Genetic signatures in a perennial ryegrass (*Lolium perenne*) population following recurrent selection for compatibility with an endophyte (*Epichloë* spp.) from tall fescue

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

Master of Science

in

Plant Breeding

At Massey University, Palmerston North, New Zealand

Arnel E. Pocsedio

2019

Abstract

Perennial ryegrass or *Lolium perenne* L. (Poaceae) is the most extensively grown forage especially in the temperate regions of the world, including New Zealand. The development of forage cultivars is important to New Zealand since the livestock industry depends on perennial ryegrass for its nutrition needs. Among forage breeding objectives, persistence is particularly complex. It refers to the stability of dry matter yield over time. It is economically important because reseeding and cultivation can be capital-intensive. Persistence is partly modulated by the interaction of perennial ryegrass with *Epichloë* spp*.* as these fungal endophytes confer insect resistance for a more stable yield. Genetic factors in the host influence fungal biomass, alkaloid concentration, and endophyte vertical transmission frequency. The symbiotic relationship is therefore exploited in perennial ryegrass breeding. Thus, the objective of this study is to investigate a perennial ryegrass breeding population under recurrent selection (RS) for compatibility with an endophyte sourced from tall fescue. Specifically, this study aims to (1) investigate the transmission of the *Epichloë* sp. FaTG-3 strain AR501 in the breeding population PGG04, and (2) to examine how genetic variation changes during RS in terms of population differentiation. Since the selection program targets endophyte compatibility, signatures of selection that may be associated with the grass-endophyte interaction were also determined. It was hypothesized that: there will be a reduction of diversity, and an excess of rare alleles. Furthermore, it was hypothesized that loci under positive selection will have higher fixation index (F_{ST}), and their genotypes will be more correlated with the components of PCA-based population structure analysis compared to neutral loci.

The presence of AR501 was examined in seeds, in the growing tillers, and by microsatellite genotyping for both the early and late generations of PGG04. The seed squash assay revealed that more than 90% of PGG04 seeds harboured the endophyte, regardless of the generation. Viable endophyte detection using tissue-print immunoblotting showed an increase in infection from ca. 5% to 33% between the early and late generations. Thus, the results suggest that positive selection for endophyte compatibility increased the proportion of viable endophyte in the population. This study provides evidence supporting host genetic control of the association in grass-endophyte interaction, and that this can be exploited in plant breeding programs.

Changes in the genetic variation of PGG04 was investigated by comparing GBS data of the early and late generations. Results showed that selection enriched the late generation with rare alleles (0.02 - 0.08) compared with the early generation. Also, selection reduced expected heterozygosity from 0.3069 in PGG04-C2 to 0.3033 in PGG04-C6. Further, selection changed the population structure based on UPGMA dendrogram, PCA, and the model-based clustering method implemented in STRUCTURE. A few single nucleotide polymorphisms (SNPs) have relatively larger contribution to the population structure changes hence, they have relatively high F_{ST}, and their genotypes correlated with principal components. Logistic regression of these SNPs with infection data identified nine SNPs to be associated with the trait. Depending on the allele frequency, these SNPs can increase the odds of favourable infection by more than five times. Annotation of these SNPs identified S7_160751877 to be tagging an ABCG transporter gene. Since some ABC transporters mediate plant-microbe interactions, it is possible that the identified SNPs are tagging a gene involved in the host genetic control of grass-endophyte interaction.

Keywords: *Lolium perenne*, *Epichloë* spp., perennial ryegrass-endophyte interaction, transmission, genotyping-by-sequencing, selection signature, recurrent selection, F_{ST}, PCA, ABCG transporters

Acknowledgements

This work will not be possible without the help of several people and institutions.

First, I thank my supervisors: Vaughan Symonds, for all his help in my studies and for his guidance in my research work; and Marty Faville, for allowing me to work in AgResearch, facilitating the materials and equipment that I need, and connecting me to the right people.

Also, I am truly thankful to the people of LoST lab: to Prashant who helped me in my lab work at Massey; to Sofie who shared her knowledge on topics that we both covered; to Jen; and to the rest of the people in the lab for the learning sessions in and outside our regular journal club.

My appreciation to the SFS team, especially to Ann Tuter for being approachable about SFS matters; and to Matthieu Vignes for accommodating me every time I have questions about data analysis.

My warmest gratitude to the folks of Agresearch Grasslands: to Won who taught me GBS library preparation apart from the extensive assistance he has given me, from sowing seeds, to leaf collection, and to DNA extraction; to Craig and Anna who assisted me in the lab; to Anouck who taught me the seed squash assay and verified my scoring; to Dr. Stuart Card who helped me photograph my seed cross sections; to Andrew and Stephen who helped me grow my plants in the glasshouse; and to Mingshu Cao for conducting SNP calling and annotation of my GBS data.

I am also very grateful to the PGG Wrightson Seeds Ltd., especially Keith Salisbury, who developed the plant materials in my experiment, and for allowing me to work in your materials.

To the Filipino community in Palmerston North, especially to Kapatiran Prayer Fellowship and to Kuya Ricky and family, thank you for providing a home away from home.

To my batchmates, Kuya Allan and his family, Carla, and Cynthia, I am grateful for all your support and for being our extended family. To FAMUS friends – Gian, Tinay, Norma, Adarlo sisters, Dacillo Siblings, Tita Baby and family, Marc, Spencer, Dominic, Fem, Ayaquil couple, Tracey, Reez and Jean – thank you for making home sickness coupled with study anxiety bearable.

To ISSO, especially to Sylvia, Jamie, Logan and Saba, thank you for the constant support for me and my family's welfare. To NZAid and other international students who became my friends, thank you for taking me around the world by experiencing your cultures and personalities.

To MFAT and to the people of New Zealand for supporting my studies, I will be forever indebted to all of you.

To my family in the Philippines – Mama Paz, Ate Maan, Ate Joy, and Kuya Archie – thank you for being my inspiration and role models. To all my family and friends in the Philippines, thank you also.

To my loving wife Pau who has shared all the burdens and challenges in my academic journey, thank you and I love you. To my kids, Zach and Missy, whose smiles and laughter take my stress away, I love you.

To all the people I have not mentioned but helped in any way, I also express my gratitude to all of you.

Lastly, I would like to thank my Creator, without Whom I will be nothing. I hope that this work will give glory to my saviour, Lord Jesus Christ, who has been my rock during the tough times.

Abbreviations

ABC – ATP-binding cassette

AMOVA – analysis of molecular variance

AP – ancestral population (hypothetical)

bp – base pairs

BLAST – basic local alignment search tool

df – degrees of freedom

DR – defence related (i.e. genes)

FaTG – *Festuca arundinacea* taxonomic group

F_{IS} – inbreeding coefficient

F_{ST} – fixation index

GBS – genotyping-by-sequencing

H_e – expected heterozygosity

H_o – observed heterozygosity

LD – linkage disequilibrium

LpTG – *Lolium perenne* taxonomic group

MAF – minor allele frequency

MAS – marker-assisted selection

MM – mycelial mass

NZ – New Zealand

PCA – principal component analysis

PCoA – principal coordinate analysis

PCR – polymerase chain reaction

QTL – quantitative trait loci

RS – recurrent selection

SM – selection mapping

SMC – seed moisture content

SNP – single nucleotide polymorphism

SS – seed squash assay

SSR – simple sequence repeats

TPIB – tissue-print immunoblotting

UPGMA – unweighted pair group method

Table of contents

List of figures

List of Tables

Chapter 1

1. Introduction

Perennial ryegrass or *Lolium perenne* L. (Poaceae) is among the many plant species that are valuable to pastoral agriculture, especially in temperate regions of Europe, New Zealand, Australia, Japan, South Africa, South America, and the United States (Humphreys et al., 2010; Thorogood, 2003). It is the most extensively grown forage in the temperate regions of the world and is favoured because of its digestibility and ability to maintain herbage yield under grazing (Humphreys et al., 2010; Wilkins, 1991). In Europe, the importance of perennial ryegrass is reflected in its seed industry. For example, about 50% of grass seed planted in the European Union is *L. perenne*. Further, in the United Kingdom, it accounts for 75% of seeds marketed (Humphreys et al., 2010). In New Zealand, pastoral agriculture depends heavily on the *Lolium* spp., especially perennial ryegrass, as they provide about 75% of the nutrition needs of livestock industries. Together, the ryegrass species, are estimated to contribute $14 billion annually to the country's economy (Nixon, 2015). The development of ryegrass cultivars that meet the needs of animal growers is important to New Zealand. The primary objectives considered in perennial ryegrass breeding include total and seasonal herbage yield, forage quality, and persistence. Other breeding objectives include tolerance to abiotic stresses, pest and disease resistance, and seed production traits (Humphreys et al., 2010; Stewart & Hayes, 2011; Wilkins, 1991). Among these traits, persistence is particularly complex. It refers to the stability of dry matter yield over time (Parsons et al., 2011) and is economically important because reseeding and cultivation can be capital-intensive. Persistence to some extent is modulated by the interaction of ryegrass with *Epichloë* spp*.* fungal endophytes. Endophytes confer insect resistance resulting in improved plant survival and yield stability (L. Johnson et al., 2013). Perennial ryegrass plants in New Zealand are naturally infected with a single endophyte strain (Simpson et al., 2012), referred to as common toxic endophyte, that causes health issues in grazing ruminants. Thus, research and discovery of novel endophytes are important to the country to provide fungal strains for forage protection from insect pests but with limited or no detrimental effects to livestock. Commercial endophytes, such as EndoSafe, AR1 and AR37, are estimated to contribute around $200 million annually to the country's economy (L. Johnson et al., 2013).

1.1. Genetic diversity in perennial ryegrass

Lolium perenne is believed to have originated from the Middle East and from there, expanded throughout Europe, consistent with the spread of agriculture (Balfourier et al., 2000). It then dispersed outside Europe together with human migrations. For example, British immigrants brought with them seeds of perennial ryegrass to New Zealand (Stewart, 2006). Perennial ryegrass populations, both natural and synthetic, have highly variable genetic composition because of its inability to self-pollinate. Genetic diversity is higher within the population than among different populations. For example, Barth et al. (2015) reported ~88% of the variation is found within cultivars and ~90% within ecotypes versus among populations using SSR markers. This is in agreement with previous reports using SSR (Brazauskas et al., 2011; Kubik et al., 2001), RAPD (Bolaric et al., 2005), AFLP (Ghesquiere et al., 2003; Guthridge et al., 2001) and recently with SNP markers (Blackmore et al., 2015; Blackmore et al., 2016) in studies involving cultivars and natural populations. Further, Blackmore et al. (2015) investigated *L. perenne* ecotypes from various regions in Europe and found that the majority of variability (68%) is due to differences within accessions as opposed to differences among regions (8%) and accessions within regions (24%). In breeding populations and cultivars, the within-population variability is influenced by the nature of its founding parents. As such, variability increases with an increasing number of founders. Populations derived from a non-restricted base, that is, from several parents, have higher diversity than those from restricted base (Auzanneau et al., 2007). Aside from the number of founders, relatedness amongst the founders is also important since related individuals have a low chance of contributing novel alleles to increase diversity. In perennial ryegrass breeding, the crossing of closely related plants is not generally favoured because it may lead rapidly to inbreeding depression. Inbreeding can lead to homozygosity of deleterious recessive alleles resulting to poor performance (Acquaah, 2012). More importantly, genetic diversity has been shown to directly affect performance (Ghesquiere et al., 2013; KöLliker et al., 2005). KöLliker et al. (2005) used AFLP markers to examine the genetic diversity of a parental germplasm and based on this, constructed narrow and wide diversity polycrosses. Synthetic populations (Syn1 and Syn2 generations) derived from wide polycrosses were shown to have higher dry matter yield compared to populations based on narrow polycrosses. This highlights the value of maintaining diversity in a perennial ryegrass breeding program or other cross-pollinated crops in general.

In general, the investigation of population structure has always been part of genetic diversity studies. The structure of genetic diversity generally reveals adaptation and/or relatedness. For example, it was shown that genetic diversity correlates very well with the geographic origin amongst European ecotypes (Blackmore et al., 2015; Blackmore et al., 2016). Principal component analysis of allele frequencies divided the *L. perenne* accessions corresponding to East and West Europe regions (PC1) as well as separating the UK and Iberian accessions (PC2) (Blackmore et al., 2015). This is consistent with a study of advanced breeding germplasm where genetic structure also corresponded to geographic origin, that is from the UK or continental Europe (Brazauskas et al., 2011). Population structure can also reflect breeding, such as the separation of forage and amenity cultivars in Blackmore et al. (2016). However, cultivar populations are not generally expected to have a structure within them, despite being characterised by high within-population diversity. This is because they are derived from a series of panmictic bulking generations (Auzanneau et al., 2007) and departure from panmixis or random mating is what causes stratification in the population.

Germplasm collections and breeding populations are routinely characterized both in terms of morphological (Bothe et al., 2016) and genotypic variation (as cited in the examples above). Diverse germplasm collections are not only sources of breeding materials but can also be used as a platform to study marker-trait associations to identify useful genomic regions for breeding, as well as to understand the genetic basis of traits of interest. LD or the non-random association of alleles between two or more loci is exploited in association mapping (or LD mapping); that is, the molecular markers (i.e. SNPs) are in linkage disequilibrium (LD) with the causal variant (i.e. gene of interest). LD in perennial ryegrass is expected to be low because of its outbreeding nature (Blackmore et al., 2016). However, LD is affected by several factors such as selection, demographic, and genomic history. It is therefore dependent on the nature of the population in which it is measured. Skøt et al. (2005) found three AFLP markers related to heading date through LD mapping. These markers are also linked to a major quantitative trait locus (QTL) explaining about 70% of the variation in heading date (Armstead et al., 2004), a contribution expected for a highly heritable trait. Similarly, LD mapping has been conducted to study complex traits such as submergence tolerance of perennial ryegrass (Yu et al., 2011). The study identified 15 SSR markers tagging 3 QTLs for correlated traits, namely, reduction in leaf colour, chlorophyll fluorescence, maximum plant height, and relative growth rate, under submergence stress. These association studies utilized high diversity germplasm collections. This type of mapping population is generally structured and thus can lead to spurious association results if the structure is not accounted for in the analysis. The heading date study identified the presence of hybrid ryegrass accessions driving population structure and consequently eliminated them from the final association analysis, although doing this had little impact on the overall results (Skøt et al., 2005). Another way of dealing with this is to include population structure (i.e. Q) in the model used in association analysis, such as what was done by Yu et al. (2011). Synthetic cultivars, which typically do not exhibit population structure, can also be used for LD mapping (Auzanneau et al., 2007; Brazauskas et al., 2011). LD decays across genetic distance and this rate increases with genetic variability in the population (Auzanneau et al., 2007). Blackmore et al. (2016) reported that LD breakdown is faster in perennial ryegrass ecotypes than synthetic cultivars. However, LD breaks more rapidly in cultivars derived from a non-restricted base, but its diversity can still be exploited using a candidate gene approach of association mapping. On the other hand, cultivars derived from fewer parents would have slower LD decay, and would therefore allow for genome-wide association studies.

1.2. Genotyping perennial ryegrass populations

Most of the genetic diversity studies in perennial ryegrass, such as those mentioned above, have utilized relatively low-density marker systems such as RAPDs, AFLPs and SSRs, with the exception of Blackmore et al. (2015) and Blackmore et al. (2016). The latter studies used SNP arrays of more than two thousand markers. SNPs are more frequent in the genome than other DNA polymorphisms and are under a simple mutation model, making them useful for a wide range of applications. Investigation of genetic variation is moving towards high-density genome-wide approaches, making SNP the marker system of choice. Yet, the only publicly-available SNP array for perennial ryegrass is relatively small (Blackmore et al., 2015) compared with the wealth of genomic resources available for major cereal species, such as maize with a 600K SNP array (Unterseer et al., 2014); wheat with 820K (Winfield et al., 2016); and rice with 1M (McCouch et al., 2010). However, a major criticism of array-based SNPs is that of ascertainment bias. Traditionally, SNP sets for arrays are discovered based on a particular panel of individuals (i.e. discovery panel), but the subsequent application may be in a large range of populations with varying degrees of relatedness to the original discovery panel. This creates a systematic bias and dependency on the discovery panel used (Frascaroli et al., 2013).

Genotyping-by-sequencing (GBS) has been proposed as a low cost, high-density SNP genotyping system for high diversity species, and was first demonstrated in maize and barley (Elshire et al., 2011). In GBS, SNP marker discovery and genotyping are accomplished at the same time and in the same population, addressing the ascertainment bias issue of SNP arrays. It is also capable of generating tens to hundreds of thousands of markers and is applicable to a wide range of species (Elshire et al., 2011; He et al., 2014). More importantly, it has been utilized in perennial ryegrass in applications such as genomic selection (Faville et al., 2016; Fè et al., 2015), genome-wide association study (Fè et al., 2015), trait prediction (Byrne et al., 2017), and genetic map construction (Velmurugan et al., 2016). In addition, it was used to measure genome-wide allele frequency fingerprints which can be used to discriminate between perennial ryegrass cultivar populations (Byrne et al., 2013). To date, the use of GBS data in genetic diversity studies of perennial ryegrass is still limited. It is therefore of interest to investigate genetic diversity, population structure, and linkage disequilibrium in perennial ryegrass populations using the more powerful GBS technology.

GBS relies on genome complexity reduction through the use of restriction enzymes (RE) to divide the large genome into smaller DNA fragments. The choice of RE is therefore of crucial importance. The first GBS report utilizes ApekI (Elshire et al., 2011) and was subsequently used in perennial ryegrass (Byrne et al., 2013). Byrne et al. (2013) also tested three other enzymes, namely AgeI, EcoRI, and PstI and found that aside from ApekI, PstI can also be suitable for GBS. In addition to the Byrne et al. (2013) study, a literature search (scanning 150 articles returned by Google Scholar using keywords: GBS and perennial ryegrass) on RE used in GBS in perennial ryegrass returned four more studies utilizing ApeKI (Arojju et al., 2016; Byrne et al., 2017; Faville et al., 2018; Fè et al., 2016), and three more studies utilizing PstI (Faville et al., 2016; Rabier et al., 2016; Velmurugan et al., 2016). Thus, both enzymes have proven to be amenable for GBS in perennial ryegrass, with ApeKI being the more common choice. However, Byrne et al. (2013) showed that better sequencing depth with a relatively small number of loci can be achieved with PstI at lower sequencing budget (i.e. less than 10 million reads), although ApeKI can generate more loci as sequencing budget increases. Further, both enzymes were used in GBS for perennial ryegrass genomic selection studies in New Zealand (Faville et al., 2018; Faville et al., 2016). The two-enzyme system developed by Poland et al. (2012), is another interesting protocol. It combines the use of rare-cutting enzyme PstI (CTGCAG) and MspI, which has relatively more common recognition site (CCGG). It also uses a common Y-adapter that leads to amplification of a set of uniform fragments. This system is being explored at AgResearch as it can provide better sequencing depth than using ApeKI and generate a larger set of markers than using PstI alone in perennial ryegrass GBS (Dr. Mingshu Cao, unpublished data, 2018). In the end, the choice of restriction enzyme is dependent on the goals of the experiment. For example, genomewide association studies, especially with low LD populations, would require very dense marker coverage. On the other hand, a few thousands SNP may be sufficient for characterizing genetic diversity.

1.3. Endophytic *Epichloë* species of perennial ryegrass and tall fescue

The importance of endophytes in agriculture has been recognized largely due to their detrimental effects on herbivores. It has been known that *Epichloë festucae* var. *lolii*

(Leuchtmann et al., 2014) (previously *Neotyphodium lolii*) causes a nervous disorder known as ryegrass staggers, as well as heat stress in sheep or cattle feeding on endophyte-infected perennial ryegrass (Fletcher & Harvey, 1981). On the other hand, it has been known also to provide resistance to insect pests, such as Argentine stem weevil (Prestidge et al., 1982). In tall fescue, *Epichloë coenophialum* (Leuchtmann et al., 2014) was found to cause summer fescue toxicosis, when cattle feeding on the endophyte-infected cultivar Kentucky 31 showed symptoms of toxicity (Schmidt et al., 1982). Similarly, this endophyte species is known to confer insect resistance such as deterrence to aphid feeding (M. Johnson et al., 1985). Later, it was discovered that these reactions were underpinned by alkaloids produced *in planta* by the endophytes. For example, lolitrem B (an indole-diterpene) is the main neurotoxin responsible for ryegrass staggers (Gallagher et al., 1981), while a different alkaloid, called peramine, acts as a feeding deterrent to Argentine stem weevil (Rowan & Gaynor, 1986). Ergovaline (an ergot alkaloid), the compound causing fescue toxicosis (Lyons et al., 1986) and heat stress, is also present in endophyte-infected perennial ryegrass, and similarly confers insect resistance to pests including African black beetle (Ball et al., 1997). Another class of alkaloids called lolines have a wide range of insecticidal properties and are generally safe for livestock (Bush et al., 1993; Schardl et al., 2007); however, these are present only in endophytes derived from fescue species. In New Zealand, the prospect of endophyte-free perennial ryegrass cultivars was forwarded to solve animal health issues, but it was discovered early on that this has negative impacts on persistence (Prestidge et al., 1982). The discovery of a large diversity of endophytes in European and Northern African collections of ryegrasses and fescues, including bioactive alkaloids, stimulated the search for endophyte strains conferring insect resistance (i.e. with peramine) and that were safe for ruminants (i.e. without or with only low levels of lolitrem B and/or ergovaline) (L. Johnson et al., 2013). Combined with the development of a methodology to transfer these novel endophytes into cultivars, this eventually led to the commercialization of novel endophytes including Endosafe, AR1 and AR37, and MaxQ and MaxP (for tall fescue) (L. Johnson et al., 2013). Easton et al. (2001) presented a comprehensive review of the history and progress of endophyte research in New Zealand and more recently L. Johnson et al. (2013) expanded this review.

Early isozyme analysis revealed separate taxonomic groupings (TG) in endophytes forming symbioses with tall fescue (i.e. *E. coenophiala*, *Festuca arundinacea* (Fa) TG-2 and FaTG-3) meadow fescue (i.e. *E. uncinata*) and perennial ryegrass (i.e. *E. festucae* var. *lolii* and *Lolium perenne* (Lp) LpTG-2) (Christensen et al., 1993). In addition, the isozyme phenotypes correspond well to alkaloid profiles. A more recent study on a global collection of perennial ryegrass endophytes reported similar findings (van Zijll de Jong et al., 2008). For example, “ruminant-safe” endotypes form a separate group from common-toxic endophytes based on

SSR marker data. The study of van Zijll de Jong et al. (2008) similarly found the distinct group of LpTG-2. However, they discovered three possible distinct groups within *E. festucae* var. *lolii* (i.e. A, B and C). These groups are characterized, roughly, by their alkaloid profiles, that is, moderate to high levels of lolitrem B and peramine (i.e. group A); moderate to high levels of lolitrem B, ergovaline and peramine (B); and moderate to high levels of ergovaline and peramine (C). Further, this grouping also corresponds to their geographic origin and correlates well with the chloroplast SNP genotype of host plant populations. Alkaloid diversity profiles of different *Epichloë* species and strains can be explained by the underlying genetic diversity. Alkaloid biosynthesis is associated with single loci such as LOL for lolines, LTM for indole-diterpenes, EAS for ergot alkaloids and PER for peramines (Schardl et al., 2012). Except for the peramine gene called *perA,* these alkaloid loci are found in gene clusters. Variation in alkaloid profiles can be explained by the presence/absence of genes in the cluster or variation in the genes that are present although, in some cases, gene expression can be silenced (Saikkonen et al., 2016). Schardl et al. (2012) reviewed the chemotypic diversity of endophytes and its underlying genetics. Similarly, L. Johnson et al. (2013) reviewed *Epichloë* alkaloid biosynthesis and genetics in relation to agriculturally important forage species. Further, Saikkonen et al. (2016) presented a succinct summary on the evolution of endophytic *Epichloë* species and their genetic diversity influencing phenotype and ecology.

1.4. Novel and commercial endophytes of forage grasses

Synthetic associations were the basis of the exploitation of endophytes in grass population improvement programs, albeit within the natural associations. For example, the commercial endophytes of perennial ryegrass cultivars, namely, Endosafe, AR1, and Endo5 are strains of *E. festucae* var. *lolii*. Similarly, the endophyte of tall fescue cultivars, AR542 (MaxQ or MaxP) is an *E. coenophiala* strain (L. Johnson et al., 2013).

The benefits of commercial endophytes can only be realized by end users when the endophytic seed product is of high quality. In New Zealand, quality control measures are in place to ensure that: the seed is infected with correct endophyte strain with no contamination; there is a high percentage of viable endophytes; and the beneficial alkaloids are present (Rolston & Agee, 2007). Therefore, endophyte detection techniques are important. This includes molecular techniques such as SSR-based DNA fingerprinting for strain identification and enzyme-linked immunosorbent assay (ELISA) for alkaloid detection (Rolston & Agee, 2007). In determining endophyte infection rate one common technique in New Zealand is the seed squash assay especially for newly harvested seeds (Card et al., 2011). The seed squash method involves direct observation of the endophyte through histological staining and microscopy (Latch & Vaughan, 1995). However, endophytes in the seeds could either be dead or alive which cannot be distinguished in the seed squash method. Therefore, seed industry practices also employ the tissue print immunoblotting (TPIB) to detect viable endophytes (Card et al., 2011). This is an indirect method that uses antibodies that bind with endophyte proteins for detection (Hahn et al., 2003; Simpson et al., 2012). TPIB technique itself does not discriminate viable endophyte. However, it is conducted in growing seedlings and thus, only live endophytes growing in synchrony with the plant is detected.

Novel association of endophytes with forage grasses have been commercialized (Table 1.1). As described above, common toxic endophytes have been replaced with strains that offer insect resistance with little or no mammalian toxicity. In general, this has been limited to the naturally forming associations such as perennial ryegrass with *E. festucae* var. *lolii* and tall fescue with *E. coenophiala.* Nevertheless, it also involves the use of endophytes from the uncommon taxonomic groups, although still within the natural association, such as *Lolium perenne* Taxonomic group 3 (LpTG-3) in perennial ryegrass. Cross-species associations have been investigated but have not been fully exploited as compared to the current commercial endophytes (Table 1.1). One of the reasons for this is the problem of host-endophyte incompatibility.

Table 1.1. Commercial and wild-type endophyte strains and some of their key properties. Adapted from L. Johnson et al. (2013).

Commercial or common name	Fungal species	Notable alkaloids produced	Key traits	Key regions of use
Common-toxic (wild-type)	*E. festucae* var. *lolii*	Lolitrems Peramine Ergovaline	Ryegrass staggers; negative impacts on animal health. Good ASW & black beetle resistance	Ryegrass pastures and turf New Zealand Australia and South America
Common-toxic (wild-type)	*E. coenophiala*	Peramine Ergovaline Lolines	Fescue toxicosis. Broad spectrum insect resistance	Tall fescue pastures and turf USA
Common-type (wild-type)	*E. uncinate*	Lolines	Broad spectrum insect resistance	Meadow fescue pastures USA Europe

Table 1.1. continued.

Commercial or common name	**Fungal species**	**Notable alkaloids produced**	**Key traits**	**Key regions of use**
Endosafe	*E. festucae* var. *lolii*	Peramine Ergovaline Peramine	No ryegrass staggers. Good ASW resistance Broad spectrum insect resistance	Ryegrass pastures New Zealand
MaxQ	*E. coenophiala* strain AR542 and AR584 (MaxQII)	Lolines	No fescue toxicosis.	Tall fescue pastures USA
MaxP	*E. coenophiala* strain AR542 and AR584	Lolines Peramine	No fescue toxicosis. Broad spectrum insect resistance	Tall fescue pastures New Zealand Australia
AR1	*E. festucae* var. *lolii*	Peramine	No ryegrass staggers and good ASW resistance	Ryegrass pastures New Zealand Australia and South America
Endo5	*E. festucae* var. *lolii*	Peramine Ergovaline Peramine Ergovaline	Good ASW and black beetle resistance. No ryegrass staggers	Ryegrass pastures Australia

Table 1.1. continued.

Commercial or common name	Fungal species	Notable alkaloids produced	Key traits	Key regions of use
NEA2	Mix of *E. festucae* var. *lolii* strains	Lolitrems	Good black beetle resistance	Ryegrass pastures New Zealand Australia
AR37	*E. festucae* var. *lolii* (LpTG-3)	Epoxy-janthitrems	Broad spectrum insect pest resistance; Excellent animal performance but some ryegrass staggers	Ryegrass pastures New Zealand Australia
Avanex	*E. coenophiala* strain AR601	Ergovaline Lolines	Bird and wildlife deterrent	Tall fescue pastures Airports
Avanex	*E. festucae* var. *lolii* strains AR94/95	Peramine Ergovaline Lolitrem B (only for AR95)	Bird and wildlife deterrent	Ryegrass Sport fields, recreational parks

1.5. Grass-endophyte interaction as influenced by the host's genetics

The genetic control of the grass-endophyte association has been studied in the host, although not as exhaustively as the contribution of fungal genetics. The genetic background of the host is known to influence the ability of the endophyte to produce alkaloids (Adcock et al., 1997; Easton et al., 2002). This was exploited in tall fescue breeding by selecting against the concentration of toxic ergot alkaloid in endophyte-infected populations. This trait was found to be highly heritable and after two cycles of selection, an 86% reduction in alkaloid concentration was achieved (Adcock et al., 1997). In infected perennial ryegrass populations, concentrations of ergovaline and peramine, as well as mycelial mass, were also found to be highly heritable (Easton et al., 2002). These traits were found to be highly correlated and alkaloid concentration was at least partly a function of mycelial mass (MM). This relationship is stronger with peramine than ergovaline (Faville et al., 2015). Easton et al. (2002) found some half-sib families with a mycelial mass close to the mean and therefore have moderate peramine (ruminant-safe) level, but low ergovaline (ruminant-toxic) level. This suggests that it is possible to select for perennial ryegrass plants with a moderate level of the beneficial alkaloid and low level of the toxic alkaloid (Easton et al., 2002). Faville et al. (2015) reported QTLs for peramine concentration that were mostly co-located with mycelial mass (MM) QTLs, while for ergovaline concentration, MM-independent loci were also detected. It therefore confirmed that bi-directional selection for the concentration of the two alkaloids could be achieved. Furthermore, QTL detection suggested that the host genetic control of alkaloid concentration and mycelial mass of fungal endophytes are controlled predominantly by additive genetic effects. Diallel analysis of Easton et al. (2002) also supported this hypothesis. Their results showed a large contribution of general combining ability effects (i.e. mostly additive) and smaller specific combining ability effects (i.e. non-additive gene actions). In addition, even though the endophyte is transmitted via the female parent, no significant maternal effects were detected (Easton et al., 2002).

In Faville et al. (2015), QTLs were detected in two populations infected with the common endophyte and AR501, respectively. Some QTL co-aligned in the two populations (Fig. 1.1). The detection of common QTLs across genetic backgrounds (i.e. two mapping populations) and across endophyte strains, with multi-year data, strongly suggests that there may be conserved genes in the host that regulate grass interaction with fungal endophytes (Faville et al., 2015). Although these QTLs were relatively large genomic regions with hundreds of genes, this discovery nevertheless leads to a closer understanding of the underlying genetic basis in the host of the grass-endophyte interaction. Faville et al. (2015) suggested that defence-related (DR) genes in the host may play a role in the symbiosis, similar to fungal genes that regulate antagonism or mutualism. This hypothesis is supported by their result of multiple QTL detection in linkage group 7 (Fig. 1.1), a region that is co-linear with a wheat genomic region (Jones et al., 2002; Sim et al., 2005) that has been shown to be relatively rich with DR genes (Li et al., 1999). DR genes have also been identified on perennial ryegrass LG 7 (Faville et al., 2004; Faville et al., 2015). This hypothesis is further supported by their result of relative similarity of the QTL region in LG4 (Fig. 1.1) with the sequence of pathogenesis-related class 10 protein (Zhang et al., 2011) via BLAST search against the syntenic rice chromosome 3 (Faville et al., 2015). Further, comparing endophyte-infected and endophyte-free tall fescue plants, genes encoding WRKY transcription factors were found to be among those differentially expressed (Dinkins et al., 2017). A WRKY gene homologue was also reported to be upregulated in endophyte-infected perennial ryegrass by Dupont et al. (2015). WRKY transcription factors are known to play a role in the pathogenesis response, for example, WRKY1 and WRKY2 (*A. thaliana* homologues AtWRKY18 and AtWRKY40, respectively) are involved in powdery mildew response in both barley and Arabidopsis (Shen et al., 2007). In another transcriptome study in perennial ryegrass, it was also found that a majority of the upregulated genes play a role in plant protection (Schmid et al., 2016). The second most important group includes genes involved in hormone signalling which is also related to defence-response through the "priming" of the host to biotic and abiotic stresses. Consequently, hormone signalling is compatible with an endophyte-mediated insect resistance host response model (Schmid et al., 2016). This suggests that the grass-endophyte symbiosis and its reaction to herbivores may be interrelated.

Host genetic control also affects the vertical transmission of fungal endophytes from generation to generation via seed. This was observed in perennial ryegrass populations infected with the *E. festucae* var. *lolii* strain AR37 (Gagic et al., 2018). As expected, populations that had undergone selection for improved AR37 transmission showed a higher percentage of viable endophyte than those that did not. Individual plants within half-sib families were also found to have significantly different transmission frequencies. In the present study, endophyte transmission was also investigated although in this case, with a non-native endophyte of perennial ryegrass namely, AR501. Analysis of variance in Gagic et al. (2018) with transmission frequencies using clones in two sites for two years showed that environment significantly affects endophyte transmission but also a large and significant (host) genotypic variance component, and only a small genotype-by-environment interaction. Strong evidence for host genetic control of the association was reported from this study in terms of the potential to predict the AR37 transmission trait using information from host genomic SNP markers (genomic selection). A relatively high genomic prediction accuracy (r = 0.54) (Gagic et al., 2018), estimated as the correlation between predicted and observed trait values, was observed. As with MAS, genomic selection may therefore be a promising breeding approach to improve endophyte compatibility for novel associations.

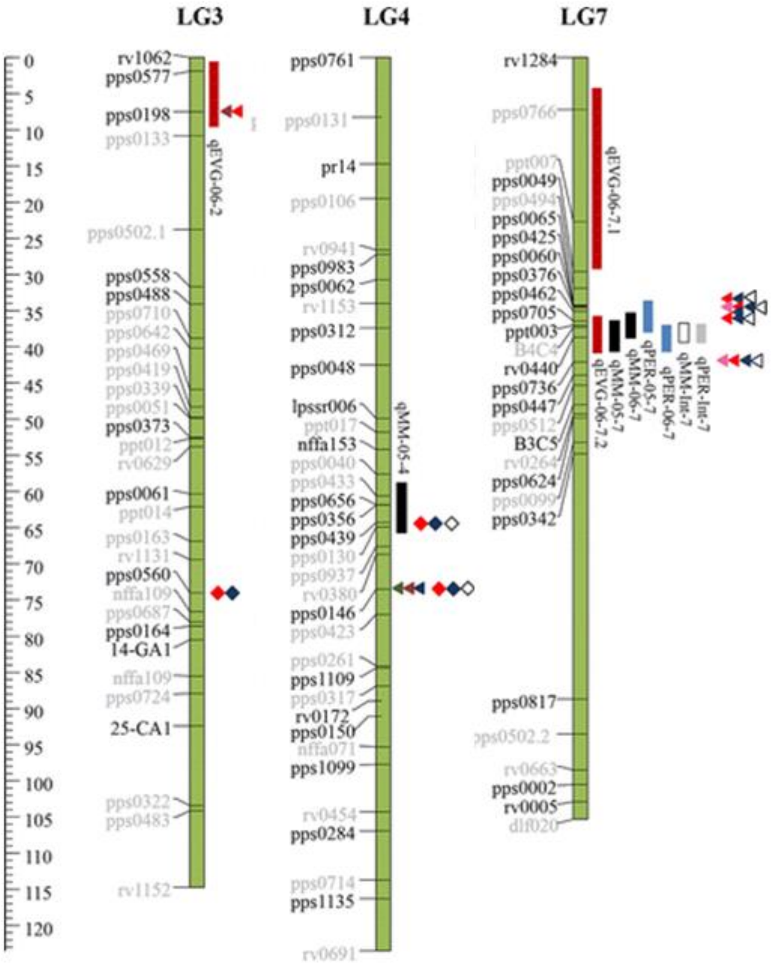


Figure 1.1. QTLs for endophyte biomass and alkaloid expression co-detected in two perennial ryegrass mapping populations. Green long bars, with marker names on the left, represent linkage groups; other bars (or rectangles) are QTLs detected from population 1; and triangles and diamonds are from population 2. The scale on the leftmost represents centimorgan distances. Adapted from Faville et al. (2015).

In the present study, the host genetic control of the ryegrass-endophyte association was investigated in a perennial ryegrass breeding population developed through recurrent selection for improved endophyte compatibility. Endophyte compatibility refers to two component traits: viability and transmission. Viability describes the ability of the endophytes to remain viable in seeds under storage, while transmission refers to the ability of the endophyte to be transmitted from parents to offspring or from one generation to the next (M.

J. Faville, personal communication, 2018). Genomic regions that show signatures of selection are likely to contribute to genetic control of the association in the host because of the directional selection for endophyte compatibility applied in the population under investigation. Genomic variation that characterizes loci under selection differs from neutral loci. This pattern differs across the genome and the variation that characterizes selection is termed as selection signatures (Kreitman, 2000). A QTL/gene discovery approach, called selection mapping (SM), has been proposed to discover genomic regions under selection and are related to the trait being selected in a breeding population (Wisser et al., 2008). SM has some advantages over the more conventional QTL mapping (Wisser et al., 2008). SM can be applied to commercial breeding populations, such as PGG04, which can bridge the gap between QTL discovery and utilization. QTLs discovered with SM can be used readily in the advanced generation of the breeding population through marker-assisted selection. Furthermore, because breeding populations used in outcrossing species are generally derived from multiple parents, they potentially contain several allelic variants that can be explored. In contrast, linkage mapping utilizes biparental mapping population which contains limited genetic variation (i.e. only from the two parents) and is generally different from the breeding population. Population for linkage mapping may be different from the breeding population thus QTL incorporation will not be immediate. However, QTL mapping populations allow for the partitioning of individuals into different genotypic groups (i.e. alleles or haplotypes), and directly test whether significant phenotypic differences exist between groups. This makes hypothesis testing fairly straightforward. In a population undergoing RS, changes in allele frequency are not only influenced by selection but other factors as well, notably, random genetic drift. In SM therefore, a statistical test of the null hypothesis of genetic drift is considered (Wisser et al., 2008). Gene flow or migration and mutation, other than selection and genetic drift, also influence allele frequency shifts. These factors were assumed negligible for the current study. Mutation rates are generally slow, and the RS populations in the study have a short period of time between them (four cycles of selection), therefore mutation can be largely discounted. Gene flow is also likely to be insignificant, if not totally absent, as polycrossing of selected individuals was done meticulously in isolation, with little or no opportunity for non-selected pollen or volunteer plants to enter. With these confounding factors, SM results will be relatively less direct, than say, linkage mapping. Nevertheless, SM has been carried out in perennial ryegrass (Brazauskas, Pašakinskienė, et al., 2013) and in conjunction with linkage mapping (Brazauskas, Xing, et al., 2013). Similarly, SM results in the present study could provide supporting evidence for linkage mapping results of Faville et al. (2015) or identify new regions of importance. For example, it is possible to compare the physical location of the SNPs detected in SM and QTLs in Faville et al. (2015). QTLs in Faville et al. (2015) represent large genomic regions and while the present study has the advantage of higher marker density, GBS generally samples only small proportion of the genome, not to mention that perennial ryegrass reference genome is incomplete (i.e. in scaffolds). Therefore, more studies are needed, such as fine mapping experiments, to identify the precise location of QTLs and/or genes. Selection signatures in the genome detected in this study will nonetheless give insights into the genetic variation influencing the grass-endophyte interaction. Furthermore, the combined use of strategies such as SM, linkage mapping, LD mapping, and functional genetics studies, will collectively offer a more comprehensive genetic dissection of the host-endophyte interaction.

