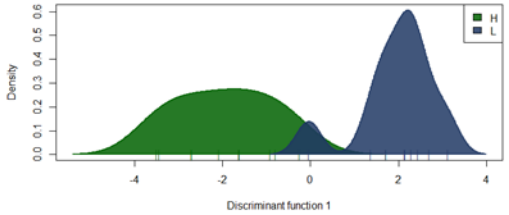
	Pop	WH	WH	WH	WH	WH	FH	FH	FH	FH	FH	WL	WL	WL	WL	WL	FL	FL	FL	FL	FL
	Ind	36-12	36-14	36-18	36-20	36-24	73-12	73-16	73-18	73-20	73-28	37-12	37-16	37-18	37-26	37-28	72-12	72-14	72-24	72-28	72-30
Chr	Pos																				
12	10584005	0	0	1	0	0	0	2	2	1	2	2	1	0	0	0	1	2	1	2	1
12	10584515	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
12	10586425	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0
12	29991602	0	1	1	1	1	0	1	1	1	1	1	2	2	?	0	0	2	1	2	0
12	29994073	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	29994419	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

NO SNPs

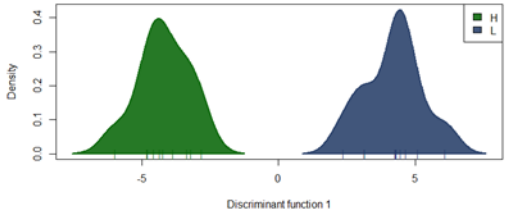
G *glgA* chr16.jg6075.t1 (106)



H *glgB* chr12.jg7613.t1 (120)



I *glgC* chr4.jg5408.t1 (99)

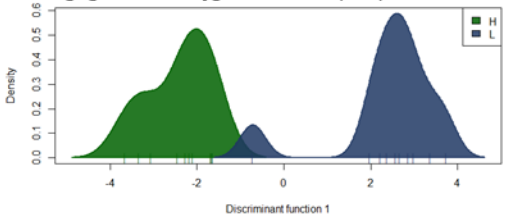


	Pop	WH	WH	WH	WH	WH	FH	FH	FH	FH	FH	WL	WL	WL	WL	WL	FL	FL	FL	FL	FL
	Ind	36-12	36-14	36-18	36-20	36-24	73-12	73-16	73-18	73-20	73-28	37-12	37-16	37-18	37-26	37-28	72-12	72-14	72-24	72-28	72-30
Chr	Pos																				
16	43506357	0	2	?	?	0	2	2	?	2	2	1	0	0	?	2	0	0	?	0	0

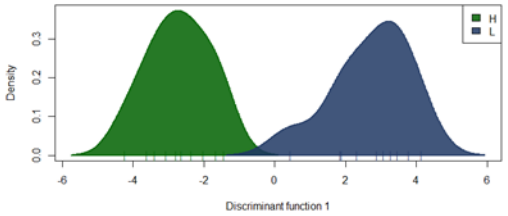
NO SNPs

4	39413411	0	?	0	0	?	0	0	?	0	0	2	0	2	2	2	?	0	0	0	0
4	39415408	0	0	0	0	0	0	2	?	0	?	?	0	2	?	2	1	1	2	?	2
4	39418688	1	0	0	1	1	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0
4	39418704	1	0	0	1	1	0	1	2	1	0	0	1	0	0	0	0	0	0	0	0
4	39418746	2	0	2	2	2	2	2	2	2	2	2	2	1	1	1	0	0	1	0	1

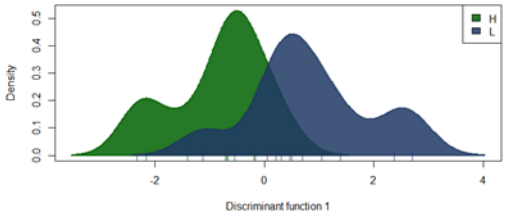
J *glgC* chr5.jg7106.t1 (89)



K *glgC* chr11.jg7807.t1 (93)



L *ISA3* chr3.jg8461.t1 (71)



	Pop	WH	WH	WH	WH	WH	FH	FH	FH	FH	FH	WL	WL	WL	WL	WL	FL	FL	FL	FL	FL
	Ind	36-12	36-14	36-18	36-20	36-24	73-12	73-16	73-18	73-20	73-28	37-12	37-16	37-18	37-26	37-28	72-12	72-14	72-24	72-28	72-30
Chr	Pos																				
5	47905750	1	0	0	1	1	0	0	0	0	1	1	1	1	1	1	0	1	1	1	1
5	47905854	1	0	0	1	1	0	0	0	0	1	1	1	1	1	1	0	1	1	1	1
5	47905856	1	0	0	1	1	0	0	0	0	1	1	1	1	1	1	0	1	1	1	1
5	47906242	0	0	0	1	1	0	0	0	0	0	1	1	1	1	1	0	1	1	1	1
5	47906251	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	0	1	1	1	1
5	47906267	0	0	0	0	1	0	1	0	0	0	0	1	1	1	1	0	1	1	1	1
5	47906791	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	0	1	1	1	1
5	47907602	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
11	52421391	0	0	0	0	0	0	2	2	1	1	0	0	0	0	0	1	0	0	0	0
3	55478919	2	?	1	1	1	0	2	2	1	0	0	1	0	0	0	0	0	0	0	0
3	55479019	?	?	0	0	1	1	0	0	0	2	1	1	1	1	2	?	2	2	2	2

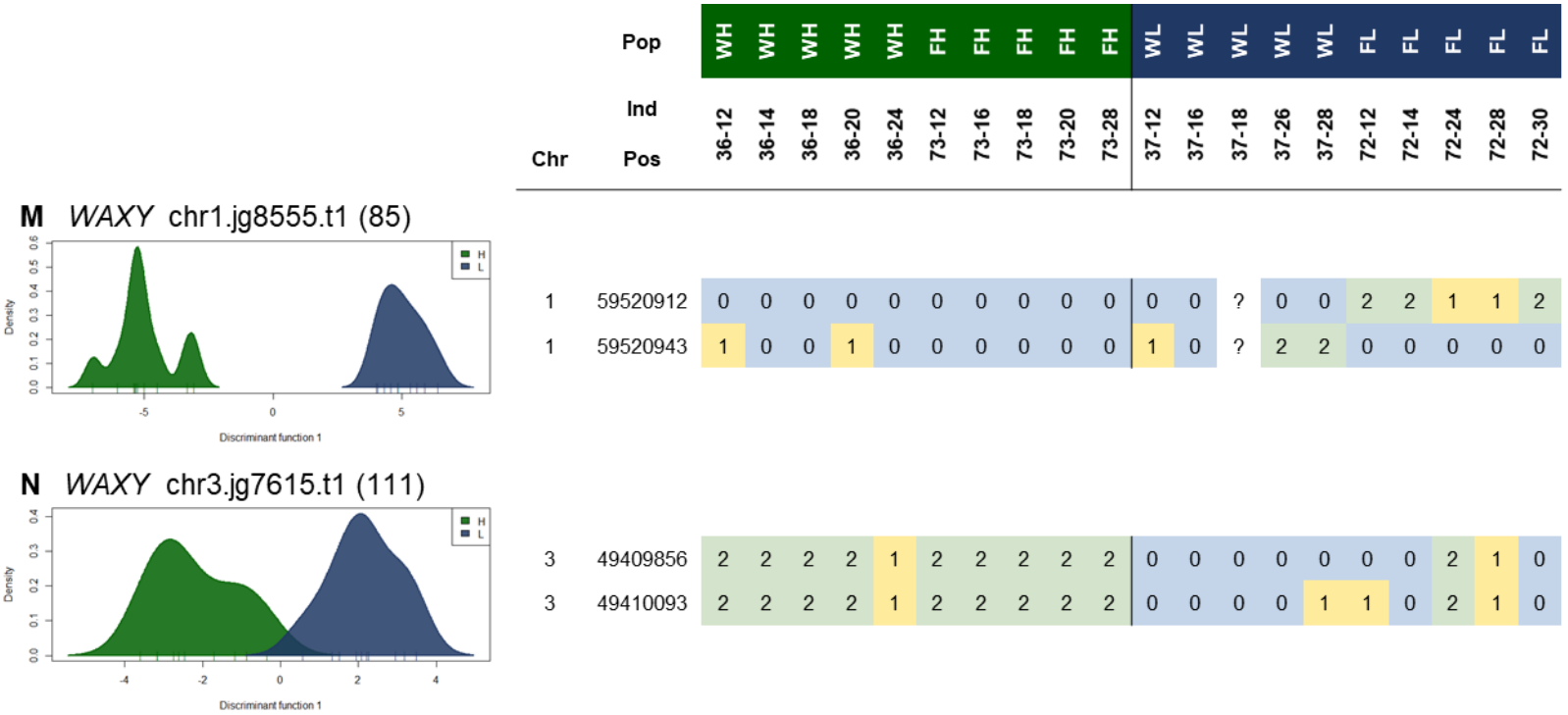
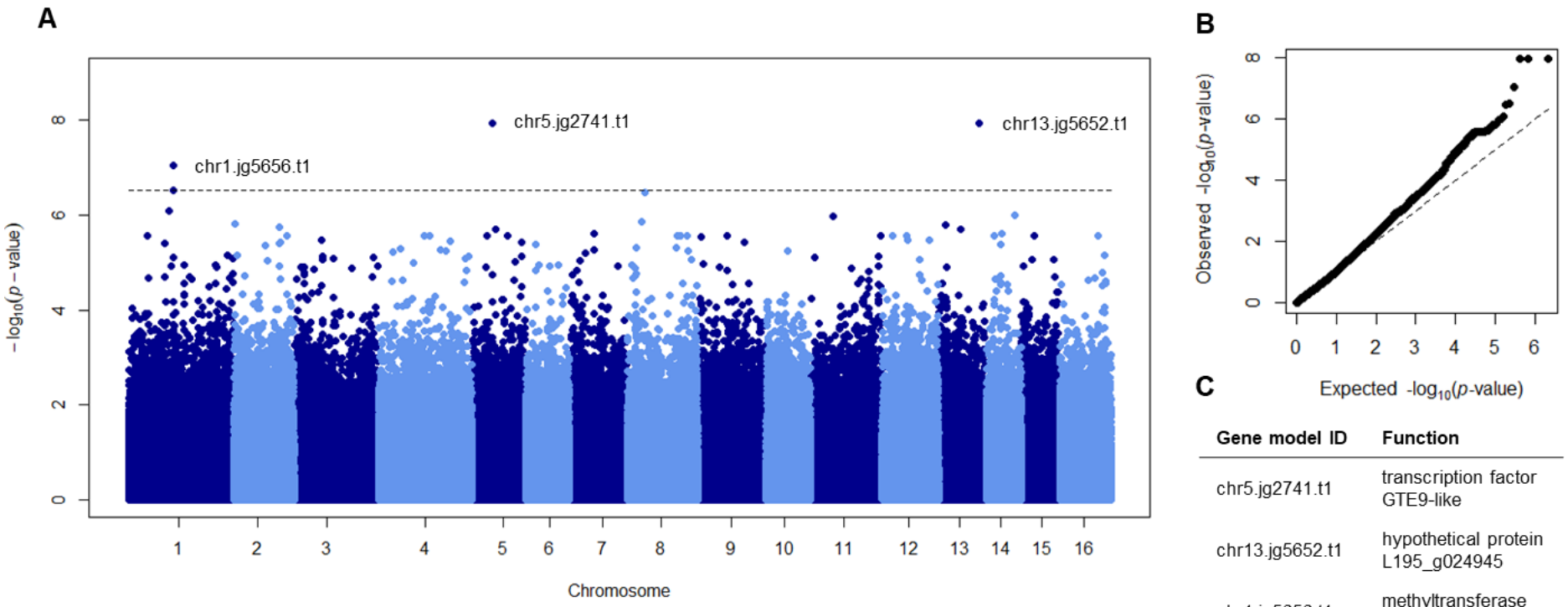


Figure S4.7 Single nucleotide polymorphism (SNP) variation driving separation between high and low water-soluble carbohydrate (WSC) populations for 14 candidate genes (**A**-**N**). On the left-hand side of the figure are the discriminant analysis of principal components (DAPC) individual density plots. These plots display the density of individuals for the first discriminant function (DF), with dark green representing individuals from the high WSC populations (H) and dark blue representing individuals from the low WSC populations (L). On the right-hand side of the figure are the associated SNPs that contribute the greatest to separating the H and L individuals from the first DF for the given gene model ID. "NO SNPs" indicates there were no SNPs within the gene with loadings above 0.03. Each column corresponds to an individual and each row corresponds to a SNP locus. Genotypes are colour and number coded, where: 0 = homozygous for the reference allele (blue), 1 = heterozygous (yellow), 2 = homozygous for the alternate allele (green), and ? = missing data. Gene model IDs are alphabetically ordered based on gene name. The number of SNPs used in the DAPC are presented inside parentheses next to gene model ID. *Note*: Chr = chromosome, Pos = position, Ind = individual, Pop = population. FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high WSC, Low = low WSC, End = End generation, W = Widdup.



C

Gene model ID	Function
chr5.jg2741.t1	transcription factor GTE9-like
chr13.jg5652.t1	hypothetical protein L195_g024945
chr1.jg5656.t1	methyltransferase type 11

Figure S4.8 Manhattan plot (**A**) from the genome-wide association study (GWAS) using 1,025,071 SNPs called from transcriptomic dataset and 20 individuals. -$\log_{10}$(*p*-values) are plotted against physical map position of SNPs on chromosomes. Significant loci associated with water-soluble carbohydrate levels lie above the false discovery rate threshold as denoted by the black dashed line (α = 0.05). Quantile-Quantile plot (**B**) of *p*-values from **A**. The function annotations of three highly significant gene model IDs (**C**).