1.6. Perennial ryegrass breeding and signatures of selection

Perennial ryegrass is the most extensively grown forage in the temperate world therefore, it represents a relatively large seed industry (Humphreys et al., 2010; Wilkins, 1991). The main breeding objectives for perennial ryegrass are: total and seasonal herbage yield; forage quality; and persistence. Other breeding objectives include: abiotic stress tolerance; pest and disease resistance; and seed production traits (Humphreys et al., 2010; Stewart & Hayes, 2011; Wilkins, 1991). Population improvement is the primary breeding strategy employed in perennial ryegrass because of its out-crossing nature. Nevertheless, several techniques have been employed in breeding perennial ryegrass including induction of polyploidy to create tetraploid varieties and interspecific hybridization (Humphreys et al., 2010; Lee et al., 2012). Molecular breeding is also exploited in perennial ryegrass although in a limited capacity. For example, a meta-analysis reported at least 560 published QTLs (Shinozuka et al., 2012) and marker-assisted selection has been carried out (Dolstra et al., 2007).

Population improvement in ryegrass typically starts with pair crossing amongst cultivar and/or ecotype germplasm followed by recurrent selection (RS) through the polycross method, with phenotypic or half-sib family selection. Finally, synthetic varieties are generated by intermating selected desirable individuals (i.e. mother plants) (Wilkins, 1991). RS is described as the repeated cycles of selection with the aim of increasing the frequency of the favourable allele for a particular trait of interest (Sleper & Poehlman, 2006). There are three types of selection, and combinations of these are practiced. The three types of selection are: phenotypic (i.e. individual performance); genotypic (i.e. based on the performance of progenies); and marker-assisted selection (including genomic selection) (Conaghan & Casler, 2011). Genomic selection based on joint marker effects has been gaining some ground in perennial ryegrass breeding (Faville et al., 2016; Fè et al., 2015; Grinberg et al., 2016) but will take some time before being fully implemented in the majority of breeding programs. On the other hand, phenotypic or mass selection is described as visual plant selection for desirable traits followed by bulking seeds from selected plants without any form of progeny testing (Sleper & Poehlman, 2006). The frequency of the trait of interest can be shifted using mass selection provided that heritability is high, and the trait can be measured easily. Conversely, in genotypic selection, the breeding value of an individual is assessed based on the performance of its progenies. Progeny testing can be based on half-sib or full-sib families and the genetic component of the phenotypic variance can be estimated based on replicated trials. It also allows for multi-location testing leading to more accurate heritability estimates and therefore genetic gain (Conaghan & Casler, 2011). In the UK, a combination of phenotypic and half-sib family selection conducted for 12 years was shown to improve dry matter yield and water-soluble carbohydrate content in perennial ryegrass (Wilkins & Humphreys, 2003). In population improvement, it is also important to avoid inbreeding depression. High-intensity selection, leading to the greater superiority of the selected individuals, is directly proportional to genetic gain. On the other hand, fewer selected individuals can lead to greater inbreeding and low genetic variability. Thus, consideration of the effective population size is important in population improvement as it affects selection intensity on one side, and inbreeding depression on the other (Posselt, 2010).

The goal of RS is to increase the frequency of favourable alleles. If these frequency changes can be monitored during RS, genetic signatures of artificial selection can be investigated. More generally, how genetic diversity is shaped by artificial selection can be monitored. Siol et al. (2010) reviewed some of the methods on how signatures of selection can be inferred from molecular data. Positive directional selection leads to fixation of a favourable allele and hitchhiking effect leads to the fixation of linked neutral alleles (Smith & Haigh, 1974). Selection signature, in this case, is the reduction of diversity around the locus of interest (Siol et al., 2010). On the other hand, balancing selection due to overdominance or negative frequency-dependent selection will cause increased diversity (Siol et al., 2010). Positive selection will also influence the relative proportion of alleles, that is, skewing the allele frequency distribution to high and low frequencies (Fay & Wu, 2000). A selection signature can also be inferred by comparing synonymous and nonsynonymous mutations (i.e. substitutions in the coding region leading to amino acid replacement [McDonald & Kreitman, 1991]) (Siol et al., 2010). An excess in LD also can be a sign of selection (Smith & Haigh, 1974). However, McVean (2007) showed that selection can also decrease or eliminate LD depending on the relative position of the locus under selection and neutral loci near it. Selection shapes population structure and in a subdivided population, selection may favour different or similar alleles leading to increased or decreased population differentiation (i.e. F_{ST}), respectively, because of allele frequency changes (Siol et al., 2010). In the present study, the effects of selection in terms of shaping the population structure was investigated.

Aside from understanding selection, selection signatures can be linked to traits to discover QTLs influencing trait variation. Wisser et al. (2008) described an approach called selection mapping (SM) that, "identifies alleles, loci, and epistatic interactions using populations that have been subjected to iterative cycles of recombination and selection." It has the advantage of utilizing the breeding population in QTL discovery. Rare alleles can also be detected in SM by virtue of its enrichment in RS. However, it cannot be used to estimate marker effects since selection is generally not limited to one primary trait of interest. The nature of LD as a result of the characteristics of the base population and selection intensity can also confound the detection of selection (Wisser et al., 2008). Using an RS population, primarily for the improvement of northern leaf blight resistance in maize, they identified loci under selection based on allele frequency shifts. These loci influence resistance as confirmed by cosegregation analysis (Wisser et al., 2008). There are limited studies that utilized SM in trait discovery; nevertheless, it was reported in perennial ryegrass (Brazauskas, Pašakinskienė, et al., 2013). Prospecting temporal allele frequency changes of five candidate genes in an RS population for axillary tiller development, the researchers identified selection in *LpIAA1*. This gene is a member of an auxin-regulated gene family, which is known to control shoot morphology in Arabidopsis and rice (Brazauskas, Pašakinskienė, et al., 2013). In another study, utilizing a population that was divergently selected for crown rust resistance, Brazauskas, Xing, et al. (2013) identified seven loci under selection. From the seven loci, one QTL was verified based on linkage mapping. In relation to linkage mapping and LD mapping, SM studies have not been utilized extensively and therefore merits further investigation.

1.7. Summary and research gaps

Perennial ryegrass breeding is important in pastoral agriculture. Persistence, partly modulated by fungal endophytes in a symbiotic relationship with the ryegrass, is among the most important breeding objectives. Endophytes support persistence by the production of alkaloids toxic to pest herbivores. While the grass-endophyte interaction has been studied largely in fungal mutants, host genetic control of the association is more relevant in cultivar development. Genetic factors in the host influence fungal biomass, alkaloid concentration and endophyte vertical transmission frequency. The symbiotic relationship is therefore exploited in perennial ryegrass breeding by exploring perennial ryegrass cultivar and fungal strain combinations. Being vertically transmitted, the compatibility of grasses to novel endophytes is important in maintaining the beneficial interaction. Most, if not all commercial endophytes, are associated with a native grass host. Therefore, the generation of cross-species associations results in interesting interactions. The investigation of novel endophyte transmission in a non-native host is necessary if breeders are to successfully explore and exploit cross-species associations.

Several studies on the genetic diversity of perennial ryegrass populations have been reported. However, current molecular marker technologies (e.g. GBS) are stimulating the re-examination of genetic variation in advanced breeding populations which may contribute to the progress of perennial ryegrass breeding through trait discovery and prediction. Cultivar development is accomplished mainly through population improvement using RS. The investigation on how genetic variation at the molecular level is shaped by directional selection represents an important avenue of plant breeding research. Transfer of tall fescue endophytes to perennial ryegrass, followed by RS to improve the association, is desirable because it is a route to bringing highly effective loline alkaloids into NZ ryegrass pastures. Monitoring genetic variation in breeding populations developed for endophyte compatibility will provide more insights on the genetic basis of grass-endophyte interactions, and on changes in genetic diversity due to RS.

1.8. Hypothesis and Objectives

The objective of this study is to investigate a perennial ryegrass breeding population under recurrent selection (RS) for compatibility with an endophyte sourced from tall fescue. There are two specific objectives. First, to investigate the transmission of the endophyte, AR501, in the breeding population, PGG04, to provide evidence that positive selection improves endophyte compatibility. It was hypothesized that an increase in endophyte infection attributed to selection will be observed. The second objective is to investigate how genetic variation changes during RS. Further, this study aims to examine how allele frequencies change in an RS program specifically in terms of population differentiation, that is, with fixation index (F_{ST}) and PCA-based population structure analysis. Since the selection program targets endophyte compatibility, signatures of selection that may be associated with the grass-endophyte interaction is also determined. It was hypothesized that substantial changes in genetic diversity will be observed between the early and late generation of PGG04. A reduction of diversity and an excess of rare alleles is also expected. Although, it is also possible to detect an increase in diversity around the loci under selection. This is observed under balancing selection and if for example, selection favours advantageous heterozygous loci (i.e. they promote compatibility), increased diversity may be observed around these loci relative to the rest of the genome. In general, deviation from expected levels of diversity and allele frequencies will point to signatures of selection. Furthermore, loci under positive selection will have a larger contribution to genetic differentiation than neutral loci which are affected mostly by random genetic drift. Therefore, the null hypothesis of genetic drift is tested. Loci under selection pressure are expected to have extremely high F_{ST} values and to be correlated with the components of PCA-based population structure analysis. Among regions under selection, those that are related to endophyte compatibility were further determined through an association test with infection data. These genomic regions represent putative QTLs influencing the grass-endophyte symbiosis and, although these must be verified independently, represent a molecular breeding tool for this important trait.

A conceptual framework, relating recurrent selection and changes in genetic diversity, is shown below (Fig. 1.2). The left part of the image describes the recurrent selection process. From a base population (not shown), the first cycles of selection produce the early generation. Subsequently, desirable individuals (purple) are selected and crossed. In this case, selection is based on endophyte compatibility (i.e. transmission and viability). With repeated selection and polycrossing, the number of individuals with good compatibility increases through time. Analysis and comparison of the SNP genotypes of samples from the early and late generations will reveal changes in genetic diversity. Reduction in diversity is exemplified in the highlighted (enclosed in blue rectangles) genomic regions.

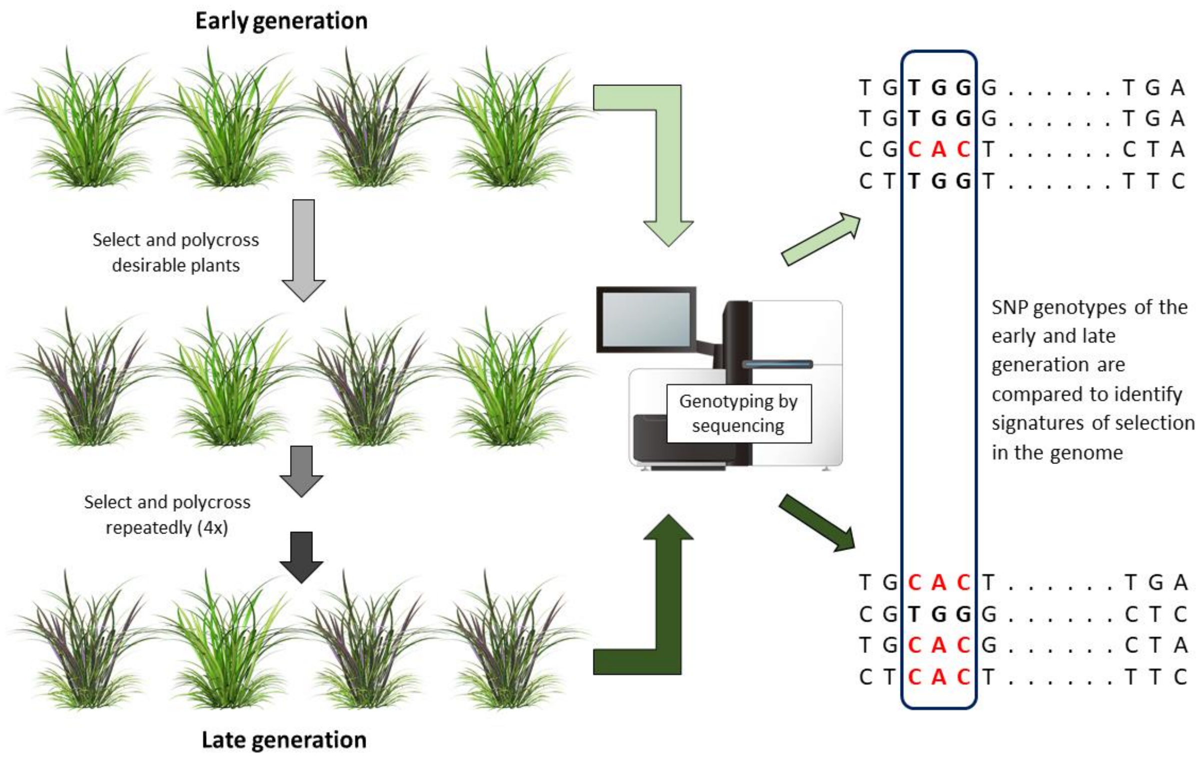

Early generation
Select and polycross desirable plants
Select and polycross repeatedly (4x)
Late generation
Genotyping by sequencing
T G T G G G . . . . . . T G A
T G T G G G . . . . . . T G A
C G C A C T . . . . . . C T A
C T T G G T . . . . . . T T C
SNP genotypes of the early and late generation are compared to identify signatures of selection in the genome
T G C A C T . . . . . . T G A
C G T G G G . . . . . . C T C
T G C A C G . . . . . . C T A
C T C A C T . . . . . . T T C

Figure 1.2. Changes in genetic variation between early and late generations of a population that undergone recurrent selection. The image of a grass plant was accessed in the web ("grass-22", n. d. Retrieved from https://photoshop-kopona.com/49594-klipart-grass.html) and sequencer from Database Center for Life Science (2013). Purplish plants represent individuals compatible with the endophyte.

1.9. References

Acquaah, G. (2012). *Principles of plant genetics and breeding. [electronic resource]*: Chichester, West Sussex, UK ; Hoboken, NJ : Wiley-Blackwell, 2012, 2nd ed.

Adcock, R., Hill, N., Bouton, J., Boerma, H., & Ware, G. (1997). Symbiont regulation and reducing ergot alkaloid concentration by breeding endophyte-infected tall fescue. *J Chem Ecol, 23*(3), 691-704.

Armstead, I. P., Turner, L. B., Farrell, M., Skøt, L., Gomez, P., Montoya, T., . . . Humphreys, M. O. (2004). Synteny between a major heading-date QTL in perennial ryegrass (Lolium perenne L.) and the Hd3 heading-date locus in rice. *Theor Appl Genet, 108*(5), 822-828. doi: 10.1007/s00122-003-1495-6

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 Biol, 16*(1), 160. doi: 10.1186/s12870-016-0844-y

Auzanneau, J., Huyghe, C., Julier, B., & Barre, P. (2007). Linkage disequilibrium in synthetic varieties of perennial ryegrass. *Theor Appl Genet, 115*(6), 837-847. doi: 10.1007/s00122-007-0612-3

Balfourier, F., Imbert, C., & Charmet, G. (2000). Evidence for phylogeographic structure in Lolium species related to the spread of agriculture in Europe. A cpDNA study. *Theor Appl Genet, 101*(1), 131-138. doi: 10.1007/s001220051461

Ball, O., Miles, C., & Prestidge, R. (1997). Ergopeptine Alkaloids and Neotyphodium lolii-Mediated Resistance in Perennial Ryegrass Against Adult Heteronychus arator (Coleoptera: Scarabaeidae). *J Econ Entomol, 90*(5), 1382-1391. doi: 10.1093/jee/90.5.1382

Barth, S., McGrath, S. K., Arojju, S. K., & Hodkinson, T. R. (2015). An Irish perennial ryegrass genetic resource collection clearly divides into two major gene pools. *Plant Genetic Resources, 15*(3), 269-278. doi: 10.1017/S1479262115000611

Blackmore, T., Thomas, I., McMahon, R., Powell, W., & Hegarty, M. (2015). Genetic–geographic correlation revealed across a broad European ecotypic sample of perennial ryegrass (Lolium perenne) using array-based SNP genotyping. *Theor Appl Genet, 128*(10), 1917-1932. doi: 10.1007/s00122-015-2556-3

Blackmore, T., Thorogood, D., Skøt, L., McMahon, R., Powell, W., & Hegarty, M. (2016). Germplasm dynamics: the role of ecotypic diversity in shaping the patterns of genetic variation in Lolium perenne. *Scientific Reports, 6*, 22603. doi: 10.1038/srep22603

Bolaric, S., Barth, S., Melchinger, A. E., & Posselt, U. K. (2005). Genetic diversity in European perennial ryegrass cultivars investigated with RAPD markers. *Plant Breeding, 124*(2), 161-166. doi: 10.1111/j.1439-0523.2004.01032.x

Bothe, A., Nehrlich, S., Willner, E., & Dehmer, K. J. (2016). Phenotyping Genetic Diversity of Perennial Ryegrass Ecotypes (Lolium perenne L.). In I. Roldán-Ruiz, J. Baert, & D. Reheul (Eds.), *Breeding in a World of Scarcity: Proceedings of the 2015 Meeting of the Section "Forage Crops and Amenity Grasses" of Eucarpia* (pp. 21-27). Cham: Springer International Publishing.

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

Brazauskas, G., Pašakinskienė, I., & Lübberstedt, T. (2013). Estimation of Temporal Allele Frequency Changes in Ryegrass Populations Selected for Axillary Tiller Development *Breeding strategies for sustainable forage and turf grass improvement* (pp. 81-87): Springer.

Brazauskas, G., Xing, Y., Studer, B., Schejbel, B., Frei, U., Berg, P., & Lübberstedt, T. (2013). Identification of genomic loci associated with crown rust resistance in perennial ryegrass (Lolium perenne L.) divergently selected populations. *Plant Sci, 208*, 34-41.

Bush, L., Fannin, F., Siegel, M., Dahlman, D., & Burton, H. (1993). Chemistry, occurrence and biological effects of saturated pyrrolizidine alkaloids associated with endophyte-grass interactions. *Agric, Ecosyst Environ, 44*(1-4), 81-102.

Byrne, S., Conaghan, P., Barth, S., Arojju, S., Casler, M., Michel, T., . . . Milbourne, D. (2017). Using variable importance measures to identify a small set of SNPs to predict heading date in perennial ryegrass. *Scientific Reports, 7*(1), 3566. doi: 10.1038/s41598-017-03232-8

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., & Asp, T. (2013). Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. *PloS One, 8*(3), e57438.

Card, S., Rolston, M., Park, Z., Cox, N., & Hume, D. (2011). Fungal endophyte detection in pasture grass seed utilising the infection layer and comparison to other detection techniques. *Seed Science and Technology, 39*(3), 581-592.

Christensen, M. (1995). Variation in the ability of Acremonium endophytes of Lolium perenne, Festuca arundinacea and F. pratensis to form compatible associations in the three grasses. *Mycol Res, 99*(4), 466-470. doi: https://doi.org/10.1016/S0953-7562(09)80647-3

Christensen, M., Leuchtmann, A., Rowan, D., & Tapper, B. (1993). Taxonomy of Acremonium endophytes of tall fescue (Festuca arundinacea), meadow fescue (F. pratensis) and perennial ryegrass (Lolium perenne). *Mycol Res, 97*(9), 1083-1092. doi: https://doi.org/10.1016/S0953-7562(09)80509-1

Conaghan, P., & Casler, M. (2011). A theoretical and practical analysis of the optimum breeding system for perennial ryegrass. *Irish Journal of Agricultural and Food Research*, 47-63.

Database Center for Life Science. (2013). *genomesequencer4* [Electronic image]. Retrieved from http://togotv.dbcls.jp/ja/togopic.2011.34.html

Dinkins, R. D., Nagabhyru, P., Graham, M. A., Boykin, D., & Schardl, C. L. (2017). Transcriptome response of Lolium arundinaceum to its fungal endophyte Epichloë coenophiala. *New Phytol, 213*(1), 324-337. doi: doi:10.1111/nph.14103

Dolstra, O., Denneboom, C., de Vos, A. L., & van Loo, E. (2007). Marker-assisted selection for improving quantitative traits of forage crops. *Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish. FAO, Rome*, 59-65.

Dupont, P.-Y., Eaton, C. J., Wargent, J. J., Fechtner, S., Solomon, P., Schmid, J., . . . Cox, M. P. (2015). Fungal endophyte infection of ryegrass reprograms host metabolism and alters development. *New Phytol, 208*(4), 1227-1240. doi: doi:10.1111/nph.13614

Easton, H., Christensen, M., Eerens, J., Fletcher, L., Hume, D., Keogh, R., . . . Popay, A. (2001). *Ryegrass endophyte: a New Zealand Grassland success story.* Paper presented at the Proceedings of the conference-New Zealand Grassland Association.

Easton, H., Latch, G., Tapper, B., & Ball, O.-P. (2002). Ryegrass host genetic control of concentrations of endophyte-derived alkaloids. *Crop Sci, 42*(1), 51-57.

Easton, H., Lyons, T., Mace, W., Simpson, W., De Bonth, A., Cooper, B., & Panckhurst, K. (2007). *Differential expression of loline alkaloids in perennial ryegrass infected with endophyte isolated from tall fescue.* Paper presented at the Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses. New Zealand Grassland Association, Dunedin, New Zealand.

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), e19379.

Faville, M., Briggs, L., Cao, M., Koulman, A., Jahufer, M., Koolaard, J., & Hume, D. (2015). A QTL analysis of host plant effects on fungal endophyte biomass and alkaloid expression in perennial ryegrass. *Mol Breed, 35*(8), 161.

Faville, M., Ganesh, S., Cao, M., Jahufer, M., Bilton, T., Easton, H., . . . Barrett, B. (2018). Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theor Appl Genet*. doi: 10.1007/s00122-017-3030-1

Faville, M., Ganesh, S., Moraga, R., Easton, H., Jahufer, M., Elshire, R., . . . Barrett, B. (2016). Development of genomic selection for perennial ryegrass *Breeding in a World of Scarcity* (pp. 139-143): Springer.

Faville, M., Vecchies, A., Schreiber, M., Drayton, M., Hughes, L., Jones, E., . . . Spangenberg, G. (2004). Functionally associated molecular genetic marker map construction in perennial ryegrass (Lolium perenne L.). *Theor Appl Genet, 110*(1), 12-32.

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics, 155*(3), 1405-1413.

Fè, D., Ashraf, B., Pedersen, M., Janss, L., Byrne, S., Roulund, N., . . . Jensen, J. (2016). Accuracy of Genomic Prediction in a Commercial Perennial Ryegrass Breeding Program. *The Plant Genome, 9*(3). doi: 10.3835/plantgenome2015.11.0110

Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B., Pedersen, M., . . . Jensen, C. (2015). Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics, 16*(1), 921.

Fletcher, L., & Harvey, I. (1981). An association of a Lolium endophyte with ryegrass staggers. *New Zealand veterinary journal, 29*(10), 185-186.

Frascaroli, E., Schrag, T. A., & Melchinger, A. E. (2013). Genetic diversity analysis of elite European maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. *Theor Appl Genet, 126*(1), 133-141.

Gagic, M., Faville, M. J., Zhang, W., Forester, N. T., Rolston, M. P., Johnson, R. D., . . . Hudson, D. (2018). Seed transmission of Epichloë endophytes in Lolium perenne is heavily influenced by host genetics. *Frontiers in Plant Science, 9*, 1580.

Gallagher, R., White, E., & Mortimer, P. (1981). Ryegrass staggers: isolation of potent neurotoxins lolitrem A and lolitrem B from staggers-producing pastures. *New Zealand veterinary journal, 29*(10), 189-190.

Ghesquiere, A., Calsyn, E., Baert, J., & Riek, J. (2003). Genetic diversity between and within ryegrass populations of the ECP/GR collection by means of AFLP markers. *Czech Journal of Genetics and Plant Breeding, 39*(Special issue), 333.

Ghesquiere, A., Muylle, H., & Baert, J. (2013). Use of Molecular Marker Information in the Construction of Polycrosses to Enhance Yield in a Lolium perenne Breeding Programme *Breeding strategies for sustainable forage and turf grass improvement* (pp. 63-67): Springer.

grass-22, (n. d.). Retrieved from https://photoshop-kopona.com/49594-klipart-grass.html

Grinberg, N. F., Lovatt, A., Hegarty, M., Lovatt, A., Skøt, K. P., Kelly, R., . . . Armstead, I. (2016). Implementation of genomic prediction in Lolium perenne (L.) breeding populations. *Frontiers in plant science, 7*.

Guthridge, K. M., Dupal, M. P., Kölliker, R., Jones, E. S., Smith, K. F., & Forster, J. W. (2001). AFLP analysis of genetic diversity within and between populations of perennial ryegrass (Lolium perenne L.). *Euphytica, 122*(1), 191-201. doi: 10.1023/a:1012658315290

Hahn, H., Huth, W., Schöberlein, W., Diepenbrock, W., & Weber, W. (2003). Detection of endophytic fungi in Festuca spp. by means of tissue print immunoassay. *Plant Breeding, 122*(3), 217-222.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., & Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in plant science, 5*.

Humphreys, M., Feuerstein, U., Vandewalle, M., & Baert, J. (2010). Ryegrasses. In B. Boller, U. K. Posselt, & F. Veronesi (Eds.), *Fodder Crops and Amenity Grasses* (pp. 211-260). New York, NY: Springer New York.

Johnson, L., de Bonth, A., Briggs, L., Caradus, J., Finch, S., Fleetwood, D., . . . Card, S. (2013). The exploitation of epichloae endophytes for agricultural benefit. *Fungal Diversity, 60*(1), 171-188. doi: 10.1007/s13225-013-0239-4

Johnson, M., Dahlman, D., Siegel, M., Bush, L., Latch, G., Potter, D., & Varney, D. (1985). Insect feeding deterrents in endophyte-infected tall fescue. *Appl Environ Microbiol, 49*(3), 568-571.

Jones, E. S., Mahoney, N. L., Hayward, M. D., Armstead, I. P., Jones, J. G., Humphreys, M. O., . . . Balfourier, F. (2002). An enhanced molecular marker based genetic map of perennial ryegrass (Lolium perenne) reveals comparative relationships with other Poaceae genomes. *Genome, 45*(2), 282-295.

KöLliker, R., Boller, B., & Widmer, F. (2005). Marker assisted polycross breeding to increase diversity and yield in perennial ryegrass (Lolium perenne L.). *Euphytica, 146*(1), 55-65. doi: 10.1007/s10681-005-6036-8

Kreitman, M. (2000). Methods to Detect Selection in Populations with Applications to the Human. *Annual Review of Genomics and Human Genetics, 1*(1), 539-559. doi: 10.1146/annurev.genom.1.1.539

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 Sci, 41*, 1565-1572. doi: 10.2135/cropsci2001.4151565x

Latch, G., & Vaughan, D. (1995). Search for seedborne endophytic fungi in rice. *International Rice Research Notes*.

Lee, J. M., Matthew, C., Thom, E. R., & Chapman, D. F. (2012). Perennial ryegrass breeding in New Zealand: a dairy industry perspective. *Crop and Pasture Science, 63*(2), 107-127. doi: https://doi.org/10.1071/CP11282

Leuchtmann, A., Bacon, C. W., Schardl, C. L., White, J. F., & Tadych, M. (2014). Nomenclatural realignment of Neotyphodium species with genus Epichloë. *Mycologia, 106*(2), 202-215. doi: 10.3852/13-251

Li, W., Faris, J., Chittoor, J., Leach, J., Hulbert, S., Liu, D., . . . Gill, B. (1999). Genomic mapping of defense response genes in wheat. *Theor Appl Genet, 98*(2), 226-233.

Lyons, P., Plattner, R., & Bacon, C. (1986). Occurrence of peptide and clavine ergot alkaloids in tall fescue grass. *Science, 232*(4749), 487-489. doi: 10.1126/science.3008328

McCouch, S. R., Zhao, K., Wright, M., Tung, C.-W., Ebana, K., Thomson, M., . . . Bustamante, C. (2010). Development of genome-wide SNP assays for rice. *Breeding Science, 60*(5), 524-535. doi: 10.1270/jsbbs.60.524

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature, 351*, 652. doi: 10.1038/351652a0

McVean, G. (2007). The Structure of Linkage Disequilibrium Around a Selective Sweep. *Genetics, 175*(3), 1395-1406. doi: 10.1534/genetics.106.062828

Nixon, C. (2015). *How valuable is that plant species*. (MPI Technical Paper No: 2016/62). Retrieved from https://www.mpi.govt.nz/dmsdocument/14527-how-valuable-is-that-plant-species-application-of-a-method-for-enumerating-the-contribution-of-selected-plant-species-to-new-zealands-gdp/sitemap.

Parsons, A., Edwards, G., Newton, P., Chapman, D., Caradus, J., Rasmussen, S., & Rowarth, J. (2011). Past lessons and future prospects: plant breeding for yield and persistence in cool-temperate pastures. *Grass Forage Sci, 66*(2), 153-172.

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS one, 7*(2), e32253.

Posselt, U. K. (2010). Breeding Methods in Cross-Pollinated Species. In B. Boller, U. K. Posselt, & F. Veronesi (Eds.), *Fodder Crops and Amenity Grasses* (pp. 39-87). New York, NY: Springer New York.

Prestidge, R., Pottinger, R., & Barker, G. (1982). *An association of Lolium endophyte with ryegrass resistance to Argentine stem weevil.* Paper presented at the Proc. NZ Weed Pest Control Conf. 35th, 1982.

Rabier, C.-E., Barre, P., Asp, T., Charmet, G., & Mangin, B. (2016). On the Accuracy of Genomic Selection. *PLOS ONE, 11*(6), e0156086. doi: 10.1371/journal.pone.0156086

Rolston, M., & Agee, C. (2007). *Delivering quality seed to specification-the USA and NZ novel endophyte experience.* Paper presented at the Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses'. Christchurch, New Zealand. Grassland Research and Practice Series.

Rowan, D., & Gaynor, D. (1986). Isolation of feeding deterrents against Argentine stem weevil from ryegrass infected with the endophyteAcremonium loliae. *J Chem Ecol, 12*(3), 647-658.

Saikkonen, K., Young, C. A., Helander, M., & Schardl, C. L. (2016). Endophytic Epichloë species and their grass hosts: from evolution to applications. *Plant Mol Biol, 90*(6), 665-675. doi: 10.1007/s11103-015-0399-6

Schardl, C., Grossman, R., Nagabhyru, P., Faulkner, J., & Mallik, U. (2007). Loline alkaloids: Currencies of mutualism. *Phytochemistry, 68*(7), 980-996. doi: https://doi.org/10.1016/j.phytochem.2007.01.010

Schardl, C., Young, C., Faulkner, J., Florea, S., & Pan, J. (2012). Chemotypic diversity of epichloae, fungal symbionts of grasses. *Fungal Ecology, 5*(3), 331-344. doi: https://doi.org/10.1016/j.funeco.2011.04.005

Schmid, J., Day, R., Zhang, N., Dupont, P.-Y., Cox, M. P., Schardl, C. L., . . . Zhou, Y. (2016). Host Tissue Environment Directs Activities of an Epichloë Endophyte, While It Induces Systemic Hormone and Defense Responses in Its Native Perennial Ryegrass Host. *Mol Plant-Microbe Interact, 30*(2), 138-149. doi: 10.1094/MPMI-10-16-0215-R

Schmidt, S. P., Hoveland, C. S., Clark, E. M., Davis, N. D., Smith, L. A., Grimes, H. W., & Holliman, J. L. (1982). Association of an endophytic fungus with fescue toxicity in steers fed Kentucky 31 tall fescue seed or hay. *J Anim Sci, 55*(6), 1259-1263.

Shen, Q.-H., Saijo, Y., Mauch, S., Biskup, C., Bieri, S., Keller, B., . . . Schulze-Lefert, P. (2007). Nuclear Activity of MLA Immune Receptors Links Isolate-Specific and Basal Disease-Resistance Responses. *Science, 315*(5815), 1098-1103. doi: 10.1126/science.1136372

Shinozuka, H., Cogan, N. O., Spangenberg, G. C., & Forster, J. W. (2012). Quantitative Trait Locus (QTL) meta-analysis and comparative genomics for candidate gene prediction in perennial ryegrass (Lolium perenne L.). *BMC Genet, 13*(1), 101.

Sim, S., Chang, T., Curley, J., Warnke, S., Barker, R., & Jung, G. (2005). Chromosomal rearrangements differentiating the ryegrass genome from the Triticeae, oat, and rice genomes using common heterologous RFLP probes. *Theor Appl Genet, 110*(6), 1011-1019.

Simpson, W. R., Schmid, J., Singh, J., Faville, M. J., & Johnson, R. D. (2012). A morphological change in the fungal symbiont Neotyphodium lolii induces dwarfing in its host plant Lolium perenne. *Fungal Biology, 116*(2), 234-240. doi: https://doi.org/10.1016/j.funbio.2011.11.006

Siol, M., Wright, S. I., & Barrett, S. C. (2010). The population genomics of plant adaptation. *New Phytol, 188*(2), 313-332.

Skøt, L., Humphreys, M. O., Armstead, I., Heywood, S., Skøt, K. P., Sanderson, R., . . . Hamilton, N. R. S. (2005). An association mapping approach to identify flowering time genes in natural populations of Lolium perenne (L.). *Mol Breed, 15*(3), 233-245. doi: 10.1007/s11032-004-4824-9

Sleper, D. A., & Poehlman, J. M. (2006). *Breeding field crops*: Blackwell publishing.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research, 23*(1), 23-35.

Stewart, A. (2006). Genetic origins of perennial ryegrass (Lolium perenne) for New Zealand pastures. *Grassland Research and Practice Series, 12*, 55-62.

Stewart, A., & Hayes, R. (2011). Ryegrass breeding-balancing trait priorities. *Irish Journal of Agricultural and Food Research*, 31-46.

Thorogood, D. (2003). Perennial ryegrass (Lolium perenne L.). *Turfgrass biology, genetics, and breeding. John Wiley & Sons, Inc., Hoboken, New Jersey*, 75-105.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., . . . Schön, C.-C. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics, 15*(1), 823. doi: 10.1186/1471-2164-15-823

van Zijll de Jong, E., Dobrowolski, M. P., Bannan, N. R., Stewart, A. V., Smith, K. F., Spangenberg, G. C., & Forster, J. W. (2008). Global Genetic Diversity of the Perennial Ryegrass Fungal Endophyte Neotyphodium lolii. *Crop Sci, 48*(4), 1487-1501. doi: 10.2135/cropsci2007.11.0641

Velmurugan, J., Milbourne, D., Connolly, V., Heslop-Harrison, J. S., Anhalt, U. C. M., Lynch, M. B., & Barth, S. (2018). An Immortalized Genetic Mapping Population for Perennial Ryegrass: A Resource for Phenotyping and Complex Trait Mapping. *Frontiers in Plant Science, 9*(717). doi: 10.3389/fpls.2018.00717

Velmurugan, J., Mollison, E., Barth, S., Marshall, D., Milne, L., Creevey, C. J., . . . Milbourne, D. (2016). An ultra-high density genetic linkage map of perennial ryegrass (Lolium perenne) using genotyping by sequencing (GBS) based on a reference shotgun genome assembly. *Ann Bot, 118*(1), 71-87. doi: 10.1093/aob/mcw081

Wilkins, P. (1991). Breeding perennial ryegrass for agriculture. *Euphytica, 52*(3), 201-214. doi: 10.1007/bf00029397

Wilkins, P., & Humphreys, M. (2003). Progress in breeding perennial forage grasses for temperate agriculture. *The Journal of Agricultural Science, 140*(2), 129-150.

Winfield, M. O., Allen, A. M., Burridge, A. J., Barker, G. L. A., Benbow, H. R., Wilkinson, P. A., . . . Edwards, K. J. (2016). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol J, 14*(5), 1195-1206. doi: 10.1111/pbi.12485

Wisser, R. J., Murray, S. C., Kolkman, J. M., Ceballos, H., & Nelson, R. J. (2008). Selection Mapping of Loci for Quantitative Disease Resistance in a Diverse Maize Population. *Genetics, 180*(1), 583-599. doi: 10.1534/genetics.108.090118

Yu, X., Bai, G., Luo, N., Chen, Z., Liu, S., Liu, J., . . . Jiang, Y. (2011). Association of simple sequence repeat (SSR) markers with submergence tolerance in diverse populations of perennial ryegrass. *Plant Sci, 180*(2), 391-398. doi: https://doi.org/10.1016/j.plantsci.2010.10.013

Zhang, N., Zhang, S., Borchert, S., Richardson, K., & Schmid, J. (2011). High levels of a fungal superoxide dismutase and increased concentration of a PR-10 plant protein in associations between the endophytic fungus Neotyphodium lolii and ryegrass. *Mol Plant-Microbe Interact, 24*(8), 984-992.

Chapter 2

Plant selection leads to improved endophyte compatibility in terms of the percentage of viable endophyte in the population

2.1. Abstract

The mutualistic association between *Lolium perenne* L. and *Epichloë* spp. fungal endophytes is fundamentally important for the success of pastoral agriculture in New Zealand. Compatibility between the endophyte and its host grass is influenced by a host genetic component. The improved understanding of the vertical transmission of novel endophytes in perennial ryegrass populations is important in breeding for endophyte compatibility. The objective of this study, therefore, is to characterise the transmission of a tall fescue endophyte in a perennial ryegrass breeding population which had undergone recurrent selection for endophyte compatibility. It was hypothesized that an increase in endophyte infection attributed to selection will be observed. To meet the objectives, the early and late generation of the breeding population PGG04, which was infected with *Epichloë* sp. FaTG-3 strain AR501, were examined. Endophyte detection was conducted using three methods, namely, seed squash, tissue print-immunoblot, and endophyte genotyping using a microsatellite marker. More than 90% of the seeds of PGG04 harboured the endophyte, regardless of generation. Tiller immunoblotting, which in contrast to seed squash detects viable endophyte only, showed an increase in infection from ca. 5% to 33% between the early and late generations. These results indicate that positive selection for endophyte compatibility lead to an increase in the proportion of viable endophyte in the population. This study provides evidence supporting host genetic control of the association in the grass-endophyte interaction, and that this can be exploited in plant breeding programs.

2.2. Introduction

The symbiotic relationship between *Lolium perenne* and its fungal endophyte, *Epichloë* spp. is one of the most important interactions in pastoral agriculture. This is because perennial ryegrass is the most extensively grown forage in the temperate world (Humphreys et al., 2010; Wilkins, 1991), and endophytes produce toxic alkaloids that can be detrimental to both livestock and insect pest. In New Zealand, endophytes partially mediate the grass' persistence because of improved plant defences against insect pest (Popay & Hume, 2011). Ryegrass species collectively contribute $14 billion (Nixon, 2015) while commercial endophytes are valued to provide an additional $200 million (Johnson et al., 2013) annually to the New Zealand economy.

For farmers to benefit from commercial endophytes, quality control measures are observed in the production of endophytic seed products (Rolston & Agee, 2007). For example, production fields are monitored for the percentage of viable endophyte as well as for contamination of the toxic endophyte (Rolston & Agee, 2007). In New Zealand, it is common to test tiller samples with the tissue-print immunoblot (TPIB) technique to measure the proportion of viable endophyte in cultivar populations (Card et al., 2011). TPIB detection is based on the ability of endophyte proteins to bind to specific antibodies (Hahn et al., 2003; Simpson et al., 2012). It detects live endophyte developing simultaneously with the growing seedlings. Because it is a "grow-out" test, it requires time and physical resources (Card et al., 2011). Therefore, a quick endophyte detection assay commonly used in freshly harvested is the seed squash assay (Card et al., 2011) which relies on histological staining and microscopy (Latch & Vaughan, 1995). However, this assay detects both live and dead endophytes. Because it is direct observation of fungal hyphae, false positive detection occurs when there are contaminant fungi or when endophyte is present but not in the embryonic tissues, hence (Card et al., 2011) improved on this detection method.

Endophyte viability, which declines more rapidly than seed viability (Hume et al., 2011), is affected by several factors including genetics, environment, and management (Freitas, 2017). For example, during seed production, judicious use of fungicides for disease control is important because not all fungicides are safe for endophytes (Rolston & Agee, 2007). At harvest, the seed moisture content is an important factor both for seed and endophyte viability. During seed storage, high temperature and moisture can negatively affect not only the seed but the endophyte as well (Hume et al., 2013). The effects of these factors vary depending on the grass-endophyte combinations; hence endophyte compatibility has genetic component both in the fungi and the host. For example, a seed storage experiment reported a higher proportion of viable tall fescue endophyte (i.e. AR501) in tall fescue seeds than in perennial ryegrass, under different temperature and seed moisture levels (Freitas, 2017). This observation is expected since endophytes exhibit host specificity (Karimi et al., 2012), hence, breeding for endophyte compatibility is conducted. In this study, genetic factors in the host were investigated.

The plant-fungal interaction is partially under genetic control from the host and this is exploited in perennial ryegrass breeding. The genetic background of the host is known to influence the ability of the endophyte to produce alkaloids (Adcock et al., 1997; Easton et al., 2002) and both endophyte biomass (i.e. amount of mycelium) and alkaloid concentration have been shown to be highly heritable traits and is largely controlled by additive genetic effects (i.e. general combining ability) (Easton et al., 2002). These traits are also correlated with alkaloid concentration as a function of mycelial mass (MM) (Easton et al., 2002). Although, this relationship is stronger in some alkaloids, such as peramine, than others. These findings are supported by a quantitative trait locus (QTL) study (Faville et al., 2015) that identified common QTL (genomic regions) controlling both traits, which suggests that they may be pleiotropic. Not all of the alkaloid and MM QTLs co-located which suggests that there are unique genetic factors influencing alkaloid concentration. Furthermore, MM-peramine and MM-independent ergovaline QTLs were detected indicating that bi-directional selection for alkaloid concentration could be achieved (Faville et al., 2015); i.e. selection for an increased level of the animal-safe peramine and selection against the animal-toxic ergovaline may be effectively accomplished at the same time.

Host genetic control of the association also affects the vertical transmission of fungal endophyte. This was observed in perennial ryegrass populations infected with the *E. festucae* var. *lolii* strain AR37 (Gagic et al., 2018). As expected, populations that have undergone selection for improved AR37 infection showed a higher percentage of viable endophyte than those that have not. The same study showed that there is strong genotypic effect on endophyte transmission which was also reflected in the relatively high predictive ability of a genomic prediction model developed for this trait (Gagic et al., 2018). Genomic selection, therefore, may be a promising breeding approach to improve endophyte compatibility for novel associations. In the present study, the endophyte transmission efficiency was also investigated but in the context of a cross-species interaction. The transmission of a tall fescue endophyte in an *L. perenne* breeding population that has undergone recurrent selection for improved endophyte compatibility was examined. The objective of the study was to investigate the impact of positive recurrent selection on endophyte compatibility traits, principally by assessing changes in endophyte infection frequency over selection cycles. It was expected to observe an increase in endophyte infection attributed to selection.

2.3. Materials and methods

2.3.1. Plant and endophyte materials

PGG04, a perennial ryegrass breeding population that has undergone recurrent selection (RS) for endophyte compatibility, was chosen for experimentation. The polycross method, and a combination of phenotypic and half-sib family selection were employed to improve the population. Polycrossing is a hybridization procedure employed in outcrossing crops that are vegetatively propagated. It takes advantage of natural hybridization to intermate a number of parental genotypes (Fehr, 1991). PGG04 population was bred by PGG Wrightson Seeds Limited and was derived from a restricted base population. The population was initiated with a cross between plants of two known commercial cultivars namely Extreme and Arrow (K. Saulsbury, personal communication, 2018). The Extreme plants were the maternal parents in the cross (seed was harvested from these plants alone) and had previously been artificially infected (Fig. 2.1) with the tall fescue endophyte *Epichloë* sp. FaTG-3 strain AR501 by AgResearch (K. Saulsbury, personal communication, 2018) following the protocol of Latchs and Christensen (1985). Therefore, the progeny from this cross was infected with AR501. A combination of phenotypic and half-sib family selection was employed for six cycles (C6) of recurrent selection. At each cycle of 1-2 years, there was phenotypic selection for agronomic ‘fitness’. In addition, each polycross was harvested as single plants, and each half-sib family was checked for the transmission of viable endophyte. On average, a blend of the top 10% transmitters was selected and sown for the following cycle, although the actual selection intensity varied (K. Saulsbury, personal communication, 2018). Endophyte compatibility was assessed in terms of viability and transmission. Viability describes the ability of the endophytes to remain viable in seeds under storage, while transmission refers to the ability of the endophyte to be transmitted from parents to offspring or from one generation to the next (M. J. Faville, personal communication, 2018).

The endophyte used in this study represents a class of novel endophytes that is not naturally associated with perennial ryegrass. Endophyte strains for perennial ryegrass have been studied in the common species, *E. festucae var. lolii* (i.e LpTG-1) and the less common *Epichloë* sp. LpTG-2 (*Lolium perrene* Taxonomic group 2) and LpTG-3 (Johnson et al., 2013; van Zijll de Jong et al., 2008). By contrast, AR501 is not an *L. perrene* endophyte but of tall fescue (*Festuca arundinacea*) and is relatively less common, i.e. it belongs to FaTG-3 (*F. arundinacea* Taxonomic group 3), as opposed to the more common *E. coenophiala,* i.e. FaTG-1. It was isolated from tall fescue collected in Southern Spain and classified initially based on isozyme phenotypes (Christensen et al., 1993). FaTG-3 is believed to have originated from the hybridization of an *E. typhina*-like endophyte and an endophyte in the *Lolium*-associated clade (Moon et al., 2004). While novel associations of endophytes with forage grasses have been exploited, it is generally limited to associations occurring within the normal host species range, specifically perennial ryegrass with *E. festucae* var. *lolii* strains and other LpTGs; or tall fescue with *E. coenophiala* and other FaTGs (Johnson et al., 2013; Young et al., 2014). Nevertheless, exploitation of novel association of AR501 has been reported in a few studies (Easton et al., 2007; Fletcher, 2012; Freitas, 2017). AR501 does not produce the typical alkaloid profile found in infected perennial ryegrass (Young et al., 2014). Most importantly, it produces lolines, a class of alkaloids not produced in perennial ryegrass infected with its natural endophyte. Lolines are reported to have wide insecticidal properties but are safe for livestock (Schardl et al., 2007). Similar to perennial ryegrass endophytes, AR501 also produces peramine, an alkaloid known to deter Argentine stem weevil (Rowan & Gaynor, 1986). AR501 has also been reported to offer some deterrence against black beetle adults (Popay et al., 2005), pasture mealy bug (Pennell & Ball, 1999), root aphid (Popay & Jensen, 2005), corn flea beetle (Ball et al., 2011), and fall armyworm (in infected perennial ryegrass but not tall fescue) (Ball et al., 2006).

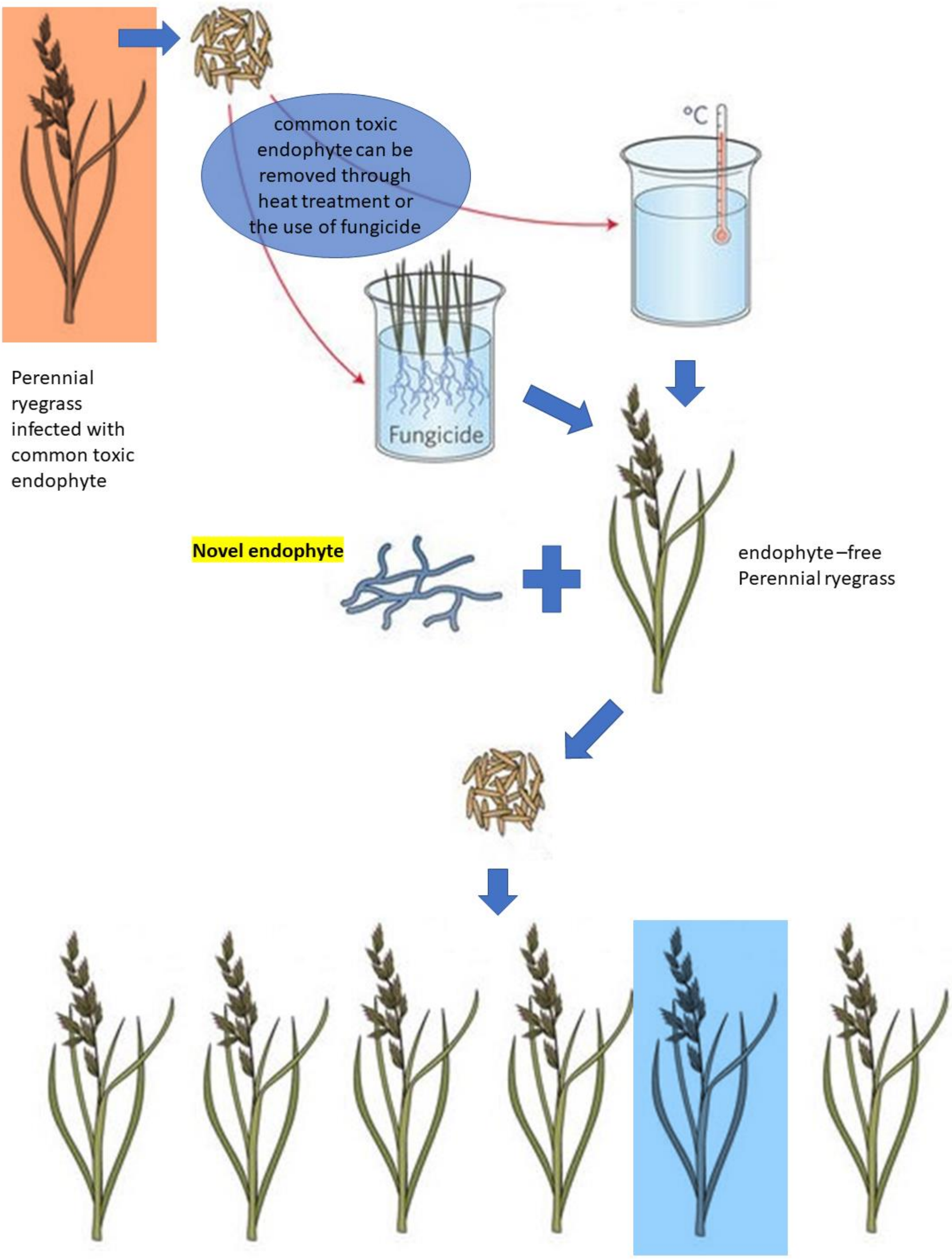

common toxic endophyte can be removed through heat treatment or the use of fungicide
°C
Fungicide
Perennial ryegrass infected with common toxic endophyte
Novel endophyte
endophyte –free Perennial ryegrass

Figure 2.1. Generating novel grass-endophyte associations. Adapted from Kauppinen et al. (2016). Plants infected with common toxic endophytes (highlighted in orange-filled box) can be subjected to heat or fungicide treatment to remove the native endophyte, after which a novel endophyte can be introduced. The resulting association will not always exhibit compatibility, as exemplified by just one highlighted plant (blue-filled box) at the bottom (i.e. successful infection).

2.3.2. Growing and maintaining the plants

Two generations of PGG04 were grown for the experiment. These were an early and a late generation, specifically the second (C2) and sixth (C6) selection cycles. For each generation, 144 seeds were sown in seedling trays containing seed raising soil mix. The soil comes from a two cubic metre mix of 6 parts of fine bark, 1 part of fine pumice, and 1 part of coir fibre. This mix was also fertilized with 6 kg of Osmocote® Pro 3-4M, 3 kg of dolomite lime, 6 kg of agricultural lime, 2 kg of gypsum, and 2 kg of a soil conditioner (e.g. Permawet^{TM}). Osmocote® Pro 3-4M is a slow release synthetic fertilizer that contains 17% N, 11% P, 10% K, 2% MgO and trace elements: Fe (chelated), Mn, B, Cu, Mo, and Zn. In addition, the soil mix naturally contains (i.e from other components) Ca, Mg, and Na. The plants were kept in a glasshouse on the AgResearch Palmerston North campus with controlled irrigation. The plants showed high germination rates: 78.47% for PGG04-C2, and 81.25% for PGG04-C6. Pests and disease symptoms were monitored, and control measures were applied when necessary. The plants showed signs of leaf drying especially in the older leaves. Nevertheless, examination of plant samples under the microscope did not show any signs of pathogens (Dr. Stuart Card, personal communication), and no insect damage was observed. Four weeks after sowing, 94 plants, corresponding to a 96-well sequencing plate (2 wells for control), were collected for each generation. Leaf and leaf sheath materials in the pseudostem were collected from individual plants. The pseudostem, which contains the endophytes, allowed for endophyte genotyping. After about four months, the plants have used almost all the nutrients in the soil (Osmocote® Pro 3-4M has longevity of 3-4 months), therefore they were transferred into larger pots containing a new soil mix. The mixture was composed of 1-part peat, 4-parts screened bark, 2-parts fine pumice, and 1-part coir fibre. It also contained the same amount of dolomite lime, agricultural lime, gypsum and soil conditioner as the seed raising mix. However, it had 10 kg of Osmocote® Pro 8-9M instead of the 3-4M type. This Osmocote® has the same nutrient as 3-4M except that it has 16% N and lasts longer (i.e. 8-9 months). The larger pot allowed the plants to grow bigger and produce more tillers, which were needed for an immunoblot assay.

2.3.3. DNA extraction

DNA was extracted from endophyte-rich pseudostems of perennial ryegrass plants following the protocol of Anderson et al. (2018). Two plates corresponding to the two generations namely, PG004-C2 and PG004-C6, were processed.

The first step of extraction involved grinding of leaf samples. For each sample, approximately 50 milligrams (mg) of leaf material was collected in a well of a Corning 1-millilitre (mL) deep well plate. Two 3.97-millimetre (mm) stainless steel beads were placed in each well of the plate containing leaf samples. The plates were then heat-sealed with Thermo Bond seal at 175 °C for 2.5 seconds (s). The sealed plates were floated in liquid nitrogen for 5 minutes (min) to freeze the samples. After this, they were ground in a tissue lyser (Qiagen TissueLyser II) at 20 Hertz (Hz) for 30 s. This was done twice to finely grind the samples. Samples were temporarily stored in a -80 °C freezer. The samples were then recovered for the succeeding procedures.

The second step was homogenization. The plates were spun down and the heat-sealed lid removed. Five hundred microliters (µL) of homogenization buffer and 1.8 µL of Proteinase K were added to each well. The homogenization buffer contains NaCl, Tris, EDTA, sodium sulphite, and sodium dodecyl sulphate (SDS). SDS dissolves the cell wall and cell membrane to free the cell nucleus containing genomic DNA. It also denatures the proteins in the cell. Proteinase K then degrades the unfolded proteins. After addition to the ground leaf materials, the plate was heat-sealed again. Then, the samples were mixed by shaking manually followed by centrifugation at 4000 x g (standard acceleration due to gravity) for 10 min.

The third step involved precipitation. After spinning, 300 µL of the solution from the previous step was transferred to a new Axygen 1mL 96-well plate. The same amount of precipitation buffer (potassium acetate and acetic acid) was then added and mixed in the new plate by shaking and inversion. Then, it was incubated in ice water for 15 min and centrifuged at maximum speed (8595 x g) for 30 min. The buffer facilitates the precipitation of cellular debris and SDS. It contains potassium acetate that precipitates dodecyl sulfate (DS) and DS-bond proteins leading to the removal of proteins from the DNA. Potassium acetate also acts as salt in the ethanol precipitation of DNA. Moreover, cell lysis occurs under alkaline conditions leading to denaturation of proteins and nucleic acids, and this is neutralized by potassium acetate to allow DNA precipitation.

In the fourth step, DNA was bound to a silica medium using a filter plate. Four hundred µL of the centrifuged solution and 600 µL of binding buffer were transferred to a Pall AcroPrep Advance 96 1 mL filter plate with a collection plate fitted at the bottom. The filter plate contains a silica-based quartz glass fibre media that allows for efficient binding of DNA. The binding buffer contains guanidium chloride, TE (Tris EDTA), and ethanol. Guanidium chloride is a chaotropic salt that disrupts the association of nucleic acids with water which affords their transfer to silica. It also contains ethanol which enhances binding of DNA in the filter media. It was then spun down at 4000 x g for 2 min and the flow-through in the collection plate discarded.

The next step involved washing away protein and salt residues as well as other impurities. Three hundred µL of binding buffer was added to the filter plate and centrifuged at 4000 x g for 2 min. Next, 300 µL of wash buffer (NaCl, Tris-HCL, ethanol and water) was added and the plate centrifuged again at 4000 x g for 2 min. This buffer contains NaCl and ethanol that makes SDS soluble and thus removes the SDS with the flow-through. Lastly, 300 µL of absolute ethanol was added for the final wash and centrifuged again at 4000 x g for 2 min. The flow-through in the collection plate was likewise discarded.

The filter media was dried from ethanol by spin-drying at 4000 x g for 5 min. The final step was the elution of DNA. The collection plate was replaced by a new 1 mL 96-well Axygen plate and fitted in the filter plate. One hundred fifteen µL of Tris/RNAse A mixture was added to the filter plate. The mixture was prepared by mixing 4 µL RNAse A to 12 mL of 10 millimolar (mM) Tris, for every plate. The filter plate fitted with the collection plate was centrifuged at 4000 x g for 1 min. The flow-through in the collection plate contained the final extracted DNA. DNA was released from the silica with 10 mM Tris, which is more basic than water. DNA will dissolve in Tris and is more stable under a slightly alkaline environment.

2.3.4. Microsatellite analysis of endophyte DNA

Endophyte genotyping was conducted with the polymerase chain reaction (PCR) using the microsatellite B11 (Moon et al., 1999). This was conducted to ensure that the correct endophyte strain was present in PGG04 and that there was no contamination (i.e. from the common toxic endophyte). B11 was shown to resolve endophyte groupings complementary to isozyme phenotypes reported by Christensen et al. (1993) (Moon et al., 1999). The expected PCR product size was 127 bp (Faville et al., personal communication, 2018) for the B11 locus in AR501, although results can vary by a few base pairs depending on the size standards, the dye used, the conditions of capillary electrophoresis, and the capillary electrophoresis instrument itself. AR501 and AR106 DNA were included to serve as positive controls, and a blank (water) to serve as a negative control. One of the primers of B11 (i.e. B11.1) was fluorescently labelled with 6-carboxyfluorescein (6-FAM) (Faville et al., personal communication, 2018) as opposed to its old label HEX in Moon et al. (1999). The label allows for automated microsatellite analysis using an ABI 3730 DNA Analyzer (Applied Biosystems).

After extraction, DNA was diluted in water at 1:10, and 10 μL of PCR reaction mixture was prepared for each sample. Each reaction was composed of 4 μL of water, 1 μL of 10x PCR buffer (Mg^{2+} free), 0.3 μL of $MgCl_2$ (25 mM), 0.1 μL of dNTPs (2.5 mM each), 0.4 μL of B11 primer mix, and 0.2 μL of *Taq* polymerase combined with 4 μL of the diluted DNA. The B11 locus was then amplified using the thermal cycling protocol of: initial denaturation at 94 °C for 4 min; 35 cycles of denaturation, annealing and extension at 94 °C, 60 °C and 72°C respectively, for 30 s each; and a final extension at 72 °C for 7 min. The PCR results were validated by running 12 samples in 2% agarose gel electrophoresis at 75 Volts for 1.5 hours. The expected PCR products (and negative) is shown below (Fig. 2.2). After amplification, it was prepared for microsatellite analysis by combining the PCR product with Hi-Di formamide and size standard. First, 1.5 μL of CASS size standards (Symonds & Lloyd, 2004): 100 bp; 200 bp; 300 bp; and 400 bp, were combined with 6 μL of water. Second, the size standard mixture was mixed with 500 μL of Hi-Di formamide. Mixing was accomplished by flicking followed by quickly spinning in the centrifuge. Next, 9 μL of this mixture was added to 1 μL of the PCR product for each sample. It was then spun down and finally sent to the Massey Genome Service for microsatellite analysis. The results were visualized using Genemapper. The expected product size of B11 for AR501 was 127 bp and thus, a positive infection is the presence of a peak signal at around this size. The actual size was based on the peak detected in the control, that is, the AR501 DNA.

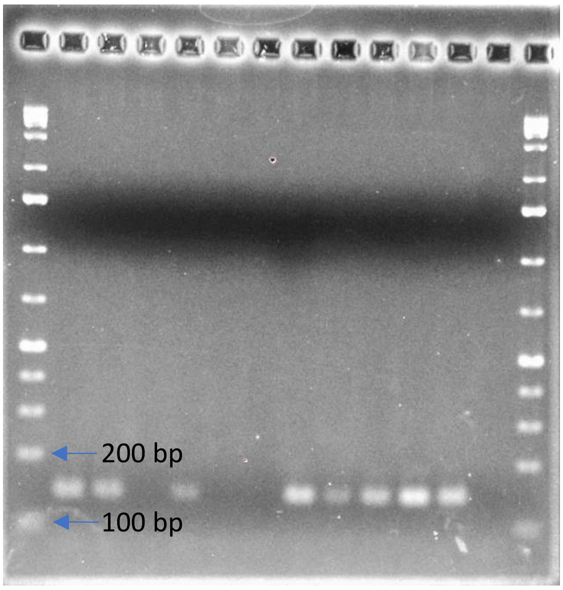


Figure 2.2. PCR products of the B11 locus in some PGG04 samples. The gel confirmed successful PCR as well as the infection of the samples based on tissue print-immunoblot (TPIB) assay. As expected, TIPB-positive samples, namely, lanes 1, 2, 4, 7 – 11 showed the expected product size while the rest were negative.

2.3.5. Endophyte detection by immunoblotting

The performed tissue print-immunoblot (TPIB) assay was based on the protocols of Hahn et al. (2003) and Simpson et al. (2012). Individual grass plants were cut at the base ca. 5 mm, leaving the pseudostem exposed. Tillers were cleaned by removing dirt and necrotic tissues before another transverse cut. The cut side of the pseudostem was then placed on a nitrocellulose membrane (0.45 μm), imprinting plant sap of the tissue into the blotting medium, and leaving a circular mark. This was done for three tillers of each plant in the two generations. The blotted medium was sent to AgResearch Ruakura Agricultural Centre for processing.

TPIB makes use of antibodies that bind with endophyte proteins for detection. The primary antibody used was a rabbit anti-endophyte produced at AgResearch in conjunction with Massey University's Small Animal Production Unit. A secondary antibody, goat anti-rabbit IgG-AP, was also used. Chromogen label was utilized to detect the signal of antibodies binding to endophyte proteins. It was based on a Fast Red TR/Naphthol AS-MX and TR phosphate substrate system that produces deep red for a positive reaction. In addition, a milk protein blocking solution was used to block the surfaces with no bound protein.

2.3.6. Endophyte detection in the seeds

In addition to the microsatellite and TPIB methods, the seed squash (SS) method was also performed to visualize endophytes in the seed (Latch & Vaughan, 1995). Since this is a destructive method, a different set of seeds from the previous two methods were used. Nevertheless, they were similarly random samples of the same generations, PGG04-C2 and PGG04-C6. The seeds were first soaked overnight in 5% aqueous sodium hydroxide to kill the seeds and fix the cells. After this, they were washed with water to remove alkali. Garner's stain was then added to the rinsed seeds. The stain was prepared by combining 0.325 g of aniline blue, 50 mL of 85% lactic acid in 100 mL of water. The seeds soaked in Garner's stain were heated until boiling and then allowed to cool down. The seeds were then separated on a glass slide, for each generation. Each seed was deglumed using a scalpel and a pair of tweezers under a stereo microscope and mounted in a glass slide. A small drop of aniline blue stain was added to each seed and slide cover was added on top. The aniline blue stain was prepared with 1-part lactic acid, 2-parts glycerol, 1-part water, and 0.05% aniline blue. The cover was then gently pressed to squash the seed for a flat-mounted specimen. The slides were viewed under a compound microscope at 10x and 40x stage objective. Endophyte can be found mostly in the embryo and nucellus, which is just above the aleurone layer, in the seed. Since the aleurone layer stains distinctively as large, square, blue cells, they were used as a guide. Endophyte mycelium growing in the nucellus can be observed as a convolute stained filament, just above the aleurone cells. Infection frequency was then estimated based on the endophyte-positive seeds (as described above) as a proportion of the total number of seeds examined.

2.3.7. Data analysis

The three endophyte detection assays generated binary data, that is, positive or negative infection. In the microsatellite analysis, a positive infection is a peak fluorescence signal as determined by the positive control, the AR501 DNA. It was also determined whether other fluorescence signals were present. For example, B11 reportedly have PCR product sizes from 115 – 240 bp (Moon et al., 1999) and so fluorescence signals were examined at this size range. For the specific detection of the common toxic endophyte, the signal was based on the AR106 DNA control. For the TPIB, a positive infection was considered if at least 2 out of 3 tillers showed a deep red blot in the media. While for the SS method, a positive infection was the presence of fungal hyphae at the region of the aleurone cells as observed using microscopy. All statistical analyses were conducted using R (R Core Team, 2017). The TPIB and the SS data for the early and late generation were compared using a chi-squared test for equality of proportions using the prop.test function. In addition, the TPIB results were analyzed in terms of odds ratio with Fisher's exact test (i.e. fisher.test function) (R Core Team, 2017). A statistical test was not performed with microsatellite data since the individuals sampled were deliberately chosen, and therefore do not represent a random sample of each generation of PGG04. It was, however, compared with the TPIB data in terms of the percentage of agreement and Cohen's kappa, which is an inter-rater agreement statistic that takes into account agreement due to chance. This was computed using the R/irr package (Gamer et al., 2012).

2.4. Results

2.4.1. Immunoblotting

To perform immunoblotting, 144 seeds were sown for each generation of PGG04. Both have relatively high germination rate at 78.47% (113) for PGG04-C2 and 81.25% (117) for PGG04-C6. This was then thinned into 94 plants per generation. Based on the immunoblot, the early generation, PGG04-C2, has an infection rate of 5.38% (5 out of 93; one plant died before blotting) while the late generation, PGG04-C6, has an infection rate of 32.98% (31 out of 94). As mentioned in the methods section, an individual was scored as infected if at least two out three tillers had the positive signal of red staining in the blotting paper. For PGG04-C2, 3 plants had 3 tillers with a positive signal; 2 plants had 2 positive tillers; and 1 plant had 1 positive tiller. For PGG04-C6, 29 plants had 3 positive tillers; 2 plants had 2 positive tillers; and 2 plants had 1 positive tiller. For both generations, there were some plants that showed staining but were not scoreable (i.e. faint or very little stained area), for example, PGG04-C2-H10 and PGG04-C6-A6. The result of TPIB was presented in a bright background as well as an 8-bit grayscale format for ease of scoring (Figures 2.3 and 2.4). The rate of infection was significantly higher in PGG04-C6 than in PGG04-C2 based on a chi-squared test of equality of proportion (p-value = 2.1e-06).

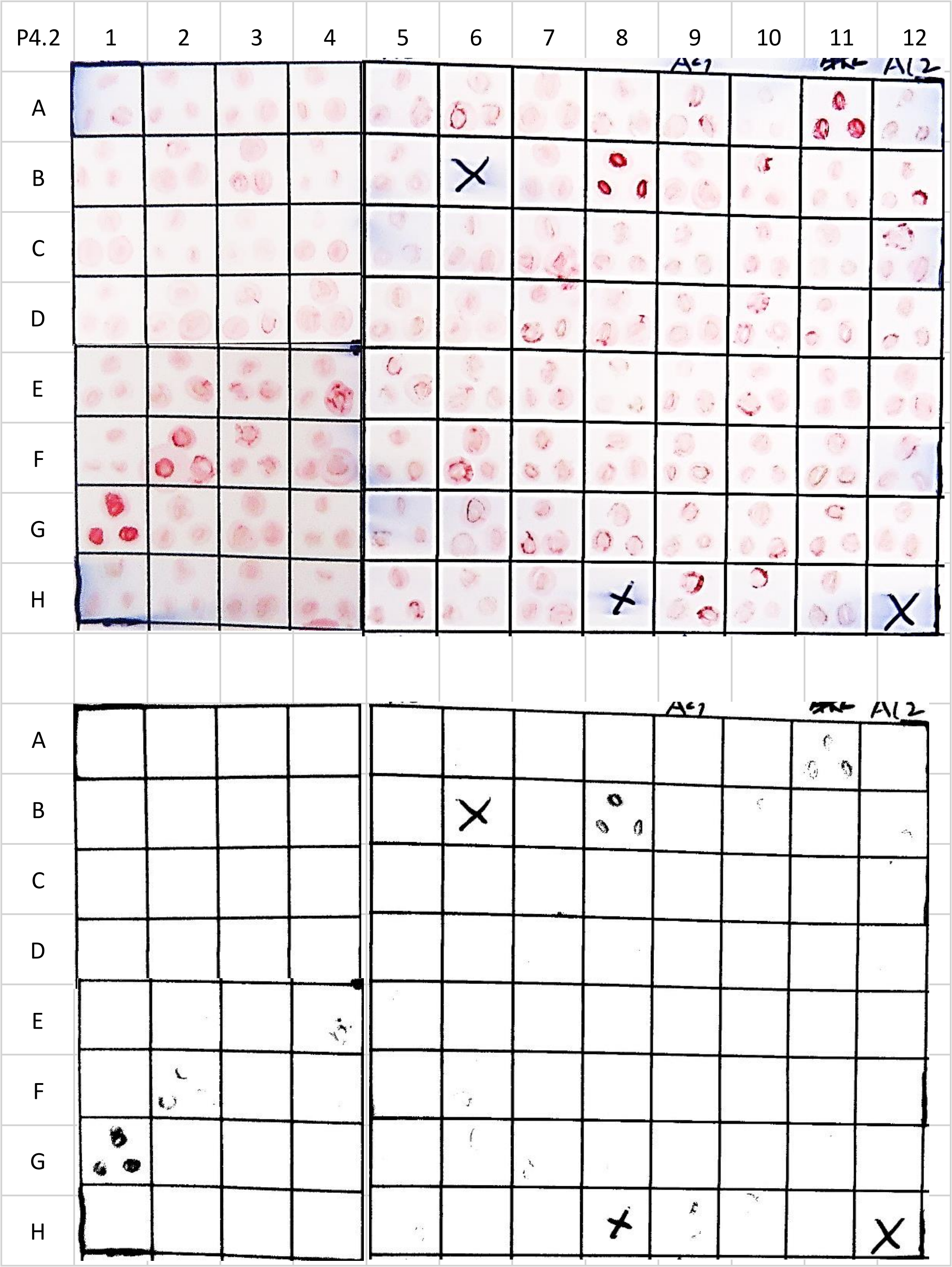


Figure 2.3. Immunoblot endophyte detection in PGG04-C2. Three dots correspond to the three tillers blotted and the red (or black in grayscale) staining corresponds to positive infection. The top image has been improved for sharpness, contrast and brightness while the bottom image was an 8-bit grayscale version.

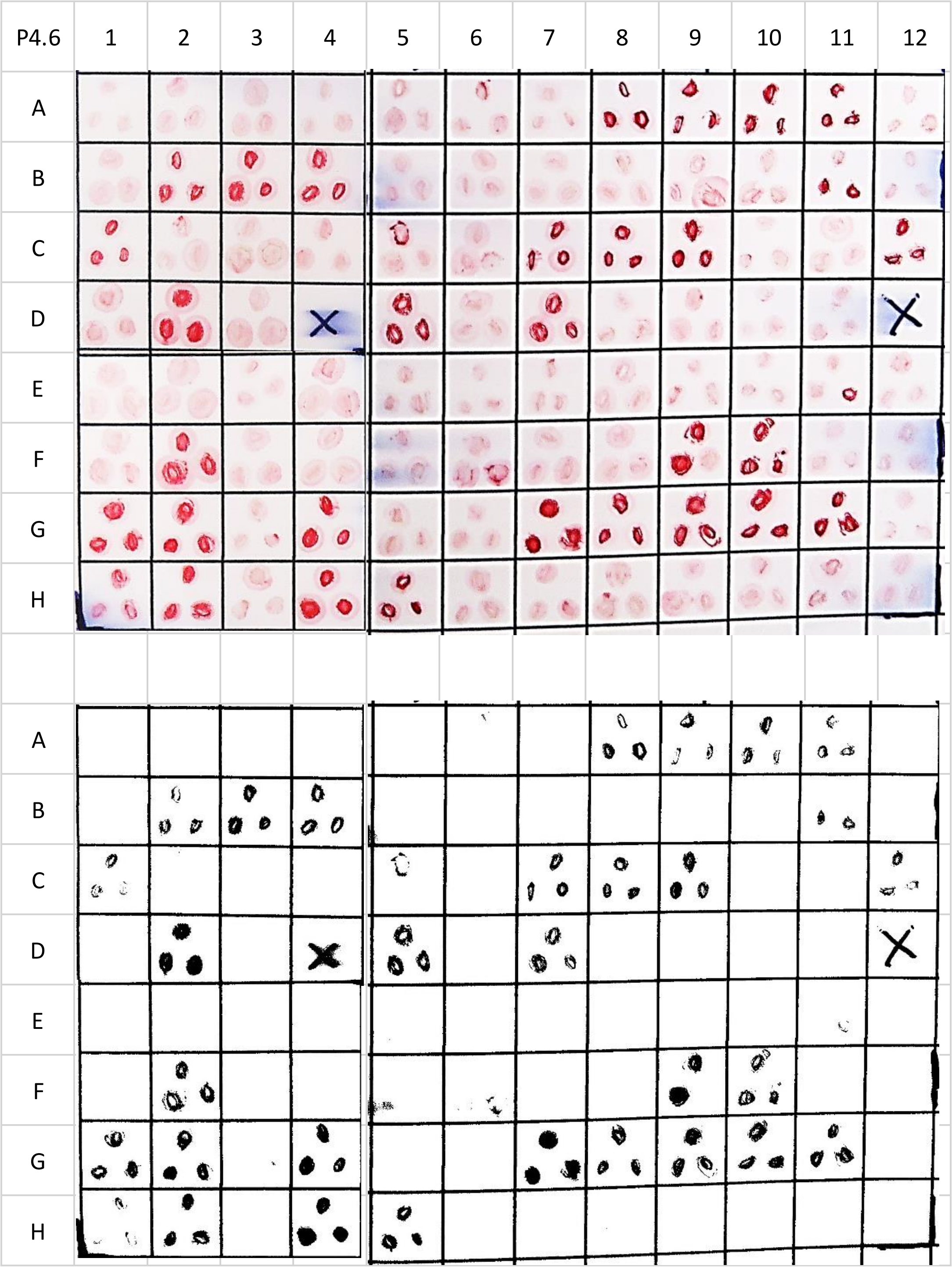


Figure 2.4. Immunoblot endophyte detection in PGG04-C6. Three dots correspond to the three tillers blotted and the red (or black in grayscale) staining corresponds to positive infection. The top image has been improved for sharpness, contrast and brightness while the bottom image was an 8-bit grayscale version.

2.4.2. Endophyte genotyping

The endophyte-plant associations were genotyped using the endophyte-specific B11 microsatellite for two reasons. First, genotyping results were used to corroborate the immunoblotting results. Second, it can confirm the identity of the endophyte strain. The individual plants to be genotyped were preselected because it was not practical to genotype all the uninfected plants (based on TPIB) as there will be no PCR product. A total of 90 plants, which includes 19 plants from PGG04-C2 and 71 from PGG04-C6 were genotyped. All plants that had at least one immunoblot-positive tiller were included. Plants with apparent staining or those that were relatively lightly stained and at a smaller area, were also included. AR501 DNA was also included as a control as well as the NZ common toxic endophyte, namely AR106, which have expected PCR product sizes of 127 and 177 bp, respectively, based on LIZ500 size standard. Using the CASS ladder, about 3-6 bp decrease in the expected sizes were observed. Specifically, a peak signal between 124-125 bp was detected in the AR501 DNA control and PGG04 infected plants, while the peak signal was between 171-174 bp for AR106 (Fig. 2.5). Although a secondary peak at 125-126 bp in AR501 and two other signals around 168 and 169 bp in AR106 were observed, these secondary signals were of lower intensity compared to the main peaks. A clear peak signal (about 20000-30000 fluorescence units) around 124-125 bp were observed in almost all TPIB-positive plants for both generations. A total of 35 plants were found to be positive by microsatellite: 32 of which were from PGG04-C6; and three from PGG04-C2. Examples of genotyping results were shown in Fig. 2.6. There were some cases of a relatively weak signal (20,000) in the 124-125 bp region, such was the case of the sixth graph (PGG04-C6-G8) in Fig. 2.6, but a peak was still evident, and the plant was consequently called as endophyte positive. In general, the genotyping results were consistent with that of TPIB with only few instances of mismatching scores. Two plants in the PGG04-C2 were classified as infected by TPIB but did not show the 124-125 bp signal with B11 genotyping. These were PGG04-C2-H9 (1), which showed two positive tillers by TPIB and PGG04-C2-F2 (2), which was reported to have three positive tillers based on TPIB (Fig. 2.3). For PGG04-C6, there were three notable plants. First, the PGG04-C6-F6, which showed only ‘apparent’ TPIB staining of its tiller, and did not show the 124 bp signal, confirming that it does not have the endophyte. Second, the PGG04-C6-C5 which clearly had only one positively stained tiller, therefore classed as “uninfected” by our criteria, and was also negative for genotyping. Thirdly, PGG04-C6-E11, which was scored as “uninfected” in TPIB since it only has one positive tiller, but genotyping results tell that it harbors the endophyte. Considering the scoring criteria, the two methods agreed in 87 out of 90 scores or a 96.67% match. Because that agreement could also be due to chance, an inter-rater agreement statistic, kappa (κ) was also computed. The results showed an almost complete agreement (i.e. when κ =1.0) with inter-rater agreement of 0.93. Thus, endophyte genotyping validated the result of TPIB. Similarly, the former strengthens the findings of the latter. At about 110 – 180 bp, no other clear signals, such as at 171-174 bp (AR106), was observed for any of the tested plants. Thus, it's generally either the 124-125 bp signal or none at all. The only exception was a very weak signal observed in PGG04-C6-F11. There were two small peaks between 155-160 bp with 6000 fluorescence units (71st graph in Appendix Fig. 2.1.), which was very weak compared to the AR501 signal (124-125 bp) at more than 30,000 units or the relatively weaker signal of more than 20,000 units (i.e. sixth graph in Fig. 2.6). Therefore, plants generally exhibited either a nil endophyte or AR501, and there was no evidence of contamination by any other endophyte strain especially the NZ common toxic endophyte.