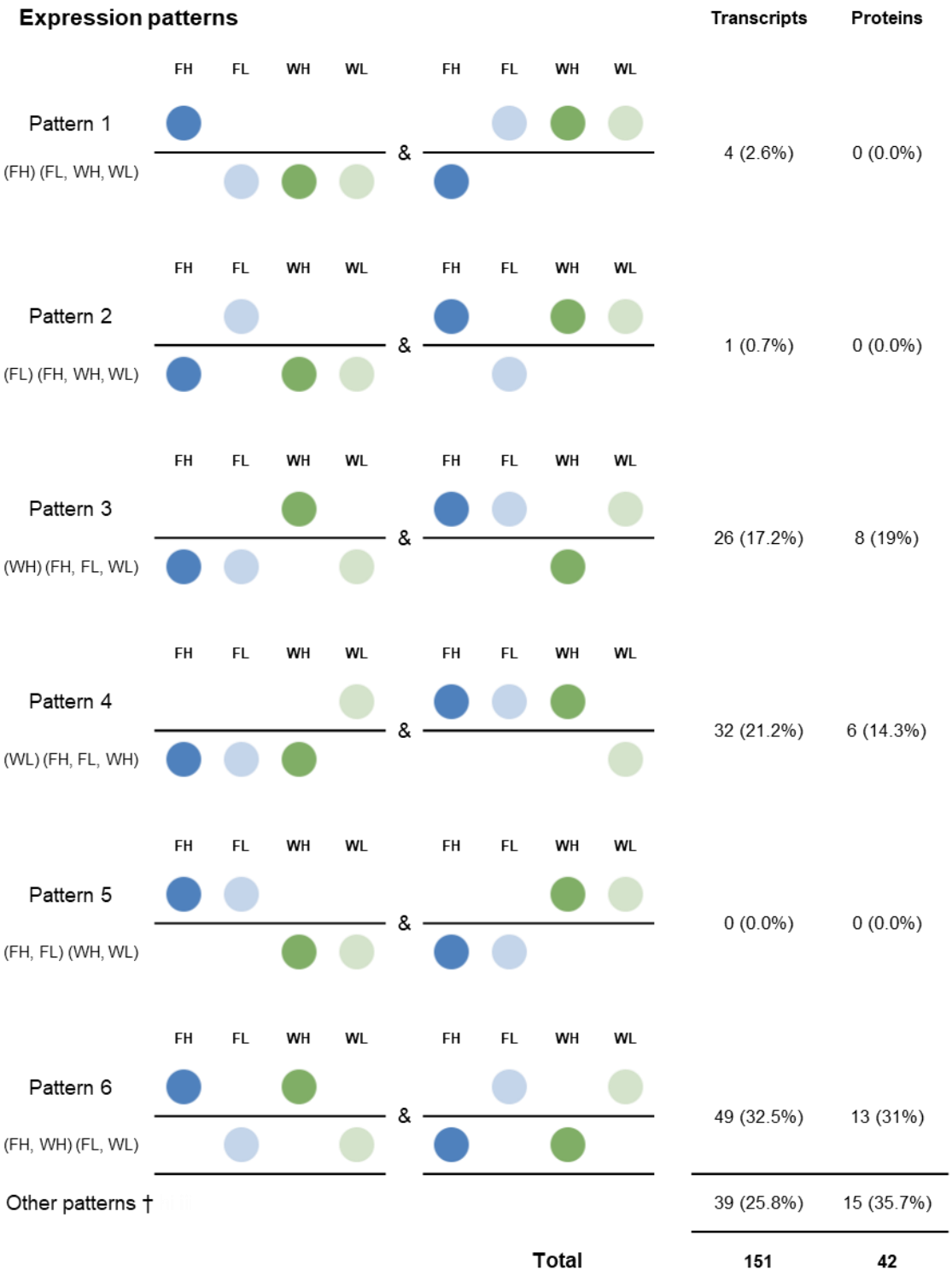


Figure S4.9 Summary of the number of differential expression pattern types of the transcripts and proteins identified in GO enrichment in four white clover populations that could be mapped to branches of the phylogenetic tree, as well as those patterns that are possible but could not be mapped. Comparisons of the number of transcripts and proteins belonging to six different expression patterns are presented to the right of the pattern figures. The percentages of the transcripts and proteins that mapped to a convergent selection response (same pattern of change in both pools; Pattern 6) was consistent for both datasets (32.5% and 31% for transcriptome and proteome, respectively. No gene model IDs were found for Pattern 5 as the 151 transcripts and 42

proteins were identified as significant for the WH-WL PWC, so all gene model IDs had to have expression differences between WH and WL.
† Other patterns include all those which do not follow patterns 1 to 6.
Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

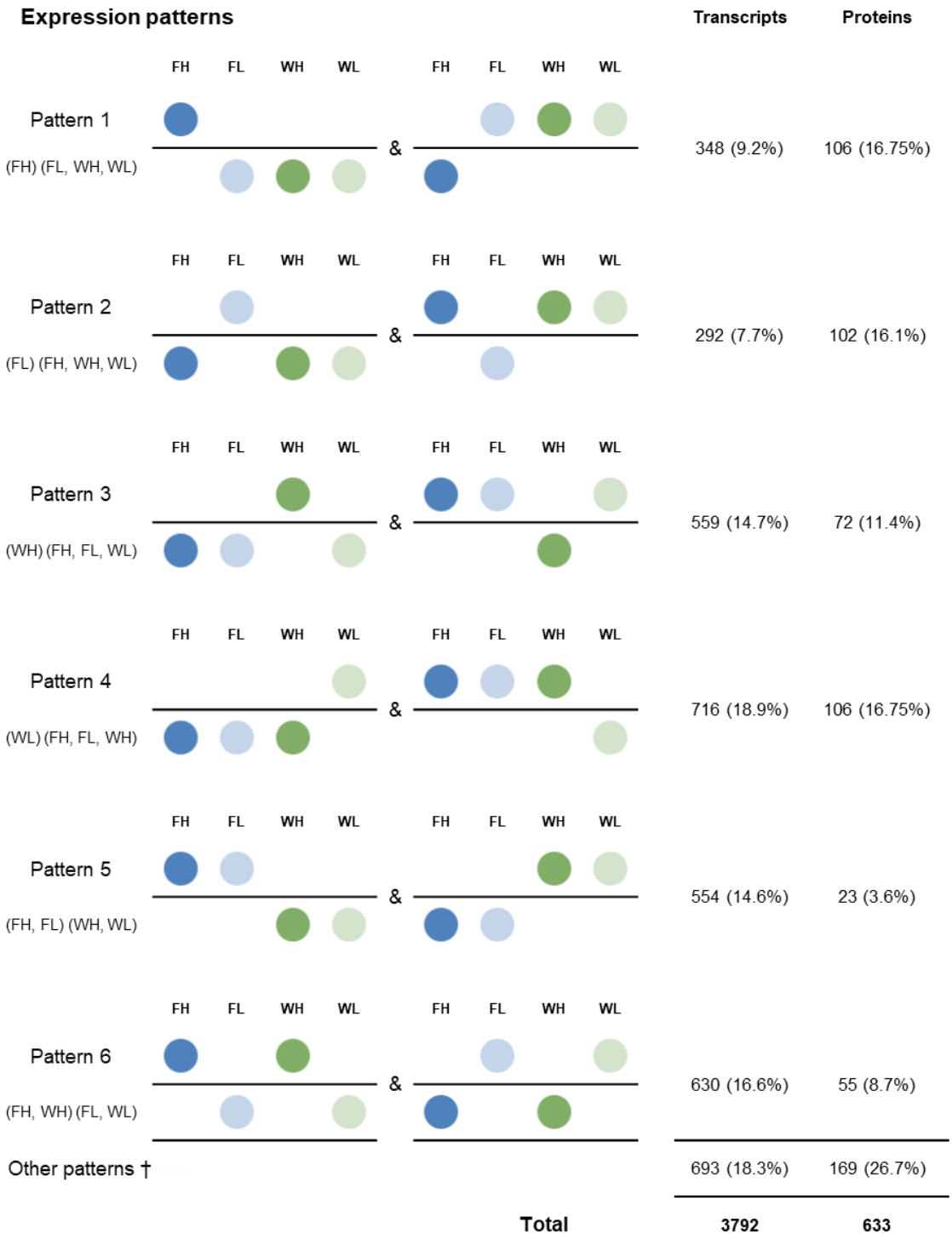


Figure S4.10 Summary of the number of differential expression pattern types of the transcripts and proteins identified in weighted gene correlation network analysis (WGCNA) in four white clover populations that could be mapped to branches of the phylogenetic tree, as well as those patterns that are possible but could not be mapped. Comparisons of the number of transcripts and proteins belonging to six different expression patterns are presented to the right of the pattern figures.

† Other patterns include all those which do not follow patterns 1 to 6.

Note: FH = FNZLL-High-End, FL = FNZLL-Low-End, WH = WNZLL-High-End, WL = WNZLL-Low-End; where: F = Ford, NZ = New Zealand/Aotearoa, LL = large leaf, High = high water-soluble carbohydrate (WSC), Low = low WSC, End = End generation, W = Widdup.

APPENDIX 4

Chapter 5 Supplementary Material

SUPPLEMENTARY EXPERIMENT – Sample size determination

Introduction

Many population genetic analyses are allele frequency based e.g. analysis of molecular variance, AMOVA (Excoffier, Smouse & Quattro, 1992), various estimates of genetic diversity and deviation from Hardy-Weinberg equilibrium. Therefore, it is imperative that the true population allele frequencies are represented when sampling from the population. The accuracy of allele frequency estimates will always increase with increasing sample size, but this rate of increase is not linear. Instead, the rate at which accuracy increases is expected to plateau as the sample number increases, but the cost of genotyping more samples increases in a linear fashion. A key question that needs to be addressed is identifying the sample size at which the increase in allele frequencies is too small to justify the cost of sampling more individuals. Because this is a statistical sampling problem, the required sample size that accurately reflects allele frequencies of the population should be congruent over multiple populations and taxa. Therefore, to determine the number of individuals required from each white clover pasture population, estimates of genetic diversity for two white clover and two perennial ryegrass (hereafter referred to as ryegrass) populations, with contrasting characteristics (including pedigree and expected level of genetic diversity), were assessed for variation in genetic diversity across a range of sample sizes. Due to the time and cost of preparing genotyping by sequencing (GBS) libraries to generate single nucleotide polymorphisms (SNPs), an alternative marker system was used to rapidly assess genetic diversity estimates in the small sample set.

Microsatellite markers, also known as simple sequence repeats (SSRs), are commonly found throughout the genome and consist of tandemly repeating DNA sequences only a few base pairs (1 – 6 bp) in length, with the di-nucleotide repeat the most abundant. SSRs are repeated multiple times and flanked by regions of non-repetitive DNA sequences (Tautz, 1989; Gupta *et al.*, 1996). These flanking conservative DNA sequences are used for designing suitable primers for amplification of the SSR loci, which is achieved using polymerase chain reaction (PCR) (Gupta *et al.*, 1996). The variation in the number of repeat motifs constitutes an SSR polymorphism, with each length representing an allele at a locus. These length differences are caused by slippage and proofreading errors during DNA replication (Selkoe & Toonen, 2006). SSRs are highly polymorphic, co-dominant, undergo Mendelian inheritance, easy to genotype and typically require a small number of loci, which makes them ideal for use

in population genetic, and evolutionary biology studies (Ashley & Dow, 1994; Sunnucks, 2000). SSRs allow the use of degraded or minute amounts of DNA (50 – 100 ng) (Queller, Strassmann & Hughes, 1993; Gupta *et al.*, 1996). However, there are large start-up costs involved in the development of SSR primers for a given species. There is a prerequisite to first identify SSR loci throughout the genome, which is achieved by screening available DNA sequences from databases or genomic libraries. Primers are then developed using the DNA template from the flanking regions around the SSR loci. Once primers are developed, they must be tested on a range of individuals from multiple populations to ensure they work efficiently (Gupta *et al.*, 1996; Barrett *et al.*, 2004; Zhang *et al.*, 2008). SSRs are generally species-specific and can provide information on the genetic structure of populations and on species delimitation (Duminil & Di Michele, 2009). Allele sizing can then be achieved either manually through the use of high-resolution acrylamide sequencing gels or using an automatic capillary sequencer, which greatly increases genotyping output (Boutin-Ganache *et al.*, 2001). Primers can be labelled with a fluorescent dye (e.g. FAM, NED or VIC) for automatic capillary sequencing. Multiple loci can be pooled for genotyping depending on the fluorescent dye used and the number and size of the repeat, which reduces the genotyping cost (Schuelke, 2000; Boutin-Ganache *et al.*, 2001). The advantage of co-dominant markers, like SSRs, is the ability to differentiate between heterozygotes and homozygotes, which is important in population genetic studies (Duminil & Di Michele, 2009).

SSRs have been used in a variety of plant genetic analyses, for example to measure population structure (Sunnucks, 2000; Blair, Soler & Cortés, 2012), assess genetic diversity (Leberg, 2002; Guan *et al.*, 2010), measure the effects of natural selection (Rodrigues & Santos, 2006), integrate genetic, physical and sequence-based maps (Córdoba *et al.*, 2010; McClean *et al.*, 2010) and for marker assisted selection (Benchimol *et al.*, 2007; Chen *et al.*, 2011). Expressed sequence tag derived SSRs have been implemented in combination with genomic SSRs to construct genetic linkage maps and estimate linkage disequilibrium in white clover (Barrett *et al.*, 2004; Isobe *et al.*, 2012; Griffiths *et al.*, 2013) as well as assess genetic diversity in white clover cultivars (George *et al.*, 2006). White clover populations are highly genetically heterogeneous due to the obligate outbreeding nature of the species, hence minimal population structure (assessed through AMOVA) has been identified in white clover through the use of SSRs (George *et al.*, 2006; Zhang *et al.*, 2010; Kooyers & Olsen, 2012). George *et al.* (2006) used 15 SSRs on a range of white cultivars from nine countries to assess the population diversity. Limited clustering of individuals from the same cultivar and substantial overlap between each cultivar was observed, suggesting that population

structure among white clover cultivars is present but limited. A larger number (or a different set) of SSR markers would be required for association genetics analyses to detect subtle population structure (George *et al.*, 2006). A study by Jahufer *et al.* (2003) utilised 39 SSRs and assessed the relatedness among 32 white clover cultivars from NZ and overseas. Significant distinction between cultivars was observed and the phylogenetic tree had a strong correlation with geographic origin and known pedigrees. Hence, SSRs are an excellent marker system for use in population genetics but their use for identifying causative genes underpinning traits is limited. For this pilot study, SSRs were used to provide a rapid and cheap estimate of allele frequencies in white clover and ryegrass populations.