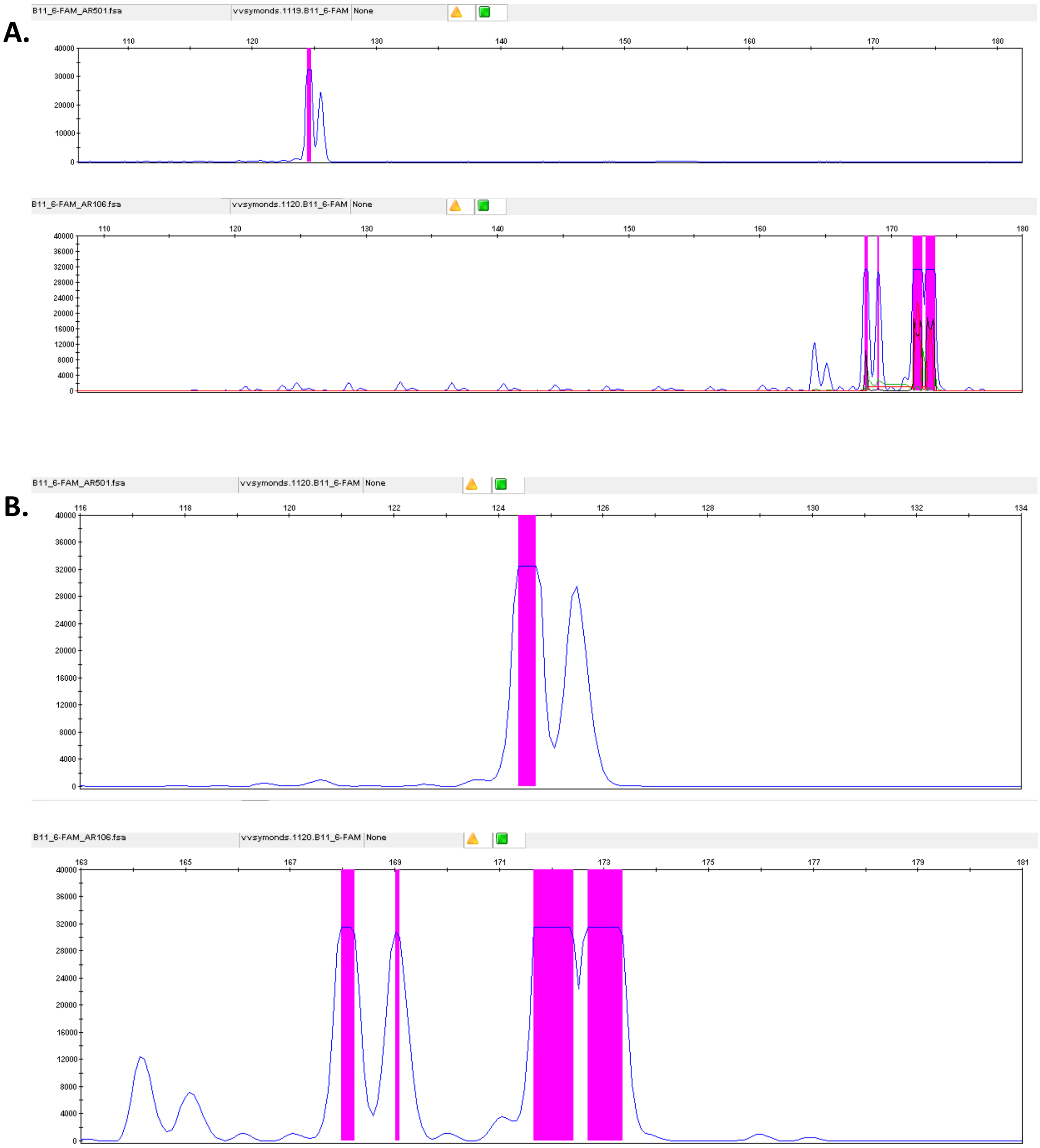


Figure 2.5. Fluorescence signals of the endophytes AR501 and AR106 (NZ common toxic strain). A shows discrete peaks for AR501 (top) and AR106 (bottom) within the 110 – 180 bp range. B shows peaks for the respective endophyte strains at a closer scale. These are 124-125 bp for AR501 (top) and 171-174 bp for AR106 (bottom). Note the presence of lower intensity, secondary peaks.

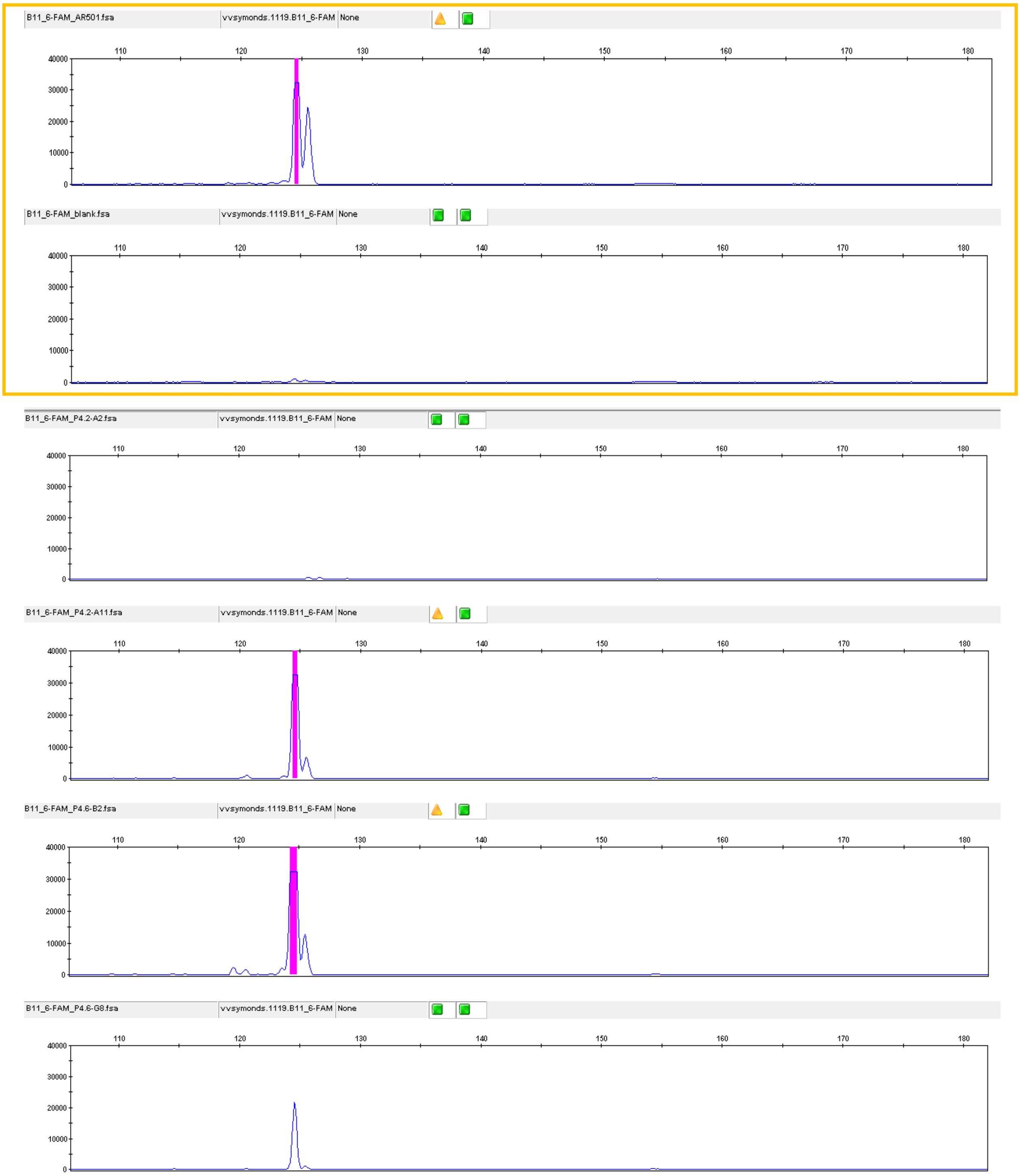


Figure 2.6. Fluorescence signals at 124-125 bp (AR501) in four plant samples and controls. The positive (first) and negative (second) controls are highlighted by a yellow box. Overall, plants exhibited either a nil endophyte or an AR501-positive signal by microsatellite.

2.4.3. Seed squash assay

The presence of AR501 in PGG04 seeds was also determined to compare its occurrence in the tillers. Endophytes in the growing tillers represent viable or living endophyte, whereas endophytes in the seed could either be alive or non-viable. The endophyte's hyphae were observed as convolute stained filaments above the blue blocky aleurone cells in the seeds. Representative seed sample cross sections showing positive and negative infection are shown in Fig.2.7 A. Also shown is a 400x magnification of aleurone cells in the seeds colonized by endophytic hyphae (Fig. 2.7 B.) The results of the seed squash (SS) assay were 93.55% (58 out of 62) infection rate in PGG04-C2 seeds and 96.36% (106 out of 110) in PGG04-C6 seeds. These infection rates were statistically similar with a p-value = 0.6421 based on a chi-squared test for equality of proportions with continuity correction. Thus, almost all the PGG04 seeds harbor endophytes, regardless of the generation. In most cases of looking for the endophytes under the microscope, they can be easily observed because they were widespread in the seed. There were a few cases where the endophyte had limited distribution. Therefore, in these cases, it took a relatively longer time to scan the entire aleurone layer to find the endophytes. Nevertheless, the scores were confirmed by an expert independently.

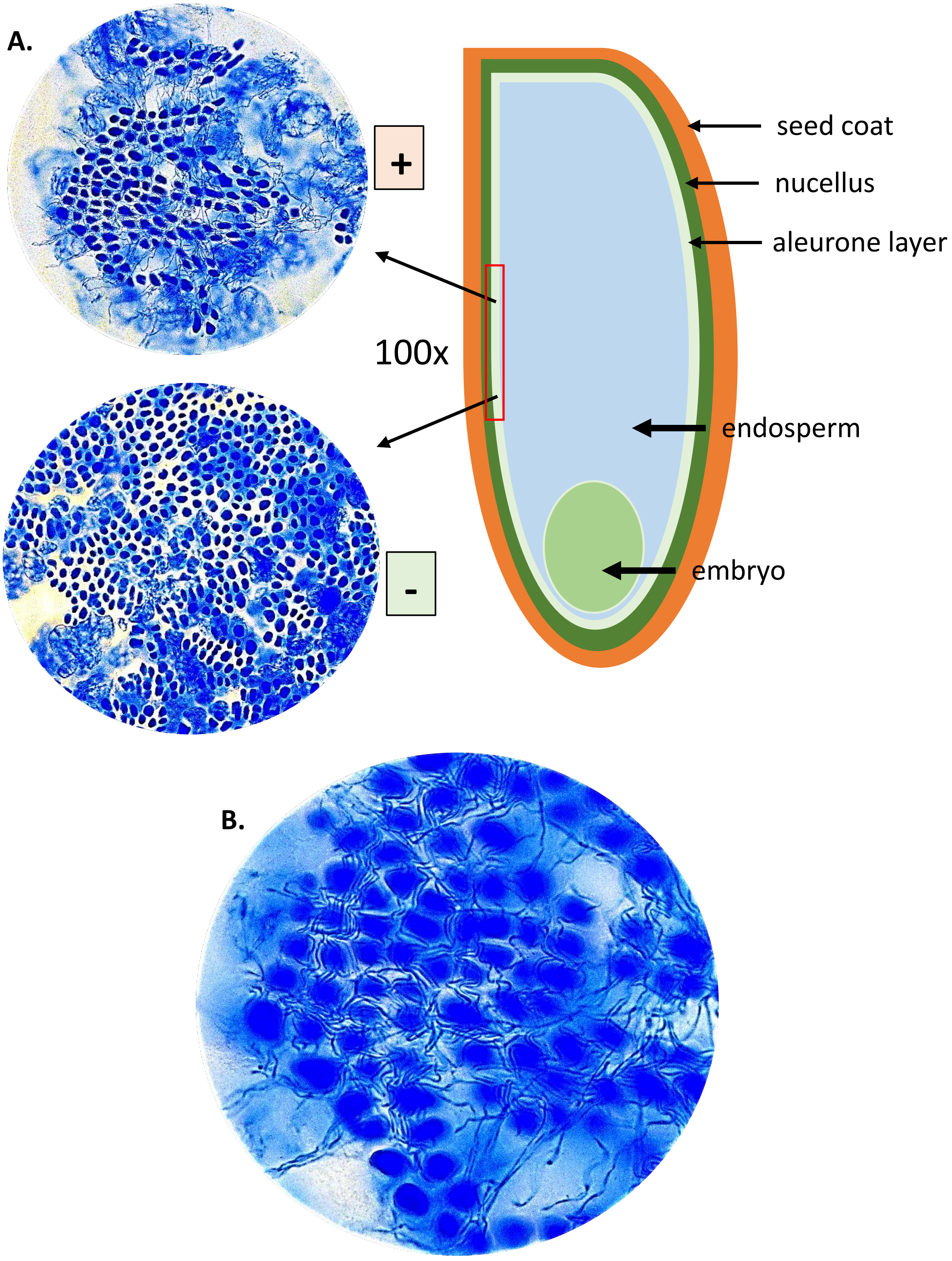


Figure 2.7. Aleurone cells under the microscope at 100x magnification showing both infected (+) and uninfected (-) seeds (A) and 400x magnification of the positive infection (B). Endophytic hyphae (convoluted filaments) sitting above aleurone cells (blue "blocks") can be more clearly observe in B.

2.4.4. Endophyte detection in PGG04 using three methods

The results of three endophyte detection methods are summarized in the table below (Table 2.1). A high proportion of PGG04 seeds, both the early (93.55%) and late (96.36%) generations, harbor AR501. This includes both viable and dead endophytes as seed squash assay cannot distinguish either. In contrast, the results of tissue print-immunoblotting (TPIB) showed an increase in the proportion of viable AR501 in PGG04 after four cycles of selection. This was supported by a significant chi-squared test for equality of proportions considering an alternative hypothesis that infection rate was greater in PGG04-C6 than in PGG04-C2. The infection rate can also be analyzed in terms of odds-ratio (OR), that is, a null hypothesis that the odds of infection in C6 is equal to that of C2 or OR = 1. The result of Fisher's exact test rejects this null hypothesis (p-value = 1.48e-06). It was found that the odds of endophyte infection among plants in the late generation was more than 8.57 (95% confidence interval of 3.07 – 29.80) times higher than the odds of infection among plants in the early generation. In other words, four cycles of selection for endophyte compatibility favorably increased the odds of endophyte infection by more than eight-fold.

Table 2.1. AR501 infection in PGG04 using three detection methods. Almost all the PGG04 seeds, regardless of generation, harbor the endophyte. In contrast, immunoblotting showed an increase in endophyte infection in the growing tillers from early to late generation of PGG04.

Detection method	**PGG04-C2** (early generation)	**PGG04-C6** (late generation)	**Chi-squared test**	
			Hypothesis	**P-value**
Immunoblot	5.38% (5 out of 93)	32.98% (31 out of 94)	C6 > C2	2.10e-06
B11 genotyping	3 out of 19	32 out of 71	not applicable	
Immunoblot corrected[1]	3.23% (3 out of 93)	32.98% (31 out of 94)	not conducted	
seed squash	93.55% (58 out of 62)	96.36% (106 out of 110)	C2 = C6	0.6421

[1] Corrected with genotyping results but still needs at least two TPIB-positive tillers.

As stated earlier, endophyte genotyping results generally agree with immunoblotting results. The few inconsistencies between TPIB and endophyte genotyping results can be explained by imperfect scoring in TPIB. In PGG04-C2, mismatching scores were from PGG04-C2-H9 and PGG04-C2-F2. They were scored as infected with the TPIB and showed 2 and 3 positive tillers, respectively (Fig. 2.3). However, upon closer inspection, we can see that relative to other positive tillers, such as PGG04-C2-G1, the stained tillers of H9 and F2 were fainter as opposed to deep red and were smaller, being limited only to some part of the circular outline. In the late generation, PGG04-C6-C5 and PGG04-C6-E11 were interesting (Fig. 2.4). PGG04-C6-C5 has one tiller harboring the endophyte based on TPIB but it was negative for genotyping. On the other hand, PGG04-C6-E11 has one blot-positive tiller and was positive based on genotyping. The second case was not surprising and was scored "uninfected" in TPIB simply because of an arbitrary criterion of at least 2 positive tillers. The first case could be a "real" inconsistency but can be explained by sampling. While it is common for endophytes to colonize all the tillers of its host, endophyte-free tillers are known to occur both in natural (Hinton & Bacon, 1985) and synthetic associations (Christensen, 1995). In the present study, cross-species association was investigated, hence, it is highly likely that not all tillers were colonized. Since the sampling procedure was conducted for blotting and genotyping separately, it is possible that the positive tillers were missed during leaf sampling for DNA extraction. While both plants harbor the endophyte, based on blotting and genotyping results, they provide weak support to the trait of interest, that is endophyte compatibility. This is because it was expected that the majority of tillers will be infected in an "endophyte-compatible" plant.

2.5. Discussion

Four cycles of selection in PGG04 resulted in a significant increase in the proportion of viable AR501 in the population based on immunoblotting. This result is further supported by a high level of agreement of TPIB and endophyte genotyping. As mentioned previously, the inconsistent results between the two methods can be explained by their imperfect scoring and sampling. Genotyping using B11, a highly informative microsatellite marker, also confirmed that AR501 is present in the population and that there was no contamination, especially with the common toxic endophyte. However, the infection rates, based on TPIB and confirmed by genotyping, were relatively low. For example, commercial cultivars are required to have at least 70% infection rate (Card et al., 2014). One explanation for the low infection rate is host incompatibility since perennial ryegrass is not a natural host for a tall fescue-derived endophyte such as AR501. It is widely known that most *Epichloë* spp. exhibit strong host-specificity (Karimi et al., 2012). The presence of endophyte in the seeds was also evaluated to determine if a similarly lower infection occurred. On the contrary, the result of the SS assay was that more than 90% of PGG004 seeds harbored the endophyte, regardless of generation. In contrast to TPIB, which will detect only live endophyte, the endophyte detection method in the seed does not distinguish viable and dead endophytes. Therefore, while the majority of the seeds contained endophyte, not all of them lead to successful colonization. One implication of this result is that selection leads to improvement in the proportion of viable endophyte in the breeding population. It suggests that even when host-specificity barriers are present, it is possible to improve compatibility by exploiting genetic variation in the host population. The result of TPIB was related to compatibility, hence the increase in the proportion of viable endophyte from C2 to C6 was evidence favoring improvement in endophyte transmission, after a few cycles of selection.

The discrepancy of endophytes detected in the seeds and seedlings could be affected by several other factors. The SS assay is imperfect and may overestimate the endophyte infection rate. This can be due to inadvertent detection of contaminant fungal species or the observation of endophyte hyphae in the aleurone layer, yet embryonic tissues remain uncolonized (Card et al., 2011) and therefore transmission does not occur. However, this was an unlikely explanation for our results as the discrepancy between the methods is too large to be explained by chance overestimation of one method over the other. For example, while Card et al. (2011) found that the infection rate in tall fescue with SS was significantly higher than with TPIB, they reported a discrepancy of only 6-7%. In the present study, the difference in infection rates between the seeds (SS) and seedlings (TPIB/genotyping) were about 88%, and 63% for PGG04-C2 and PGG04-C6, respectively. More importantly, the observation of fungal endophyte was verified by a trained expert.

The relative survival rate of endophyte-infected and non-infected seeds will determine the infection rate observed in the growing seedlings (Gundel et al., 2010). Endophytes may reduce or improve seed survival, widening the discrepancy of infection rate between seeds and seedlings. Germination failure of few infected seeds due to the presence of endophytes would certainly lead to a measurement of low infection rate in the grown seedlings. In this study, relatively high germination rates were observed with 78.47% for PGG04-C2 and 81.25% for PGG04-C6. Thus, the differential survival of infected and non-infected seeds was highly unlikely to have occurred.

Endophyte losses could be due to failure to colonize the growing tillers (i.e. pre-zygotic) or endophyte mortality in the seed (i.e. post-zygotic), both of which can be affected by factors including the environment, management, and genetics (Freitas, 2017). In our experiment, the plants were grown under uniform glasshouse conditions. If these conditions were unfavorable, we would therefore expect a uniform reduction in endophyte colonization of the growing tillers. Furthermore, seed crop management factors such as nitrogen fertilization, fungicide treatment and application of plant growth regulators, were reported to have no effect on AR501 transmission in perennial ryegrass (Freitas, 2017). Therefore, the failure of the endophyte to colonize the growing tillers, if this was the case, was most likely due to genetic factors. Any genetic differences in the population of host plants after four cycles of selection were likely to be related to endophyte infection rate. This is because direct selection applied in the population targeted improvement of endophyte compatibility.

Seed storage factors such as length, temperature, and relative humidity (RH) which is in equilibrium with seed moisture content (SMC), are known to influence endophyte mortality (Hume et al., 2013). In particular, AR501 viability decreases with high temperature and relative humidity, both in infected perennial ryegrass and tall fescue seeds (Freitas, 2017). These factors are likely to have been important in the present study. PGG04-C6 was from a 2017 seed harvest, while PGG04-C2 was extracted from cold storage after being harvested in 2013. It is possible, therefore, that the low proportion of viable endophyte in PGG04-C2 was due to the death of endophytes while in storage. Newly harvested seeds, such as the case of PGG04-C6, were likely to maintain a high proportion of viable endophytes for months. In the case of native endophytes, it was found that most perennial ryegrass seeds sold in Australia maintained the standard proportion of viable endophytes within two years after harvest (Wheatley et al., 2007). In addition, tall fescue in the US planted in autumn within three months of harvest were found to have a similar percentage of viable endophyte to what was observed in their seed production fields. This was the case for the tall fescue cultivar Jesup infected with the commercial endophyte MaxQ (*E. coenophiala*) (Rolston & Agee, 2007). More importantly, in New Zealand, it was reported that AR37 infection rate in newly harvested perennial ryegrass seeds and that of grown seedlings have a strong positive correlation, suggesting efficient seed-to-seedling endophyte transmission (Gagic et al., 2018). However, this is not necessarily true for all host-endophyte interactions, especially in this study considering a non-native endophyte. The early generation PGG04-C2 seedlings tested for viable endophyte were grown from seeds stored under a temperature controlled cold storage. Temperature is more economical to manage than RH, therefore cold storage is coupled with moisture-resistant packaging (Hume et al., 2013). This was how the PGG04-C2 seeds were stored. This storage strategy was proven to be effective as seeds stored inside aluminium-polyethylene laminated bags were found to maintain a high proportion of viable endophyte even up to 15 years of cold storage (-15 or 0 °C) (Rolston et al., 2002). A seed storage experiment (Freitas, 2017) found that a high level of viable AR501 can be maintained in perennial ryegrass seeds stored at low temperature (i.e. 5 – 10 °C) and 14% SMC**,** which is within the usual range of moisture content at harvest. This experiment however only lasted for a year while PGG04-C2 seeds were stored for nearly six years. Considering that cold storage coupled with moisture-resistant packaging is an effective long-term storage strategy (Rolston et al., 2002) and AR501 infected perennial ryegrass seeds respond positively to low temperature (Freitas, 2017), it is reasonable to assume that the storage conditions of PGG04 seeds were suitable for endophyte survival, although it was not optimal.

Under ideal conditions, host genetics play a bigger role than storage factors. For example, under low temperature (5 °C or 10 °C), AR501 survival was affected more by the host (tall fescue or perennial ryegrass) than by SMC (14% or 10%) (Freitas, 2017). Further, endophyte mortality is negatively correlated with seed weight and/or size (Card et al., 2014), which also have underlying genetic bases. The size and weight of the seed relate to the available energy. Moreover, in terms of resource use allocation, more expendable energy affords supporting endophyte maintenance. Without immediate benefit for the endophyte, such as during the absence of herbivores, the seed ought to invest more energy for plant development. In our experiment, the minimal stress level in the growing plants could possibly have resulted in the tradeoff between plant and endophyte development in favor of the former. However, in the present study, it is possible that the low infection rate in the growing seedlings was affected by endophyte mortality due to suboptimal seed storage conditions. Hence, the observed difference in infection rate between the seeds and seedlings may be equally explained by the effects of storage factors.

Our results indicate that the difference in infection rate was likely related to the genetic differences between generations and also to seed storage factors; teasing these two apart is difficult but the results seem to partly support the hypothesis that the selection regime improved endophyte transmission. The early generation seeds were harvested from infected C1 plants while C6 seeds were harvested from infected C5 plants. Selection and polycrossing of compatible plants meant that all seeds were harvested from mostly infected maternal plants. In addition, the pollen source would be mostly infected plants. Therefore, one possible explanation of the abundance of endophytes in the seeds was the successful colonization of the endophyte from the vegetative tillers of infected maternal plants to reproductive tillers and eventually, the seeds. However, infection does not necessarily translate to compatibility as infection may fail any time during host plant development. Infection data can be an indicator of compatibility (i.e. high infection) but may be only as good as the last time the endophyte was checked. The cross-pollination of the parents will result in a variety of host genetic backgrounds that may or may not be compatible for endophyte development, thus would affect the success of transmission. Some seed environments could have been hostile which would lead to endophyte mortality while the growing seedlings could have suppressed infection in their meristematic tissues. The higher number of individuals with viable endophytes observed in the late than the early generation is evidence for an improvement in endophyte transmission. Although most C2 seeds showed apparent infection, most of the endophytes were non-viable or have died. Aside from storage factors, this reflects the ability of C1 plants to transmit viable endophytes, which was low. For C6, a higher infection rate was observed; which similarly reflected the improvement in endophyte transmission of the C5 plants. This was likely the result of a few cycles of selection for endophyte compatibility. Results of this study were consistent with recent reports on host genetic control of endophyte transmission (Gagic et al., 2018). They found that AR37 transmission is higher in perennial ryegrass populations that have undergone selection for high seed infection than those that did not.

The association between grasses and their non-native endophyte is expected to be imperfect. Nevertheless, cross-speciation associations are known to occur, especially within the forage species namely, perennial ryegrass, Italian ryegrass, and tall fescue, and their endophytes. The interaction can be mutualistic or antagonistic, but before a favourable interaction is achieved, the endophyte must be able to overcome plant defenses, at the very least. Failure to do so could result to low infection rate. In parallel, the host should provide a favourable environment to the fungi, where a similarly low infection rate can result from an unreceptive host. Selection for a more receptive host should, therefore, improve compatibility and consequently an increase in infection rate in the population.

2.6. Conclusion

Four cycles of selection for endophyte compatibility in PGG04 may have resulted in an increased proportion of viable AR501 in the population. The effect of positive selection on infection rate was investigated by growing early and late generations of PGG04, and testing for endophyte infection using three methods. Tissue-print immunoblotting detected higher percentage of viable endophytes in the late generation than in the early generation of PGG04. Using the microsatellite B11, endophyte genotyping confirmed the results of immunoblotting. Genotyping also identified AR501 as the endophyte strain present in the population, and that there was no contamination of the common toxic endophyte. Endophyte detection in the seed showed that AR501 is present in the seeds at a high level, both in the early and late generations of PGG04. The difference in the infection rate in the seed (i.e. seed squash) and seedling (i.e. immunoblotting) of PGG04 can be due to several factors such as seed storage conditions and genetics. Seeds from the early generation (PGG04-C2) was stored for nearly six years before the study while the late generation (PGG04-C6) was a few months after harvest. Therefore, it is possible that storage conditions of PGG04-C2 contributed to the decline in endophyte viability. Nevertheless, the early generation seeds were stored under low temperature and in moisture-resistant packaging. Since the effects of storage conditions were not formally tested in the present study, their effects cannot be totally discounted. Endophytes detected in the seed are composed of viable and non-viable endophytes. Thus, it is also likely that endophyte loss from seed to seedlings was due to host incompatibility. Hence, endophytes are known to exhibit a high degree of host specificity. The present study deals with cross-species association, that is, a tall fescue endophyte infected in perennial ryegrass. Overall under ideal conditions, host genetics largely influence endophyte viability. Based on the results of this study, it is concluded that recurrent selection in PGG04 could possibly improve AR501 compatibility in terms of transmission, which was reflected by an increase in the proportion of viable endophyte from early to the late generation.

2.7. References

Adcock, R., Hill, N., Bouton, J., Boerma, H., & Ware, G. (1997). Symbiont regulation and reducing ergot alkaloid concentration by breeding endophyte-infected tall fescue. *J Chem Ecol, 23*(3), 691-704.

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.

Ball, O., Coudron, T., Tapper, B., Davies, E., Trently, D., Bush, L., . . . Popay, A. (2006). Importance of host plant species, Neotyphodium endophyte isolate, and alkaloids on feeding by Spodoptera frugiperda (Lepidoptera: Noctuidae) larvae. *J Econ Entomol, 99*(4), 1462-1473.

Ball, O., Gwinn, K., Pless, C., & Popay, A. (2011). Endophyte isolate and host grass effects on Chaetocnema pulicaria (Coleoptera: Chrysomelidae) feeding. *J Econ Entomol, 104*(2), 665-672.

Card, S. D., Rolston, M., Park, Z., Cox, N., & Hume, D. (2011). Fungal endophyte detection in pasture grass seed utilising the infection layer and comparison to other detection techniques. *Seed Science and Technology, 39*(3), 581-592.

Card, S. D., Rolston, M. P., Lloyd-West, C., & Hume, D. E. (2014). Novel perennial ryegrass-Neotyphodium endophyte associations: relationships between seed weight, seedling vigour and endophyte presence. *Symbiosis, 62*(1), 51-62. doi: 10.1007/s13199-014-0271-5

Christensen, M. (1995). Variation in the ability of Acremonium endophytes of Lolium perenne, Festuca arundinacea and F. pratensis to form compatible associations in the three grasses. *Mycol Res, 99*(4), 466-470. doi: https://doi.org/10.1016/S0953-7562(09)80647-3

Christensen, M., Leuchtmann, A., Rowan, D., & Tapper, B. (1993). Taxonomy of Acremonium endophytes of tall fescue (Festuca arundinacea), meadow fescue (F. pratensis) and perennial ryegrass (Lolium perenne). *Mycol Res, 97*(9), 1083-1092. doi: https://doi.org/10.1016/S0953-7562(09)80509-1

Easton, H., Latch, G., Tapper, B., & Ball, O.-P. (2002). Ryegrass host genetic control of concentrations of endophyte-derived alkaloids. *Crop Sci, 42*(1), 51-57.

Easton, H., Lyons, T., Mace, W., Simpson, W., De Bonth, A., Cooper, B., & Panckhurst, K. (2007). *Differential expression of loline alkaloids in perennial ryegrass infected with endophyte isolated from tall fescue.* Paper presented at the Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses. New Zealand Grassland Association, Dunedin, New Zealand.

Faville, M., Briggs, L., Cao, M., Koulman, A., Jahufer, M., Koolaard, J., & Hume, D. (2015). A QTL analysis of host plant effects on fungal endophyte biomass and alkaloid expression in perennial ryegrass. *Mol Breed, 35*(8), 161.

Fehr, W. (1991). *Principles of cultivar development: theory and technique*: Macmillian Publishing Company.

Fletcher, L. (2012). *Novel endophytes in New Zealand grazing systems: The perfect solution or a compromise?* Paper presented at the Epichloae, endophytes of cool season grasses: implications, utilization and biology. Proceedings of the 7th International Symposium on Fungal Endophytes of Grasses, Lexington, Kentucky, USA, 28 June to 1 July 2010.

Freitas, P. (2017). *Crossing the species barrier: investigating vertical transmission of a fungal endophyte from tall fescue within a novel ryegrass association* (Doctoral dissertation, Lincoln University, Christchurch, New Zealand). Retrieved from https://researcharchive.lincoln.ac.nz/bitstream/handle/10182/8385/Freitas_PhD.pdf?sequence=3&isAllowed=y

Gagic, M., Faville, M. J., Zhang, W., Forester, N. T., Rolston, M. P., Johnson, R. D., . . . Hudson, D. (2018). Seed transmission of Epichloë endophytes in Lolium perenne is heavily influenced by host genetics. *Frontiers in Plant Science, 9*, 1580.

Gamer, M., Lemon, J., & Singh, I. F. P. (2012). irr: Various Coefficients of Interrater Reliability and Agreement. [R package] (Version 0.84). Retrieved from https://CRAN.R-project.org/package=irr

Gundel, P., Martínez-Ghersa, M., Batista, W., & Ghersa, C. (2010). Dynamics of Neotyphodium endophyte infection in ageing seed pools: incidence of differential viability loss of endophyte, infected seed and non-infected seed. *Ann Appl Biol, 156*(2), 199-209.

Hahn, H., Huth, W., Schöberlein, W., Diepenbrock, W., & Weber, W. (2003). Detection of endophytic fungi in Festuca spp. by means of tissue print immunoassay. *Plant Breeding, 122*(3), 217-222.

Hinton, D., & Bacon, C. (1985). The distribution and ultrastructure of the endophyte of toxic tall fescue. *Canadian journal of botany, 63*(1), 36-42.

Hume, D. E., Schmid, J., Rolston, M., Vijayan, P., & Hickey, M. (2011). Effect of climatic conditions on endophyte and seed viability in stored ryegrass seed. *Seed Science and Technology, 39*(2), 481-489.

Hume, D. E., Card, S. D., & Rolston, M. P. (2013). *Effects of storage conditions on endophyte and seed viability in pasture grasses.* Paper presented at the Proc 22nd Intern Grassland Congress.

Humphreys, M., Feuerstein, U., Vandewalle, M., & Baert, J. (2010). Ryegrasses. In B. Boller, U. K. Posselt, & F. Veronesi (Eds.), *Fodder Crops and Amenity Grasses* (pp. 211-260). New York, NY: Springer New York.

Johnson, L., de Bonth, A., Briggs, L., Caradus, J., Finch, S., Fleetwood, D., . . . Card, S. (2013). The exploitation of epichloae endophytes for agricultural benefit. *Fungal Diversity, 60*(1), 171-188. doi: 10.1007/s13225-013-0239-4

Karimi, S., Mirlohi, A., Sabzalian, M. R., Sayed Tabatabaei, B. E., & Sharifnabi, B. (2012). Molecular evidence for Neotyphodium fungal endophyte variation and specificity within host grass species. *Mycologia, 104*(6), 1281-1290.

Kauppinen, M., Saikkonen, K., Helander, M., Pirttilä, A. M., & Wäli, P. R. (2016). Epichloë grass endophytes in sustainable agriculture. *Nature Plants, 2*, 15224. doi: 10.1038/nplants.2015.224

Latch, G., & Vaughan, D. (1995). Search for seedborne endophytic fungi in rice. *International Rice Research Notes*.

Latchs, G., & Christensen, M. (1985). Artificial infection of grasses with endophytes. *Ann Appl Biol, 107*(1), 17-24.

Moon, C. D., Craven, K. D., Leuchtmann, A., Clement, S. L., & Schardl, C. L. (2004). Prevalence of interspecific hybrids amongst asexual fungal endophytes of grasses. *Mol Ecol, 13*(6), 1455-1467. doi: 10.1111/j.1365-294X.2004.02138.x

Moon, C. D., Tapper, B. A., & Scott, B. (1999). Identification of Epichloë endophytes in planta by a microsatellite-based PCR fingerprinting assay with automated analysis. *Appl Environ Microbiol, 65*(3), 1268-1279.

Nixon, C. (2015). *How valuable is that plant species*. (MPI Technical Paper No: 2016/62). Retrieved from https://www.mpi.govt.nz/dmsdocument/14527-how-valuable-is-that-plant-species-application-of-a-method-for-enumerating-the-contribution-of-selected-plant-species-to-new-zealands-gdp/sitemap.

Pennell, C., & Ball, O. (1999). *The effects of Neotyphodium endophytes in tall fescue on pasture mealy bug (Balanococcus poae).* Paper presented at the Proc. 52nd NZ Plant Protection Conf.

Popay, A., & Hume, D. (2011). Endophytes improve ryegrass persistence by controlling insects. *Pasture persistence*.

Popay, A., & Jensen, J. (2005). Soil biota associated with endophyte-infected tall fescue in the field. *New Zealand Plant Protection, 58*, 117.

Popay, A., Jensen, J., & Cooper, B. (2005). *The effect of non-toxic endophytes in tall fescue on two major insect pests.* Paper presented at the Proceedings of the New Zealand Grassland Association.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rolston, M., & Agee, C. (2007). *Delivering quality seed to specification-the USA and NZ novel endophyte experience.* Paper presented at the Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses'. Christchurch, New Zealand. Grassland Research and Practice Series.

Rolston, M., Stewart, A., Latch, G., & Hume, D. (2002). Endophytes in New Zealand grass seeds: occurrence and implications for conservation of grass species. *N Z J Bot, 40*(3), 365-372.

Rowan, D., & Gaynor, D. (1986). Isolation of feeding deterrents against Argentine stem weevil from ryegrass infected with the endophyteAcremonium loliae. *J Chem Ecol, 12*(3), 647-658.

Schardl, C., Grossman, R., Nagabhyru, P., Faulkner, J., & Mallik, U. (2007). Loline alkaloids: Currencies of mutualism. *Phytochemistry, 68*(7), 980-996. doi: https://doi.org/10.1016/j.phytochem.2007.01.010

Simpson, W. R., Schmid, J., Singh, J., Faville, M. J., & Johnson, R. D. (2012). A morphological change in the fungal symbiont Neotyphodium lolii induces dwarfing in its host plant Lolium perenne. *Fungal Biology, 116*(2), 234-240. doi: https://doi.org/10.1016/j.funbio.2011.11.006

Symonds, V. V., & Lloyd, A. M. (2004). A simple and inexpensive method for producing fluorescently labelled size standard. *Mol Ecol Notes, 4*(4), 768-771.

van Zijll de Jong, E., Dobrowolski, M. P., Bannan, N. R., Stewart, A. V., Smith, K. F., Spangenberg, G. C., & Forster, J. W. (2008). Global Genetic Diversity of the Perennial Ryegrass Fungal Endophyte Neotyphodium lolii. *Crop Sci, 48*(4), 1487-1501. doi: 10.2135/cropsci2007.11.0641

Wheatley, W., Kemp, H., Simpson, W., Hume, D., Nicol, H., Kemp, D., & Launders, T. (2007). Viability of endemic endophyte (Neotyphodium lolii) and perennial ryegrass (Lolium perenne) seed at retail and wholesale outlets in south-eastern Australia. *Seed Science and Technology, 35*(2), 360-370.

Wilkins, P. (1991). Breeding perennial ryegrass for agriculture. *Euphytica, 52*(3), 201-214. doi: 10.1007/bf00029397

Young, C. A., Charlton, N. D., Takach, J. E., Swoboda, G. A., Trammell, M. A., Huhman, D. V., & Hopkins, A. A. (2014). Characterization of Epichloë coenophiala within the US: are all tall fescue endophytes created equal? *Frontiers in Chemistry, 2*(95). doi: 10.3389/fchem.2014.00095

Chapter 3

Selection for compatibility with the tall fescue endophyte AR501 shaped the structure of genetic variation in the perennial ryegrass breeding population PGG04

3.1. Abstract

The grass-endophyte interaction is influenced by host genetics, hence compatibility of the host plant with endophyte can be exploited in plant breeding. The objective of this study is to investigate changes in genetic variation in an advanced breeding population recurrently selected for endophyte compatibility, and to identify signatures of selection that are related to the trait of interest. Genetic variation between early (C2) and late generations (C6) of PGG04 were compared using a relatively high-density marker set from genotyping-by-sequencing (GBS). Results showed that selection led to an excess of rare alleles, with the late generation enriched with alleles with frequencies between 0.02 - 0.08 compared with the early generation. Selection also led to the reduction of genetic diversity, from an expected heterozygosity of 0.3069 in PGG04-C2 to 0.3033 in PGG04-C6. Selection also changed the population structure as illustrated by UPGMA dendrogram, principal components analysis (PCA), and the model-based clustering method implemented in STRUCTURE. The effect of selection was not uniform across the genome since selection was targeted towards improvement of one trait. A few SNPs under selection pressure exhibited extremely high F_{ST} and influenced PCA-based population structure considerably more than other SNPs. Five genomic regions tagged by one to five SNPs under selection were verified using logistic regression with infection data. Depending on the allele frequency, these SNPs can increase the odds of AR501 infection by a maximum of over five times. Annotation of one of these regions tagged by S7_160751877 identified an ABCG transporter gene that may be related to the host genetic control of the association, as ABC transporters are known to be involved in plant-microbe interactions.

3.2. Introduction

Perennial ryegrass breeding is important in pastoral agriculture especially in the dairy industry. In New Zealand, the relative economic estimate of genetic gains in ryegrass breeding has been valued at about \$12 – \$18 per hectare per year to dairy farmers. This is on par with that of animal breeding at about \$11 per cow per year (Chapman et al., 2017). However, direct benefit in the dairy industry can be established clearly with livestock genetic improvement, whereas this is not the case in forage breeding. Therefore, investment priorities for perennial ryegrass breeding should be geared towards traits that contribute much to economic gains (Lee et al., 2012). Persistence, that is, the stability of dry matter yield (Parsons et al., 2011), is among the most important breeding objectives since reseeding and cultivation can be capital intensive. In New Zealand, grass persistence was observed to be improved by the interaction of *Epichloë* endophytes which increases defence against pest herbivores (Popay & Hume, 2011). Not all *Epichloë* endophytes are beneficial because some are not only toxic to insects but also to livestock. The native endophyte (common toxic endophyte) found in perennial ryegrass in New Zealand causes livestock health issues. Research and discovery of novel endophytes is therefore important to the forage industry. These endophytes produce alkaloids that are non-toxic to mammals but offers protection against insect pests. Perennial ryegrass breeding for improved compatibility with novel endophytes is a promising approach for improving persistence. It can be shown directly that insect resistance improves grass survival and stabilises herbage yield (Popay & Hume, 2011). Further, this mode of pest control is favourable to the environment as it limits the use of synthetic insecticides that could contaminate soil and water. Research on the grass-endophyte interaction including in the context of cultivar development is, therefore, a worthy investment.

Improving grass persistence has proved to be difficult as most studies are conducted in a five-year period or less, which is perhaps not enough to evaluate or predict long-term stability (Lee et al., 2012). An exception would be the study of Chapman et al. (2015) who reported that earlier measurements (i.e. years 1 to 3) of dry matter yield can reasonably predict long-term yield such as after seven or eight years, although not at the tenth year. The decline in persistence is attributed to three factors: (1) plant mortality of the sown cultivar, (2) death of plants with important traits, and (3) decline in yielding ability of the sown population (Parsons et al., 2011). Alkaloids produced by fungal endophytes are known to provide protection against insect pests, improving the survival of the sown cultivar (i.e. 1, above) which leads to stable yield (3, above). For example, endophyte-free perennial ryegrass suffered from a decline in persistence due to Argentine stem weevil and was replaced by infected ryegrass plants or other plant species (Popay & Hume, 2011). More importantly, increasing yield differences in perennial ryegrass cultivars with time has been shown to be affected by different endophyte strains (Hume et al., 2009; Hume et al., 2007). In the cultivar, Samson, infection with the AR37 showed an increasing yield advantage through time compared with Samson infected with the common toxic endophyte or AR1 (Hume et al., 2007). Observations such as this can be explained by the pest protection provided by the endophyte and, in the case of AR37 in the study cited above, protection is provided against root aphid and black beetle (Hume et al., 2007). Yet, novel endophytes do not easily form associations with perennial ryegrass populations, especially the non-native endophyte species. Thus, breeding for improved endophyte compatibility is an important goal. This strategy can be used to improve grass persistence by enabling the introduction of novel endophyte strains that produce highly effective alkaloids. At the very least, novel endophytes can improve insect resistance of grass populations and reduce insect damage that cause production losses in the short term.

Recurrent selection (RS), as practiced in the population under investigation, aims to increase the frequency of favourable alleles for the trait of interest. In the current case, selection should improve endophyte compatibility – specifically improvement in the transmission of the endophyte via seed, across generations. In tall fescue, two cycles of recurrent selection successfully reduced the concentration of an unfavourable endophyte alkaloid (Adcock et al., 1997), demonstrating that host genetic control of the grass-endophyte association can be exploited in breeding. In perennial ryegrass, quantitative trait loci (QTLs) controlling mycelial mass and alkaloid concentration were discovered (Faville et al., 2015). This could be used potentially in marker-assisted selection for host endophyte compatibility. In the QTL study, biparental mapping populations were created to enable dissection of the genetic basis of the trait of interest. In mapping populations, individuals are partitioned into different genotypic groups (based on marker alleles or haplotypes) and tested to determine whether phenotypic differences between groups are significant. This type of population, in most cases, is different from the breeding populations which typically have contributions from multiple parents, not just two. Genomic regions associated with a trait of interest can also be studied in breeding populations. Selection mapping (SM) is a method described as “a range of approaches that identifies alleles, loci, and epistatic interactions using populations that have been subjected to iterative cycles of recombination and selection” (Wisser et al., 2008). In a population undergoing RS, changes in allele frequency are not only influenced by selection but other factors as well, notably, random genetic drift. In SM, therefore, a statistical test of the null hypothesis of genetic drift must be considered. SM has been successfully carried out in perennial ryegrass breeding populations to identify genes/QTLs associated with tiller development (Brazauskas, Pašakinskienė, et al., 2013) and crown rust resistance (Brazauskas, Xing, et al., 2013).

Similarly, the current study aims to map selection signatures in the genome that indicate a response to selection for endophyte transmission. Selection results in allele frequency changes in genes responsible for trait variation, as well as in loci physically linked to such genes. Allele frequencies in the breeding population will be highly variable but loci for which one allele is favoured (i.e. selected) would be more genetically differentiated than neutral loci when comparing different generations of RS. This is the basis of the F_{ST} outlier approach employed in detecting loci responsible for local adaption in evolution studies (Whitlock & Lotterhos, 2015). Aside from selection and genetic drift, gene flow or migration and mutation may also influence allele frequency shifts. These factors were assumed negligible for the current study. Mutation rates are generally slow, and the two generations of RS population used in the study have a short period of time between them (four cycles of selection and recombination), therefore mutation can be largely discounted. Gene flow is also likely to be insignificant, if not totally absent, as polycrossing of selected individuals was done meticulously in isolation (Keith Saulsbury, personal communication, 2018), with little or no opportunity for non-selected pollen or volunteer plants to enter. Selection signatures in the genome detected in this study will give insights into the genetic variation influencing the grass-endophyte interaction. Furthermore, SM combined with other strategies such as linkage mapping, LD mapping, and functional genetics studies will collectively offer a more comprehensive genetic dissection of the host-endophyte interaction towards its utilization in breeding programs.

The objective of this study is to investigate genetic variation in an advanced perennial ryegrass breeding population recurrently selected for compatibility with a novel endophyte, specifically the transmission of endophyte via seed between generations. Using a relatively high-density marker set from GBS, this study aims to describe genetic variation in a perennial ryegrass population in active development. It also aims to characterize changes in genetic variation following recurrent selection to identify genetic signature of selection. Furthermore, the study aims to understand the relationship of these selection signatures to endophyte compatibility.

3.3. Materials and methods

3.3.1. Breeding population

PGG04, a perennial ryegrass breeding population developed by PGG Wrightson Seeds Limited, was chosen for experimentation. It was also briefly described in Chapter 2. PGG04 was derived from a restricted base population and developed through recurrent selection (RS) with the polycross method. Initially, plants of mixed origins in New Zealand were pollinated by the cultivar Bronsyn to create the cultivar Extreme (K. Saulsbury, personal communication, 2018). This was later inoculated with the tall fescue endophyte *Epichloë* sp. FaTG-3 strain AR501 by AgResearch based on the protocol of Latch and Christensen (1985) (Fig. 2.1.). Afterwhich, it underwent seed multiplication and agronomic selection in the subsequent generation. Lastly, three plants selected for agronomic performance were used as female parents in three pair crosses with the cultivar Arrow, generating the progenitors of PGG04 (K. Saulsbury, personal communication, 2018). The population was selected for compatibility with AR501. A combination of phenotypic and half-sib family selection was employed for six cycles (C6) of selection. At each cycle of 1-2 years, there was phenotypic selection for agronomic 'fitness'. In addition, each polycross was harvested as single plants, and each half-sib family was checked for the transmission of viable endophyte. On average, a blend of the top 10% transmitters was selected and sown for the following cycle, although the actual selection intensity varied (K. Saulsbury, personal communication, 2018). Endophyte compatibility was assessed in terms of viability and transmission. Viability describes the ability of the endophytes to remain viable in seeds under storage. Transmission, on the other hand, refers to the transmissibility of the endophyte from parents to offspring or from one generation to the next via seed (M. J. Faville, personal communication, 2018).

3.3.2. DNA quality and quantity

DNA quality is one of the most important factors affecting the quality of GBS data (Anderson et al., 2018). Genomic DNA extraction was accomplished by adapting a protocol specifically optimized for genotyping-by-sequencing (Anderson et al., 2018). This extraction protocol was described in detail in Chapter 2. Briefly, cell lysis was based on sodium dodecyl sulphate (SDS) and precipitation utilized potassium acetate and acetic acid. Precipitated DNA was bound to a silica media and finally eluted using a Tris solution. After extraction, DNA quality was assessed using an agarose gel-based method and concentration was quantified using a fluorometric method involving Hoechst dye, which are described in further detail below.

The quality of the extracted DNA was assessed using 0.8% lithium borate agarose gel. The gel was prepared by combining 4 g of ultrapure agarose powder with 1x lithium borate (LB) and heating it in the microwave oven until all agarose powder dissolved completely. The agarose solution was cooled and 2.5 µL of ethidium bromide was mixed for every 140 mL of the solution. The solution was allowed to cool further and casted with combs for DNA loading. Each DNA sample was mixed with loading dye at 2:18 µL ratio in a 96-well plate. For each sample, 10 µL of the mixture was loaded on to the gel. A 1kb+ DNA ladder was loaded into two wells at the edges of the gel. The gel was submerged in 1x LB buffer in an electrophoresis machine and run at 100 volts for 30 min. The negatively charged DNA migrates in the gel towards the positive electrode. After the run, the gel was visualized in a Bio-Rad Gel Documentation system. Ethidium bromide interacts with the DNA and, when exposed to UV light fluoresces, enables DNA visualisation. All samples except the blank controls (e.g. lane 24, P4.6-D3 in Fig. 3.1.) showed high molecular weight DNA of reasonable quality in terms of intensity and clarity as exemplified in Figure 3.1 (gel image for other samples can be found in the Appendix Fig. 3.1 and 3.2.). No smearing or unwanted bands (i.e. unexpected size) were observed, indicating minimal DNA sample degradation. Variation in sample intensity was observed, indicating differences in DNA yield.

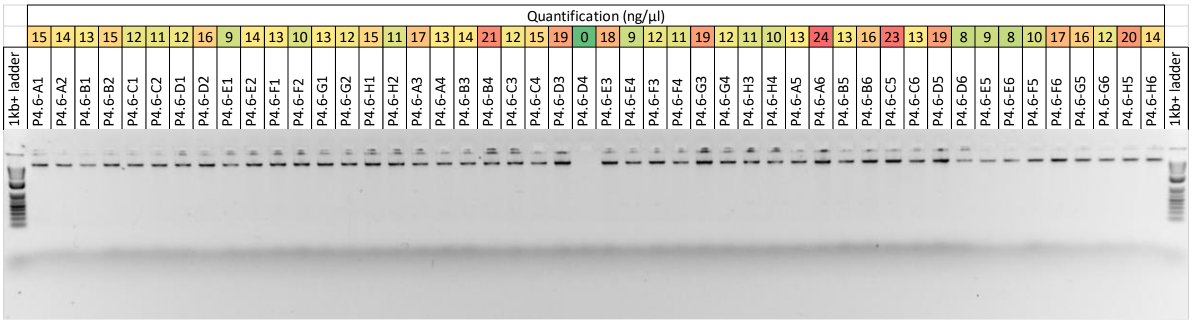


Figure 3.1. DNA quantity and quality of 48 PGG004-C6 plant samples (i.e. P4.6). Quality was assessed based on the appearance of the bands (i.e. absence of smears, size, intensity, etc.). Lane 24 (middle) is the blank control and shows no band. Quantity was computed based on the fluorescence of DNA size standards as described below.

The DNA quantification method involved the use of Hoechst dye, similar to the protocol of Rago et al. (1990). Two hundred µL of Hoechst dye was mixed with 17.8 mL of TNE. TNE was prepared by combining Tris (5 mL of 1 M Tris, pH 7.4), NaCl (58.45 g of 2M NaCl) and EDTA (1 mL 0.5 M EDTA, pH = 8) in water. The samples were prepared for plate reading using a Hamilton liquid handling robot. First, DNA was diluted in TE (Tris and EDTA) at a rate of 1:3. Then the dye reaction mixture was distributed into a 384-plate. Five µL of the diluted DNA was then added to the dye in the plate. Each sample was quantified in triplicate and distributed in three wells of the 384-plate. λ phage DNA standards of concentrations: 0, 5, 10, 15, 20, 25, 30 and 35 nanograms (ng) per µL were also added to the plated dye mixture in triplicate. The 384-plate was inserted in a Biotek plate reader for measuring fluorescence. The reader was set to 360/40 nm bandwidth excitation filter and 460/40 nm emission filter. DNA samples were quantified based on their fluorescence and comparison with the standards. The fluorescence values of the standards were regressed against their concentration. The resulting regression model was used to compute the DNA concentration of the samples based on their fluorescence values (Fig. 3.2). The linear models used have a coefficient of determination (R^2) of more than 0.99 and quantification values ranged from ~4 – 35 ng/ µL (concentration vs fluorescence values plot for PG004-C6 is in Appendix Fig. 3.3).

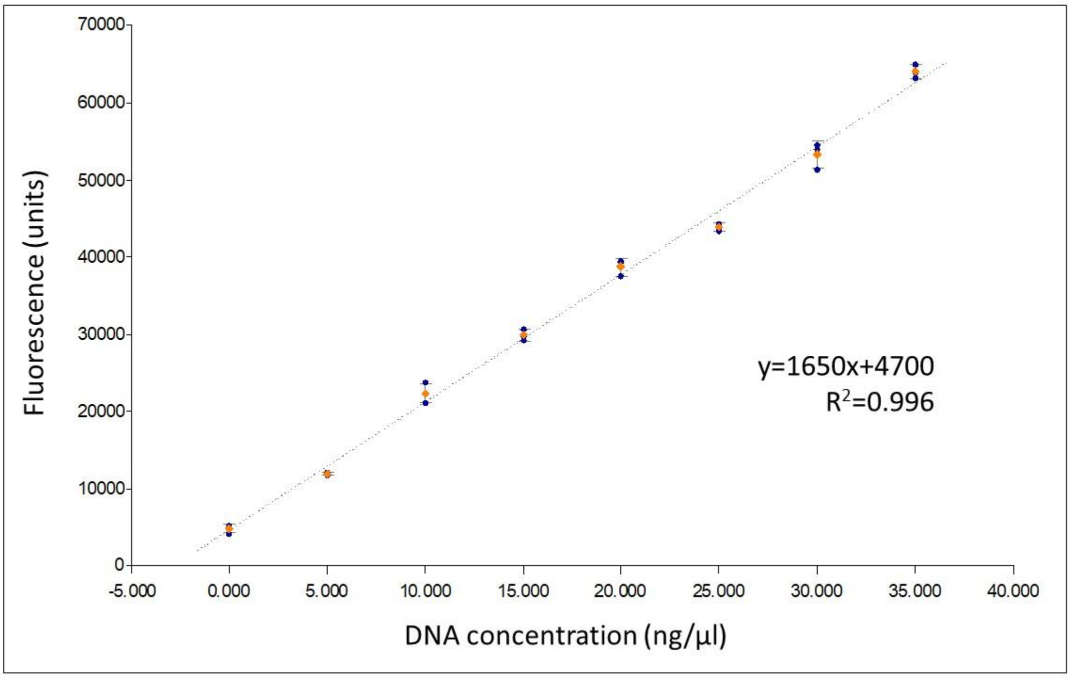


Figure 3.2. Standard curve used for DNA quantification of PGG004-C2 plant samples, showing a positive linear relationship between DNA concentration and fluorescence. The linear model explains more than 99% of the observed variation.

3.3.4. GBS library preparation

Genotype data was obtained through genotyping-by-sequencing (GBS) based on the maize protocol (Elshire et al., 2011) and as adapted to perennial ryegrass (Faville et al., 2018). A schematic diagram of the library preparation step in GBS is shown in Fig. 3.3. (Elshire et al., 2011). The protocol optimised by AgResearch was followed. Two libraries were prepared at the AgResearch Grasslands Research Centre which were then sent to the AgResearch Invermay Agricultural Centre for sequencing.

GBS is made possible because of genome complexity reduction strategies such as the use of restriction enzymes (RE). The digestion of DNA into smaller fragments is based on the recognition of the restriction sites by the enzymes. These sites were also complementary to the adapter sequences, thus allowing ligation. In the present study, we have used two enzymes, i.e. a double digest similar to the protocol of Poland et al. (2012). The commonly used restriction enzyme for GBS, ApeKI, provides hundreds of thousands of markers in perennial ryegrass (Arojju et al., 2016; Byrne et al., 2017; Faville et al., 2018; Fè et al., 2016). In selection signature studies however, the sequencing depth is particularly important since it relies on accurate estimates of allele frequencies. With a limited sequencing budget, better sequencing depth can be obtained by using PstI instead of ApeKI in perennial ryegrass GBS (Byrne et al., 2013). Including the second enzyme, MspI, results in a combination of rare (PstI) and common cutter (MspI) which provide an effective and uniform genome complexity reduction strategy (Poland et al., 2012). This two-enzyme system has been demonstrated to provide better sequencing depth than using ApeKI and to generate a larger set of markers than using PstI alone in perennial ryegrass GBS (Dr. Mingshu Cao, unpublished data).

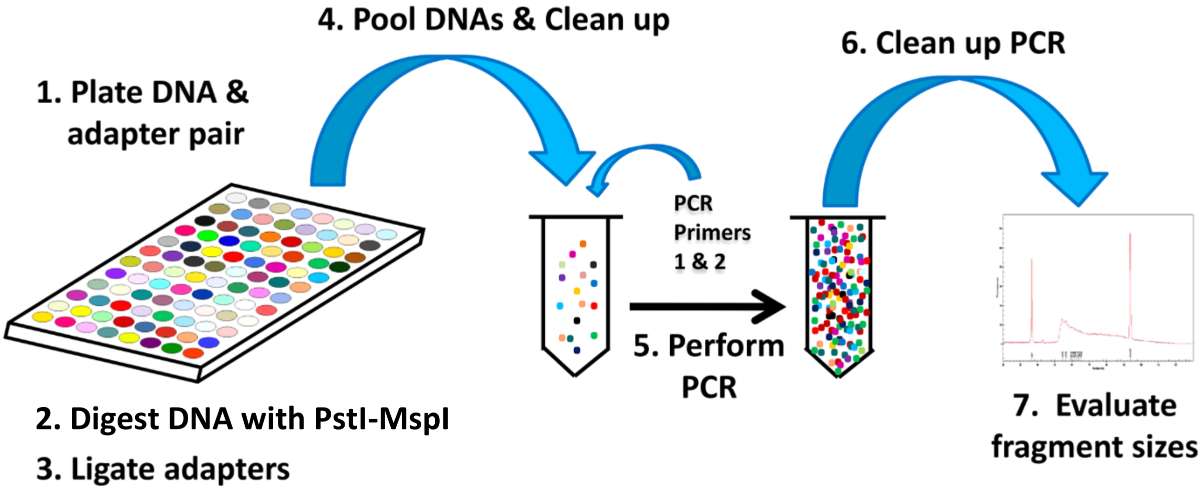
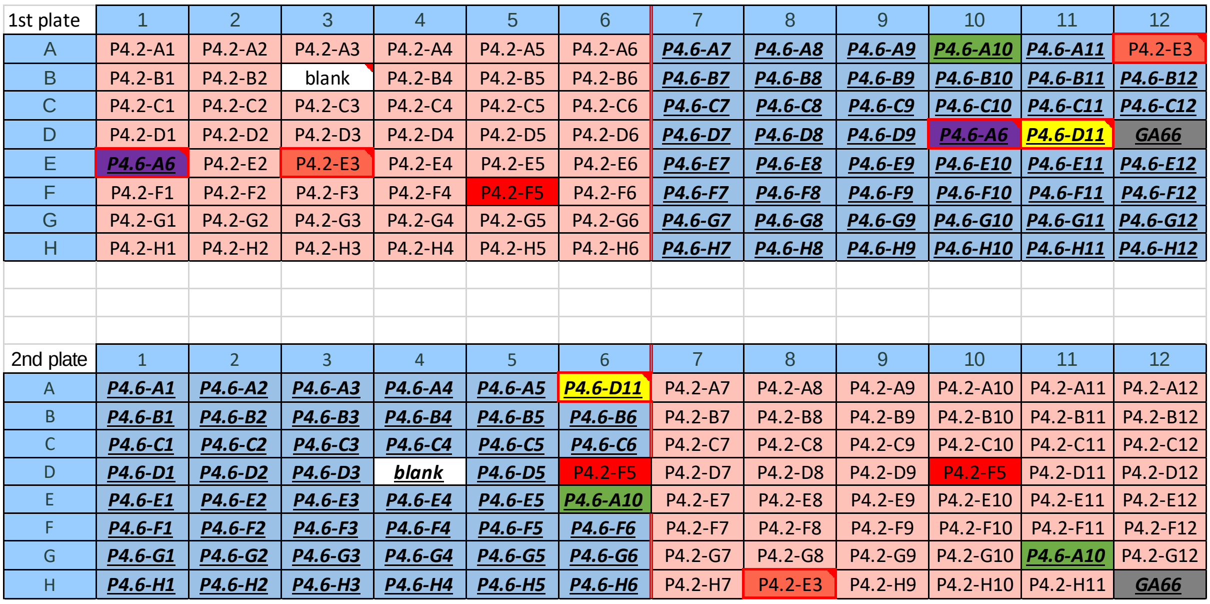


Figure 3.3. Step by step library preparation in GBS adapted from Elshire et al. (2011). First DNA is digested with the chosen restriction enzyme. Second, adapter pairs, including one barcoded adapter, are ligated to the cut ends of the DNA fragments. The samples can be then pooled for PCR. After PCR, the samples are cleaned and size selection may be conducted. Finally, the library is sequenced.

DNA plate or GBS library layout

The method of genotyping-by-sequencing entails two levels of sampling. First, the sampling of individual plants to accurately represent the genetic variation present in the population, which is critical in the genetically diverse perennial ryegrass populations. Second, the sampling of fragments to be sequenced from a pool of digested DNA. To adequately represent the recurrent selection population, we extracted DNA of 94 individuals per generations. Thus, two plates corresponding to the two generations namely, PG004-C2 (i.e. P4.2) and PG004-C6 (i.e. P4.6), were processed. After including controls and discarding individuals with low DNA quality and quantity, each library contained 90 samples to represent each generation. We employed control measures to minimize inconsistency in sequencing results. Each of the two libraries contained individuals from both PGG04-C2 and PGG04-C6. Including both in a single library minimized the influence of potential batch effects since they experienced similar preparation and sequencing conditions. For each library, we also incorporated control samples, including a blank well (negative control) and a company-wide GBS library control, namely a ryegrass genotype known as GA66. This control is included in all ryegrass GBS libraries in AgResearch and thus, the general characteristics of its sequence data are well known. In addition to this, we included individuals from each generation as technical replicates.

Within a library, two individuals (one from each generation) were replicated twice. The sequence data from these replicates gives an indication of the uniformity of sequencing within the library. Across libraries, four individuals were common, in addition to GA66. The replication across libraries facilitated joint data processing and analysis of the two libraries. The plate layout is shown below (Fig. 3.4).



1st plate	1	2	3	4	5	6	7	8	9	10	11	12
A	P4.2-A1	P4.2-A2	P4.2-A3	P4.2-A4	P4.2-A5	P4.2-A6	P4.6-A7	P4.6-A8	P4.6-A9	P4.6-A10	P4.6-A11	P4.2-E3
B	P4.2-B1	P4.2-B2	blank	P4.2-B4	P4.2-B5	P4.2-B6	P4.6-B7	P4.6-B8	P4.6-B9	P4.6-B10	P4.6-B11	P4.6-B12
C	P4.2-C1	P4.2-C2	P4.2-C3	P4.2-C4	P4.2-C5	P4.2-C6	P4.6-C7	P4.6-C8	P4.6-C9	P4.6-C10	P4.6-C11	P4.6-C12
D	P4.2-D1	P4.2-D2	P4.2-D3	P4.2-D4	P4.2-D5	P4.2-D6	P4.6-D7	P4.6-D8	P4.6-D9	P4.6-A6	P4.6-D11	GA66
E	P4.6-A6	P4.2-E2	P4.2-E3	P4.2-E4	P4.2-E5	P4.2-E6	P4.6-E7	P4.6-E8	P4.6-E9	P4.6-E10	P4.6-E11	P4.6-E12
F	P4.2-F1	P4.2-F2	P4.2-F3	P4.2-F4	P4.2-F5	P4.2-F6	P4.6-F7	P4.6-F8	P4.6-F9	P4.6-F10	P4.6-F11	P4.6-F12
G	P4.2-G1	P4.2-G2	P4.2-G3	P4.2-G4	P4.2-G5	P4.2-G6	P4.6-G7	P4.6-G8	P4.6-G9	P4.6-G10	P4.6-G11	P4.6-G12
H	P4.2-H1	P4.2-H2	P4.2-H3	P4.2-H4	P4.2-H5	P4.2-H6	P4.6-H7	P4.6-H8	P4.6-H9	P4.6-H10	P4.6-H11	P4.6-H12

2nd plate	1	2	3	4	5	6	7	8	9	10	11	12
A	P4.6-A1	P4.6-A2	P4.6-A3	P4.6-A4	P4.6-A5	P4.6-D11	P4.2-A7	P4.2-A8	P4.2-A9	P4.2-A10	P4.2-A11	P4.2-A12
B	P4.6-B1	P4.6-B2	P4.6-B3	P4.6-B4	P4.6-B5	P4.6-B6	P4.2-B7	P4.2-B8	P4.2-B9	P4.2-B10	P4.2-B11	P4.2-B12
C	P4.6-C1	P4.6-C2	P4.6-C3	P4.6-C4	P4.6-C5	P4.6-C6	P4.2-C7	P4.2-C8	P4.2-C9	P4.2-C10	P4.2-C11	P4.2-C12
D	P4.6-D1	P4.6-D2	P4.6-D3	blank	P4.6-D5	P4.2-F5	P4.2-D7	P4.2-D8	P4.2-D9	P4.2-F5	P4.2-D11	P4.2-D12
E	P4.6-E1	P4.6-E2	P4.6-E3	P4.6-E4	P4.6-E5	P4.6-A10	P4.2-E7	P4.2-E8	P4.2-E9	P4.2-E10	P4.2-E11	P4.2-E12
F	P4.6-F1	P4.6-F2	P4.6-F3	P4.6-F4	P4.6-F5	P4.6-F6	P4.2-F7	P4.2-F8	P4.2-F9	P4.2-F10	P4.2-F11	P4.2-F12
G	P4.6-G1	P4.6-G2	P4.6-G3	P4.6-G4	P4.6-G5	P4.6-G6	P4.2-G7	P4.2-G8	P4.2-G9	P4.2-G10	P4.6-A10	P4.2-G12
H	P4.6-H1	P4.6-H2	P4.6-H3	P4.6-H4	P4.6-H5	P4.6-H6	P4.2-H7	P4.2-E3	P4.2-H9	P4.2-H10	P4.2-H11	GA66

Figure 3.4. Arrangement of DNA samples of individuals from PGG04-C2 (P4.2) and PGG04-C6 (P4.6). The technical replicates are: GA66 (grey) and P4.6-D11 (yellow), which were replicated across the plates/libraries; P4.6-A6 (purple), which was replicated within the first library; and P4.2-E3 (orange), P4.2-F5 (red) and P4.6-A10 (green), which were replicated within a plate/library as well as across libraries. The position of blanks (negative controls) on each plate are also indicated.

Digestion of DNA with PstI-MspI and ligation of adapters

The library preparation started by preparing adapter plates. Oligonucleotide adapters, 9 ng of PstI barcoded adapter and 300 ng of MspI Y-adapter, were placed into each well of a 96-well plate. These concentrations had previously been optimized for perennial ryegrass genomic DNA following titrations conducted by AgResearch. The adapter plate was then dried down using a SpeedVac Concentrator (Thermo Fisher Scientific). The adapters used were complementary with PstI and MspI (New England Biolabs, NEB). Two types of adapters were used, the “barcode” adapter and the common adapter. The barcoded adapter allows for the identification of the samples after they are pooled in the succeeding steps. The forward adapter with the barcode has the PstI restriction overhang. On the other hand, the MspI overhang matches a common reverse Y-adapter. This setup makes sure that each fragment has the following form: barcoded forward adapter-genomic DNA-common reverse adapter. The two adapters form part of the anchor for DNA primers during amplification.

Twenty µL of the extracted DNA was prepared in a source plate and subsequently, 100 ng of DNA was transferred to the adapter plate using a BioNex Nanodrop™ II liquid handling robot. The liquid handling robot makes use of the quantification values generated in the previous step to aliquot the appropriate volume for 100 ng DNA. For example, for a DNA sample quantified at 20 ng/µL, the robot dispenses 5 µL to deliver 100 ng. The plate containing adapters and sample DNA was then dried down using a SpeedVac Concentrator. After this, DNA was digested with PstI and MspI. This digestion will create fragments with: PstI and MspI cut-sites; MspI-MspI cut-sites; and PstI-PstI cut-sites. PstI-PstI fragments are very rare and while MspI-MspI are common, the use of Y-adapter makes sure that only PstI-MspI fragments will be amplified in the succeeding step (Poland et al., 2012). The reagents for digestion were first slowly thawed in ice. This was also done for ligation reagents. A digestion reaction mixture for more than 96 reactions was prepared based on each single reaction requiring 2 µL of CutSmart buffer (NEB), 1 µL each of PstI-HF and MspI (both at 200 Units/µL) and 16 µL of Nuclease-free water. Using a Nanodrop II, 20 µL of the digestion mixture was dispensed in each well of the plate containing DNA and adapters. The plate was spun down and incubated at 37°C for 2 hours and 65°C for 30 min using a thermal cycler. After digestion, DNA fragments and adapters were ligated. As with the digestion step, a bulk ligation reaction mixture was prepared and 30 µL of the mixture was dispensed into each well of the DNA plate using Nanodrop II. The mixture was prepared with 5 µL of ligase buffer (NEB), 2 µL of T4 DNA ligase (NEB) and 23 µL of nuclease-free water for each reaction. The plate was spun down and placed in a thermal cycler at 22°C for ligation for 60 min and 65°C further incubation for 20 min.

Pooling and polymerase chain reaction

The DNA samples in the 96-well plates were pooled and prepared for polymerase chain reaction (PCR) using the E.Z.N.A.® cycle pure kit (Omega Bio-Tek). Using the Nanodrop II, 5 µL of DNA was taken from each well and combined into an 8-strip PCR tube. About five times the total volume of DNA, which was 2.5 mL, of CP buffer was placed in a 5-mL Eppendorf tube. The DNA samples from the 8-strip tubes were then transferred to the Eppendorf tube and mixed with the buffer by shaking and inversion. A transfer tube was prepared which consisted of a 2-ml collection tube with a HiBind DNA Mini Column inserted. The pooled DNA with CP buffer was transferred to this tube and was centrifuged at the maximum accelerative force, 14,000 x the relative centrifugal force (*g)* for 60 s. About 700 µL of DNA with CP buffer can be transferred at one time and so this step was repeated ~5 times or until all the contents of the Eppendorf tube were processed. As with the extraction protocol, DNA binds to the column and impurities are deposited in the collection tube – thus the filtrate in the collection tube was discarded after each transfer step. The column was then washed by adding 700 µL of a wash buffer diluted in absolute ethanol. Afterwards, it was spun down at maximum *g* for 60 s and the filtrate again discarded. The wash step was performed twice. The HiBind DNA Mini Column inserted in an empty collection tube was dried by centrifugation at maximum *g* for 2 min. The column was then transferred into a clean 1.5-mL collection tube for elution. Fifty µL of elution buffer was added at the center of the column. It was allowed to sit on the column for 2 min, after which it was centrifuged at maximum *g* for 60 s. The filtrate, which was purified DNA, was now ready for amplification.

After clean-up, three DNA samples of 4 µL each were transferred into three tubes of a PCR strip tube for amplification. The reaction mixture for four reactions (excess) was prepared by combining 100 µL of NEB 2X *Taq* Master Mix, 8 µL of PCR primer mix (12.5 pmol/ µL for each primer) (Poland et al., 2012) and 76 µL nuclease-free water. A total of 46 µL of the reaction mixture was combined with 4 µL of DNA in each of the three PCR tubes. It was mixed by flicking and was spun down quickly. Then, it was loaded in a PCR machine with a thermal cycling protocol of: 5 min at 72°C; 30 s at 98°C; 18 cycles of 98°C for 10 s, 65°C and 72°C for 30 s each; and a final extension at 72°C for 5 min.

The PCR product was again purified with the E.Z.N.A.® cycle pure kit, similar to the protocol described in the first clean-up above. This time, PCR product from three tubes were pooled and mixed with 750 µL CP buffer. The final elution used 35 µL of buffer.

Validation and Size selection

DNA concentration was determined using Nanodrop ND-1000 Spectrophotometer (Thermo Fisher Scientific). First, the instrument was initialized by water followed by a blank measurement (i.e. using the elution buffer) and finally by loading 1.5 µL of the library. Quantification is based on the ultraviolet light absorption of DNA, which peaks at a wavelength of 260 nm. The ratio of absorbance values at 260 and 280 nm also gives an indication of DNA purity. DNA is expected to have an A260/A280 of about 1.8 – 1.9. Lower values are indicative of protein contamination. The DNA concentration of the first library was 154.6 ng/µL while it was lower for the second library (52.1 ng/µL). The expected absorbance spectral pattern of these libraries was observed (Appendix Fig. 3.4. and 3.5.).

After the DNA concentration was determined, size selection was carried out using Pippin prep (Sage Science). Pre-Pippin sample was first obtained from the library by aliquoting 2 µL. This sample will be compared to the size-selected sample (i.e. post-Pippin). The library was subjected to size selection through automated electrophoresis. Pippin prep uses a gel that branches into an elution and separation channel. DNA moving towards the separation channel (positively charged) can be redirected to the elution channel by switching positive electrodes in that channel through a built-in computer based on the preferred size set. The timing of collection is based on the migration rate of an internal standard. Fluorescently labelled DNA marker is run ahead of the samples and their mobility is optically detected. The complete Pippin prep procedure is described in the Appendix. Size selection was validated by comparing the pre- and post-Pippin samples using an Agilent 4200 TapeStation system.

TapeStation can be used to assess DNA concentration and size distribution of the GBS library. It uses a ScreenTape which contains minute agarose gel with buffers optimized for nucleic acid separation and built-in electrodes for automated electrophoresis. The system automatically loads samples and ladder, runs the electrophoresis, captures the gel image and processes and analyses the results. The TapeStation procedure is described in the Appendix. The DNA concentration of the pre-Pippin sample was 6.310 ng/µL and 1.720 ng/µL for the first and second library respectively. After size selection, post-Pippin sample concentrations were 2.260 ng/µL for the first library and for the second library it was 0.804 ng/µL. The average fragment size of the final libraries (post-Pippin) was 276 bp for the first library and 249 bp for the second library based on the region table in Appendix Fig. In comparison, the average fragment size of the libraries before size selection were 251 bp and 250 bp for the first and second library, respectively. More importantly, size selection improved the amount of 150 - 400 bp fragments from 50.43% to 72.31% for the first library and 52.14% to 84.68% for the second library. No unusual peaks (e.g. adapter dimers) were observed and thus it was sent for sequencing. The gel image and electropherogram of before and after size selection can be found in the Appendix Fig. 3.6. and 3.7. The library was sequenced on two lanes using an Illumina HiSeq 2500 System at the AgResearch Invermay sequencing centre.

3.3.4. Variant Calling

FASTQ files containing sequencing reads from the Illumina HiSeq 2500 System were received and processed by an AgResearch specialist bioinformatician using a modified version of the GBS pipeline in the software TASSEL 5 (Glaubitz et al., 2014), version 5.0. SNP calling using TASSEL-GBS pipeline is described in greater detail in the Appendix. Briefly, the process involves two phases, namely, SNP discovery and SNP production. In the first part, polymorphisms were determined using available information from sequencing reads and a reference genome. It also determines the position of these polymorphisms in the genome. The second phase makes use of the polymorphism information generated from the first and then proceeds with generating multi-SNP genotypes for each sample. First, raw reads were processed. Those with lengths of less than 74 nucleotides were removed and the remaining reads were de-multiplexed (organised by individual barcode). Next, an existing perennial ryegrass reference genome was used for tag alignment. The reference genome had been constructed as seven pseudochromosomes by mapping published ryegrass scaffolds (Byrne et al., 2015) to the Barley genome (The International Barley Genome Sequencing et al., 2012) (release-38, https://plants.ensembl.org/Hordeum_vulgare/Info/Index). A total of 1,059,522 quality tags were aligned to this reference genome using bowtie2 (Langmead & Salzberg, 2012). Finally, SNP calling based on the alignment was performed using the TASSEL tools "DiscoverySNPCallerPluginV2" and "ProductionSNPCallerPluginV2". This resulted to 505,054 SNPs which were then subjected to initial filtering on minor allele frequency (MAF) greater than 5%, and missing data rate less than 50%. The 5% MAF threshold is a common quality control strategy and is also consistent with the recommendations and default filtering values of the analyses performed in the current study, namely OutFlank (Whitlock & Lotterhos, 2015) and pcadapt (Luu et al., 2017). MAF threshold is reported to influence population structure analysis (Linck & Battey, 2019) especially in model-based inference such as STRUCTURE (Falush et al., 2003; Pritchard et al., 2000). It is suggested to complement these analyses with non-model-based methods (Linck & Battey, 2019); in this study PCA and distance-based clustering (i.e. UPGMA dendrogram based on Provesti's distance). The initial filtering resulted in 219,870 SNPs and the genotype data were then exported using in-house scripts for statistical and genetic analysis.

3.3.5. SNP Filtering and Data processing

The initial dataset of 219K SNPs was further processed in R (R Core Team, 2017). This dataset was characterized in terms of read depth, missingness and capacity for exploratory genetic analysis (i.e. PCoA). The 219K dataset was filtered with the help of the R package vcfR (Knaus & Grünwald, 2017). First, non-biallelic SNPs and those with a mean read depth of fewer than 5x were removed. Second, SNPs with more than 20% missing data and those with ambiguous base calls (i.e. reference of any base, N) were removed. Next, the consistency of base calls across biological replicates was considered by removing SNPs that had different bases between replicates. The resulting dataset had 84K SNPs which was also characterized and compared to the initial dataset. Filtering was then applied to the samples by removing replicates 2 and 3 as well as the GA66 control. Earlier, in the variant calling step, the negative control was removed after use in quality control steps. The minimum amount of missing data for the samples were also set to 20% but all the 180 samples had already met this criterion. Finally, seven SNPs had to be discarded because previous filtering steps made them uninformative (i.e. nonpolymorphic). The final dataset had 84,402 SNPs in total. This filtered data set was utilized for genetic diversity and population structure analysis. In running STRUCTURE (Falush et al., 2003; Pritchard et al., 2000), a smaller dataset was used for computational efficiency, a commonly used strategy. The 84K dataset was pruned based on an LD threshold (maximum correlation coefficient) of 0.20 using the R/SNPRelate (Zheng et al., 2012). After pruning, 20,237 SNPs remained. LD pruning was consistent with STRUCTURE's assumption that loci within the population are in linkage equilibrium, although it can deal with weakly linked markers. Before conducting the selection signature analysis, additional filtering was performed to ensure only high-quality SNPs were included. Markers with extremely low or high read depth were removed. A minimum read depth ensures heterozygotes are called correctly and extremely high read depth can be indicative of misalignment of tags from paralogous regions which may result in inaccurate locations in the genome. SNP genotype data with associated read depth below the 5th percentile (4x) and those with read depth above the 95th percentile (158x) were deleted. This deletion increased the amount of missing data and thus, finally, SNPs with more than 5% missing data were filtered out. This resulted in a final total of 31,093 SNPs available for the detection of signatures of selection.