Two clover populations ('Grasslands Huia' and 'Crau') were grown from seed and screened at seven single-locus microsatellite loci. For ryegrass, an existing data set (Faville *et al.*, 2020) of 180 ryegrass individuals from two populations ('Grasslands Nui' and 'Alto'), with data at 14 microsatellite loci, was used. All datasets were analysed using GenAlEx v6.502 (Peakall & Smouse, 2012) and genetic diversity parameters were estimated. If the accuracy of allele frequency estimates was found to converge on a similar sample size for all the datasets, then the results presented here should be applicable to other studies for sampling decisions for population genetic studies based on microsatellite data. The results presented here will give an accurate estimate of the number of samples required for landscape genomics studies in white clover throughout New Zealand/Aotearoa.

Materials and methods

Plant material

A total of 91 individuals were sampled from two white clover accessions ('Grasslands Huia' and 'Crau') that were grown from 0.3 g of seed each from the Margot Forde Forage Germplasm Centre (AgResearch Grasslands Research Centre, Palmerston North, New Zealand) following the protocol in Chapter 2, section 2.3.1. These seed were planted in trays containing a mix of peat and sand with a three-month slow release Osmocote fertiliser and placed in a glasshouse for 12 weeks before leaf tissue was harvested. Populations were selected based on date of commercial release as Huia is an older accession, developed more than 30 years ago (1964) and is expected to show high genetic diversity, whereas Crau (1966) and is expected to show less genetic diversity.

DNA extraction and genotyping

Leaf tissue from 43 Crau and 48 Huia individuals was collected into one 96-well plate and genomic DNA was extracted using the Anderson *et al.* (2018) method. All samples were resolved by electrophoresis on a 0.8% lithium borate agarose (w/v) gel containing 25 µg ethidium bromide for 40 minutes at 3.3 v cm^{-1} and visualised under UV (Gel Doc™, Bio-Rad, CA, USA) to assess DNA quality, molecular weight and quantity. A total of 91 individuals were screened at seven single-locus microsatellite markers described in Griffiths *et al.* (2013) (**Table SE5.1**). The PCR amplification was performed in a 10 µL volume containing 3.89 µL $_sH_2O$, 1 µL 10x PCR (-$MgCl_2$) buffer (Invitrogen), 0.5 µL 50mM $MgCl_2$ (Invitrogen), 0.4 µL 5mM each dNTPs, 10 µM M13 primer and fluorescent dye, 10x SSR primer, 0.06 5U μL^{-1} Platinum Taq (Invitrogen), and 2 µL of 1:100 diluted DNA:H_2O (5 – 5 ng). Amplification by PCR was carried out using a 2720 Thermo cycler (Applied Biosystems) and included an initial denaturation at 94°C for 4 minutes; then 30 cycles at 94°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds; then 8 cycles at 95°C for 30 seconds, 53°C for 30 seconds, and 72°C for 30 seconds; followed by a final extension at 72°C for 30 minutes.

FAM fluorophore was incorporated into PCR products using a three-primer system as described in Sartie *et al.* (2011) and a PIG tail to the 5' end of the reverse primer (GTTTCTT) to promote non-template (A) addition at the 3' terminus of PCR product (Brownstein, Carpten & Smith, 1996). Two loci were pooled after PCR for genotyping (**Table SE5.1**). PCR products (2 µL) were co-loaded and added to 9 µL Hi-Di formamide (Applied Biosystems, Warrington, UK) and 1 µL GeneScan 500 LIZ size standard (Applied Biosystems, Warrington, UK), denatured at 94°C for 5 minutes then 4°C for 5 minutes and subsequent fragment sizing on an ABI 3130xl genetic analyser (Applied Biosystems, Warrington, UK) at AgResearch Grasslands Research Centre (Palmerston North, New Zealand). Alleles were visualised and scored manually in GeneMarker v 2.4.0 (SoftGenetics).

Table SE5.1 Primer sequences and characteristics of seven single-locus specific SSR loci developed from *Trifolium repens* (Griffiths et al. 2013).

Locus	LG	Primer sequences (5' to 3')[a]	Pooling group	Repeat motif	eSize (bp)	Size range in Huia (bp)	Size range in Crau (bp)	Ta (°C)
gtrs299	Tr_8-1	F: AATCCTATCCAAATACGGACAAT R: TGTAGTGTTGAAGCAAACTCAAAT	1	$(TAT)_{10}$	184	197-230	200-218	55
gtrs427	Tr_7-2	F: GGGTCTGGTTTGGTTTACTTTG R: CAAGATATAGTGCCAACCCTTCA	2	$(AG)_{40}$	359	215-331	265-325	55
gtrs588	Tr_2-1	F: TGGTGTCTGTGTATCTAATCTAAACAT R: GAAGTGGAATTGGAGGATTGTTA	1	$(TAT)_{7}$	134	146-161	146-164	55
gtrs600	Tr_6-2	F: TGATTCTTAGGATGACTCGTTTGA R: TTCTCTACGATTGTGTGTGGAAA	2	$(TTA)_{7}$	154	171-183	174-183	55
gtrs625	Tr_4-2	F: GGCCATTAGCAAACAAAGTCATA R: AAACCGAAGAATTGACCACTACA	3	$(TA)_{9}$	215	230-254	230-248	55
gtrs654	Tr_3-1	F: ATCAAACACTCCTCAACTCTGCT R: GTGTAACCGTCAATCTCGGAATA	4	$(ATG)_{8}$	184	180-297	186-324	55
gtrs789	Tr_1-2	F: CAAACCCTTACAATTCAAACCAG R: ATCACTTCTTCGTCATCATCGTT	3	$(TTC)_{7}$	324	344-380	344-356	55

Note: LG = Linkage group; Pooling group = markers pooled together after PCR for genotyping, Repeat motif = repeat motif and the number of times the motif was repeated in the source SSR array; eSize (bp) = predicted amplicon size (base pairs) based on *in silico* data; Ta = annealing temperature used in PCR.

[a] M13 tail (CACGACGTTGTAAAACGAC) added to the 5' end of each forward primer and a PIG tail (CTTTCTT) added to the 5' end of each reverse primer.

Statistical analysis

For all four populations, commonly used diversity statistics were calculated using GenAlEx 6.503 (Peakall & Smouse, 2012), including the average observed heterozygosity (H_O), unbiased average expected heterozygosity (H_E), number of alleles per locus (Na), and genetic distances (Fixation index, F_{ST}, and Nei's genetic distance). These values were calculated from the full sample sizes of 43 – 90 individuals for each population and were treated as the known values for their respective population. Data sets of different sample sizes (10, 20, 30, 40, 50, 60, 70, 80 for ryegrass; and 10, 20, 30, 40 for white clover) were created by randomly sampling (without replacement) individuals from each population data set using Microsoft Excel 2016 (Microsoft Co., Seattle). The mean of 10 simulations for each data set and the standard errors for each simulated data set were calculated. The mean and standard error was plotted against sample size using Minitab (v 18.1). Nei's genetic distance was also calculated at each sample size (data not shown), but generated an almost identical pattern to the pairwise F_{ST} values, so only the pairwise F_{ST} data is presented.

Results

Genetic variation

For the more recent ryegrass cultivar (Alto), actual values of genetic diversity estimates based on 90 individuals were slightly higher than those in the older cultivar (Nui) (**Table SE5.2**). The average number of alleles per locus for Nui was 5.43 and for the Alto population 6.14. For the low diversity white clover cultivar (Crau *n* = 43), the average number of alleles per locus was half the number of the higher diversity cultivar (Huia *n* = 48): 4.86 and 8.86, respectively. Expected heterozygosity was lower for Nui (H_E = 0.640, H_O = 0.584) than Alto (H_E = 0.692, H_O = 0.580) while observed heterozygosity was similar. Both heterozygosity values for Crau (H_E = 0.589, H_O = 0.416) were lower than for Huia (H_E = 0.733, H_O = 0.510). Pairwise F_{ST} values for the ryegrass and clover populations were very similar (0.047 and 0.088, respectively) and indicate little to moderate differentiation between populations within the two species (**Table SE5.2**).

Table SE5.2 Population information and genetic diversity estimates for two cultivars of ryegrass (Nui and Alto) and two cultivars of white clover (Huia and Crau).

Species	Cultivar	*n*	Nm	Na	H_O	H_E	F_{ST}
Ryegrass	Nui	90	14	5.429	0.584	0.640	0.047
	Alto	90	14	6.143	0.580	0.692	
White clover	Huia	48	7	8.857	0.510	0.733	0.088
	Crau	43	7	4.857	0.416	0.589	

Note: *n* = number of samples genotyped, Nm = number of microsatellite markers, Na = number of alleles per locus, H_O = observed heterozygosity, H_E = expected heterozygosity, F_{ST} = estimation of Wright's fixation index for microsatellite markers.

Sample size

Based on measures estimated for data sets of different sample size, H_O and H_E plateaued at approximately 30 individuals in each ryegrass and clover population (**Figures SE5.1** and **SE5.2**). Na continued to increase but at a lower rate after n = 40. In both species, Na flattens out faster in cultivars (Crau and Nui) with lower gene diversity (H_E). A typical trend observed was the reduction in standard error with increasing dataset size. Although there was no significant change in Ho and H_E values from 20 – 40 individuals in white clover and for sample sizes greater than 30 in ryegrass, the standard error was minimal for sample sizes above 30 for both datasets. This suggests that increasing the sample size above 30 individuals will have little impact on the accuracy of H_E and H_O estimates, while increasing the sample size above 40 individuals will have little impact on the accuracy of Na. The genetic distance is expected to decrease as the sample size increases but not in a linear fashion. Therefore, the point at which the increase in accuracy gained by adding extra samples (non-linear) is outweighed by the increase in cost (generally linear) of obtaining those additional samples needs to be identified. The incremental decrease in pairwise F_{ST} is also small beyond a sample size of approximately 30 for all data sets, until almost all individuals are sampled (sample size of 43 for Crau, 48 for Huia and 90 for Alto and Nui; **Figure SE5.3**). Therefore, 30 plants from each species from each population will be sufficient to accurately assess these genetic variability measures based on microsatellite allele frequencies.

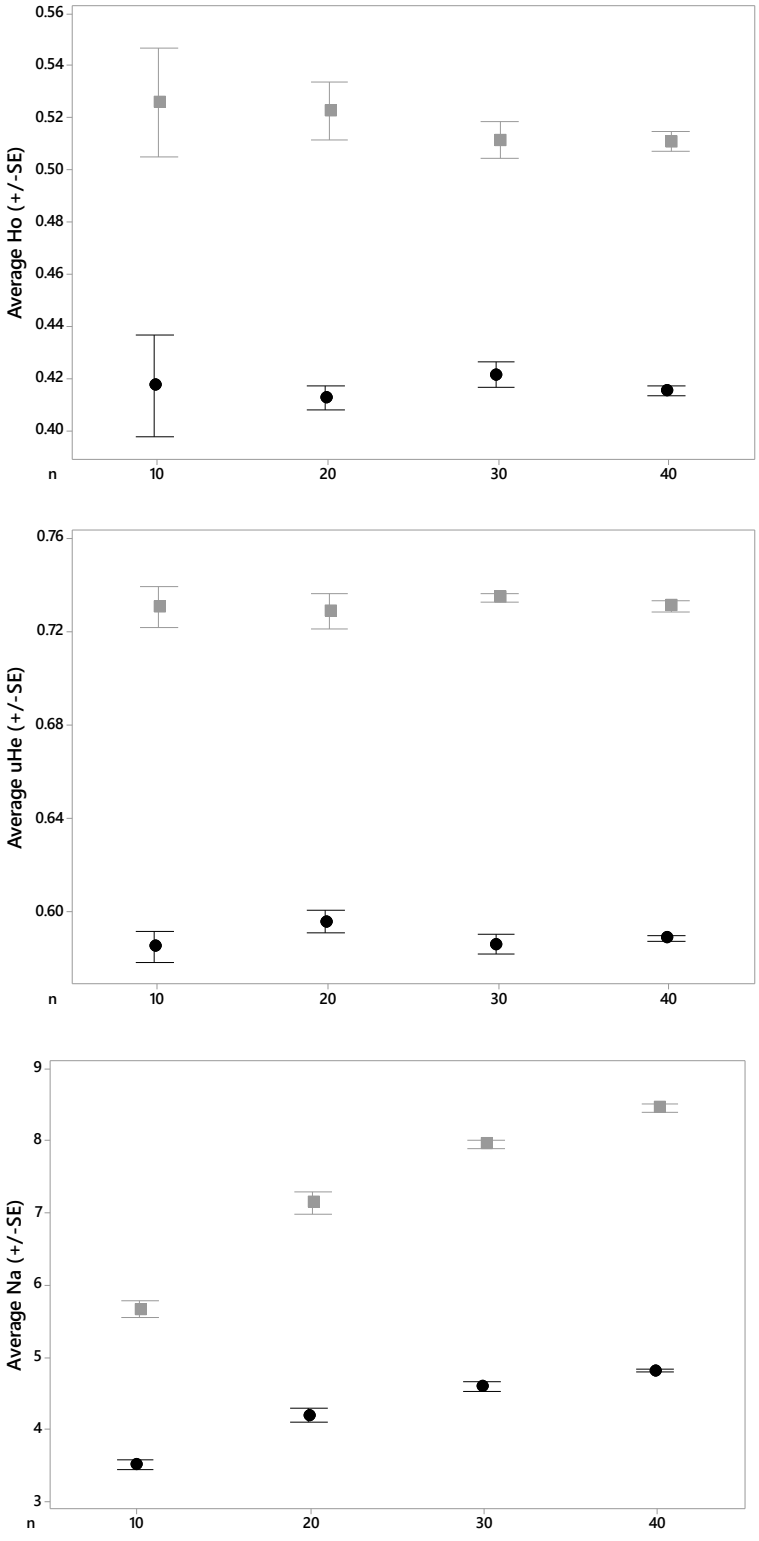


Figure SE5.1 Mean and standard error (SE) for estimates of averaged observed heterozygosity (Ho), average unbiased expected heterozygosity (uHe), and average alleles per locus (Na) for Huia (grey squares) and Crau (black circles) white clover populations. Values are based on 10 random data sets for various sample sizes (n).