3.3.6. Analysis of genetic diversity and population structure

All statistical analyses were performed in R (R Core Team, 2017) using various population genetics packages. Minor allele frequency (MAF), expected and observed heterozygosity were calculated for each SNP using the R/poppr (Kamvar et al., 2014) and then summarized. Inbreeding coefficient (F_{IS}) and fixation index (F_{ST}) values were calculated using the R/hierfstat

(Goudet, 2005). The 95% confidence interval of the mean of the F-statistics were also estimated by bootstrapping with 10000 replicates. Population structure was investigated with dendrograms of genetic distance and principal components analysis (PCA). UPGMA trees were constructed from absolute genetic distance, also called Provesti's distance (Prevosti et al., 1975). First, a bootstrap supported dendrogram was constructed by randomly sampling with replacement the loci 1000 times and considering a threshold of 0.50 using the aboot function in the R/poppr. In addition, 1000 UPGMA trees from sampled (also with replacement) loci and randomized sample order were created. The majority consensus tree was created from 1000 trees using the consensus function in the R/ape (Paradis et al., 2004). Dendrograms were drawn also using the R/ape. PCA scores were computed using the procedure of (Patterson et al., 2006) as implemented in the R/pcadapt (Luu et al., 2017). The scores were used to visualize structure in a biplot, i.e. PC1 vs PC2. Analysis of molecular variance (AMOVA) was conducted to determine the partitioning of genetic variation in the population using the R/poppr. Sources of variation in AMOVA were tested for significance based 1000 permutation tests. A Bayesian method for population structure analysis was also conducted using STRUCTURE (Pritchard et al., 2000). The model settings were set at default including an assumption of an admixed population and correlated allele frequencies (Falush et al., 2003). The program was run for K = 1 to 10 with burn-in period and MCMC steps of 50,000 each.

3.3.7. Identification of genomic regions under selection

Two approaches were utilized to identify genomic regions under selection. These were F_{ST} - based and PCA-based approaches. The F_{ST}-based method is an outlier detection approach based on an inferred null distribution of F_{ST} values. On the other hand, the PCA-based method is a regression analysis involving the SNP genotypes and PC loadings.

For the F_{ST}-based method, the R-package OutFlank (Whitlock & Lotterhos, 2015) was utilized. The distribution of neutral F_{ST} can be approximated by a chi-squared distribution (Whitlock & Lotterhos, 2015). Using the Weir and Cockerham (1984) formula two measures of F_{ST} were calculated. The first one is the original formulation which could result in negative values when the computed F_{ST} is already low and is then corrected for sample size (i.e. low sampling per population) (Whitlock & Lotterhos, 2015). However negative values are not possible in a chi-squared distribution expected for F_{ST} values. Therefore, F_{ST}' (i.e. FSTNoCorr) was also calculated without correcting for the sample size. The F_{ST}' was used in fitting the chi-squared distribution, and while all loci are expected to deviate between F_{ST} and F_{ST}' it was assumed that these deviations were similar for each locus (Whitlock & Lotterhos, 2015). This was investigated by plotting F_{ST} values against F_{ST}' values, which is expected to show a positive linear relationship. The relation of heterozygosity and F_{ST} was also visualized to check for unusual loci with relatively low heterozygosity but high F_{ST} values. Loci with low heterozygosity (< 0.1) were excluded from the analysis since they have a different F_{ST} distribution (Whitlock & Lotterhos, 2015). The mean value of F_{ST}', needed for estimation of the chi-squared distribution, was also computed. The null distribution was estimated using a subset of the SNP data. The genotype data was first pruned based on a LD threshold (maximum correlation coefficient) of 0.20 using the R/SNPRelate resulting to a total of 9417 SNPs used in the distribution inference. This subset was used because deviation from a chi-squared distribution can result from non-independent representation of the genome. Although the trimmed SNP set was not truly independent but quasi-independent, trimming minimizes the overrepresentation of regions with several loci exhibiting the same signal which may be due to relatively low recombination or extensive sweep. In addition, the shape of the distribution was inferred based only using the core F_{ST}' values, that is, the bottom and top 5% values were trimmed. Chi-squared distribution of df = 2 was inferred from the quasi-independent and trimmed data using a likelihood approach. The log-likelihood of df was maximized given the observed F_{ST}' values (i.e. FSTNoCorr). The F_{ST}' values of the whole dataset were then compared with the inferred distribution. The ratio of the product of F_{ST}' and df to the mean F_{ST}' was calculated together with their associated P-values based on the inferred null distribution. False discovery was controlled using the correction method of Storey and Tibshirani (2003). Q values were determined for each SNP and outliers were declared based on a false discovery rate of 5%, i.e. q value of less than 0.05.

For the PCA-based method, SNPs associated with population structure were investigated using the R package pcadapt (Luu et al., 2017). First, the number of relevant principal components (PCs) was determined using a scree plot. For this, the number of PCs (K) were plotted against the percentage of variance explained, with the 1st PC explaining the highest amount of the variation and the percentage dropping as the number of PCs increased. Important PCs are typically identified as part of a steep decline to the left-hand side of the plot whereas random variation by a straight line that approaches horizontal, to the right of the plot. It is recommended that PCs to the left of this straight line are retained. There was no obvious point in the graph where the declining amount of variance explained equilibrated into a straight, nearly horizontal line, although the rate of decline was decreasing. Therefore, non-graphical solutions to the scree plot were employed. The elbow of the plot was determined using a numerical solution by computing for the acceleration factor using the nScree function in the R package nFactors (Raîche, 2010). The acceleration factor identified the first three PCs to be retained and this was verified further using an independent Bayesian model selection method (Auer and Gervini, 2008) as implemented in the R package PCDimensions (Wang et al., 2018) (Appendix Fig. 3.8.). SNP genotypes were then regressed with the three components and associated z-scores were obtained using the R/pcadapt. Outliers were detected based on Mahalanobis distance. Vector of z-scores of the three PCs for each SNP were used to obtain the squared Mahalanobis distance which when corrected by the genomic inflation factor follows a chi-squared distribution with three (the number of PCs retained) degrees of freedom, assuming no outlier. The first PC was also investigated solely as it the most important and has a clear biological interpretation. This was accomplished by conducting a component-wise procedure in pcadapt. The test statistic is simply based on the correlation between the 1st PC and each SNP. Gaussian approximation was used, and the standard deviation of the null distribution was estimated to compute for P-values.

SNPs that were detected by both the F_{ST}- and PCA-based methods were further validated by associating phenotypic endophyte infection data of individuals in the population (from Chapter 2) with the genotype information. Logistic regression was conducted with infection as the response variable. The Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995) was used to correct for p-values accounting for multiple SNPs tested since there were not enough SNPs to carry out a similar control of false discovery rate with q values. Significant SNPs were declared when the corrected p-value was less than 5% alpha. Logistic regression was also conducted for all the SNPs in the dataset (i.e. 31K) to determine other possible candidate loci associated with selection and the trait of interest. In this case, a false discovery rate of 5% was considered and q values were computed for each SNP. Further, for significant SNPs detected by the F_{ST} - and PCA-based methods, the linear relation of genotype and endophyte infection rate was investigated with the Cochran-Armitage trend test (Armitage, 1955; Cochran, 1954) using the independence_test function in the R/coin (Hothorn et al., 2008).

3.3.8. SNP annotation

The important SNPs, i.e. those commonly detected by several methods, were investigated for possible candidate genes controlling the trait of interest – endophyte transmission. This was accomplished with the assistance of an AgResearch bioinformatician. Briefly, the *L. perenne* scaffold containing the SNP of interest was first determined, along with the length of the scaffold and the position of the SNP in the scaffold. Then, the sequence information of the scaffold was used in a BLAST search to identify candidate genes. Since a scaffold is usually large, they may contain several genes. The position of the SNP was therefore used to identify the gene closest to that position. In summary, the candidate genes were determined based on sequence similarity of a particular *L. perenne* scaffold and the position of the SNP in the scaffold.

3.4. Results

3.4.1. Characterization of SNP dataset and filtering

The initial working dataset with 219,870 SNPs was characterized. It contained 15.61% missing data distributed across all SNPs and 190 individuals, which included representatives of the early (C2; n = 90) and late generation (C6; n = 90) of population PGG04 plus biological replicates (8) and a company-wide control (2 reps). Read depth ranged from 1x (excluded zero read depth) to 1612x and large proportion of variants had low depth (i.e. 1 to 5x) while relatively few had extremely high depth (i.e. above 200x) (Appendix Fig. 3.8). Nevertheless, majority of the variants had a read depth between 5 to 150x. This was the case between two batches of GBS libraries between the early (PGG04-C2) and late generation (PGG04-C6) population (Appendix Fig. 3.8). The utility of the dataset was tested by running a principal coordinate analysis (PCoA) (Appendix Fig. 3.9). The dataset clearly separated the two generations into two groups and that there was no obvious clustering due to batch. In addition, the labelled biological replicates (“0”, “+”, “x”) clustered together, at almost the same position, except for replicates of PGG04-C6-A10 and PG04-C6-D11. Nevertheless, the genetic distances separating their replicates are still very small.

After filtering for missing data, read depth, and consistency across biological replicates, there was an improvement in the dataset (Table 3.1). Although, there were only 84,402 remaining SNPs, missing data was now only 4.26% (Appendix Fig. 3.10) and SNPs with mostly low read depth were removed. In addition, filtering resulted in a decrease in genetic distance estimates between biological replicates as depicted in PCoA (Fig. 3.5).

Table 3.1. Parameters pre- and post-filtering, for the GBS dataset used for population genetic analysis. The starting dataset composed of 219K SNPs. Filtering resulted to 84K SNPs and missing data improved from about 16% to 4%.

Parameters		Original VCF[1]	Post-filter VCF[1]
General	No. of SNP	219,870	84,402
	% missing	15.61%	4.26%
	% Heterozygote	20.96%	26.37%
	Average MAF[2]	0.22604	0.22288
Alleles[3]	C	19.34%	21.58%
	G	19.25%	21.52%
	T	12.40%	13.14%
	A	12.37%	13.14%
	R	6.39%	7.77%
	Y	6.39%	7.70%
	S	2.53%	4.06%
	K	1.99%	2.58%
	M	1.96%	2.51%
	W	1.48%	1.74%
	N	15.61%	4.26%
	heterozygous indel	0.21%	-
	deletion	0.07%	-
Depth	Min	1	1
	1st Quartile	4	23
	median	19	50
	3rd Quartile	56	85
	Max	1612	1610

[1] variant call file; [2] minor allele frequency

[3] nucleic acid notation where R = A or G; Y = C or T; S = G or C; K = G or T; M = A or C; W = A or T; and N = any base

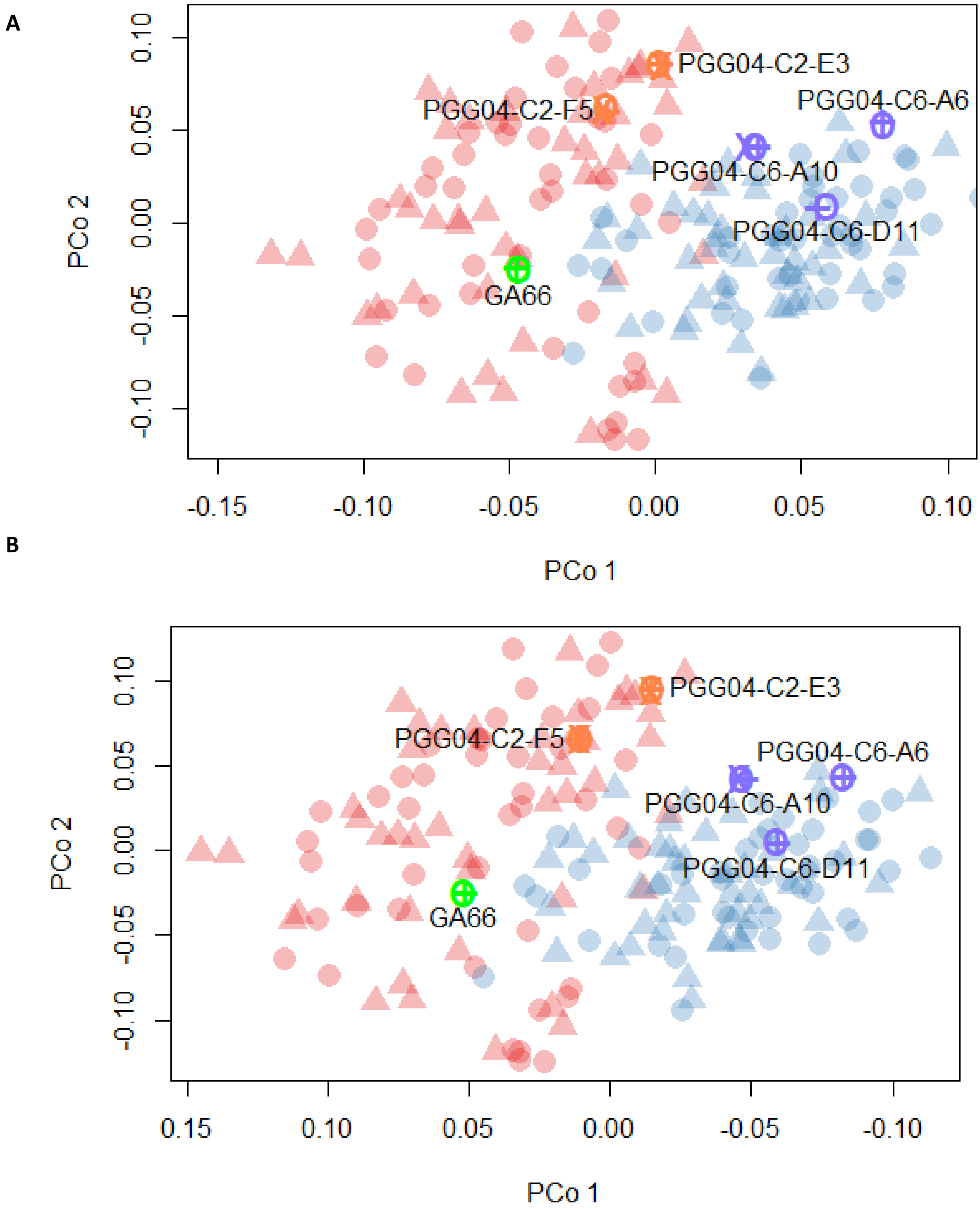


Figure 3.5. PCoA before and after (B) filtering shows improvement in distance-based clustering of the replicates ("0", "+", "x") especially with PGG04-C6-A10 and D11. The PCo1 axis in B is in reverse for consistency but the general relationship remains – PCo1 separates the generations of PGG04 and two library batches (circle or triangle) are spread randomly. In addition, the early generation (transparent red) still showed a larger spread in the plot while the clustering of the late generation (transparent blue) was closer, which again suggest a reduction in diversity likely due to selection.

3.4.2. Genetic diversity and population structure

The minor allele frequency (MAF) tended to occur most commonly at lower frequencies (i.e. between 0.0 to 0.20) resulting in a right-skewed frequency distribution for both generations (Fig. 3.6). The average minor allele frequency (MAF) of PGG04-C2 was 0.2209, while for PGG04-C6 it was 0.2191. At lower frequencies, the late generation was enriched with alleles at MAF of 0.02 – 0.08 (i.e. 2nd to 4th bin in Fig. 3.6), while for the early generation this occurred at MAF 0.10 – 0.20. The excess of alleles at relatively low MAF is consistent with a population under selection. Observed heterozygosity (H_o) was also positively skewed for both PGG04-C2 and C6 (Fig. 3.7 A). However, average H_o is not significantly different (t-test p-value of 0.6067) between the two generations of PGG04, with an average of 0.2712 for the early generation and 0.2717 for the late generation. Nevertheless, in contrast to the early generation, the late generation had several SNPs at a relatively lower H_o of 0.05 to 0.15 (2nd bin in Fig. 3.7 A). On the other hand, the distribution of expected heterozygosity (H_e) or gene diversity was almost flat in the middle, has relatively fewer SNPs with low heterozygosity (less than 0.1), and a lot of SNPs with the maximum heterozygosity of 0.5 (Fig. 3.7 B). From an average of 0.3069 in PGG04-C2, H_e slightly declined to 0.3033 in PGG04-C6 (t-test p-value of 4.85e-07), which was expected after selection. There was also a higher number of SNPs with lower diversity (H_e = 0.04 to 0.16) in the late generation compared with the early generation.

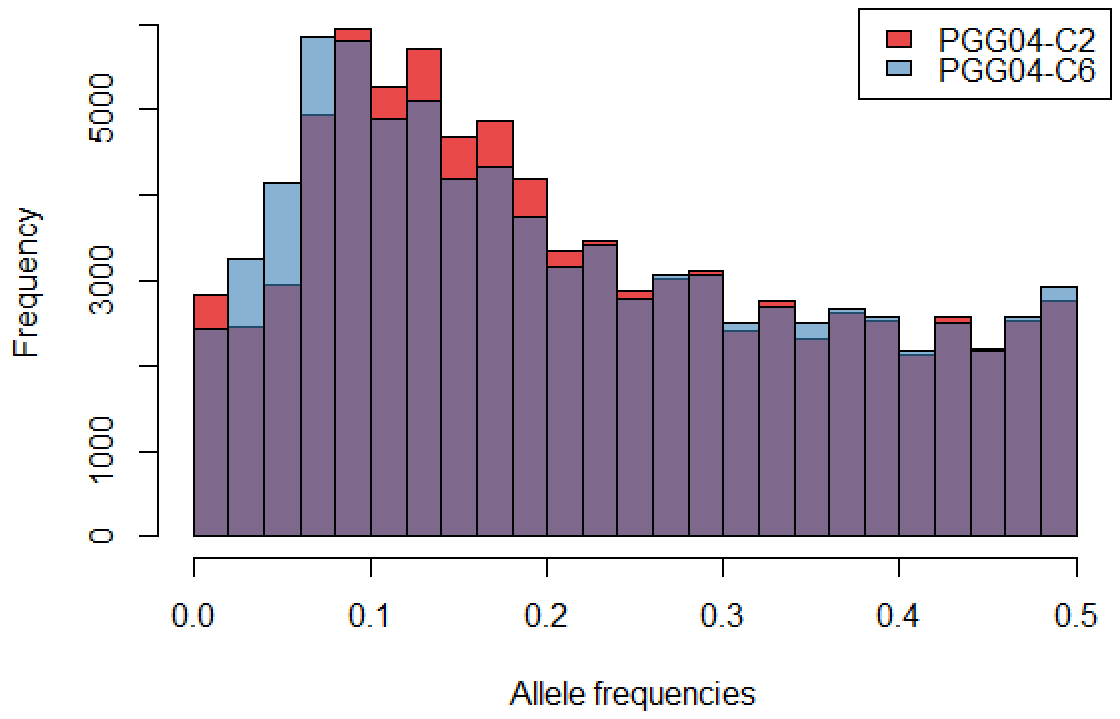


Figure 3.6. Minor allele frequency (MAF) distribution of the early and late generations of PGG04. Loci in the mode have MAF around 0.1. PGG04-C6 (late) has more loci with MAF lower than 0.1 while PGG04-C2 (early) has more loci higher than 0.1.

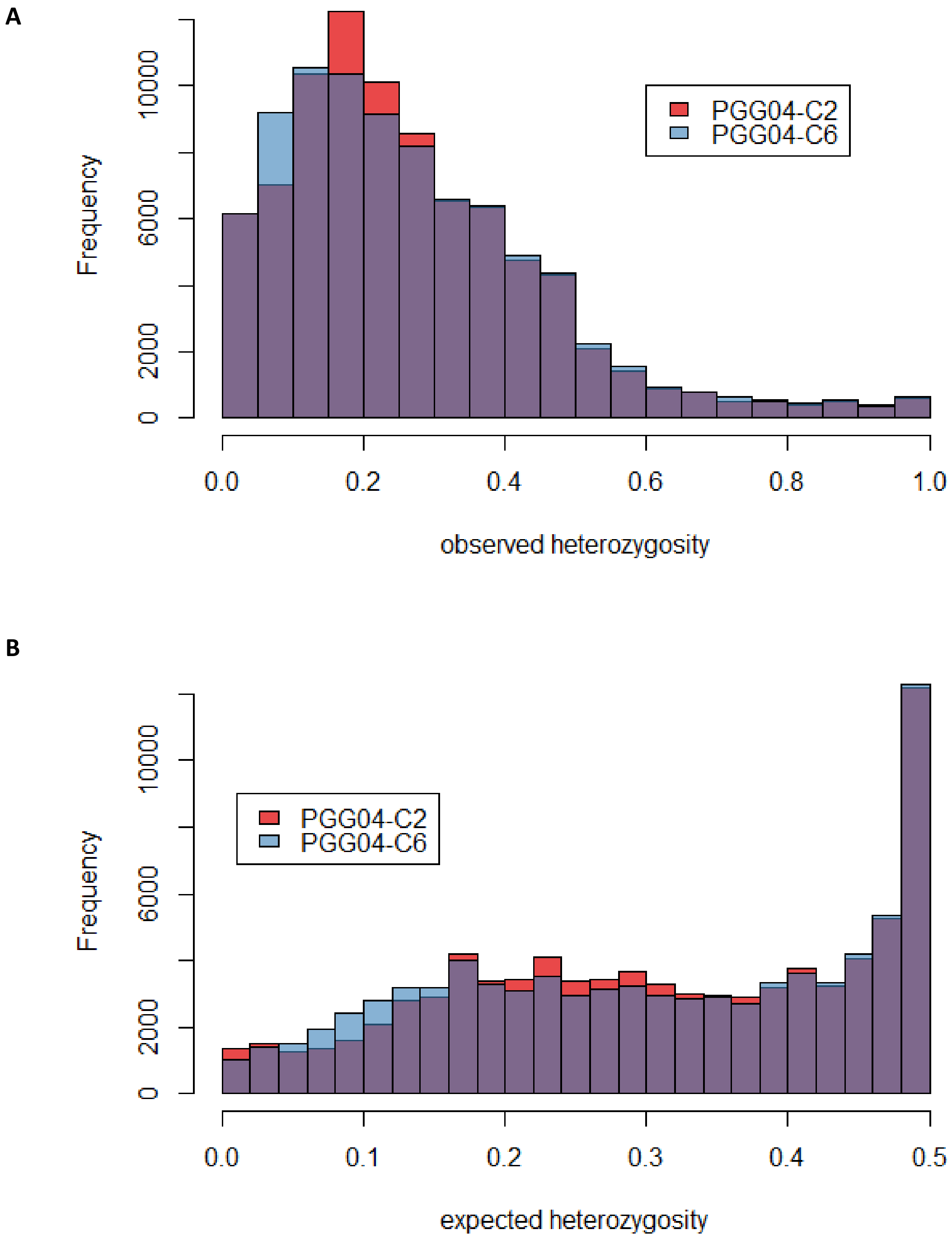


Figure 3.7. Distribution of observed (A) and expected heterozygosity (B) in early (PGG04-C2) and late (PGG04-C6) generations of population PGG04. The shape of the distribution of the observed heterozygosity (H_o) was similar to that of MAF (Fig. 3.6). The distribution for the expected heterozygosity (H_e) was generally flat in the middle, with few loci having H_e lower than 0.1 and an excess of H_e = 0.5.

The overall fixation index, F_{IS}, decreased significantly from 0.1292 in the early generation to 0.1174 in the late generation of PGG04. This reduction was supported by non-overlapping 95% confidence interval. In both generations, there were a few loci that were completely heterozygous (i.e. F_{IS} of -1) and this was higher in PGG04-C6 (80) than C2 (67). In contrast, there were loci that were completely homozygous (F_{IS} of +1). These occurred more frequently in C2 (653) than in C6 (468) (Table 3.2). This was consistent with the observed reduction in inbreeding between the two generations of PGG04. In general, selection would be expected to increase inbreeding which was not the case for this data. Reduction in inbreeding is a result of an increase in heterozygote plants. It is possible that selection has favoured heterozygous plants because these plants have better compatibility compared to the homozygous form.

Comparing the early and late generations of PGG04, overall genetic differentiation, F_{ST}, at 95% confidence interval was from 0.0307 to 0.0315 which was evidence that the PGG04-C6 differentiated from C2, that is, F_{ST} was not zero (Table 3.2). However, F_{ST} values less than 0.05 typically correspond to little genetic differentiation. F_{ST} values for individual SNPs ranged from 0 to 0.5691; the latter indicates very high genetic differentiation in that locus. The actual minimum F_{ST} was -0.0132. However, the theoretical limit only allows up to zero, hence negative F_{ST} values were considered as zero. Negative values result from correcting for bias in sample size, that is, very low F_{ST} values (i.e. near zero) when corrected for sample size reduce to lower than zero estimates. In the dataset, there were 25,490 SNPs or 30.64% with negative F_{ST} values.

Table 3.2. Inbreeding coefficients (F_{IS}) of the early (C2) and late (C6) generations of PGG04 and fixation index (F_{ST}) between the two generations.

Parameters	F_{IS}		F_{ST}
	PGG04-C2	PGG04-C6	
Min (no. of loci) [a]	-1 (67)	-1 (80)	-0.0132 (25490)
Max (no. of loci) [b]	+1 (653)	+1 (468)	0.5691
Median	0.1055	0.0896	0.0111
Mean	0.1292	0.1174	0.0311
95% confidence interval (mean)	0.1265 - 0.1320	0.1147 - 0.1203	0.0307 - 0.0315

[a] For F_{IS} the number of loci with min F_{IS}; for F_{ST} the number of loci with F_{ST} values less than zero (effectively zero).

[b] For F_{IS} only, the number of loci with maximum F_{IS}

Population structure was investigated with two dendrogram construction approaches. First, a bootstrap supported UPGMA tree was constructed using Provesti's distance. High bootstrap support values were obtained, for example, more than a quarter of all the nodes had support above 90%. The only exception were eight nodes (5%) with support values of less than 60% but above 52%. In addition, the trees at higher cutoff (i.e. 70%) were almost identical to that of 50% cutoff (tree not shown). The unrooted tree appeared as one relatively undifferentiated group and the separation of the two generations was not evident, for example by relatively long branch or branches. Nevertheless, each generation still formed largely separate clusters. The tree formed roughly 12 clusters with groups on top (i.e. clusters 1-5) composed mainly of PGG04-C2. On the other hand, groups at the bottom (i.e. clusters 6-12) were mainly PGG04-C6 (Fig. 3.8 A). The only exemptions were cluster 3 (PGG04-C6) on top and cluster 11

(PGG04-C2) below, as well as PGG04-C2-C12 grouping with the late generation (cluster 9). A second approach was conducted to provide a different perspective of the population structure using dendrograms. The majority-consensus of 1000 UPGMA trees generally supported the results of the first tree, although they presented different structures. The second tree can also be described as a single group. Moreover, the divergence of the two generations was not quite so evident, with exception of the apparent divergence of some individuals from PGG04-C6 (bottom left in blue brackets) (Fig. 3.8 B). These plants represent the most genetically dissimilar in PGG04. These were plants in clusters eight and nine based on cluster labels in the first tree (Fig. 3.8 A). While the separation of the early and late generations of PGG04 was not evident, individuals tend to group within the same generation. In general, the groupings in the first tree was preserved. The exemptions were mostly in the early generation, in clusters 1, 2, and 5. A large part of cluster 1 (i.e. 1a in Fig. 3.8 B) is grouped together, while the rest can be found within the branches highlighted by red bracket. Similarly, cluster 2 individuals can be found in that highlighted part of the tree. Some individuals from cluster 5 can also be found in that part of the tree while the rest formed three groups, that is clusters 5a, 5b and 5c in Fig. 3.8 B.

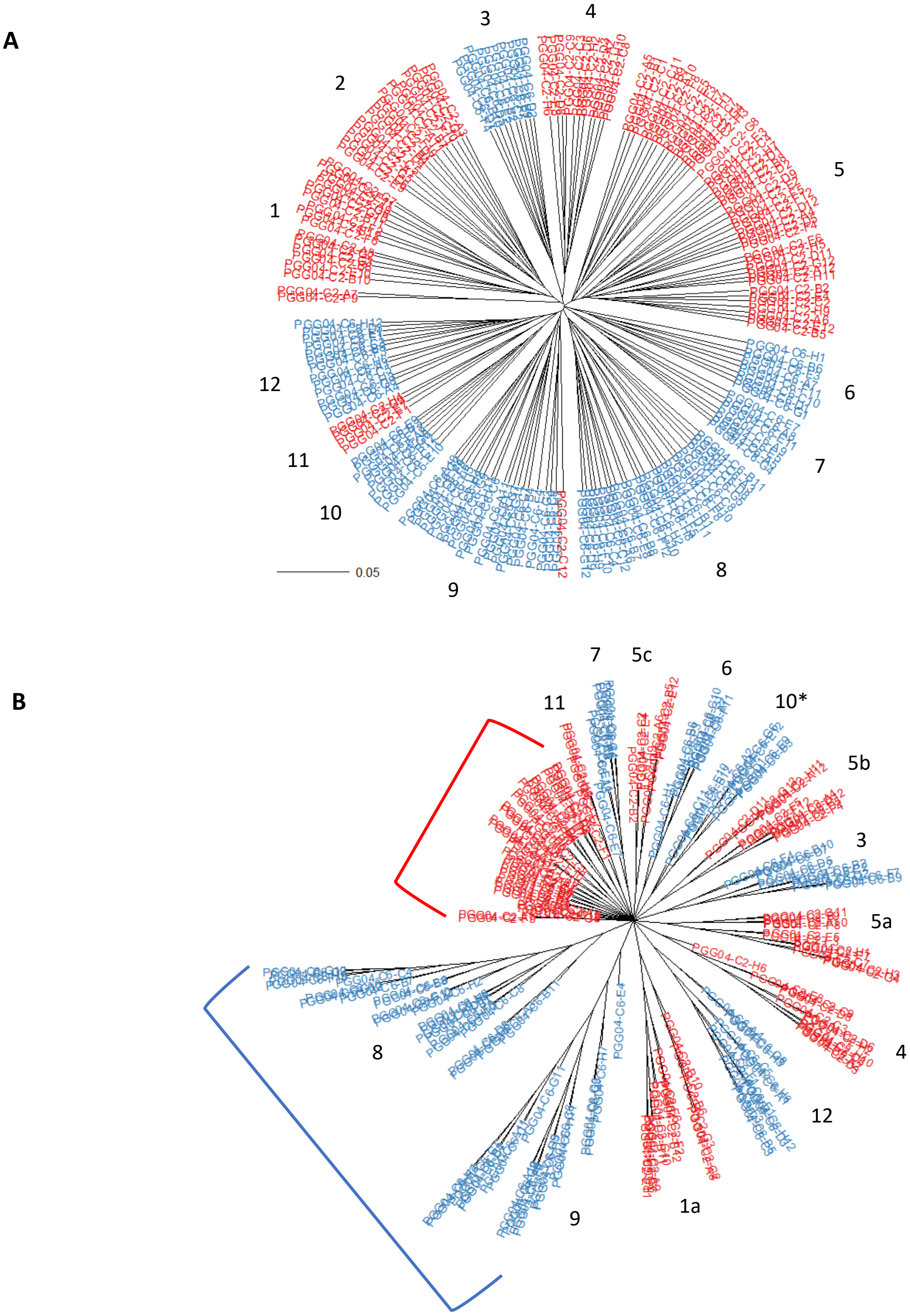


Figure 3.8. Population structure based on UPGMA unrooted tree of Provesti's distances. The bootstrap supported dendrogram (A) showed one large group of PGG04 although individuals from the early (PGG04-C2, red) and late (PGG04-C6, blue) generations tend to cluster together. Majority-consensus tree (B) also looked like one group but some PGG04-C6 plants appear to be differentiated.

The structure of genetic variation was also visualized by PCA biplot. The plot showed distinction of the generations of PGG04 at the first principal component (Fig. 3.9). PGG04-C2 (red) is on the right and is highly dispersed, while PGG04-C6 (blue) is on the left and more clustered together. Ellipses representing a 95% confidence interval were also drawn for each group. Some of the late generation plants as well as a few early generation plants are in the middle of the plot (i.e. PC1: -0.5 to 0.5 and PC2: -0.5 to 0.5), and are inside the two ellipses. These are plants across generations that share high genetic similarity. Considering cluster labels in Fig. 3.8 A, the plants inside the two ellipses in the PCA were mostly cluster 3 and cluster 5 plants. Interestingly, members of cluster 3 are PGG04-C6 plants that grouped with PGG04-C2 plants (top) in the Fig. 3.8 A UPGMA tree. On the other hand, dots at both ends represent plants that have low genetic similarity. Based on Fig. 3.8 A cluster labels, blue dots at the left most of the PCA plot (i.e. PC1 score of less than -0.1) (Fig. 3.9) were members of clusters 8 and 9. As mentioned previously, these clusters appear to be diverging from the rest of PGG04 based on Fig. 3.8 B UPGMA tree.

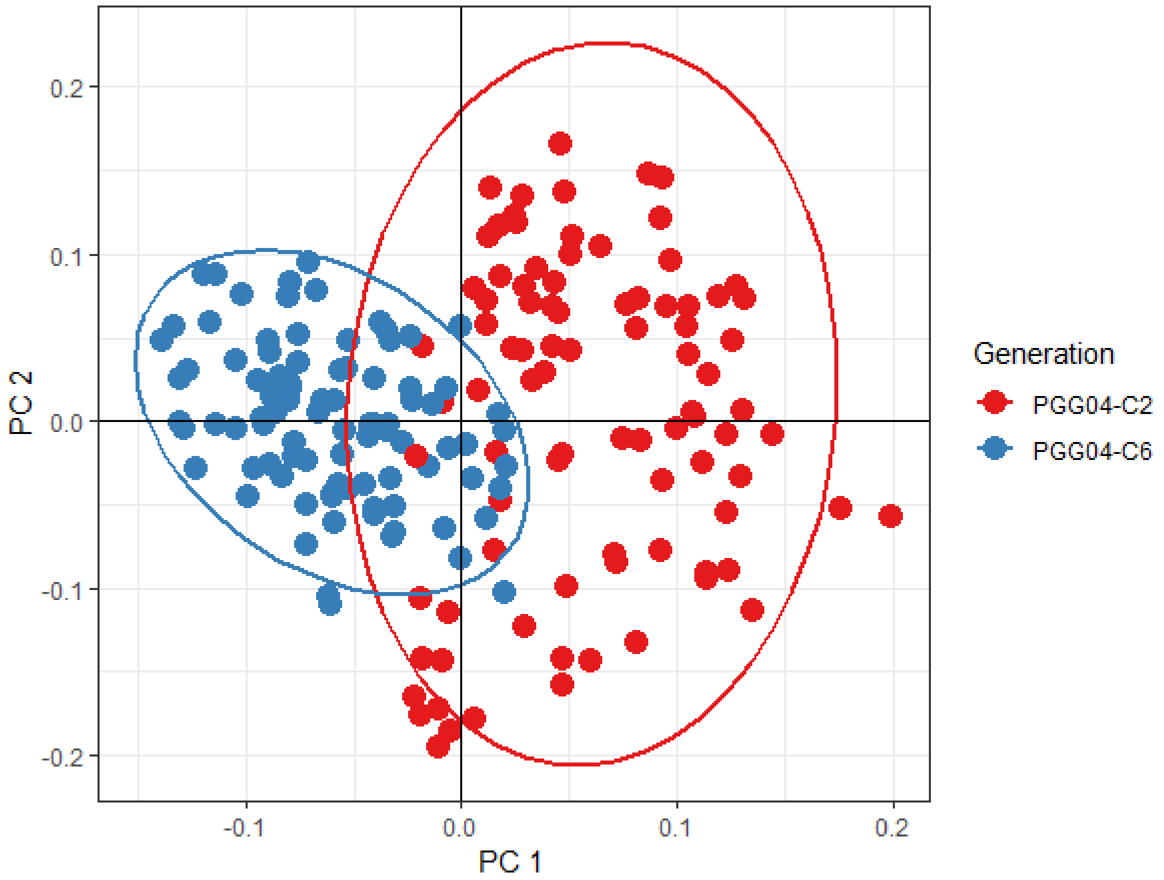


Figure 3.9. Population structure based on a PCA biplot of the first two components. Generations of PGG04 can be clearly separated at PC1. The early generation (PGG04-C2, red) was more dispersed reflecting relatively higher diversity than the late generation (PGG04-C6, blue).

The result of AMOVA showed that diversity was mainly due to genetic differences within each sample; about 10% was due to variation within generation and only about three percent can be attributed to genetic differences between generations (Table 3.3). This was consistent with the structure of the UPGMA tree (Fig. 3.8), where several branches shared by some individuals radiate from the centre as opposed to the two main branches that distinguish the two generations. It was also consistent with the relatively low F_{ST} (0.0307 - 0.0315) obtained (Table 3.2). Nevertheless, the variation between generations was significant, as were the other two sources of variation, based on permuting the AMOVA results 1000 times (Appendix Fig.

3.11). The PCA biplot shows the separation of the two generations supporting the AMOVA result of significant genetic variation between generation.

Table 3.3. Results from analysis of molecular variance (AMOVA). A relatively low, yet significant, amount of observed variation can be explained by genetic differences between the early and late generations of PGG04.

Sources of variability	df	Sum of Squares	Mean Squares	% variance	P-value[1]
Between Generation	1	85544.87	85544.87	3.01	0.000999
Within Generation	178	2529205.26	14209.02	10.87	0.000999
Within Samples	180	2042081.44	11344.90	86.12	0.000999
Total	359	4656831.58	12971.68	100.00	

[1]based on 1000 permutation test of AMOVA results

Population structure was also investigated using a model-based clustering method implemented in STRUCTURE version 2.3.4. Although, there were two groups (C2 and C6) in terms of the two generations of PGG04 population, the optimal number of subpopulations were also identified based on SNP data. Using the Evanno et al., (2005) method, delta K indicates two as well as four hypothetical ancestral populations also. Figure 3.10 below shows the probability of membership of each plant to two groups (K = 2) while Fig. 3.11 to four inferred populations (K = 4). At K = 2, the two groups generally correspond to the early and late generations of PGG04 (Fig. 3.10). The majority of the individuals were correctly classified. Also, considering membership to one or the other group as being at least 60%, 88% of PGG04-C6 were correctly placed into group 1 with no mismatches while the remaining 12% were admixtures. For PGG04-C2, almost 75% of PGG04-C2 plants were correctly placed into group 2 with the remaining plants being admixtures (9) or mismatches (14). At an 80% membership threshold, the mismatches in PGG04-C2 was halved but the matched grouping also decreased because these plants (mismatched and individuals with membership probability of less than 80% but at least 60%) were then considered as admixtures. Following the clusters labels in Fig. 3.8, a lot of admixture, especially mismatches in PGG04-C2, were plants from cluster 4. For PGG04-C6, the admixtures are mostly from cluster 3. Cluster 3 plants, being "mixture" of the two generations was consistent with the PCA biplot, where these plants can be found in the overlap of the two ellipses (Fig. 3.9). At K = 4, the two generations have a distinguishing pattern (Fig. 3.11). The early generation can be characterized by plants with high membership probability with ancestral population 2 (AP2, orange), AP3 (grey), AP4 (yellow) or a combination of any two of these APs, and low probability of membership with AP1 (blue). Almost half (44) of PGG04-C2 plants have at least 60% membership probability with either AP2, AP3 or AP4; while 17 plants have combined probability of ~80% for any two of APs 2, 3 and 4. For the remaining plants, the highest probabilities (i.e. less than 60%) were generally with the three APs as well. The late generation was the opposite and is generally characterized by relatively lower probability for APs 2, 3 and 4 in favour of AP1 (blue). A total of 28 plants from PGG04-C6 have probability of at least 60%with AP1; while 30 plants have the highest probability (33% to 59%) with AP1. In the PCA biplot (Fig. 3.9), PC1 separates the early generation in the right, and the late generation in the left hence, the direction of selection was possibly towards the left side. Indeed, the 28 plants with the highest probability of membership to AP1 were at the leftmost side in the PCA biplot. Furthermore, several of these (20 of 28) plants were classified in clusters 8 and 9, which appeared to diverge from the rest of PGG04 in Fig. 3.8 B, potentially due to selection.