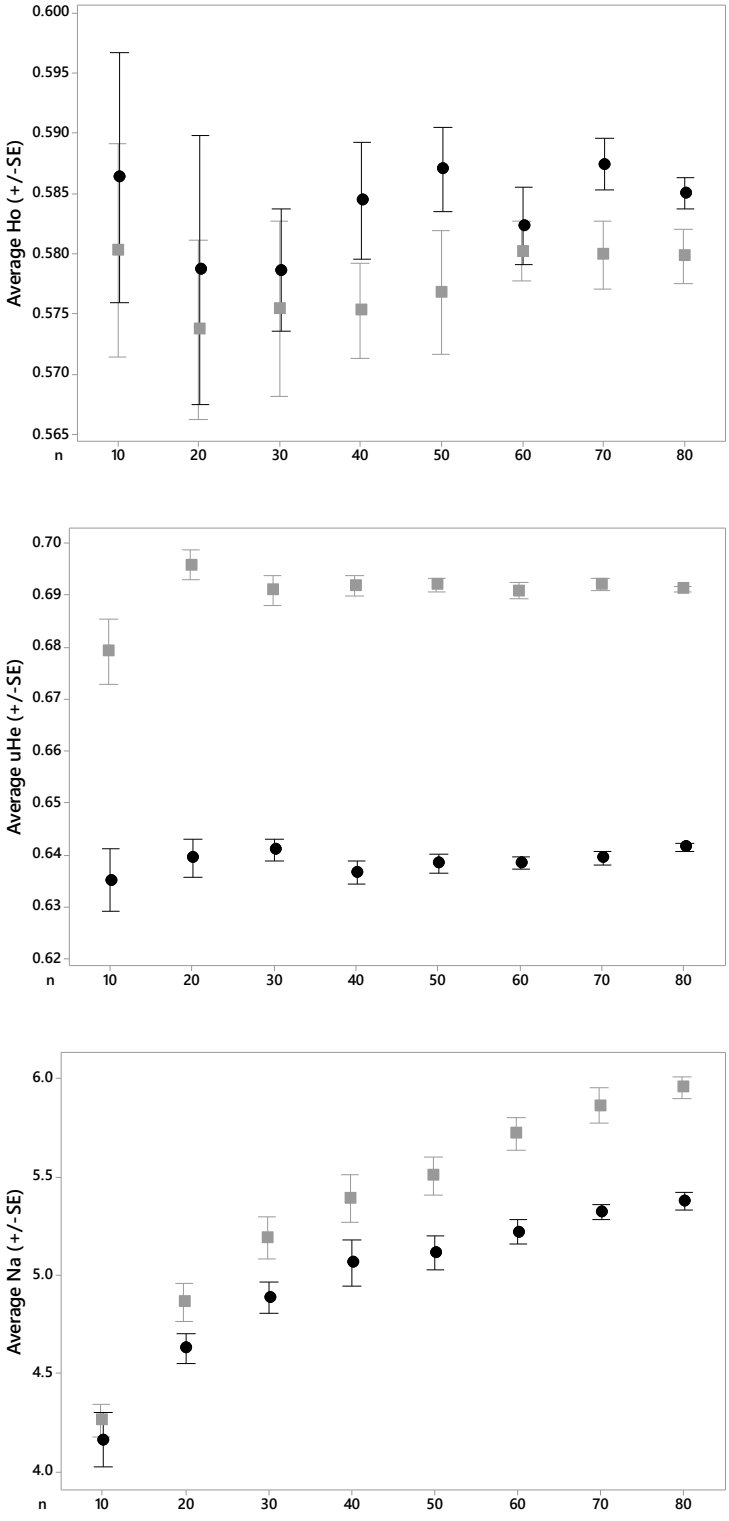
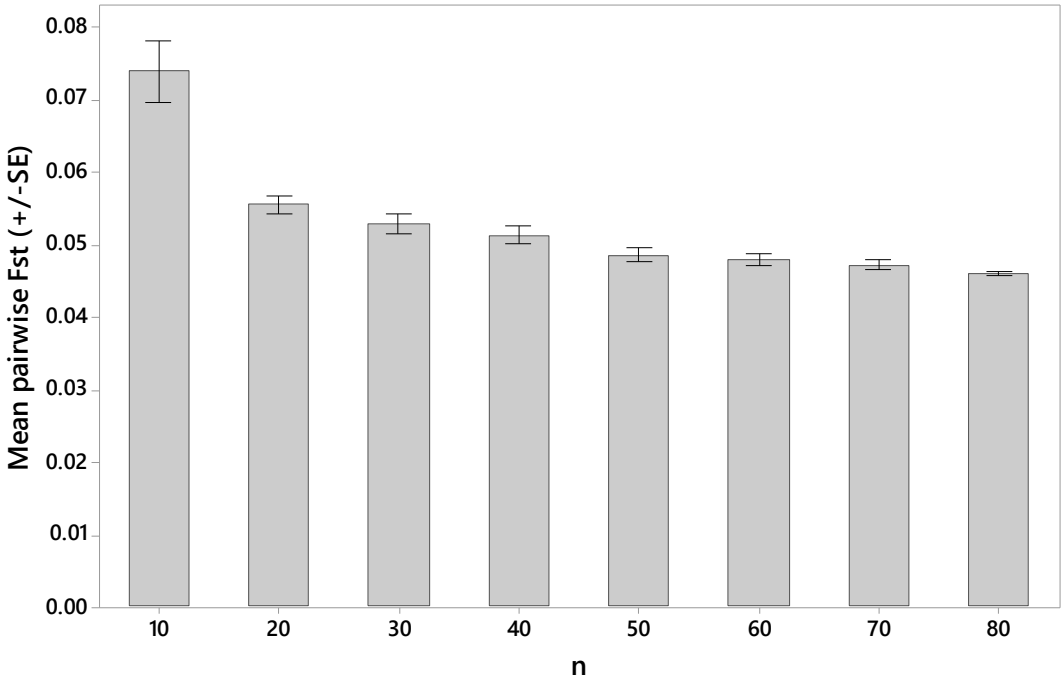


Figure SE5.2 Mean and standard error (SE) for estimates of average observed heterozygosity (Ho), average unbiased expected heterozygosity (uHe), and average alleles per locus (Na) for Alto (grey squares) and Nui (black circles) ryegrass populations. Values are based on 10 random data sets for various sample sizes (n).



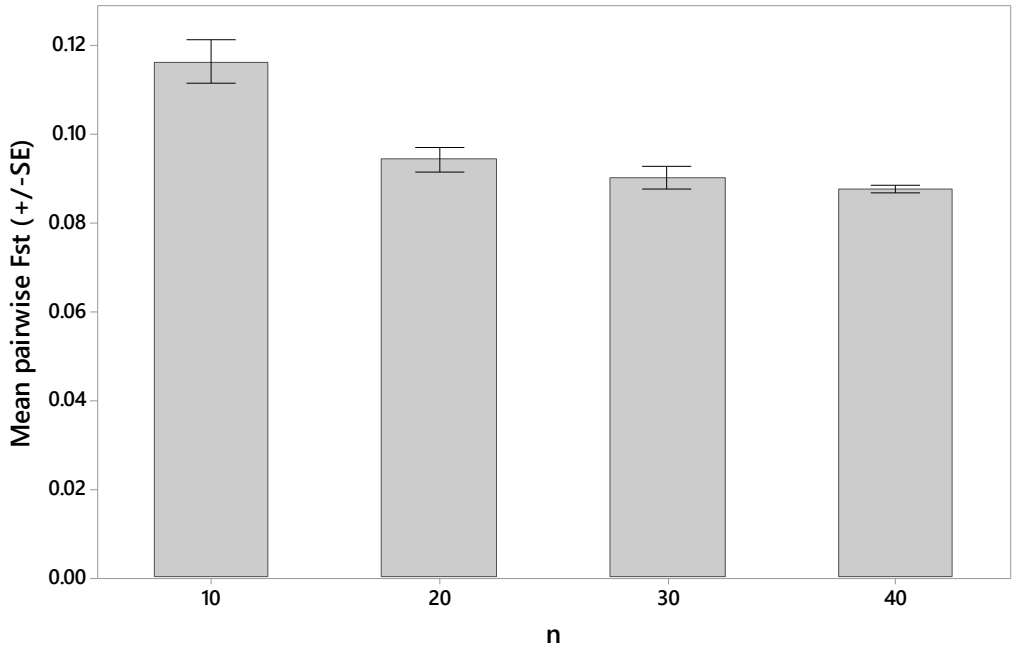


Figure SE5.3 Mean pairwise F_{ST} between the 10 random replications for the ryegrass dataset (top) and the white clover dataset (bottom) at each sample size (n). Error bars are the standard error.

Discussion

Genetic diversity estimates

Some life history characteristics are indicative of higher genetic variation such as wind pollination, both sexual and asexual reproduction, having a widespread geographic range and being a perennial; all of which have been observed in perennial ryegrass (Hamrick, Linhart & Mitton, 1979; Loveless & Hamrick, 1984; Nybom, 2004; Australian Government, 2008). White clover differs from this in some aspects, including animal (bee) pollination, and having a higher ploidy level ($2n = 4x = 32$ c.f. $2n = 2x = 14$ in ryegrass) (Kubik *et al.*, 2001; George *et al.*, 2006), although the ryegrass genome is larger (2.3 – 2.7 Gbp) compared to white clover (ca. 1 Gbp) (Bennett & Leitch, 2011; Griffiths *et al.*, 2019). Animal pollination tends to decrease genetic variability as bees are limited to near-neighbour pollen movement but a larger genome size tends to have more polymorphic loci and a greater number of alleles per locus (Hamrick *et al.*, 1979). One white clover cultivar (Huia) tended to have higher genetic variation compared with both perennial ryegrass cultivars which is to be expected given their life history traits. The mean number of alleles per locus (Na) is indicative of within-population diversity. Generally, if there is a large Na then there is a high population diversity e.g. Huia had Na of 8.9 compared with Crau, which had 4.9 (approximately half), therefore there was greater diversity in the older cultivar as anticipated. Conversely, Nui showed lower diversity compared with its younger counterpart, Alto (5.4 and 6.1, respectively).

Observed heterozygosity (H_O) is another measure of within-population diversity and indicates how many individuals were heterozygous as a proportion of the population. Alto and Nui had similar H_O, and both were higher for H_O than Huia and Crau. If there is a high proportion of heterozygotes, then that indicates there is a high level of allelic diversity in the population. If individuals were all homozygous then a narrower allelic diversity would be present. Expected heterozygosity (H_E) is the proportion of heterozygotes expected in the population based on Hardy-Weinberg equilibrium (HWE) proportions ($HWE = p^2 + 2pq + q^2$) estimated from observed allele frequencies. Observed heterozygosity was smaller for all populations compared to H_E. When there is a deficit of heterozygotes in the populations compared to what is expected under HWE, it suggests that there is potential inbreeding or that some of the markers are not in HWE. For both species, H_O was smaller for all populations compared to H_E which may reflect inbreeding. In ryegrass there was a greater difference for Alto (0.382)

than for Nui (0.056). For white clover, differences were similar for both cultivars (0.173 and 0.223 for Crau and Huia, respectively).

F_{ST} is commonly used as a measure of population differentiation. More specifically, it is a measure of sub-population differentiation as it estimates the proportion of heterozygosity in the sub-population relative to the whole population. For example, for ryegrass Alto and Nui collectively are the population and Alto is a sub-population, as is Nui. In this context it allows us to measure the degree of differentiation between the two sub populations. The pairwise F_{ST} values for both of the species were low (0.047 and 0.088) indicating little to moderate genetic differentiation for ryegrass and white clover, respectively. Low F_{ST} values are not uncommon for both white clover and ryegrass (Kubik *et al.*, 2001; Wright *et al.*, 2017; Inostroza *et al.*, 2018).

Sample size determination

Sample size is a critical issue for genetic diversity studies. Estimating sample sizes appropriate for genetic diversity parameter estimation requires previous knowledge of genotype frequencies for a given population (Miyamoto *et al.*, 2008). The aim of this experiment was to address the question of how many individuals would need to be sampled to give an accurate estimate of each locus' allele frequencies and thus the expected heterozygosity within a population. Previous papers that have addressed this issue have examined single taxa datasets (Miyamoto *et al.*, 2008; Pruett & Winker, 2008) as well as multiple taxa datasets (Hale, Burg & Steeves, 2012). Pruett and Winker (2008) suggest that at least 20 samples, but preferably 30, should give a useful measure of genetic diversity in song sparrow (*Melospiza melodia*) populations. Most genetic diversity studies in plant populations use sample sizes of about 50 individuals per population (Nybom, 2004). Miyamoto *et al.* (2008) found with common ash (*Fraxinus excelsior*) that 31 – 50 individuals on average are required for genetic diversity studies. Khanlou *et al.* (2011) used AFLP data (283 loci) from three white clover cultivars and recommend 20 – 30 samples should be used to quantify genetic diversity within and among white clover cultivars. A minimum of 20 samples maximises cost-effectiveness but 30 samples are recommended to adequately and precisely detect minor differences at the level of inter-cultivar genetic variation (Khanlou *et al.*, 2011). A diverse range of taxa including invertebrates, birds and mammals (hairy wood ant (*Formica lugubris*), black-browed albatross (*Thalassarche melanophris*), black stilts (*Himantopus novaezelandia*), and British red squirrel (*Sciurus vulgaris*) were used to create four datasets to investigate this sample size issue. The result of 25 – 30 diploid individuals

required for accurate microsatellite-based population genetic diversity studies was consistent across all four datasets (Hale *et al.*, 2012). The results presented here are consistent with the above-mentioned studies. Generally, 25 – 30 samples are required for animal taxa and 30 – 50 are required for plant taxa. Therefore, a minimum of 30 white clover plants will need to be sampled from each population to accurately assess allele frequencies.

It is also important to consider the number of populations and the size of the study area. There appears to be a trade-off between the number of samples analysed within populations and the number of populations, especially when sampling over large geographic areas. For example, Cao *et al.* (2018), Eckert *et al.* (2010) and Jones *et al.* (2013) investigated landscape genomics in populations of plant species in Japan, Korea and Taiwan; South-eastern USA, from Texas to Delaware; and the European Alps, respectively. Their sample sizes per population were small, ranging from 2 – 13 on average, but because of the large sample area, the number of populations ranged from 54 – 99. On the other hand, landscape genomic studies can also have a moderate number of populations with more individuals sampled. Some of these examples focus on sampling populations at the extremes of an environmental gradient and then replicating samples along at least two similar environmental gradients. For example Berthouly-Salazar *et al.* (2016) sampled four wild pearl millet (*Cenchrus americnus* ssp. *monodii*) populations at the extremity of two replicated aridity gradients, with 18 plants per population. They then included 11 additional populations (total of 762 samples) for validating single nucleotide polymorphisms (SNPs) that were potentially under selection. De Kort *et al.* (2014) sampled 15 black alder (*Alnus glutinosa*) individuals from 24 populations in eight European regions. Roschanski *et al.* (2016) sampled 39 – 55 silver fir (*Abies alba*) trees from high and low elevations on each of four mountains in the French Mediterranean Alps. Focusing on populations in divergently extreme habitats and replicating samples along environmentally similar gradients can reduce the effect of historical and spatial processes. We can rule out genetic drift and these neutral processes producing similar genetic patterns at a locus across independent environmental gradients (Poncet *et al.*, 2010). In this study, there are 26 known naturalised pastures throughout New Zealand/Aotearoa, with the majority located in the South Island/Te Waipounamu. Give the largest rain shadow effect occurs in the South Island/Te Waipounamu due to the Southern Alps/Kā Tiritiri o te Moana, the main focus of the landscape genomic study should concentrate on the 19 populations located there (van Ham *et al.*, 2016).

Conclusion

The results presented here suggest that a minimum of 30 plants for white clover will need to be sampled from each population in order to capture the full quantum of population genetic diversity and enable accurate assessment of genomic changes across environmentally contrasting habitats. However, 50 plants will be collected in case samples are too small or damaged or as a back-up if DNA extraction fails. Although this study examined population diversity with regards to microsatellite data, it should be able to be extended to single nucleotide polymorphism (SNP) data.

References

Anderson, C. B., Franzmayr, B. K., Hong, S. W., Larking, A. C., van Stijn, T. C., Tan, R., . . . Griffiths, A. G. (2018). Protocol: a versatile, inexpensive, high-throughput plant genomic DNA extraction method suitable for genotyping-by-sequencing. *Plant Methods, 14*(1), 75.

Ashley, M. V., & Dow, B. D. (1994). The use of microsatellite analysis in population biology: Background, methods and potential applications. In B. Schierwater, B. Streit, G. P. Wagner & R. DeSalle (Eds.), *Molecular Ecology and Evolution: Approaches and Applications* (pp. 185-201). Basel: Birkhäuser Basel.

Australian Government. (2008). The biology of *Lolium multiflorum* Lam. (Italian ryegrass), *Lolium perenne* L. (perennial ryegrass) and *Lolium arundinaceum* (Schreb.) Darbysh (tall fescue).

Barrett, B., Griffiths, A., Schreiber, M., Ellison, N., Mercer, C., Bouton, J., . . . Woodfield, D. (2004). A microsatellite map of white clover. *Theoretical and Applied Genetics, 109*(3), 596-608.

Benchimol, L. L., Campos, T. d., Carbonell, S. A. M., Colombo, C. A., Chioratto, A. F., Formighieri, E. F., . . . Souza, A. P. d. (2007). Structure of genetic diversity among common bean (*Phaseolus vulgaris* L.) varieties of Mesoamerican and Andean origins using new developed microsatellite markers. *Genetic Resources and Crop Evolution, 54*(8), 1747-1762.

Bennett, M. D., & Leitch, I. J. (2011). Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Annals of Botany, 107*(3), 467-590.

Berthouly-Salazar, C., Thuillet, A.-C., Rhoné, B., Mariac, C., Ousseini, I. S., Couderc, M., . . . Vigouroux, Y. (2016). Genome scan reveals selection acting on genes linked to stress response in wild pearl millet. *Molecular Ecology, 25*(21), 5500-5512.

Blair, M. W., Soler, A., & Cortés, A. J. (2012). Diversification and Population Structure in Common Beans (*Phaseolus vulgaris* L.). *PLOS ONE, 7*(11), e49488.

Boutin-Ganache, I., Raposo, M., Raymond, M., & Deschepper, C. F. (2001). M13-Tailed Primers Improve the Readability and Usability of Microsatellite Analyses Performed with Two Different Allele- Sizing Methods. *BioTechniques, 31*(1), 25-28.

Brownstein, M. J., Carpten, J. D., & Smith, J. R. (1996). Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *BioTechniques, 20*, 1004-1010.

Cao, Y. N., Wang, I. J., Chen, L. Y., Ding, Y. Q., Liu, L. X., & Qiu, Y. X. (2018). Inferring spatial patterns and drivers of population divergence of Neolitsea sericea (Lauraceae), based on molecular phylogeography and landscape genomics. *Molecular Phylogenetics and Evolution, 126*, 162-172.

Chen, L., Zhao, Z., Liu, X., Liu, L., Jiang, L., Liu, S., . . . Wan, J. (2011). Marker-assisted breeding of a photoperiod-sensitive male sterile japonica rice with high cross-compatibility with indica rice. *Molecular Breeding, 27*(2), 247-258.

Córdoba, J. M., Chavarro, C., Schlueter, J. A., Jackson, S. A., & Blair, M. W. (2010). Integration of physical and genetic maps of common bean through BAC-derived microsatellite markers. *BMC Genomics, 11*(1), 436.

De Kort, H., Vandepitte, K., Bruun, H. H., Closset-Kopp, D., Honnay, O., & Mergeay, J. (2014). Landscape genomics and a common garden trial reveal adaptive

differentiation to temperature across Europe in the tree species *Alnus glutinosa*. *Molecular Ecology, 23*(19), 4709-4721.

Duminil, J., & Di Michele, M. (2009). Plant species delimitation: A comparison of morphological and molecular markers. *Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology, 143*(3), 528-542.

Eckert, A. J., Bower, A. D., González-Martínez, S. C., Wegrzyn, J. L., Coop, G., & Neale, D. B. (2010). Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology, 19*(17), 3789-3805.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics, 131*(2), 479-491.

Faville, M. J., Crush, J. R., Hong, W., Phillips, H., Lee, J. M., & Chapman, D. (2020). Effects of pasture age on the genotype and phenotype of perennial ryegrass. *Grass and Forage Science, 75*(2), 135-144.

George, J., Dobrowolski, M. P., Jong, E. v. Z. d., Cogan, N. O. I., Smith, K. F., & Forster, J. W. (2006). Assessment of genetic diversity in cultivars of white clover (*Trifolium repens* L.) detected by SSR polymorphisms. *Genome, 49*, 919-930.

Griffiths, A. G., Barrett, B. A., Simon, D., Khan, A. K., Bickerstaff, P., Anderson, C. B., . . . Jones, C. S. (2013). An integrated genetic linkage map for white clover (*Trifolium repens* L.) with alignment to *Medicago*. *BMC Genomics, 14*, 388.

Griffiths, A. G., Moraga, R., Tausen, M., Gupta, V., Bilton, T. P., Campbell, M. A., . . . Andersen, S. U. (2019). Breaking Free: The Genomics of Allopolyploidy-Facilitated Niche Expansion in White Clover. *The Plant Cell, 31*(7), 1466-1487.

Guan, R., Chang, R., Li, Y., Wang, L., Liu, Z., & Qiu, L. (2010). Genetic diversity comparison between Chinese and Japanese soybeans (*Glycine max* (L.) Merr.) revealed by nuclear SSRs. *Genetic Resources and Crop Evolution, 57*(2), 229-242.

Gupta, P. K., Balyan, H. S., Sharma, P. C., & Ramesh, B. (1996). Microsatellites in plants: A new class of molecular markers. *Current Science, 70*(1), 45-54.

Hale, M. L., Burg, T. M., & Steeves, T. E. (2012). Sampling for Microsatellite-Based Population Genetic Studies: 25 to 30 Individuals per Population Is Enough to Accurately Estimate Allele Frequencies. *PLoS ONE, 7*(9), e45170.

Hamrick, J. L., Linhart, Y. B., & Mitton, J. B. (1979). Relationships between life history characteristics and electrophoretically detectable genetic variation in plants. *Annual Review of Ecology, Evolution, and Systematics, 10*, 173-200.

Inostroza, L., Bhakta, M., Acuña, H., Vásquez, C., Ibáñez, J., Tapia, G., . . . Munoz, P. (2018). Understanding the Complexity of Cold Tolerance in White Clover using Temperature Gradient Locations and a GWAS Approach. *The Plant Genome, 11*(3).

Isobe, S. N., Hisano, H., Sato, S., Hirakawa, H., Okumura, K., Shirasawa, K., . . . Tabata, S. (2012). Comparative Genetic Mapping and Discovery of Linkage Disequilibrium Across Linkage Groups in White Clover (*Trifolium repens* L.). *G3: Genes|Genomes|Genetics, 2*(5), 607-617.

Jahufer, M. Z. Z., Barrett, B. A., Griffiths, A. G., & Woodfield, D. R. (2003). DNA fingerprinting and genetic relationships among white clover cultivars. *Proceedings of the New Zealand Grassland Association, 65*, 163-169.

Jones, M. R., Forester, B. R., Teufel, A. I., Adams, R. V., Anstett, D. N., Goodrich, B. A., . . . Manel, S. (2013). Integrating Landscape Genomics and Spatially Explicit Approaches to Detect Loci Under Selection in Clinal Populations. *Evolution, 67*(12), 3455-3468.

Khanlou, K. M., Vandepitte, K., Asl, L. K., & Bockstaele, E. V. (2011). Towards an optimal sampling strategy for assessing genetic variation within and among white clover (*Trifolium repens* L.) cultivars using AFLP. *Genetics and Molecular Biology, 34*, 252-258.

Kooyers, N. J., & Olsen, K. M. (2012). Rapid evolution of an adaptive cyanogenesis cline in introduced North American white clover (*Trifolium repens* L.). *Molecular Ecology, 21*(10), 2455-2468.

Kubik, C., Sawkins, M., Meyer, W. A., & Gaut, B. S. (2001). Genetic diversity in seven perennial ryegrass (Lolium perenne L.) cultivars based on SSR markers. *Crop Science, 41*(5), 1565-1572.

Leberg, P. L. (2002). Estimating allelic richness: Effects of sample size and bottlenecks. *Molecular Ecology, 11*(11), 2445-2449.

Loveless, M. D., & Hamrick, J. L. (1984). Ecological determinants of genetic structure in plant populations. *Annual Review of Ecology, Evolution, and Systematics, 15*, 65-95.

McClean, P. E., Mamidi, S., McConnell, M., Chikara, S., & Lee, R. (2010). Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. *BMC Genomics, 11*(1), 184.

Miyamoto, N., Fernandez-Manjarres, J. F., Morand-Prieur, M. E., Bertolino, P., & Frascaria-Lacoste, N. (2008). What sampling is needed for reliable estimations of genetic diversity in *Fraxinus excelsior* L. (Oleaceae)? *Annals of Forest Science, 65*(4), 8.

Nybom, H. (2004). Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology, 13*(5), 1143-1155.

Peakall, R., & Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics, 28*(19), 2537-2539.

Poncet, B. N., Herrmann, D., Gugerli, F., Taberlet, P., Holderegger, R., Gielly, L., . . . Manel, S. (2010). Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology, 19*(14), 2896-2907.

Pruett, C. L., & Winker, K. (2008). The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. *Journal of Avian Biology, 39*(2), 252-256.

Queller, D. C., Strassmann, J. E., & Hughes, C. R. (1993). Microsatellites and kinship. *Trends in Ecology & Evolution, 8*(8), 285-288.

Rodrigues, T. B., & Santos, J. B. d. (2006). Effect of natural selection on common bean (*Phaseolus vulgaris*) microsatellite alleles. *Genetics and Molecular Biology, 29*, 345-352.

Roschanski, A. M., Csilléry, K., Liepelt, S., Oddou-Muratorio, S., Ziegenhagen, B., Huard, F., . . . Fady, B. (2016). Evidence of divergent selection for drought and cold tolerance at landscape and local scales in *Abies alba* Mill. in the French Mediterranean Alps. *Molecular Ecology, 25*(3), 776-794.

Sartie, A. M., Matthew, C., Easton, H. S., & Faville, M. J. (2011). Phenotypic and QTL analyses of herbage production-related traits in perennial ryegrass (*Lolium perenne* L.). *Euphytica, 182*(3), 295-315.

Schuelke, M. (2000). An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology, 18*(2), 233-234.

Selkoe, K. A., & Toonen, R. J. (2006). Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters, 9*(5), 615-629.

Sunnucks, P. (2000). Efficient genetic markers for population biology. *Trends in Ecology & Evolution, 15*(5), 199-203.

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research, 17*(16), 6463-6471.

van Ham, R., O'Callaghan, M., Geurts, R., Ridgway, H. J., Ballard, R., Noble, A., . . . Wakelin, S. A. (2016). Soil moisture deficit selects for desiccation tolerant *Rhizobium leguminosarum* bv. *trifolii*. *Applied Soil Ecology, 108*, 371-380.