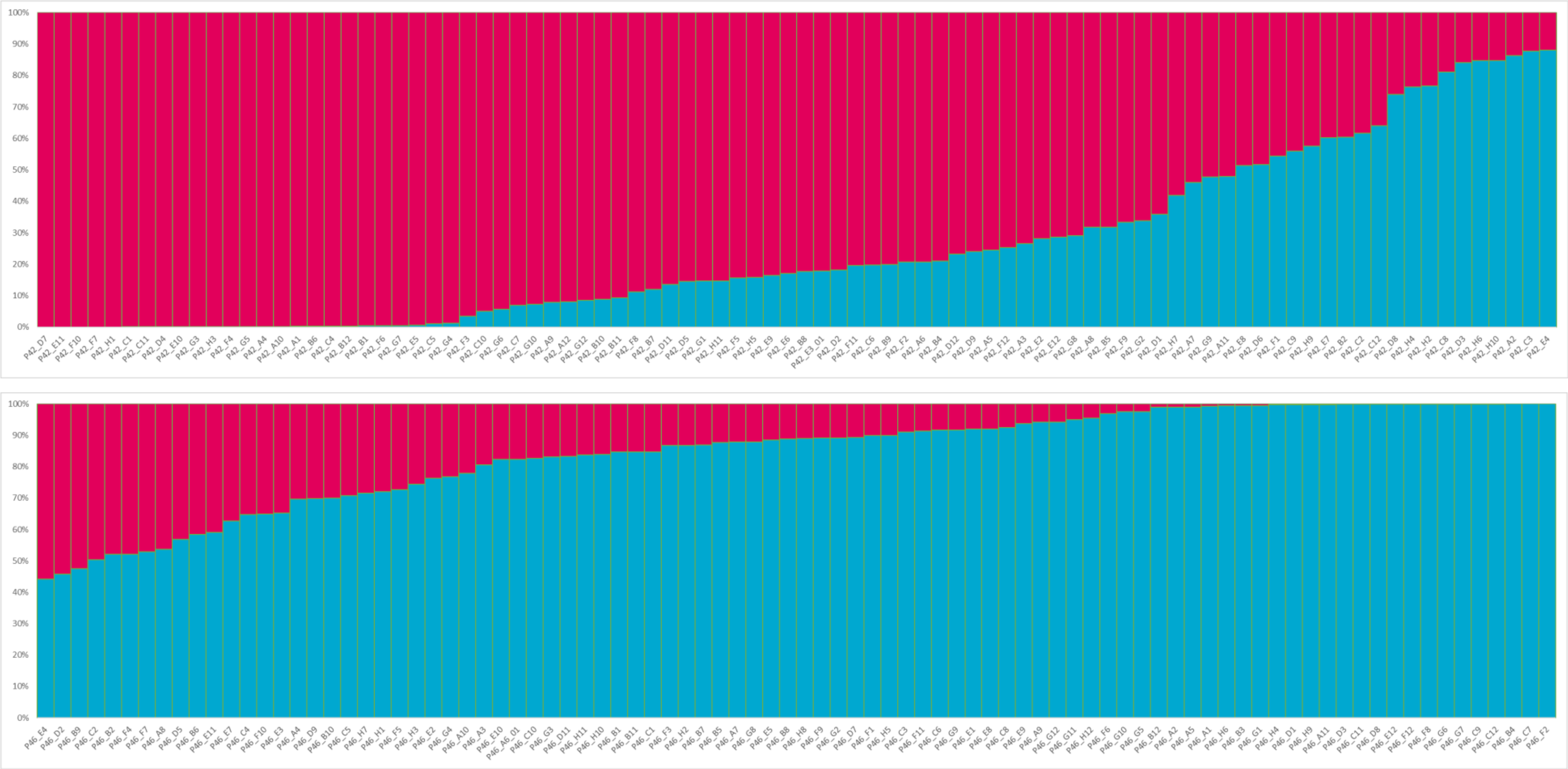


Figure 3.10. Model-based clustering method using STRUCTURE assuming two groups. The groups generally correspond to the two generations of PGG04, namely PGG04-C2 (top) and PGG04-C6 (bottom).

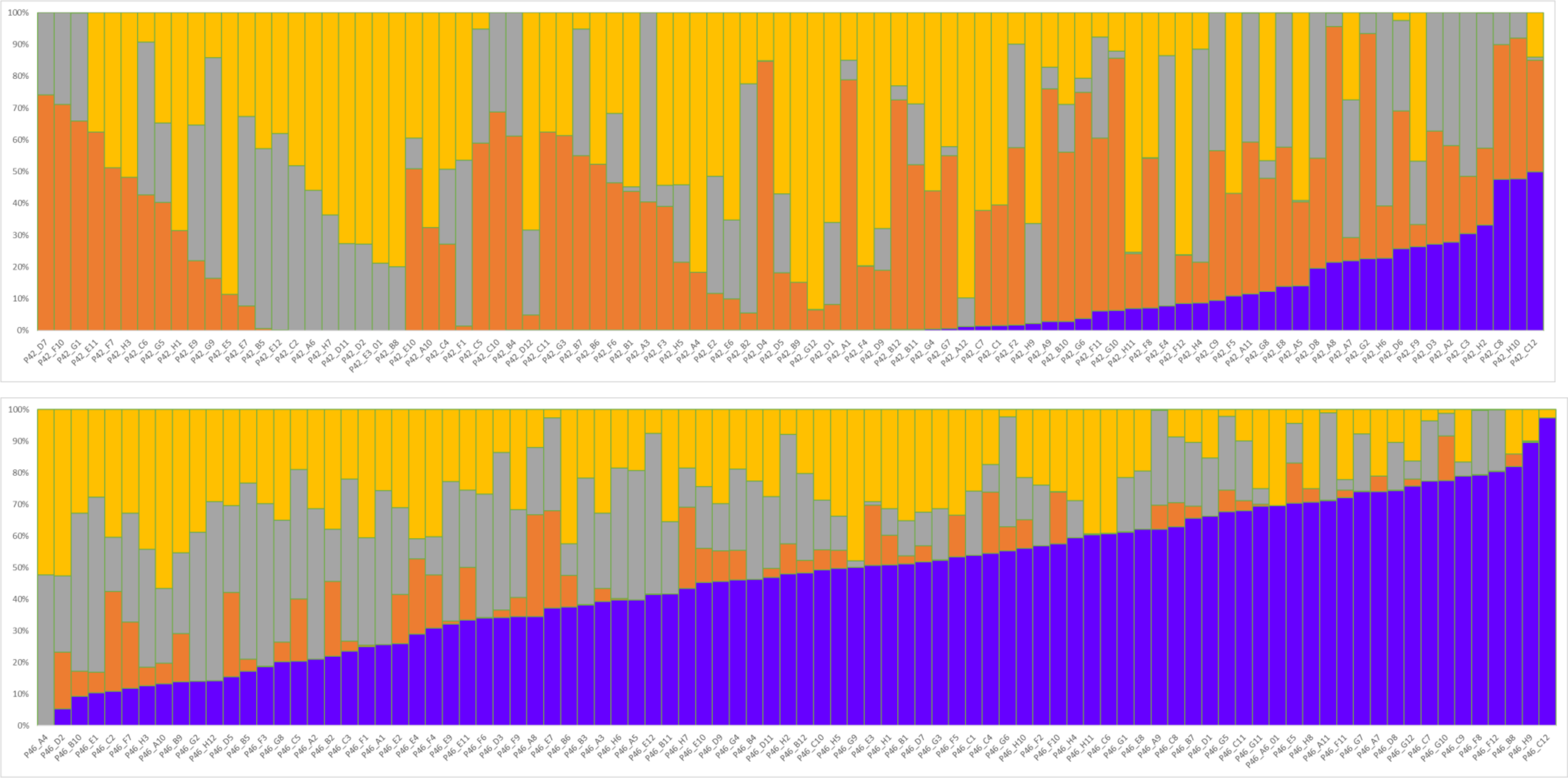


Figure 3.11. Model-based clustering method using STRUCTURE assuming four hypothetical ancestral populations. Allele frequency of AP1 (blue) was enriched in the late generation (bottom) as compared to the early generation (top). In parallel, genetic diversity related to the three other population decreased.

3.4.3. Signatures of selection

F_{ST} outlier approach

The genotype data were first investigated to satisfy the assumptions of OutFlank. The distribution of F_{ST} across expected heterozygosity (H_e) values showed a general expectation that the highest possible F_{ST} value increases with H_e (Fig. 3.12 A). Although, no loci with low H_e and relatively high F_{ST} were observed, low heterozygosity loci (i.e. $H_e < 0.1$) were removed as this was recommended for the OutFlank method. These loci have a different F_{ST} distribution especially in the case of minor allele frequency less than 5% (Beaumont and Nichols, 1996), which were filtered out early on in this study. There was a good linear relationship between the corrected and uncorrected F_{ST} (Fig. 3.12 B). This was expected because both generations have the same sample size of 90 individuals each and missing data was very low (about 1.24%). F_{ST} can be inflated when not correcting for sample size but this deviation is acceptable as long as it is similar across loci. In Fig. 3.12 B no data points deviated considerably from the linear relation.

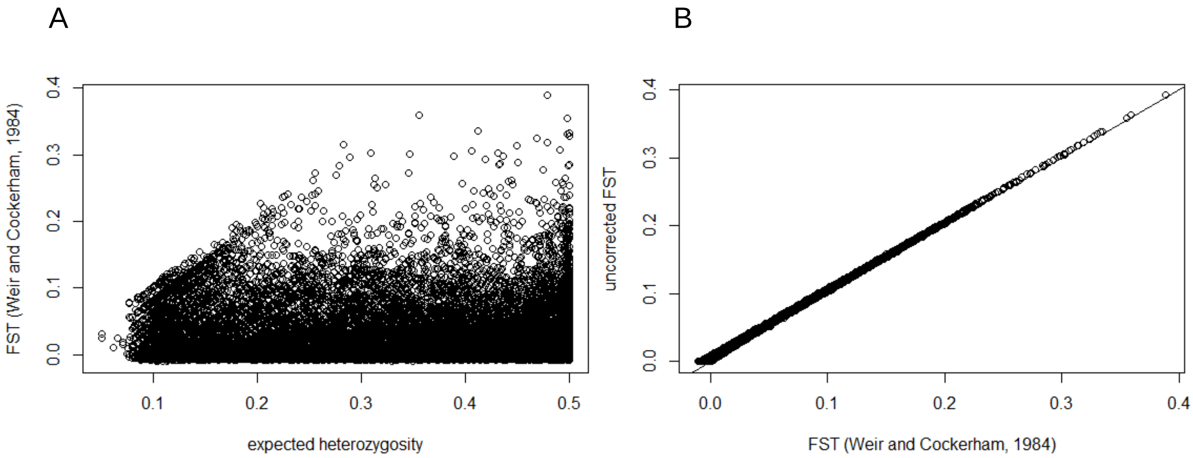


Figure 3.12. Heterozygosity vs F_{ST} plot (A) and correlation of the F_{ST} of Weir and Cockerham (1984) and its uncorrected version (B). Maximum F_{ST} values are limited by the amount of heterozygosity in A. F_{ST} when not corrected for sample size will be inflated but is strongly correlated with the corrected F_{ST} in B.

Using quasi-independent SNPs (9K), the mean F_{ST} was estimated to be 0.0237 while the uncorrected F_{ST} was 0.0290. This was lower than previously calculated from the 84K SNPs. A possible reason for this discrepancy was the non-inclusion of SNPs with relatively higher F_{ST} values because of data quality filtering (i.e. from 84K to 31K) and/or subsetting for quasi-independent SNPs (i.e. from 31K to 9K). Another possible explanation was that the iterative process of OutFlank in the estimation of mean F_{ST} successively removes "outlier" SNPs which have to the potential to pull the average F_{ST} up.

The inferred degrees of freedom of the chi-squared distribution of the uncorrected F_{ST} was 2 as opposed to that expected in the traditional (Lewontin & Krakauer, 1973) approach of number of populations minus 1 (i.e. df = 2-1). A histogram of F_{ST} (uncorrected) values was plotted together with the inferred distribution and is shown in Fig. 3.13. This figure shows that in general, the data follows the expected distribution including at the right-tail (Fig. 3.13 B) for which the test statistic was based. The inferred distribution was based on the estimated mean F_{ST}' which was calculated iteratively, removing outlier loci every time and calculating df based on a maximum likelihood approach given the observed F_{ST}' values.

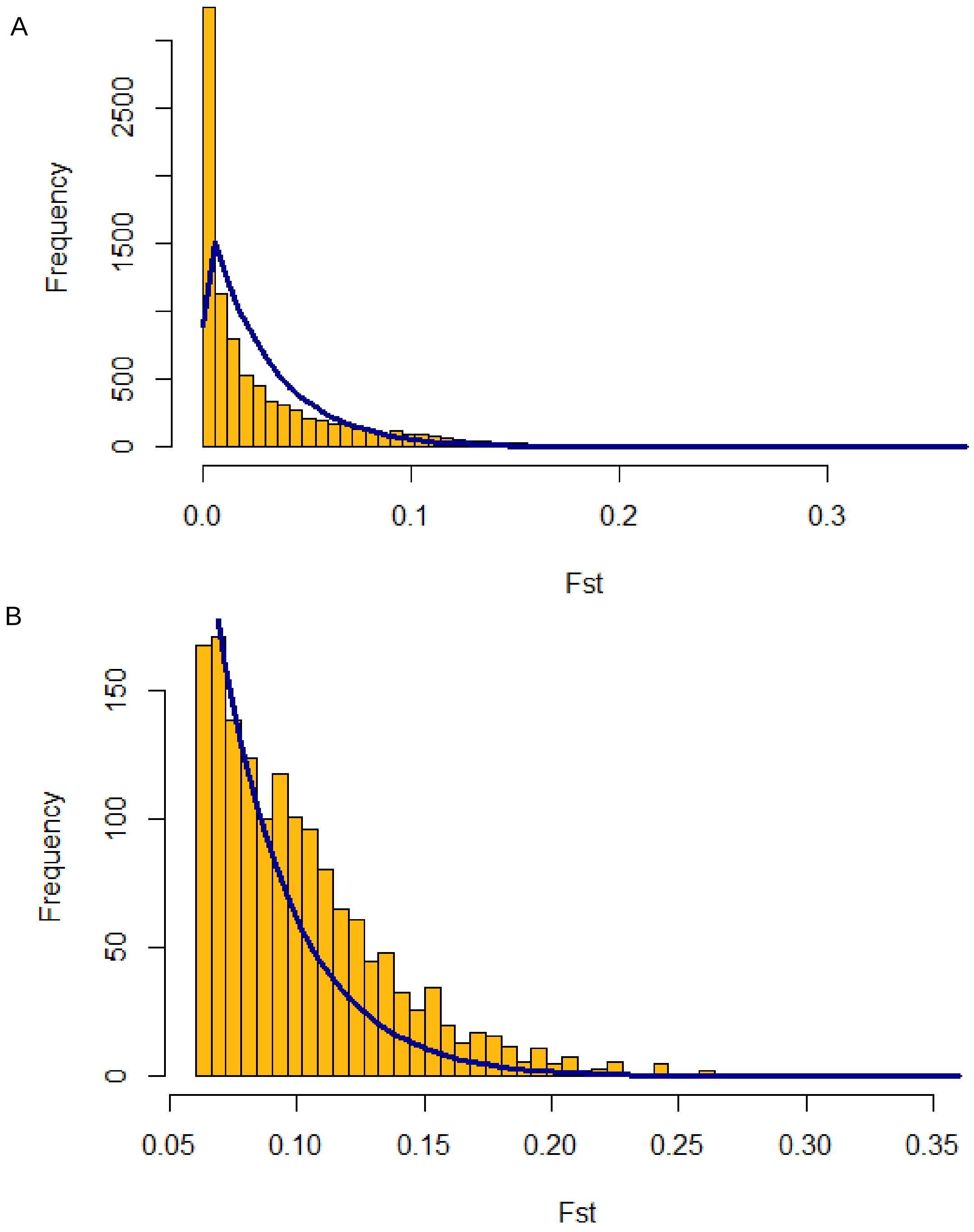


Figure 3.13. The distribution of the F_{ST} values (uncorrected) computed from quasi-independent SNPs follows a chi-squared distribution with 2 degrees of freedom (A). This df has the highest likelihood given the observed F_{ST} values. The distribution has a good relative fit also when zooming on the right tail (B).

F_{ST} values were calculated for the whole dataset (not only the 9K subset) and outlier loci were determined based on the inferred null distribution. A total of 47 SNPs was found to have relatively high F_{ST} of 0.2700 – 0.3886 (F_{ST}': 0.2758 – 0.3922) with heterozygosity of 0.2554 – 0.5000. Significant SNPs are highlighted in the H_e vs F_{ST} plot in Fig. 3.14 as well as in the Manhattan plot of p-values (right-tail) in Fig. 3.15. Outlier F_{ST} values were observed both in relatively low H_e (i.e. less than 0.3) and high H_e (i.e. 0.4 to 0.5), but only for high F_{ST} values. Selection (i.e. diversifying selection) can also lead to extremely low F_{ST} values. However, OutFlank is not recommended for this scenario.

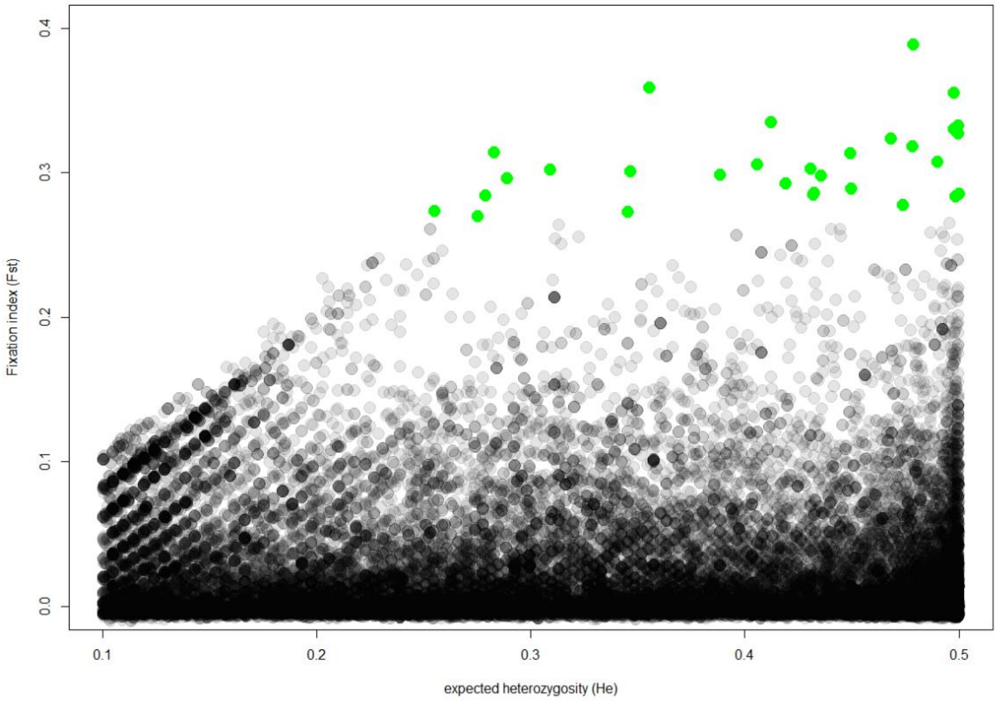


Figure 3.14. F_{ST} measures the reduction in heterozygosity of the generations relative to the total diversity in PGG04. The upper bound of F_{ST} value is determined by the available heterozygosity. Outlier F_{ST} values were highlighted in green for different levels of heterozygosity.

Outlier SNPs were detected across the seven chromosomes. Considering their relative physical position, the 47 SNPs detected roughly corresponds to 25 loci. Of interest were loci in chromosome 3 tagged by seven SNPs (S3_131682923 to S3_131682946), chromosome 4 with 5 SNPs (S4_78229071 to S4_78229108), and chromosome 7 also with 5 SNPs (S7_141199150 to S7_141199178). The highest F_{ST} was reported for a SNP in chromosome

6, namely, S6_39276213 (0.3886). This was followed by S7_154247285 (0.3585), S4_118265340 (0.3553), S4_49968467 (0.3349), and S2_128346898 (0.3324). These SNPs also have very high heterozygosity (0.4122 to 0.4997) except for S7_154247285, which has moderate to high H_e (0.3557). Among the loci tagged by multiple SNPs, the highest F_{ST} was reported for the region in chromosome 3 with seven SNPs (S3_131682923 to S3_131682946). Six out of seven SNPs have the same F_{ST} value 0.3143, while S3_131682936 have lower F_{ST} (0.2732). This locus has moderate diversity with H_e of about 0.2790.

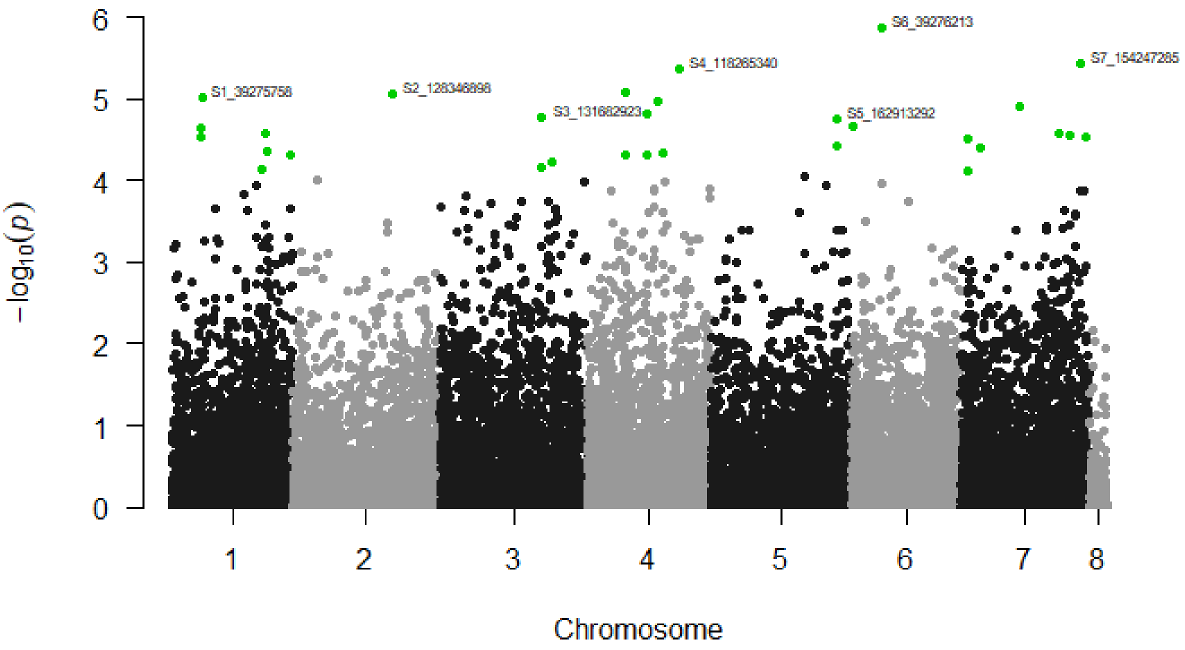


Figure 3.15. The test statistic: $\boldsymbol{FST'(df)}\big/\boldsymbol{\overline{FST'}}$, which follows a chi-squared distribution with two degrees of freedom, was computed for each locus and compared to the inferred null distribution to determine the right-tail p-values. Q-values were used to determine outliers, highlighted in green, considering a 5% false discovery rate.

PCA based analysis

The scree plot (Fig. 3.16 A) did not show a clear plateau, thus the number of principal components to retain could not be determined visually. Looking at the first 15 components, the first plateau can be observed at PC5 to PC6, although it continued to drop again after PC6 (Fig. 3.16 B). Therefore, it seemed reasonable to retain the first four components. Using non-graphical solutions as an alternative to the scree-plot, the number of important components was revised to the first three PCs only. The elbow of the plot was determined using a numerical solution, by computing for the acceleration factor. The acceleration factor or the point where the slope changes very quickly, was determined at PC = 4 (Fig. 3.17) and thus the PC before the elbow was the last to be retained. Retaining the first three principal components was verified with another strategy in determining the number of important PCs, that is, a Bayesian model selection method proposed by Auer and Gervini, 2008 (Appendix Fig. 3.12). The combined percentage of variance explained by these components was almost a quarter of the total variation (23.09%) with PC1 contributing 9.05%, followed 7.55% for PC2 and 6.48% for PC3. This is a good dimensional reduction considering 180 samples and 31K SNPs.

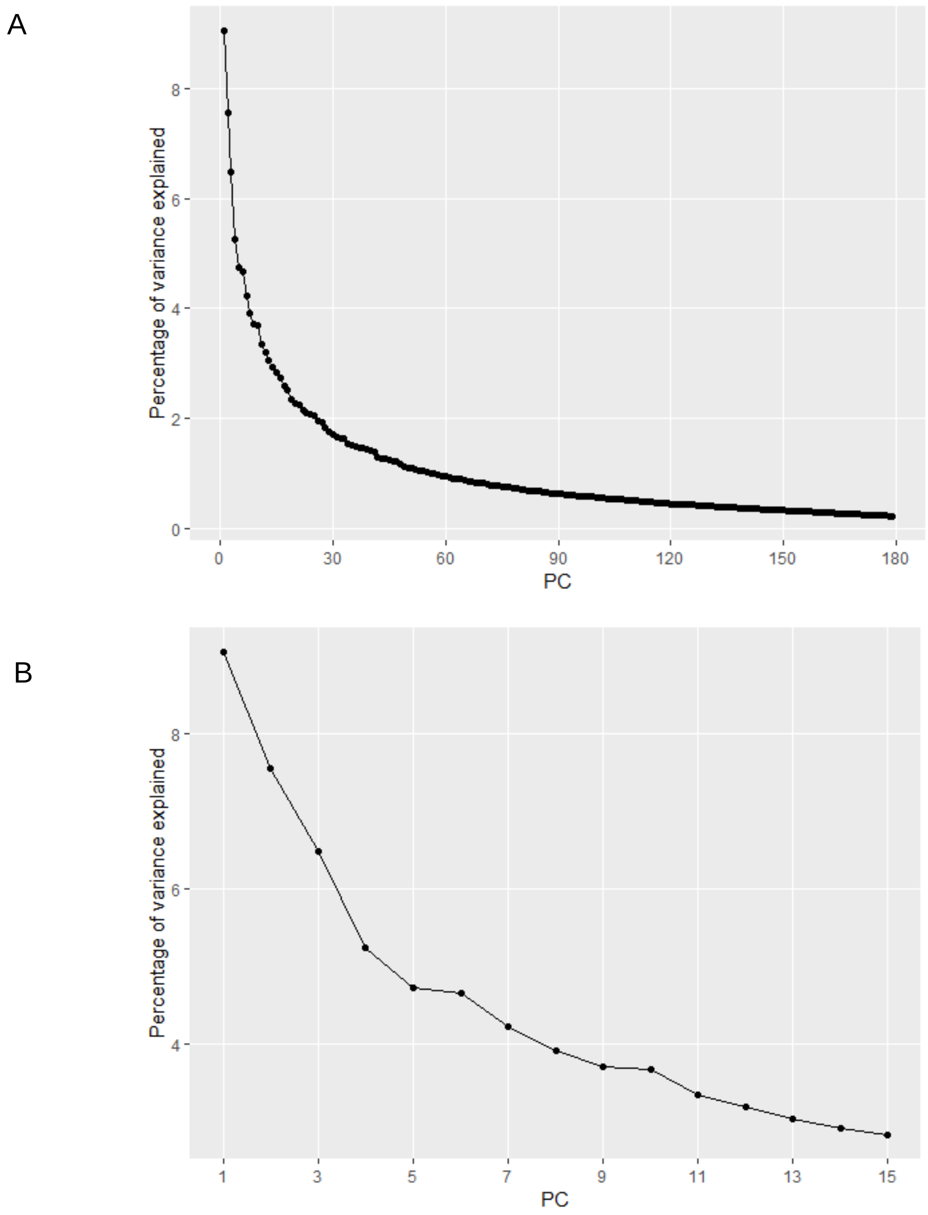


Figure 3.16 Scree plot of the first 179 principal components (A) and considering the first 15 PCs only (B). There was no obvious point in graph A where the declining amount of variance explained equilibrated into a horizontal straight line, although it seemed to level off briefly at PC = 5 (B).

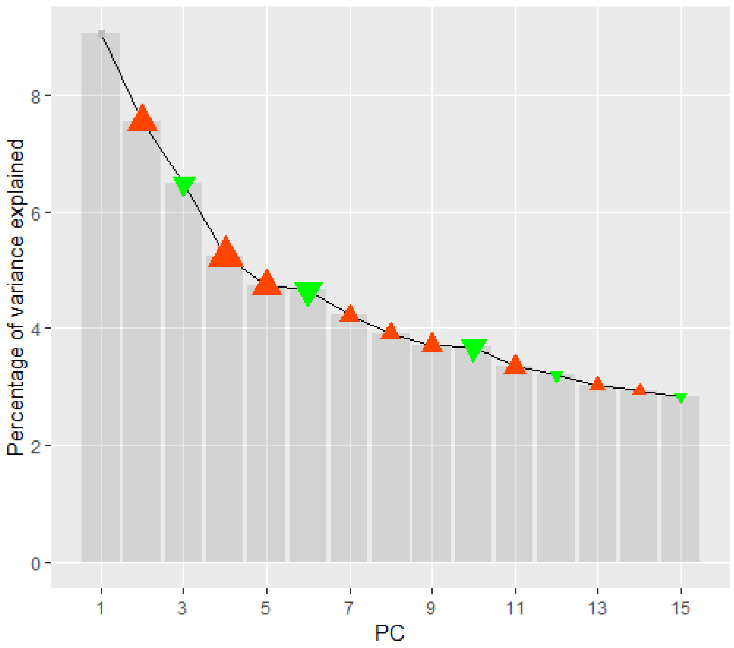


Figure 3.17. Numerical solution to the "elbow" of the scree plot. The acceleration factors, describing the rate of decline in the amount of variance explained, were represented in triangles overlaid on the barplot of percentage of variance explained for each of the principal components. The size of the triangle is proportional to the magnitude of the rate of change. Triangle pointing up (red) corresponds to positive acceleration while those pointing down (green) are negative. The highest acceleration or the elbow was found at PC = 4 and thus the components before this were retained (PC1 to PC3).

As with the PCA plot in Fig. 3.9, the first principal component separates the two generations, but population structure resolved by PC2 to PC3 was not evident. These PCs were investigated further using the results of STRUCTURE. Individual samples were assigned to the four hypothesized ancestral population when the probability was at least 0.60, otherwise, they were assigned as an admixture. PCA biplot was then used to visualize population structure using the new group memberships. The admixture, as expected, did not cluster together; whereas the four "groups" seemed to cluster across PC2 to PC3 (and also PC1), but not so much at PC4 (Fig. 3.18). In Fig. 3.18 A, the four groups clustered together around PC1 and PC2. The first population (AP1) was located at the left while AP2 was opposite and thus mostly differentiated on PC1. On the other hand, AP4 was on top while AP3 was below, and thus differentiated on PC2. In Fig. 3.18 B, PC3 roughly separates AP1 at the left and AP3 at the right. Clustering at PC3 is more obvious in biplots with PC1 (Fig. 3.18 C) and PC2 (Fig. 3.18 D). For example, in Fig. 3.18 C, the relative position of the groups can be described from top to bottom (PC3 axis) with AP3 and AP4 on top, followed by AP2, and lastly AP1. This supports the earlier findings to retain the first three components only.

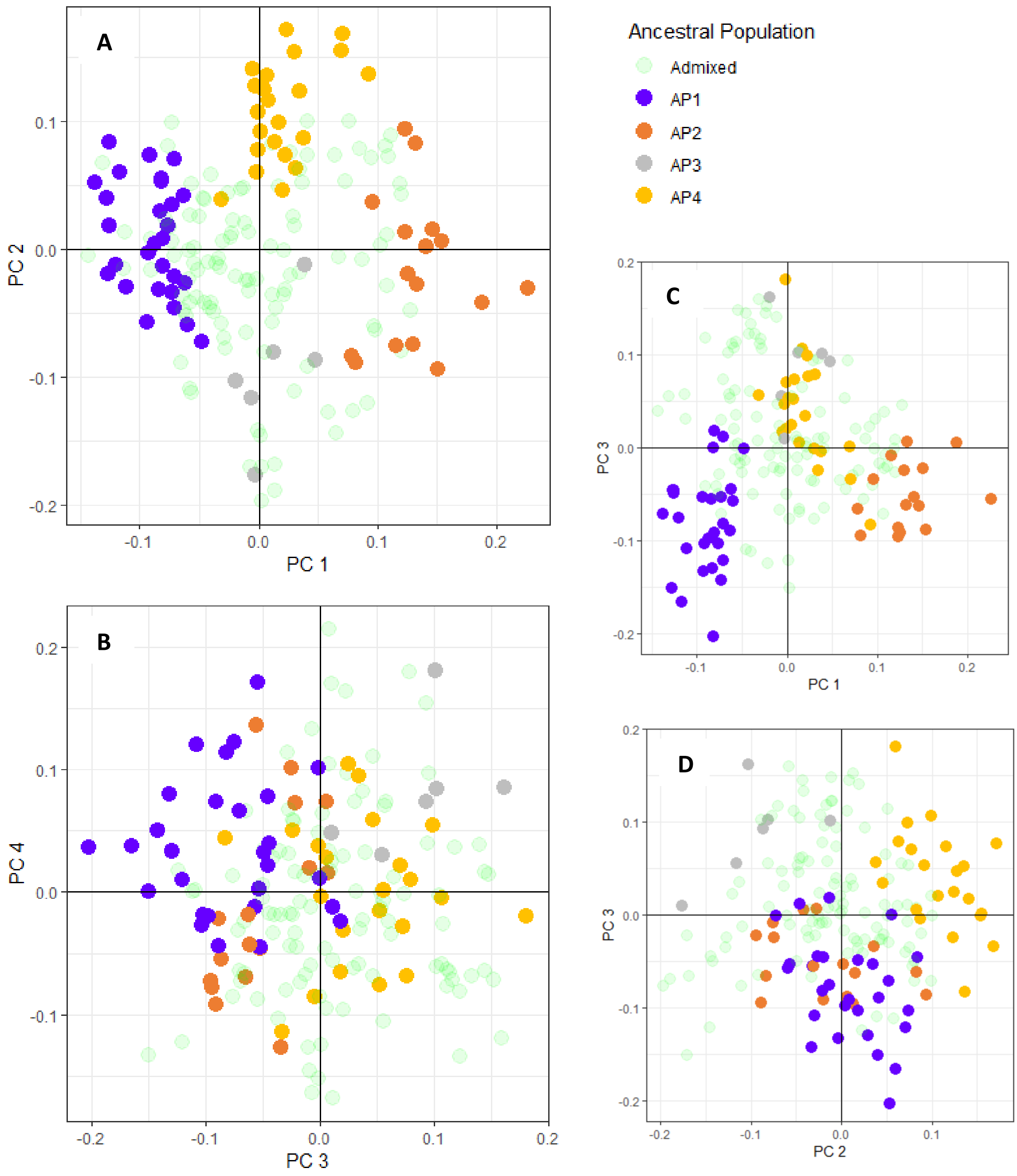


Figure 3.18. PCA biplot showing that the four hypothetical ancestral populations from STRUCTURE can be differentiated in PC1 to PC3. With PC1 and PC2, the four groups can easily be distinguished (A). The groups also cluster along PC3, although less prominently while clustering is unnoticeable along PC4 (B). PC3 grouping improved with PC1 (C) and PC2 (D).

After retaining the first three PCs, SNP genotypes were regressed with the components and the combined test statistic was Mahalanobis distance. A total of 564 SNPs were found to be significantly associated with the first three components (p and q-value < 5%): 195 of which were mostly correlated with PC1; 213 with PC2; and 156 with PC3. Figure 3.19 A shows the significant SNPs associated with PC1-3 distributed across the seven chromosomes It also shows SNPs that were arbitrarily grouped at S8 (cannot be anchored in any of the seven pseudochromosomes). Interestingly, relatively more SNPs were found on chromosomes 7 (217) and 4 (123) than other chromosomes. Regression with PC1 only identified 321 significant SNPs which were also distributed across all the chromosomes. Majority of these SNPs mapped to chromosome 3 (140) and 7 (62) (Fig. 3.19 B). Among the 195 SNPs identified in the PC1 – PC3 regression mentioned above, 68 were similarly detected in the regression with PC1 alone. The discrepancy is normal since the 195 SNPs, while mostly correlated with PC1, were also associated with PC2 and PC3 (based from the combined test statistics). Among the commonly detected SNPs, S7_160751877 is of interest because it belonged to the top 5 SNPs (based on p-value) in both PCA-based regression methods (PC1 – PC3 and PC1 only). This SNP was also identified in the OutFlank method mentioned above. In comparison with the OutFlank method, regression with PC1 – PC3 identified 12 SNPs in common. In contrast, 27 out of 47 SNPs detected in the F_{ST} -based analysis were also identified with regression with PC1 only. The relatively high agreement between the two analytical approaches was expected since F_{ST} was calculated between the generations and the two generations are also clearly differentiated at PC1 (Fig. 3. 12).

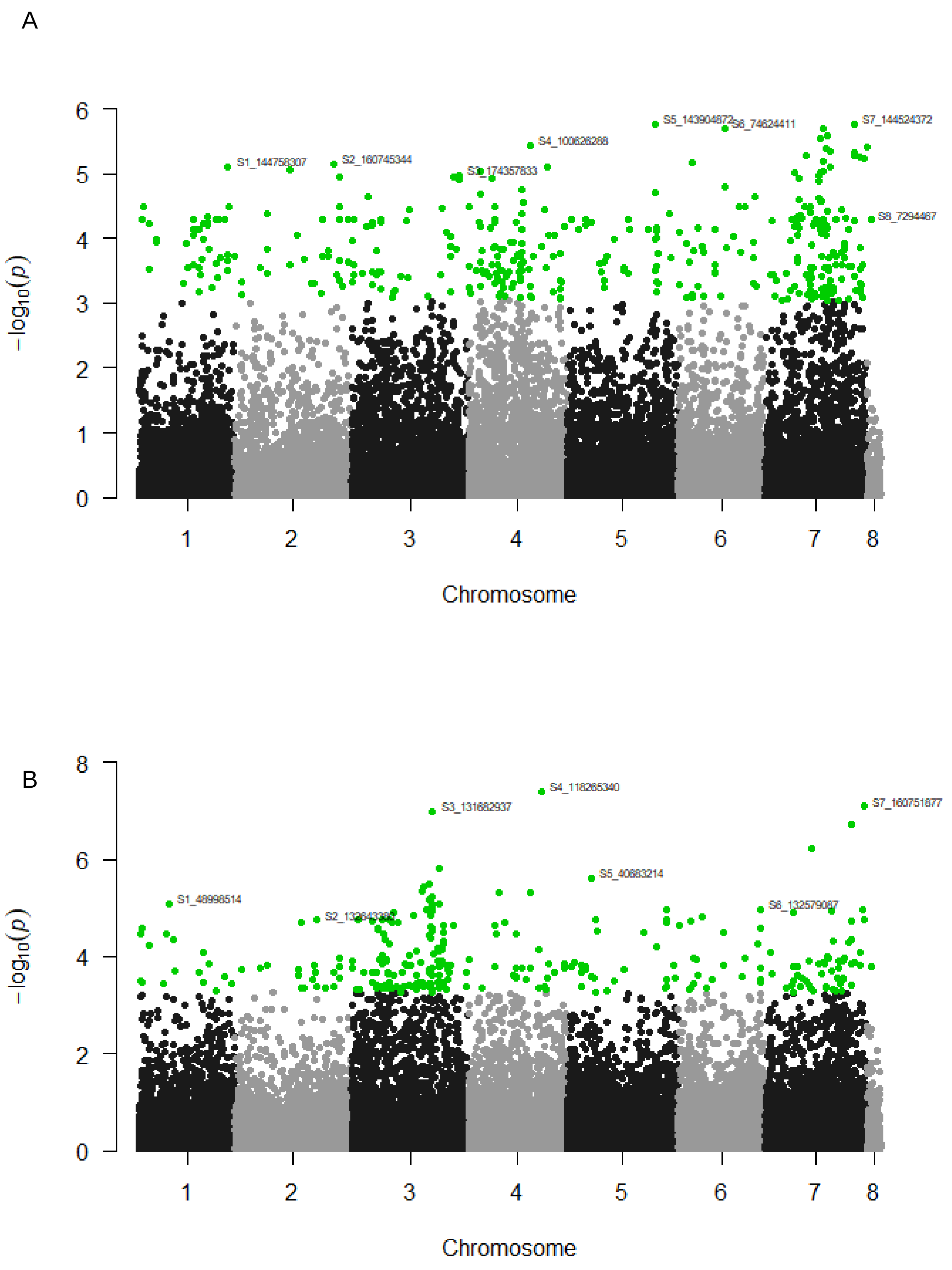


Figure 3.19. SNPs correlated with the first three components of PCA-based population structure analysis. Mahalanobis distance was computed for each SNPs and compared with a null chi-squared distribution with three degrees of freedom to determine P-values (A). The correlation of SNPs to the first principal component alone was also used to determine important loci and P-values were obtained with Gaussian approximation (B).