Wright, S. J., Cui Zhou, D., Kuhle, A., & Olsen, K. M. (2017). Continent-Wide Climatic Variation Drives Local Adaptation in North American White Clover. *Journal of Heredity, 109*(1), 78-89.

Zhang, X., Zhang, Y., Yan, R., Han, J., Fuzeng, H., Wang, J., & Cao, K. (2010). Genetic variation of white clover (*Trifolium repens* L.) collections from China detected by morphological traits, RAPD and SSR. *African Journal of Biotechnology, 9*, 3033–3041.

Zhang, Y., He, J., Zhao, P. X., Bouton, J. H., & Monteros, M. J. (2008). Genome-wide identification of microsatellites in white clover (*Trifolium repens* L.) using FIASCO and phpSSRMiner. *Plant Methods, 4*(1).

SUPPLEMENTARY TABLES

Table S5.1 Geographic co-ordinates and site details for the 18 white clover populations used in this study.

NZ Region	Site name	ID	Date Collected	Latitude	Longitude	SMD[a]	Altitude (m)	Date since resown[b]	Type of livestock	Irrigation present at site	NZSC soil order	Steepness of Slope[c]	Aspect of slope[d]
Marlborough/ Tauihu	Rai Valley	RV	11/12/18	-41.219952	173.576002	45.2	80	20+	Sheep and beef	No	Brown	Strongly rolling	SW (246º)
	Awatere Valley	AV	13/12/18	-41.667046	173.969143	125.5	185	50+	Sheep	No	Pallic	Flat	-1
Canterbury Plains/ Waitaha	Clarence	CL	12/12/18	-42.160738	173.880713	108.1	72	20+	Sheep and beef	No	Pallic	Very steep	NW (331º)
	Kaikoura	KK	12/12/18	-42.419556	173.704902	117.1	37	50+	Sheep and beef	No	Pallic	Steep	NE (54º) and N (353º)
	Waipara	WP	3/9/18	-43.0007780	172.779083	95.6	126	20+	Sheep	No	Pallic	Undulating	NE (54º)
	Waikuku	WK	3/9/18	-43.339533	172.655734	109	6	5 – 6	Calf grazing	No	Recent	Flat	-1
	Southbridge	SB	3/9/18	-43.855539	172.236023	116.3	18	45	Sheep	No	Recent	Flat	-1
Otago/ Otakou	Omarama	OM	8/9/18	-44.523397	169.808740	110.5	553	20+	Cattle	Yes	Recent	Flat	-1
	Middlemarch	MM	7/9/18	-45.542849	170.129458	127.8	193	30+	Sheep	No	Gley	Flat	-1
	Fruitlands	FL	6/9/18	-45.335434	169.306370	124.4	397	10+	Few Sheep	No	Pallic	Flat	-1
	Arrowtown	AT	6/9/18	-44.967817	168.814615	81.8	337	20 – 25+	Sheep and rabbits	No	Brown	Flat	-1
	Makarora	MR	5/9/18	-44.288453	169.201133	28.4	301	20	Few sheep	No	Recent	Undulating	W (268º)
West Coast/ Te Tai o Poutini	Haast	HA	5/9/18	-43.857637	169.039972	0	8	20 – 25+	Cattle	No	Recent	Flat	-1
	Whataroa	WR	5/9/18	-43.281808	170.383161	0	84	20+	Dairy	No	Recent	Flat	-1
	Kumara Junction	KJ	4/9/18	-42.608764	171.162131	0	61	30+	Dairy	No	Podzol	Flat	-1
	Rahu Saddle	RS	4/9/18	-42.207339	171.948646	2.6	281	20+	Cattle	No	Brown	Flat	-1
	Cape Foulwind	CF	4/9/18	-41.755667	171.527139	7.2	6	10+	Dairy	No	Gley	Flat	-1
Tasman/ Te Tai o Aorere	Lower Takaka	LT	11/12/18	-40.993595	172.817810	25.7	72	20+	Dairy	No	Brown	Flat	-1

Note: ID = site ID and NZSC = New Zealand Soil Classification.

[a] SMD = Soil moisture deficit as determined by van Ham *et al.* (2016).

[b] Date since resown relevant to date collected.

[c] Steepness of slope: Flat = 0 – 3º, Undulating = 4 – 7º, Rolling = 8 – 15º, Strongly rolling = 16 – 20º, Moderately steep = 21 – 25º, Steep = 26 – 35º, Very steep > 35º.

[d] Compass direction that a slope faces, measured in degrees (º) from North, -1 indicates no slope.

Table S5.2 Soil characteristics for 18 white clover populations located in the South Island/Te Waipounamu of New Zealand/Aotearoa.

Site name	ID	pH	CEC	VW	Olsen P	SO_4-S	Ca	ECa	Mg	EMg	K	EK	Na	ENa	CaBS	MgBS	KBS	NaBS
Rai Valley	RV	5.8	21	0.61	7	7	10	13.0	19	1.35	10	0.85	4	0.11	62	6.4	4.0	0.5
Awatere Valley	AV	6.4	19	0.71	5	4	12	13.5	30	1.87	4	0.32	6	0.17	73	10	1.8	0.9
Clarence	CL	6.7	16	0.91	8	1	14	11.9	34	1.64	10	0.59	3	0.07	76	10	3.7	0.4
Kaikoura	KK	7.8	40	0.64	7	6	30	36.3	23	1.57	13	1.06	5	0.15	92	4.0	2.7	0.4
Waipara	WP	5.9	20	0.78	10	4	13	12.6	34	1.89	7	0.47	10	0.24	63	9.4	2.3	1.2
Waikuku	WK	5.9	24	0.68	40	6	12	13.6	53	3.39	10	0.78	7	0.20	57	14	3.2	0.8
Southbridge	SB	5.7	16	0.57	10	6	5	7.1	38	2.87	10	0.92	6	0.18	45	18	5.8	1.2
Omarama	OM	5.1	12	0.84	30	21	5	4.5	12	0.64	16	0.97	2	0.05	37	5.3	8.1	0.4
Middlemarch	MM	6.7	21	0.66	12	10	12	14.5	49	3.23	10	0.80	32	0.89	68	15	3.7	4.2
Fruitlands	FL	7.2	20	0.70	52	18	12	13.4	70	4.30	21	1.56	24	0.63	65	21	7.6	3.1
Arrowtown	AT	5.6	21	0.68	3	7	11	12.4	19	1.25	2	0.19	1	0.03	59	5.9	0.92	0.2
Makarora	MR	5.3	12	0.63	10	3	4	5.2	17	1.15	4	0.35	1	0.02	44	9.7	2.9	0.2
Haast	HA	5.2	7	0.62	9	5	2	3.0	9	0.65	3	0.26	3	0.08	41	8.8	3.5	1.0
Whataroa	WR	5.7	14	0.61	30	3	7	9.1	10	0.69	3	0.27	2	0.07	63	4.8	1.9	0.5
Kumara Junction	KJ	5.2	30	0.45	23	14	8	14.1	12	1.14	4	0.49	6	0.23	47	3.8	1.6	0.8
Rahu Saddle	RS	5.9	18	0.62	25	11	10	12.9	9	0.66	5	0.40	2	0.06	71	3.6	2.2	0.3
Cape Foulwind	CF	5.6	8	0.85	18	3	5	4.7	10	0.49	2	0.12	4	0.08	58	6.0	1.5	1.0
Lower Takaka	LT	5.7	17	0.72	25	6	9	9.8	22	1.33	8	0.57	3	0.08	57	7.7	3.3	0.5
Mean		6.0	19	0.68	18	7.5	10	11.8	26.1	1.67	7.9	0.61	6.7	0.19	59.9	9.1	3.4	0.98

Note: ID = site ID; CEC = Effective Cation Exchange Capacity (cmol+/kg); VW = Volume weight (g/ml); Olsen P = Olsen Phosphorus (mg/l); SO_4-S = Sulfate Sulfur (mg/kg); Ca = Calcium MAF QT (MAF QT); ECa = Exchangeable Calcium (cmol+/kg); Mg = Magnesium MAF QT (MAF QT); EMg = Exchangeable Magnesium (cmol+/kg); K = Potassium MAF QT (MAF QT); EK = Exchangeable Potassium (cmol+/kg); Na = Sodium MAF QT (MAF QT); ENa = Exchangeable Sodium (cmol+/kg); CaBS = Calcium Base Saturation (%); MgBS = Magnesium Base Saturation (%); KBS = Potassium Base Saturation (%); NaBS = Sodium Base Saturation (%).

Table S5.3 Analysis of molecular variance (AMOVA) for different hierarchical levels of 17 white clover populations using three different formulas and based on 15,120 SNPs. **A**) all 17 populations as populations, **B**) two regions ("Dry" and "Wet"), with the 17 populations as 'population', **C**) three groups identified by cross-validation error as discrete clusters, with the 17 populations as 'population'.

Source of variation	df	SS	MS	VC	%	*p*-value
A) Seventeen populations						
Among populations	16	153976.7	9623.545	147.61	3.9	< 0.001
Within populations	657	2477783.9	3771.361	3771.36	96.1	< 0.001
B) Two geographical regions ("Dry" vs "Wet")[a]						
Among regions	1	13261.3	13261.3	11.532	0.3	0.0096
Among populations within regions	15	140715.4	9381.0	141.501	3.8	< 0.001
Within populations	657	2477784	3771.4	3771.4	95.9	< 0.001
C) Three clusters[b]						
Among clusters	2	49725.93	24862.97	78.575	2.1	< 0.001
Among populations within clusters	14	104250.8	7446.485	92.707	2.5	< 0.001
Within populations	657	2477784	3771.361	3771.361	95.4	< 0.001

Note: df = degrees of freedom, SS = Sum of squares, MS = Mean square, VC = Variance components, % = Percentage of variation.

[a] "Dry" grouping includes: Arrowtown (AT), Awatere Valley (AV), Clarence (CL), Fruitlands (FL), Kaikoura (KK), Middlemarch (MM), Omarama (OM), Southbridge (SB) and Waipara (WP). "Wet" grouping includes: Cape Foulwind (CF), Haast (HA), Kumara Junction (KJ), Lower Takaka (LT), Makarora (MR), Rahu Saddle (RS), Rai Valley (RV) and Whataroa (WR).

[b] White-coloured cluster includes: AT, FL, OM, CL and RV; Green-coloured cluster includes: AV, CF, HA, MM, MR, SB and WP; Black-coloured cluster includes: KJ, KK, LT, RS and WR.

Table S5.4 Summary of New Zealand/Aotearoa (NZ) white clover cultivars with regards to abiotic and biotic adaptations; genetic background; and locations best suitable for growing. Information collated and modified from Caradus, Hay and Woodfield (1996), Caradus *et al*. (1997), Caradus and Woodfield (1997), Woodfield *et al.* (2001), Woodfield *et al.* (2003), Widdup and Barrett (2011).

Cultivar	Leaf size class	Climate adaptation	Germplasm used to create cultivar	Land class	Stock class and management	Release date	Pest and disease tolerance
'Huia'	M	Yield and persistence under grazing	Ecotypes collected from Hawke's Bay and North Canterbury (NZ)	General	Rotational sheep	1964	
'Crau'	L	Frost tolerance and winter hardiness	Selected from a Mediterranean population from the Crau region of France	General	Rotational sheep	1966	
'Pitau'	M/L	Winter growth and seasonal production	Huia (NZ) and Spanish ecotype	Flat	Rotational cattle/dairy	1975	Leaf rust
'Prop'	S	Early and prolific flowering and yield in dry hill country	Ecotypes from summer dry hill country in Waikato, south Auckland and Coromandel (NZ)	North Island dry hill country	Set-stock sheep	1979-80	Leaf rust
'Tahora'	S	High yield and persistence in low fertility	Ecotypes from NZ moist hill country	Low fertility moist hill country, dry hill country and dry lowland	Set stock sheep	1982	
'Kopu'	L	Yield on intensive lowland dairy farms	Grasslands Pitau (NZ) and three USA ladino cultivars	Flat lowland North Island	Rotational dairy	1986	Stem nematode resistance
'Prestige'	M/S	High yield and persistence under sheep grazing	Northland ecotypes	Hill and rolling country northern North Island, and dry lowland (NZ)	Set-stock and rotational sheep	1990	Stem nematode and leaf rust
'Demand'	M/S	Early spring and summer growth and persistence in Southland	Southland ecotypes and ecotypes from NZ, Mediterranean and southern France	Hill and rolling country South Island, and moist lowland	Set-stock and rotational sheep	1990	
'Challenge'	L	Cool season growth in Northland and good spring and summer growth	NZ and Mediterranean germplasm	Flat	Rotational cattle/dairy	1994	Stem nematode and leaf rust
'Sustain'	M/L	High stolon growing point density without reducing leaf size and production and high yielding	Huia and germplasm from NZ, Mediterranean and USA	General	General	1994	
'Kopu II'	L	Persistence	World collection of white clover cultivars	Higher fertility farms	Rotational dairy/sheep	1997	Clover root weevil
'Tribute'	M/L	Drought tolerance and winter activity	Sustain, Crau, a Syrian accession, and Southern Europe II	Flat	Rotational sheep/dairy	2003	Clover root weevil, pepper spot and sclerotinia

Note: L = Large, M = Medium, S = Small.

SUPPLEMENTARY FIGURES

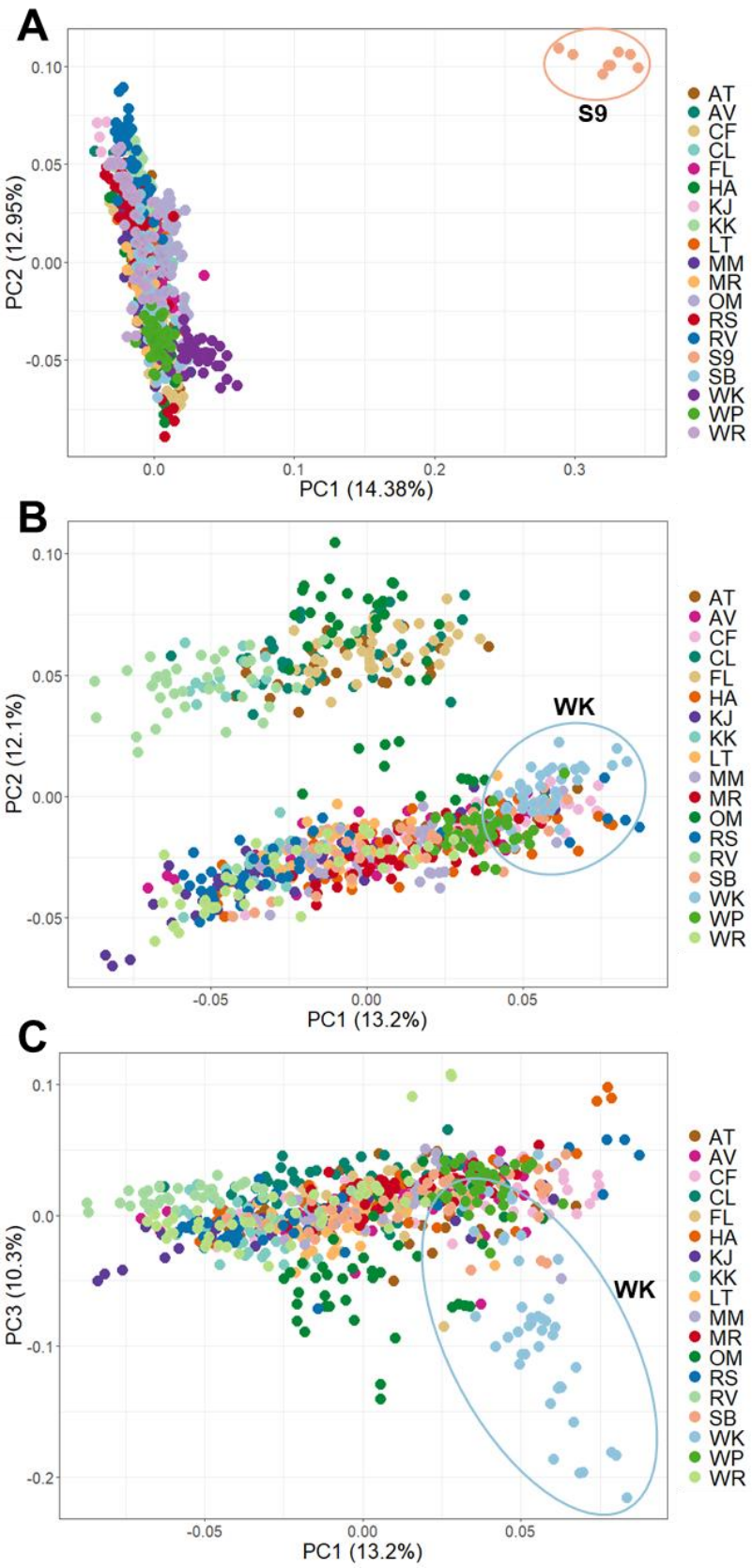


Figure S5.1 (Figure legend on next page)

Figure S5.1 PCAdapt score plots for all 18 populations. **A**) Score plot of PC1 and PC2 of 18 populations; S9 positive controls are co-located on the plot distal to all populations. **B**) As for A but with S9 positive controls and duplicated samples removed. Samples from the Waikuku population are indicated by the blue circle. **C**) As for B but PC1 and PC3. Waikuku samples are separated from the group it was associated with on the **B** plot. Population colours are identical between **B** and **C** but **A** has different population colouring. AT = Arrowtown, AV = Awatere Valley, CF = Cape Foulwind, CL = Clarence, FL = Fruitlands, HA = Haast, KJ = Kumara Junction, KK = Kaikoura, LT = Lower Takaka, MM = Middlemarch, MR = Makarora, OM = Omarama, RS = Rahu Saddle, RV = Rai Valley, SB = Southbridge, WK = Waikuku, WP = Waipara and WR = Whataroa.

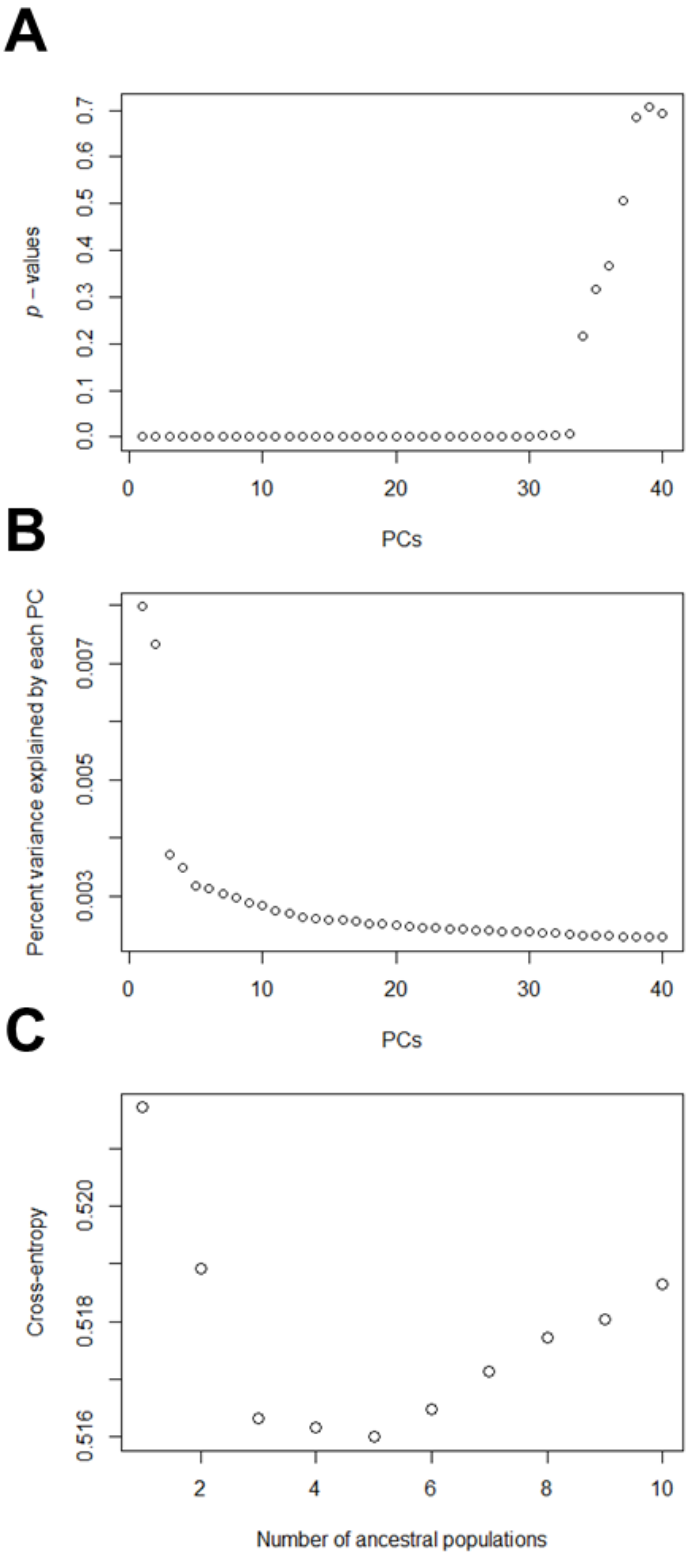


Figure S5.2 Population structure analysis performed in *LEA* to inform imputation and to determine the number of latent factors to retain in latent factor mixed model analyses. **A**) *p*-values from principal component analysis (PCA) indicating 33 principal components (PCs) best represent the data. **B**) Five PCs explain the majority of variance from the PCA analysis. **C**) sNMF cross-entropy cross validation indicate 3 – 5 putative ancestral populations.

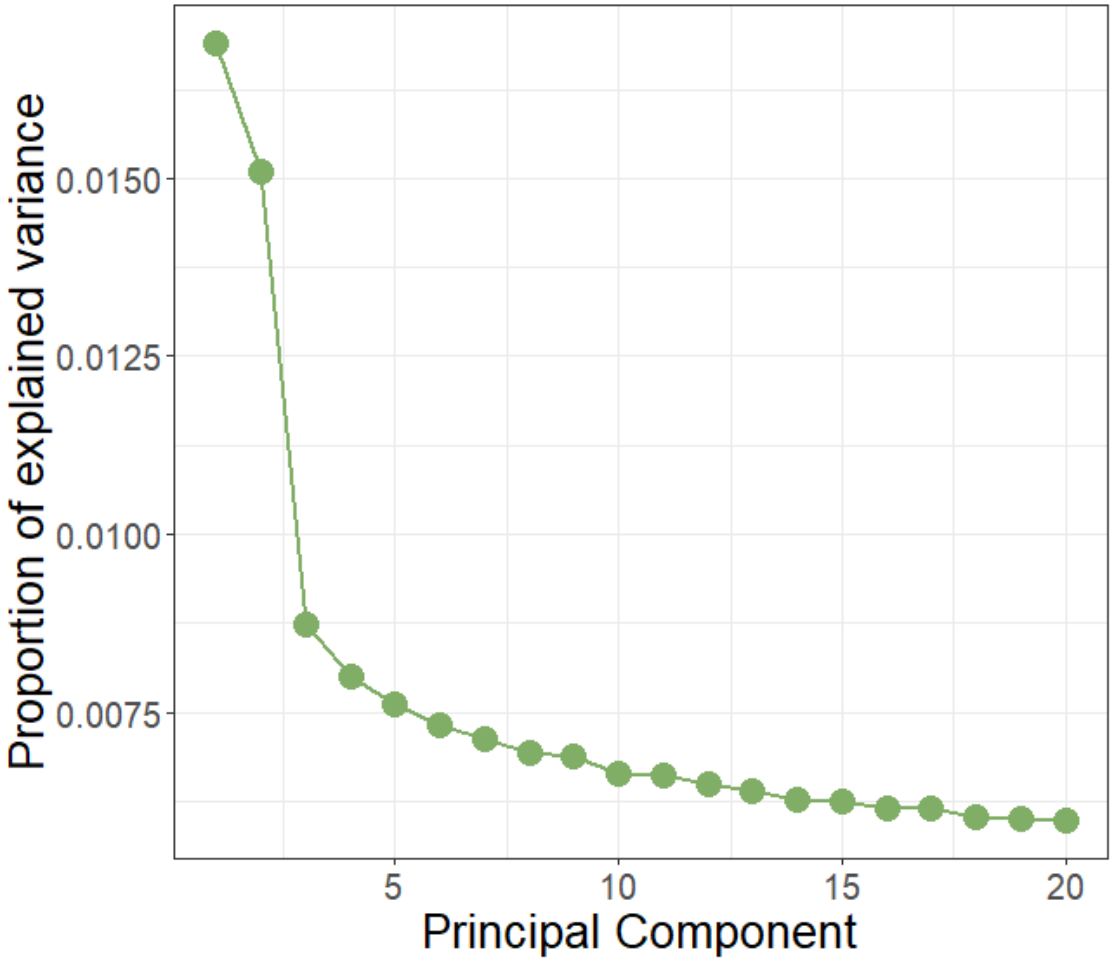


Figure S5.3 Scree plot determined in PCAdapt analysis. Proportion of explained variance is displayed on the *y*-axis with principal components from 1 to 20 displayed on the *x*-axis

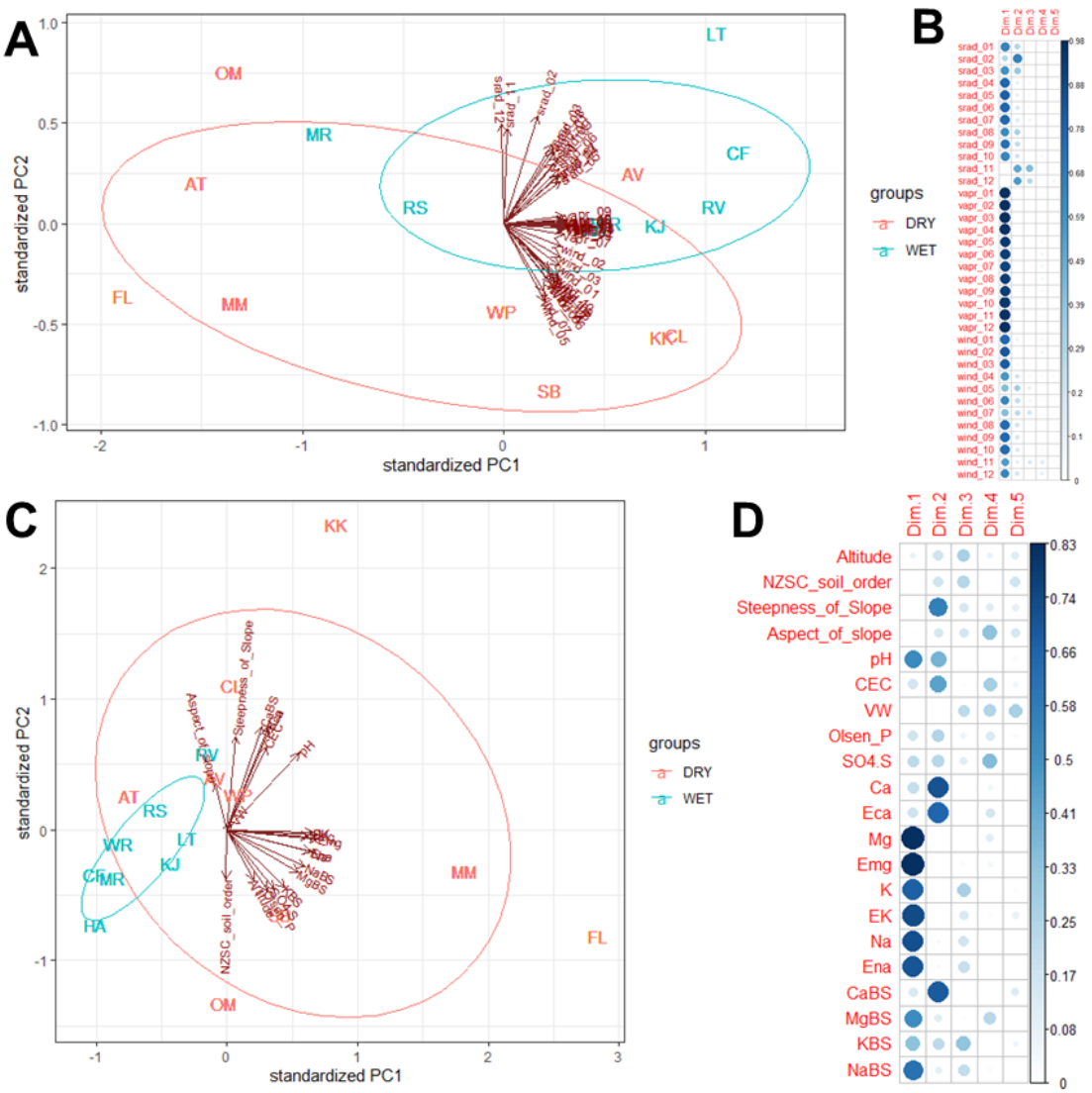


Figure S5.4 Principal component analyses (PCA) of 57 environmental variables. **A** and **C**) Biplots displaying populations grouped as either “Dry” (red) or “Wet” (blue) and the environmental variables. Percentage of explained variance for principal component (PC) 1 = 71.8% and PC2 = 15.7% for plot **A**, and percentage of explained variance for PC1 = 36.3% and PC2 = 23.3% for plot **C**. List of the 36 solar radiation (srad), water vapour (vapr) and wind speed (wind) environmental variables from each month of the year (1 = Jan – 12 = Dec) (**B**) and 21 site and soil characteristics (**D**) and their contribution to the first five PCs. The contribution of each environmental variable is displayed in a continuum where large dark blue circles indicate a greater contribution to the respective PC (1 – 5) (Dim1 – 5), while a smaller circle and pale blue indicates less contribution to the respective PCs, and no circle indicates the variable had no contribution to the respective PC. AT = Arrowtown, AV = Awatere Valley, CF = Cape Foulwind, CL = Clarence, FL = Fruitlands, HA = Haast, KJ = Kumara Junction, KK = Kaikoura, LT = Lower Takaka, MM = Middlemarch, MR = Makarora, OM = Omarama, RS = Rahu Saddle, RV = Rai Valley, SB = Southbridge, WP = Waipara and WR = Whataroa.

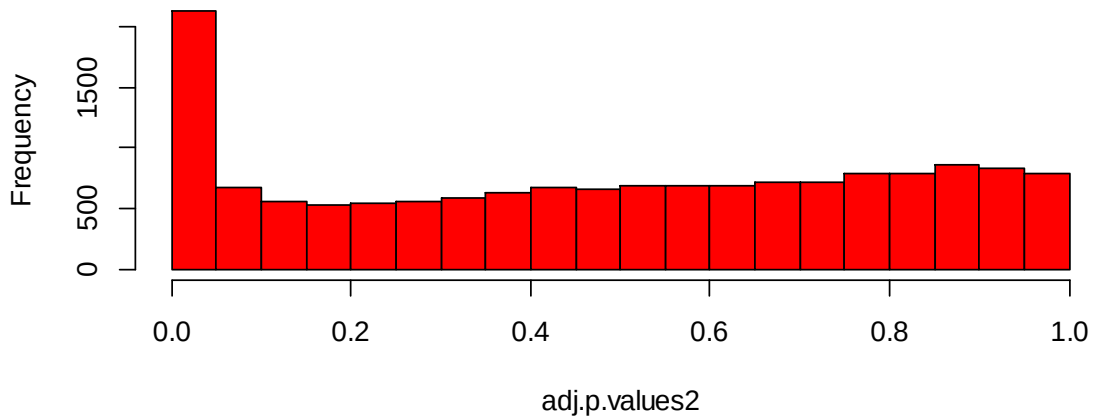


Figure S5.5 Example of the distribution of corrected *p*-values for the *LEA* latent factor mixed model with a latent factor of three for the soil moisture deficit model. The histogram displays the number of SNPs (*y*-axis) that fall into a range of *p*-values (bins) from 0 to 1 (*x*-axis). The uniform distribution and a peak of significant SNPs close to zero indicates the null hypothesis (selective neutrality for most loci) is correct. This pattern was observed for all latent factor values for all four environmental variables, so one example is presented.

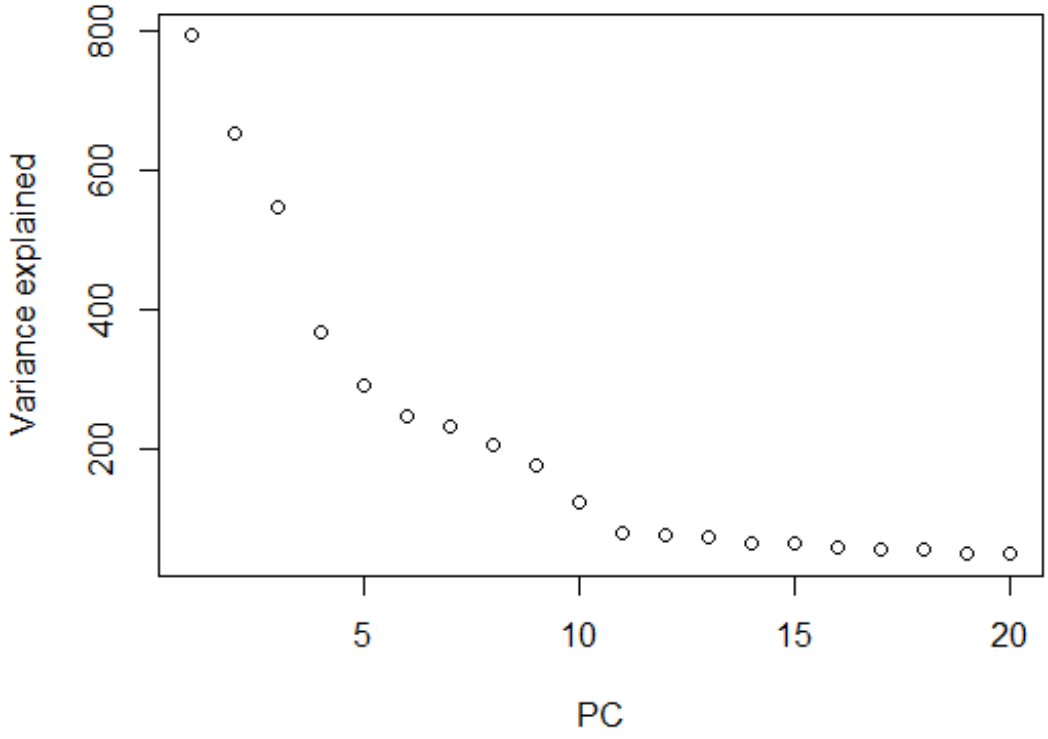


Figure S5.6 Genetic structure analysis performed in *lfmm* to determine the number of latent factors to retain for the latent factor mixed model analyses for each environmental variable. Three principal components (PC) explain the majority of variation from the principal component analysis for *lfmm*, so three latent factors were retained.

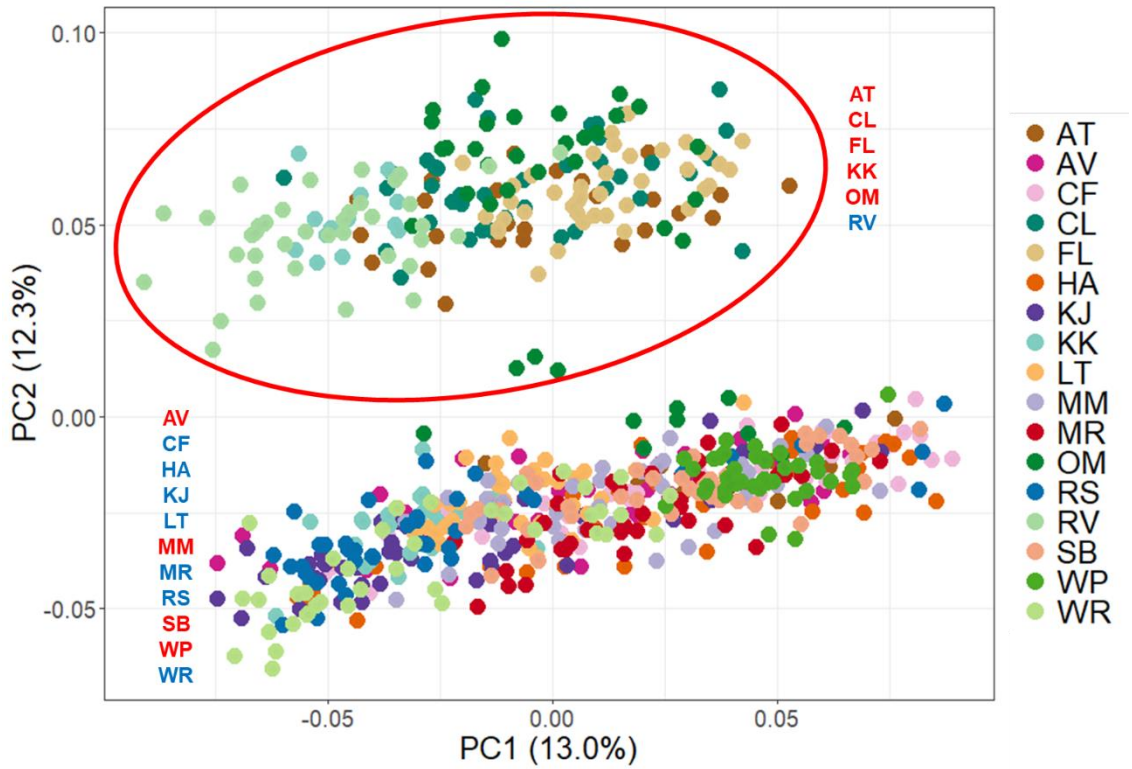


Figure S5.7 Score plot from PCAdapt analysis using the first two principal components (PC) and 17 white clover populations. Each dot represents a single individual and the colour corresponds to individuals from the same population. The red circle indicates a cluster of individuals separated by PC2; with the populations the individuals belong to presented on the right-hand side of the circle. The colour of the population names corresponds to whether the population is classified as “Dry” or “Wet” by red and blue, respectively. The remaining individuals are spatially positioned below and contain a mixture of populations from both “Dry” and “Wet”. AT = Arrowtown, AV = Awatere Valley, CF = Cape Foulwind, CL = Clarence, FL = Fruitlands, HA = Haast, KJ = Kumara Junction, KK = Kaikoura, LT = Lower Takaka, MM = Middlemarch, MR = Makarora, OM = Omarama, RS = Rahu Saddle, RV = Rai Valley, SB = Southbridge, WP = Waipara and WR = Whataroa.

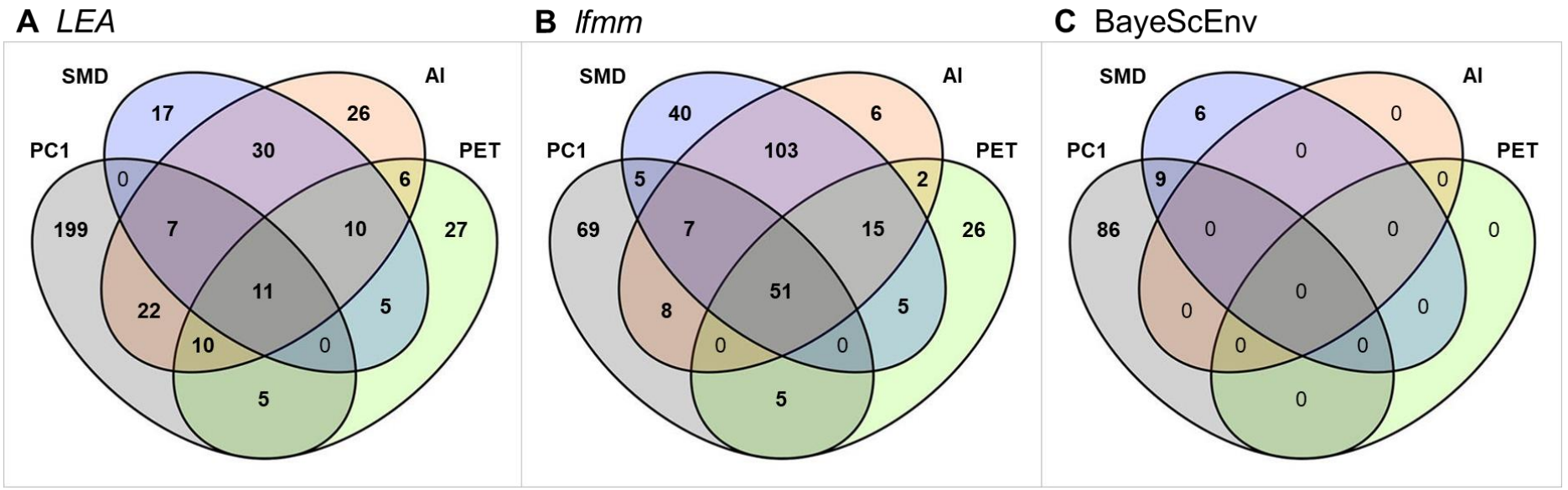


Figure S5.8 Venn diagrams of the overlap between SNP loci detected as putatively adaptive for each environmental variable using *LEA* (**A**), *lfmm* (**B**) and BayeScEnv (**C**). Loci above the false discovery (FDR) *q*-value significance threshold (α = 1e-05) for at least two of the three latent factors (K_E) values (3, 4 and 5) for each environmental variable are presented in **A**. Loci above the FDR *q*-value significance threshold (α = 0.05) for the K_E values of 3 for each environmental variable are presented in **B**. Loci above the FDR *q*-value significance threshold (α = 0.05) for each environmental variable are presented in **C**. AI = aridity index; SMD = soil moisture deficit; PC1 = PC1 co-ordinates from PCA analysis of environmental variables including SMD, AI, PET and the standard 19 bioclimatic variables; PET = potential evapotranspiration.