Genomic regions affecting endophyte infection

A total of 12 SNP markers, corresponding to approximately eight discrete loci were investigated for their association with endophyte infection. These were SNPs with relatively high F_{ST} values and were strongly related to population structure in terms of PCA. Three loci, namely, S2_128346898, S6_1750602, and S7_75550405, were found to have insufficient evidence to reject the null hypothesis of similar odds of infection among SNP genotypes (Table 3.4). Of interest was the locus on chromosome 7 consisting of multiple correlated SNPs (S7_141199150 – S7_141199178) that were found to be under selection based on both F_{ST} and correlation with principal components (especially with PC1) as well as their statistical association with the trait of interest. This locus had the highest odds ratio (5.4) and theoretically increases the odds of infection by more than five times relative to the number of copies of the alternative allele. In other words, the odds of infection increase by more than five-fold with the substitution of the reference allele with the alternative allele.

Another interesting SNP was S7_160751877, which was identified above as one of the top ranked SNPs from the PCA-based detection method. This SNP increases the odds of infection by almost thrice (i.e. 2.96), although in this case, the favourable form was the reference allele. In addition, relative to the alternative allele, two other SNPs increases the odds of infection close to 3x. These were S7_154247285 (2.81) and S6_39276213 (2.62). These two SNPs also have the highest F_{ST} estimates, second and first respectively. Finally, the last significant SNP based on logistic regression was S4_118265340, which nearly doubles (1.81) the odds of infection in terms of the reference allele. This SNP also has the 3rd highest F_{ST} , and the lowest p-value in PC1 regression analysis.

Table 3.4. Nine SNPs were found to be related to endophyte infection using logistic regression and these SNPs were also detected using F_{ST} - and PCA-based methods.

Loci No.	CHR	SNP	Frequency of the alternative allele		F_{ST} -based		PCA-based		Logistic regression	
			PGG04-C2	PGG04-C6	F_{ST}	P-value (Right-tail)	P-value (PC1)	P-value (PC1-3)	Odds Ratio	P-value
1	2	S2_128346898	0.7356	0.2849	0.3324	8.76E-06	1.51E-04	4.88E-04	1.54^{R}	6.01E-02^{ns}
2	4	S4_118265340	0.7000	0.2333	0.3553	4.23E-06	4.08E-08	1.30E-04	1.81^{R}	2.29E-02
3	6	S6_1750602	0.7753	0.3494	0.3076	2.13E-05	2.32E-05	5.14E-04	1.56^{R}	7.64E-02^{ns}
4	6	S6_39276213	0.1556	0.6389	0.3886	1.35E-06	1.51E-05	1.35E-04	2.62	1.97E-04
5	7	S7_75550405	0.5899	0.1611	0.3238	1.23E-05	5.94E-07	8.75E-04	1.71^{R}	4.81E-02^{ns}
6	7	S7_141199150	0.0227	0.3556	0.3020	2.82E-05	1.88E-05	2.64E-04	5.40	5.41E-05
	7	S7_141199152	0.0227	0.3556	0.3020	2.82E-05	1.97E-07	2.64E-04	5.40	5.41E-05
	7	S7_141199154	0.0227	0.3556	0.3020	2.82E-05	4.51E-05	2.64E-04	5.40	5.41E-05
	7	S7_141199159	0.0227	0.3556	0.3020	2.82E-05	1.88E-05	2.64E-04	5.40	5.41E-05
	7	S7_141199178	0.0227	0.3556	0.3020	2.82E-05	1.97E-07	2.64E-04	5.40	5.41E-05
7	7	S7_154247285	0.0341	0.4310	0.3585	3.73E-06	8.41E-05	1.42E-04	2.81	1.03E-04
8	7	S7_160751877	0.4500	0.0778	0.2983	2.95E-05	8.18E-08	5.83E-06	2.96^{R}	5.80E-03

^{R}Regression was performed relative to the number of reference allele; otherwise the alternative allele.

^{ns}Not significant after adjusting the p-values using the BH method (adjusted p-val > 0.05).

The number of copies of the favourable allele for infection, which could be either the reference or alternative allele, was proportional to the endophyte infection rate. These findings suggest an additive effect of the favourable alleles among the significant SNPs. In the case of S6_39276213, positive infection increased from approximately 8% in AA to 21% in AC (and/or CA); and to 38% in CC, where C is the favourable alternative allele (Fig. 3.20). Similarly, infection rate increased with the copy number of the alternative allele in S7_154247285. On the other hand, for S7_160751877, infection rate increased from about 5% in GG to 10% in AG (GA); and to 26% in AA, where A is the favourable reference allele (Fig. 3.21). Similarly, in S4_118265340 the favourable allele is the reference and the number of copies was proportional to infection rate. These linear trends were verified by Cochran-Armitage trend test with greatest evidence (to reject the null hypothesis of no linear relationship) for S7_154247285 (p-value of 3.637557e-05), followed by S6_39276213 (9.758335e-05), and S7_160751877 (3.378009e-03). For S7_141199150 to S7_141199178 (5 SNPs), infection rate increased from around 9% to 35% from the homozygous reference allele to the heterozygote (there were no plants that were homozygous for the alternative allele).

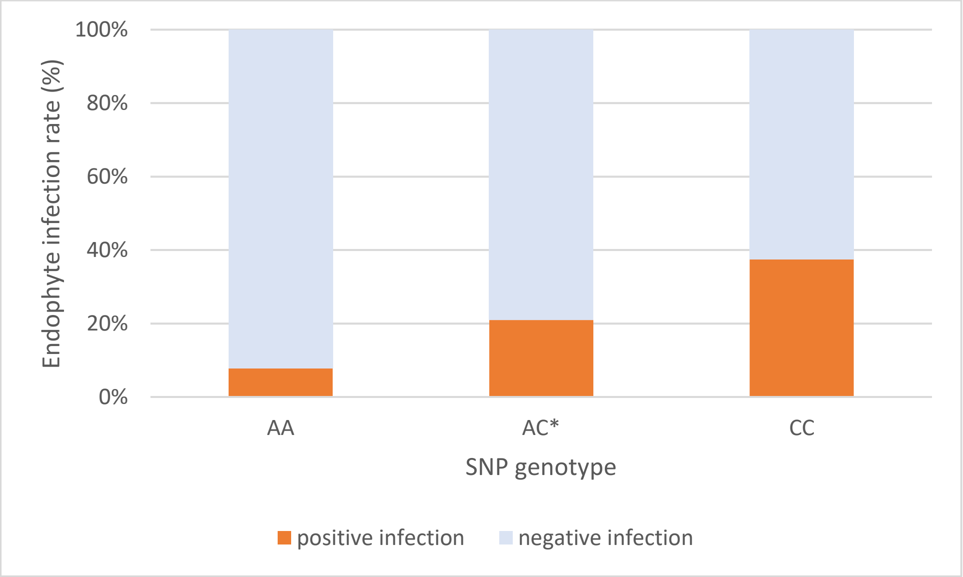


Figure 3.20. The number of alternative alleles of S6_39276213 (C) was proportional to the endophyte infection rate. Successful endophyte infection was higher in CC genotype as opposed to AA. *AC or CA

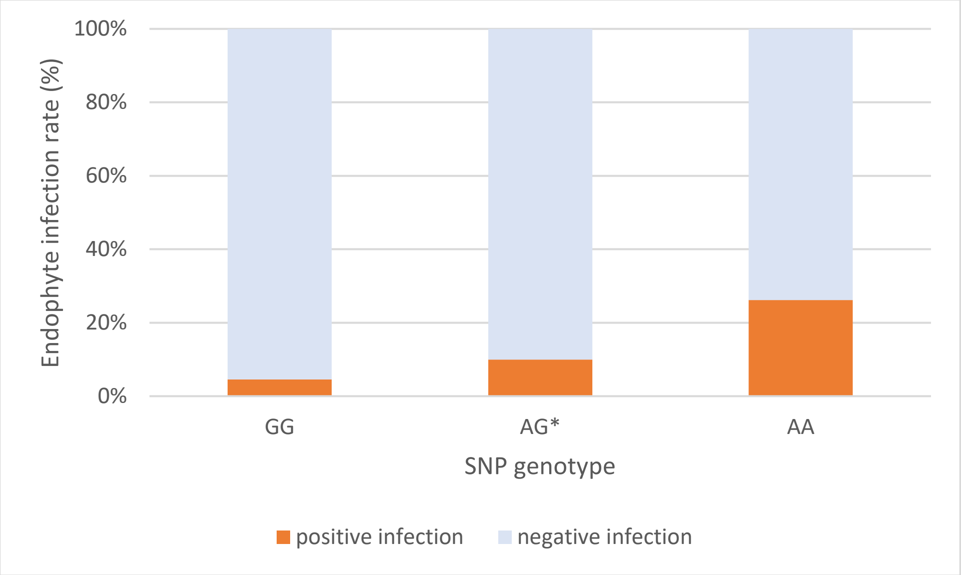


Figure 3.21. The favourable allele in S7_160751877 is the reference allele, A, and infection rate increased with the copy number of this allele. *AG or GA

For the nine SNPs of greatest interest (representing five loci), the *L. perenne* scaffolds to which they belonged were investigated in terms of potential gene candidates. A BLAST search was conducted on each scaffold, and putative genes were identified based on proximity to the SNP position. Only one SNP can be mapped to a known gene (S7_160751877). The other SNPs, namely, S4_118265340, S6_39276213, S7_154247285, were found to be located in non-coding sequences. The locus on chromosome 7 tagged by 5 SNPs (S7_141199150 to S7_141199178) was located in an uncharacterized coding sequence. S7_160751877 was found to map to a gene encoding ABC transporter G family member 6 (ABCG6). ABC transporters belong to one of the biggest protein families, and generally function as ATP-driven pumps. In plants, they were first described as being important in detoxification (Martinoia et al., 1993). Several studies reported that they have diverse functions including pathogen response and phytohormone transport, hence they are involved with processes of plant interaction with its environment, among many others (Kang et al., 2011).

Interesting SNPs detected by the four approaches are summarized in a Venn diagram below (Fig. 3.22). For all tests, a 5% alpha was considered. Moreover, the false discovery rate was also set at 5%. Logistic regression of all 31K SNPs with endophyte infection data identified the highest number of significant SNPs (760). This was followed by the PCA-based method considering PC1 to PC3 (564), PC1 only (321), and last was the F_{ST} -outlier test (47). The most interesting SNPs were the seven detected by all four methods. They were: S6_39276213, S7_141199150, S7_141199152, S7_141199154, S7_141199159, S7_141199178, and S7_154247285. These were the same SNPs listed in Table 3.4 together with S2_128346898, S6_1750602, S7_75550405, S4_118265340, and S7_160751877. Logistic regression of infection data with 31K SNPs found that S2_128346898, S6_1750602, and S7_75550405 were not significant, similar to the results reported in Table 3.4. After controlling for false discovery rate (FDR) in the 31K logistic regression, S4_118265340 and S7_160751877 were also not significant. In contrast, when only the 12 most important SNPs were considered in the regression, as in Table 3.4, the two SNPs proved significant even after multiple hypothesis correction. S4_118265340 and S7_160751877 have relatively higher p-values as compared with the seven SNPs mentioned above (Table 3.4). When accounting for multiple testing of 31K SNPs, as opposed to just the 12, these two SNPs did not pass the 5% FDR threshold. For the seven SNPs mentioned above (S6_39276213; S7_141199150 to S7_141199178; and S7_154247285), four analytical approaches provided strong evidence that these SNPs were under selection and were related to the trait of interest. In the case of S4_118265340 and S7_160751877, it is still reasonable to think that they were under selection based on F_{ST} - and PCA-based tests, and that they were related to the trait of interest based on logistic regression with modest p-value support (Table 3.4). Three SNPs, namely, S2_128346898, S6_1750602, and S7_75550405, were not significant in terms of logistic regression but it is possible that these loci were under selection or at least correlated with loci under selection, due to linkage or LD. The results of F_{ST} - and PCA-based analyses provided evidence that they were under selection. However, there is no enough evidence to support that this selection is related to the trait of interest. It is possible that these SNPs are related to other traits as the development of PGG04 indeed includes selection for “agronomic fitness”, aside from the main breeding objective of endophyte compatibility.

Two other groups of SNPs are also worth mentioning as they were commonly identified by two or three methods. These were: the 20 SNPs detected by both the PCA-based method (PC1-3 and PC1 alone) and logistic regression, but not by F_{ST} -outlier approach; and 28 SNPs commonly detected by logistic regression and PCA with PC1-3 (Fig. 3.22). The F_{ST} approach (OutFlank) identified the lowest number of significant SNPs; this the lower limit in terms of the number of commonly detected SNPs among the four methods. In addition, the F_{ST} approach was limited to detecting extremely high F_{ST} outliers but not the opposite. Therefore, while these 48 SNPs (20 and 28) have relatively lower importance as compared to the nine in Table 3.4, they are worthy of future investigation.

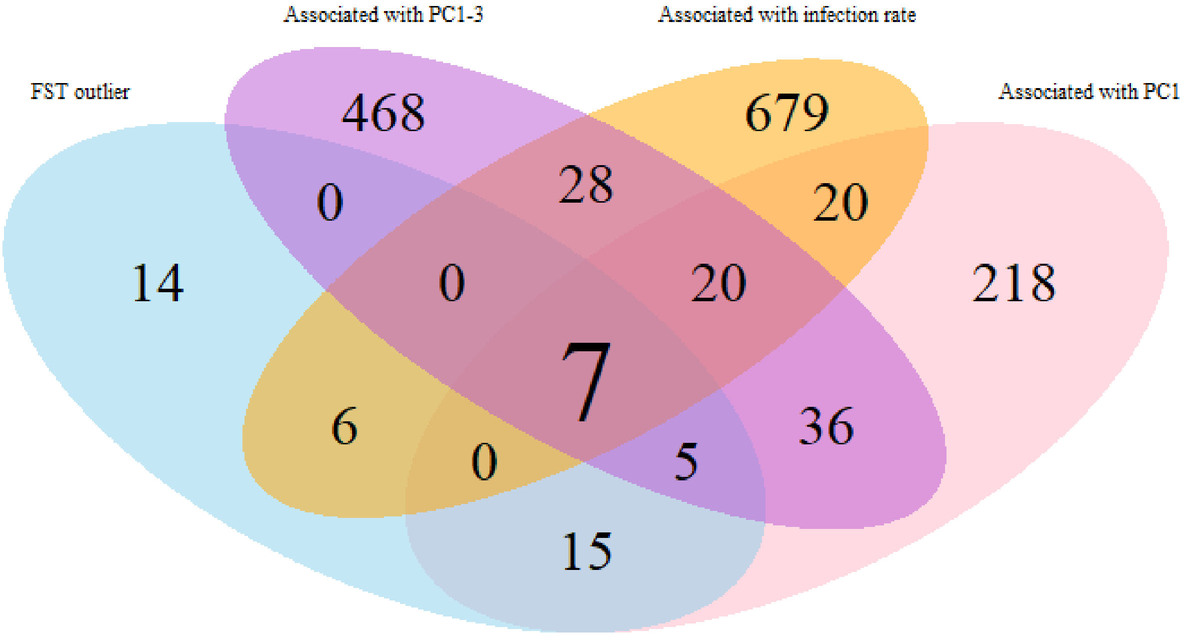


Figure 3.22. The number of significant SNPs detected by four methods namely: (1) F_{ST} outlier detection, SNP regression with the first (2) and the first three (3) principal components and (4) logistic regression with endophyte infection. Some SNPs were commonly detected by different methods.

3.5. Discussion

3.5.1. Trait selection for endophyte compatibility also shaped the structure of genetic variation

Using genotyping-by-sequencing (GBS), this study compared genetic variation between early and late generations of a perennial ryegrass breeding population selected for improved endophyte transmission. In one of the first applications of GBS data in perennial ryegrass, Byrne et al., (2013) noted that it can be used for monitoring allele frequency changes and we report here such application. Four cycles of selection from PGG04-C2 (early generation) to PGG04-C6 (late generation) resulted in an overall excess of rare alleles and a slight reduction in genetic diversity in terms of expected heterozygosity, from H_e of 03069 to 0.3033. This is consistent with a population under selection. In a breeding population divergently selected for water-soluble carbohydrate (WSC) content, H_e (as well as observed heterozygosity) declined from 0.667 in C0 to 0.641 in $C2^{S-}$ (positive selection) and 0.528 in $C2^{S+}$ (negative selection) (Gallaher et al., 2015). This reduction in H_e is comparatively greater than what was observed in the present study. The difference in the estimates of H_e decline could be due to the nature of the respective traits and breeding populations, as well as marker systems used. In particular, the current study utilized more than 219K SNPs from GBS, whereas the WSC experiment utilized 35 SSRs. In the current study, the distribution of observed heterozygosity (H_o) was skewed toward relatively low values with a mode around 0.1 (Fig. 3.10 A), while for the distribution of H_e, there was an excess of the maximum value (0.5) (Fig. 3.10 B). Several studies (Barth et al., 2015; Kovi et al., 2015; Gallaher et al., 2015; Brazaukas et al., 2011) similarly reported lower observed heterozygosity (H_o) than expected (H_e) using different marker systems in ryegrass ecotypes, cultivars, and breeding populations. In contrast, Blackmore et al. (2015) and Blackmore et al. (2016) reported higher H_o than H_e which was suggested to be due to the outcrossing nature of *L. perenne* caused by self-incompatibility. The surplus of homozygotes reported in the current study, in Kovi et al. (2015), and Gallaher et al. (2015) can be explained by selection applied in the respective breeding populations under investigation. Interestingly, the fixation index (F_{IS}) declined in the current study — from 0.1292 in C2 to 0.1174 in C6. This is contrary to what is expected from selection. The decline in F_{IS} is potentially due to selection for better endophyte compatibility favouring plants heterozygous at a number of loci (i.e. heterozygote advantage). In a perennial ryegrass breeding population negatively selected for freezing tolerance, a negative F_{IS} was reported. This suggests that selection against this trait also favoured heterozygote plants, although positive selection resulted in positive F_{IS} (Kovi et. al., 2015). In WSC selection study, F_{IS} was also relatively higher in the early (C0) than the late generation (C2), both in positive and negative selections. In addition, Blackmore et al. (2016) reported negative F_{IS} in a recurrent selection population, forage and amenity cultivars, and European ecotypes. Genetic differentiation between C2 and C6 of PGG04 was significantly different from zero (F_{ST}: 95% confidence interval of 0.0307 – 0.0315), although it suggests little differentiation (less than 0.05). Kubik et. al., (2001) reported F_{ST} values from 0.065 to 0.197 among seven cultivars, which can be described as moderate (0.05 – 0.15) to high (0.15 – 0.25) genetic differentiation. Also, when comparing four recent cultivars developed from the same breeding program, Blackmore et al. (2016) reported a mean F_{ST} of 0.154. The higher differentiation among cultivars is expected since they represent distinct populations. In contrast, PGG04-C2 and C6 can be described as two stages of the same population, hence the lower F_{ST}. In contrast to the current study, the WSC selection study reported a slightly higher significant differentiation (F_{ST} ~ 0.06) between C0 and C2 of both positive and negative selection.

Several genetic diversity studies report relatively lower genetic variability between populations or groups than within groups and/or individuals. For example, Blackmore et al. (2016) reported that a quarter of genetic variability was due to differences among four perennial ryegrass cultivars from the same breeding program versus three-quarters of genetic variability within each cultivar using SNP markers. This is in agreement with previous AMOVA results of within-cultivar variation of ~88% using SSR (Barth et al., 2015) and 67% using RAPD (Bolaric et al., 2005). In the present study, similarly high within-generation variability estimates were obtained. AMOVA results showed that only about 3% of genetic variation can be attributed to the differences between the early and late generations of PGG04 (Table 3.3). The low intergeneration variance component is consistent with the low genetic differentiation mentioned previously. Hence, this low variance component can be similarly explained as the two generations of PGG04 being the same population, only at different stages. In the freezing tolerance study, Kovi et al. (2015) reported a similar percentage of variance explained (~3%) between population selected for high frost tolerance ($C2^{S+}$) and the unselected population (Syn4), both of which also represent the same breeding population. In contrast, variance explained by genetic differences between Syn4 and negatively selected population ($C2^{S-}$) was slightly higher (~8%). The same analysis of molecular variance indicates that, although this intergenerational variance is low, it is important to the whole picture of genetic variation of PGG04 which parallels with F_{ST} estimated to be significantly different from zero.

The UPGMA tree and PCA biplot show that genetic differences between generations of PGG04 indeed caused population structure. The PCA biplot captured the reduction of diversity due to selection (Fig. 3.9). This was shown in terms of the wider dispersion of the early generation and closer clustering of the late generation. More importantly, the first principal component separated the early generation and the late generation. In the current study, the leftmost plants in the PCA biplot, being more differentiated, reflect the divergence of these same plants (i.e. clusters 8 and 9) away from the whole population in the UPGMA tree in Fig. 3.8 B. Moreover, it was supported by the results of a Bayesian clustering method from the program STRUCTURE, where at K=2, the two groups generally correspond to the two generations of PGG04 (Fig. 3.10). Previous studies also utilized PCA, genetic distance-based dendrograms, and STRUCTURE to investigate population structure. Blackmore et al. (2016) showed that selection resulted in the separation of a cultivar (Aurora, AF3) from its founder (CH6, an ecotype) based on a PCA of SNP data. Agreement between PCA and STRUCTURE results were also reported previously. For example, PCA identified an east-west geographic divide among European ryegrass ecotypes described by the first principal component (Blackmore et. al., 2015). In parallel, the same study found that membership probabilities from STRUCTURE were correlated with the longitude of the original seed sample site of the ecotypes. In another study, a related dimensionality reduction strategy, principal coordinate analysis (PCoA), was utilized and showed a separation of cultivars and advanced breeding germplasm into two groups based on their place of origin (Brazaukas et al., 2011). The same clustering was observed when using STRUCTURE in the said study. Finally, another study reported agreement between PCA, UPGMA, and STRUCTURE results detecting two genes pools among ecotypes and cultivars (Barth et al., 2015).

In the present study, the Evanno et al., (2005) method detected two and four hypothetical groups. PGG04 was generated from three paired-crosses, thus a hypothesis of four subpopulations (K = 4) is interesting (Fig. 3.11). One possible interpretation of K = 4, is that a fourth group emerged from the three ancestral population (i.e. three paired-crosses) due to selection. This interpretation is supported by the observation of an increase in the probability of a particular subpopulation (i.e. ancestral population 1, AP1, blue) from PGG04-C2 to PGG04-C6, presumably the endophyte compatibility alleles. This interpretation is also consistent with the results of the UPGMA tree and PCA. PGG04-C6 plants that have high membership probability to AP1 were generally the same cluster 8 and 9 plants in the UPGMA tree (Fig. 3.8 B) or the leftmost plants in the PCA biplot (Fig. 3.9). In Fig. 3.8 B, cluster 8 and 9 plants appear to diverge from the rest of PGG04, presumably due to selection. In Fig. 3.9, these plants were farthest (i.e. leftmost) from PGG04-C2 plants which means they have low genetic similarity to the early generation, potentially because of selection. Finally, phenotype data provided additional evidence supporting that the fourth group is related to the trait of interest. Most of the infected plants in PGG04-C6 have high membership probability with AP1. Kovi et al. (2015) put forward a similar interpretation of their STRUCTURE results. In the said study, the population positively selected for freezing tolerance ($C2^{S+}$) had a higher probability for a particular subpopulation, whereas it had a lower probability for another subpopulation. The reverse was true in the negative selection population ($C2^{S-}$). The subpopulation with high probability in the positively selected population is suggested to describe alleles for freezing tolerance.

In the current study, it is interesting that while the PCA biplot, UPGMA tree, and STRUCTURE bar plots were constructed based on genotype data alone, they captured the effect of selection in the trait of interest. In Chapter 2, it was reported that selection improved endophyte infection rate from about 4% to 33% viable endophyte. In parallel, we report here an overall shift in allele frequencies reflected by population structure (in terms of PCA, UPGMA tree, and STRUCTURE) also due to selection. More than half (58.82%) of the positive infection in PGG04 (early and late generation) belonged to the plants that showed evidence of selection. These were the diverging plants with membership in clusters 8 and 9 in the UPGMA tree, the leftmost plants in the PCA biplot, or plants with higher membership probability with AP1 based on STRUCTURE. For PGG04-C6 in particular, 31 out of 94 plants harbor AR501. Moreover, about two-thirds of these infected plants were from cluster 8 and 9.

Recurrent selection in PGG04 aims to increase the frequency of favourable alleles to shift the average endophyte compatibility (i.e. mean transmission) of the population in a positive direction. Again, in Chapter 2, we reported evidence for improvement in the trait of interest. Here, we extend further and provide evidence that the phenotypic change is underpinned by allele frequency shifts, influencing the structure of genetic variation. The use of an active breeding population, high-density SNP markers, and multiple population structure analysis procedures plus corroboration with phenotype data provided strong evidence of selection shaping genetic variation as the breeder improved the trait of interest.

3.5.2. Selection signatures in the genome underpin genotype-trait association

It was shown that four cycles of selection changed the structure of genetic variation in PGG04. The effects of selection as expected in breeding populations were not uniform with some genomic regions experienced relatively higher pressure. Genetic differentiation (FST) was relatively higher for 47 SNPs compared to all others. FST outlier detection is extensively used to detect candidate genes (Lotterhos & Whitlock, 2014) although most commonly in the context of adaptation in natural populations. For example, it was used to identify candidate genes under selective pressure, both from natural and artificial selection, during the differentiation of the *indica* and temperate *japonica* rice (Sun et al., 2015). It was also used to study domestication selection in waxy and common maize (Hao et al., 2015). In perennial ryegrass, it was used to determine loci under selection in populations divergently selected for WSC content (Gallagher et al., 2015) and freezing tolerance (Kovi et al., 2015). These studies however utilized low marker density, 35 SSRs and 278 SNPs, respectively. In contrast, the current study utilized 31K SNPs from genotyping-by-sequencing. In addition, we utilized methods that encompassed improvement in the general FST-based analytical approach. All the FST studies cited above, including the two studies in perennial ryegrass, used LOSITAN (Antao et al., 2008). LOSITAN generates an expected null distribution of FST that can deviate significantly from observed data, resulting in a large number of false positive outliers (Flanagan & Jones, 2017), and is restrictive with the island model assumption (Whitlock & Lotterhos, 2015). In contrast, the method used in this study infers the null model by likelihood approach and by trimming the FST distribution.

In the present study, some SNPs were found to influence population structure more than others. Specifically, 564 SNPs had a significantly higher correlation with the first three principal components in PCA, which describes the population structure in PGG04, than others. PCA-based gene discovery methods have been reported previously for perennial ryegrass. Blackmore et al., (2015) investigated the population structure of perennial ryegrass European ecotypes with PCA using SNP data. They found that the first principal component describes an east-west geographic divide while the second PC describes a north-south divide based on the place of origin of the ecotype. The said study identified possible candidate genes relating to this geographic adaptation based on 50 markers with the highest correlation with PC scores. Similarly, PCA was used to identify highly correlated SNPs (top 20 loadings) with PC scores relating to the differentiation of amenity and forage perennial ryegrass cultivars (Blackmore et al., 2016). Two candidate genes relating to the differences in these two functional groups of perennial ryegrasses were identified. One gene has a function in chlorophyll breakdown and senescence which could be important in maintaining color in amenity grasses. The other one relates to cell wall robustness which contrasts the ease of digestibility in forage and resistance to trampling in amenity grasses. Both PCA-based studies utilized a moderate number of SNPs (over two thousand) as compared to the high marker density in the current study. Furthermore, the two studies utilized correlation measures to determine important SNPs while the current study makes use of the more robust Mahalanobis distance (i.e. for PC1 to PC3) (Luu et al., 2017).

Proportionately more SNPs detected by FST- and PCA-based analyses mapped to chromosome 7 than to other chromosomes; 13 out of 47 SNP with high FST; 217 out of 564 SNPs correlated with PC1 – 3; and 62 out 321 SNPs correlated with PC1 (regression with PC1 only). This observation partly supports a hypothesis on the role of defence-related (DR) genes in the grass-endophyte interaction. As mentioned in the methods section, the ryegrass assembly used in the present study was aligned to the barley genome. Homoeologous chromosomes of wheat and barley, including chromosome 7, are known to generally share collinearity (Mayer et al., 2011). Wheat chromosome 7 is known to be rich in host DR genes, hence Faville et al., (2015) suggested that their detection of several QTLs for endophyte biomass and alkaloid concentration on this colinear chromosome in perennial ryegrass may support a hypothesis that variation at DR gene loci is involved in mediating the grass-endophyte interaction.

The significant SNPs detected in FST- and PCA-based methods in the current study differs from neutral loci which were mostly affected by random genetic drift. Changes in allele frequency, including in neutral loci, influence the structure of genetic variation but genetic differentiation will be higher in loci under selection than those affected by drift alone. As mentioned previously, gene flow and mutation also influence allele frequency shifts. These factors were assumed negligible for the current study. During PGG04 development, polycrossing was done meticulously in isolation hence the effects of gene flow is unlikely to be substantial. The generations of PGG04 compared in the study had only four cycles of selection between them. Since mutation rates are generally slow, the effects of mutation will be minimal also. It is, therefore, reasonable to infer that selection caused relatively higher genetic differentiation and correlation with the principal components in some SNPs. We have combined evidence from FST and PCA-based analyses that 12 SNPs were under selection and were different from neutral regions. Nine of these SNPs were related to the trait being selected in the breeding population based on logistic regression with infection data. Moreover, a linear positive trend can be established between the number of favourable alleles and infection frequency.

Trait improvement due to selection was underpinned by parallel allele frequency changes. Selection signatures in the genome can be investigated to discover important regions that potentially regulate the trait of interest. In perennial ryegrass, Brazauskas, Pašakinskienė, et al. (2013) identified candidate genes for axillary tiller development using a recurrent selection population. Similarly, selection signatures in a perennial ryegrass breeding population recurrently selected for crown rust resistance were utilized in trait discovery (Brazauskas, Xing, et al., 2013). In both studies, allele frequency changes were observed directly. In the present study, we utilized other indicators linked to changes in allele frequency during selection. These indicators were F_{ST} and association with PCA-based population structure. In Brazauskas, Pašakinskienė, et al. (2013) and Brazauskas, Xing, et al. (2013) as well as in Gallagher et al. (2015) and Kovi et al. (2015) discussed above, selection was conducted divergently, that is, selection for and against the trait of interest. Hence, alleles enriched with positive selection could be verified in the negative selection population. In the current study, the breeding population was specifically developed by a private seed company for improved AR501 compatibility, therefore selection against endophyte compatibility was not pursued as it is not a desirable trait. In the tiller development (Brazauskas, Pašakinskienė, et al., 2013) and crown rust resistance (Brazauskas, Xing, et al., 2013) studies both targeted variation at pre-identified candidate gene loci that were putatively related to the trait of interest. In contrast, the perennial ryegrass studies on WSC (Gallagher et al., 2015) and freezing tolerance (Kovi et al., 2015) mentioned above as well as the present study, took a genome-wide approach. Having prior information on possible candidate genes may simplify the dissection of the genetic basis of the trait of interest, however, it also limits the ability to discover previously unknown genes. For the present study, information on specific ryegrass genes associated with endophyte compatibility was absent, although other studies have suggested that defence-related (DR) genes may be important (Faville et al., 2015).

While the current study reports improvements from previous related works, further improvement can be employed to increase evidential support on gene discovery studies, especially in perennial ryegrass breeding populations. The F_{ST} based methods will benefit from improvement in the precision of allele frequency measurements which is especially relevant to GBS data. As a quality control in this study, we filtered out SNPs with read depth below the 5th percentile (i.e. at least 4x) and above the 95th percentile (i.e. at most 158x). From the filtered data we computed F_{ST}. However, the accuracy of allele calls was actually higher in SNPs with higher read depth, for example 4x vs. 158x. Therefore, F_{ST} estimates were consequently more accurate in SNPs with higher read depth. To overcome this, Dodds et al. (2018) proposed a depth-adjusted F_{ST} calculation that can be applied in GBS data. It would thus be interesting to calculate F_{ST} using a depth-adjusted method and from there, infer a better distribution to detect outliers at much higher confidence. Among the limitations of F_{ST} based methods is the requirement of assigning individuals to distinct groups which, in the present study, were the two generations. On the other hand, PGG04 can be considered as a continuous population with admixture genotypes being one breeding population at different stages. We overcame this problem by also investigating population structure with PCA and the program STRUCTURE which does not require prior population labels. We further used a PCA-based regression to detect important SNPs related to the observed population structure. In contrast, we have not examined STRUCTURE results further, that is, to similarly look for SNPs contributing to population structure changes. Therefore, a possible interesting improvement of this study is the application of an FST formulation based on *Q* (i.e. ancestry coefficient) and *F* (ancestral allele frequency) matrices of STRUCTURE (or other related software) as implemented in the R package LEA (Frichot & François, 2015; Martins et al., 2016). Finally, several other methods also could be implemented in the data set to provide additional evidence for (or against) the SNPs detected. For example, in the crown rust resistance study, Brazauskas, Xing, et al. (2013) used linkage mapping in conjunction with allele frequency changes in gene discovery. It would be interesting to look at the parallels of this study and the QTL study of Faville et al. (2015) which investigated the grass-endophyte interaction, yet did not investigate endophyte transmission and viability. Of immediate interest is whether the QTLs detected on that study and genomic regions in the present study were co-located and/or linked. In addition, it would be interesting to measure loline concentration in PGG04, as was done in the QTL study, and compare it with infection and genotype data. In the present study, the identification of a few genomic regions under selection, which are possibly related to endophyte compatibility, adds further evidence to the host genomic control of the association. Similar to the QTLs in Faville et al. (2015), genomic regions identified in this study are potentially useful in perennial ryegrass breeding through the development of a marker-assisted selection protocol.

3.5.3. ABC transporter genes could be important in the grass-endophyte interaction

The combined test of F_{ST} outlier detection, population structure analysis based on PCA, and logistic regression with infection data, identified five genomic regions; each containing one to five SNPs, under selection and potentially associated with AR501 transmission. Three of these were located on chromosome 7, including one locus tagged by five SNPs. The observation of a group of significant SNPs tagging a region in chromosome 7, which is suspected to be rich in DR genes, also partly supports the hypothesis that DR genes play a role in grass-endophyte interaction.

Only one of the five genomic regions returned a hit when annotation was conducted using BLAST procedures. The one hit was a region on chromosome 7 tagged by S7_160751877. The scaffold containing this SNP mapped closest to a gene encoding ATP-binding cassette (ABC) transporter G family member 6. Genes encoding ABC proteins belong to a superfamily which has diverse functions in substrate transport. More importantly, they are reported to be involved in plant–pathogen interactions (Rea, 2007).

ABCG transporters are involved in plant-microbe interactions (Kang et al., 2011). In Arabidopsis, full-size (i.e. pleiotropic drug resistance, PDR subfamily) ABCG transporters were shown to mediate phytohormones involved in biotic stress responses such as jasmonic acid and/or salicylic acid (Kang et al., 2011). Mutation in Arabidopsis ABCG30 showed altered root exudate composition which affected the diversity of soil microflora (Badri et al., 2009). ABCG involvement in plant defence is evidenced by *Np*PDR1 in *Nicotiana plumbaginifolia* which is induced by the anti-fungal diterpenoid sclareol. Silencing *Np*PDR1 increased susceptibility to the fungus *Botrytis cinerea* (Stukkens et al., 2005). In bread wheat, an interesting example is the case of *Lr34* located in chromosome 7D encoding full-size ABCG transporter. It confers durable resistance against leaf rust, stripe rust, and powdery mildew, and is thus used in breeding for disease resistance (Krattinger et al., 2011). ABCG transporters are also involved in symbiotic associations which may be more relevant to the current study. In legume-rhizobia interaction, the signaling flavonoid root exudate was reported to be mediated by a PDR-type ABC transporter (Sugiyama et al., 2007). Furthermore, STR and STR2 encoding half-size (i.e. white/brown complex, WBC subfamily) ABCG transporters were also found to be important in the symbiosis of arbuscular mycorrhizal (AM) fungi in *Medicago truncatula* (Q. Zhang et al., 2010). Finally, homologous ABCG genes in rice have been reported to be important in arbuscule formation and hence in the rice-AM fungi interaction (Gutjahr et al., 2012).

ABCG transporter genes have also been suggested to be of importance in the tall fescue- *E. coenophiala* symbiosis. Using both wheat and barley GeneChip®, Dinkins et al. (2010) identified 32 genes differentially expressed in endophyte-infected and non-infected tall fescue, including an ABC transporter. In another study, a PDR-like ABC transporter gene was shown to have higher expression levels in infected tall fescue than uninfected (L. J. Johnson et al., 2003). This was observed for tall fescue cultivars KY31 and GI-320 infected with *E. coenophiala* as well with KY31 infected *E. siegelii*. On the other hand, the same study reported similar expression levels in the endophyte-free perennial ryegrass and those infected with either *E. festucae* var. *lolii* or *E. typhina.* Interestingly, Khan et al. (2010) reported MRP-like (i.e. ABCC transporter) and PDR-like ABC (i.e. ABCG) transporter genes as two of the several differentially expressed genes in infected and endophyte-free perennial ryegrass. The PDR-like ABC transporter gene in perennial ryegrass was a homologue (based on BLASTN) of that of tall fescue's in the L. J. Johnson et al., (2003) study. ABCG transporter genes, therefore, are possible mediators of the grass-endophyte interactions.

In perennial ryegrass, studies on the function of ABC transporters are limited. Nevertheless, *Lp*ABCG5, as well as its paralogue *Lp*ABCG6 (not the same ABCG6 in the present study), were reported to be important but in a different capacity, that is, in plant architecture (Shinozuka et al., 2011). Its orthologue in rice, *Os*ABCG5 is known to control tiller number. Interestingly, an early study (Latch et al., 1985) reported that the infection of the common toxic endophyte in the perennial ryegrass cultivar Nui could have increased its tiller number, although the effect of endophytes on ryegrass plant growth has not been clearly established (L. Johnson et al., 2013). Since endophyte growth and development is in synchrony with that of the host, it is tempting to speculate that ABCG transporter-mediated tiller bud development may share similar mechanisms with that of ABCG transporter-mediated symbiosis.

ABCG transporter genes are strong candidates amongst the many genetic factors in the host that influence its association with fungal endophytes. Further studies can be conducted to confirm this, especially focusing on the gene identified in the present study. A quick and practical approach would be to design primers based on the identified gene and to carry out a marker-assisted selection (MAS) in the forward generation of PGG04 or other perennial ryegrass breeding populations infected with AR501. Although primer design will require further characterization of the SNP and the genomic region around the SNP, several options could be pursued to convert the SNP into a more practical PCR-based marker such as CAPS (cleaved amplified polymorphic sequence), AS-PCR (allele-specific PCR), TS-PCR (temperature-switch PCR), KASP (Kompetitive Allele Specific PCR by LGC Genomics), and STARP (semi-thermal asymmetric reverse PCR) (Rasheed et al., 2017). MAS would not only confirm the result of the study but, when successful, would provide a readily accessible tool to breed for AR501 compatibility. At a more fundamental level, expression analysis of ABCG transporter genes in AR501 infected vs uninfected plants can be undertaken. Another study that may be explored is on comparative genomics of perennial ryegrass and other grasses such as rice, maize, wheat and barley in terms of diversity and similarity of ABC transporter genes. Dinkins et al. (2010) took advantage of the DNA homology of tall fescue with wheat and barley, and utilized GeneChip® from these grass species in their expression study. GeneChip® of perennial ryegrass and its endophyte have been developed for symbiosis study (Voisey et al., 2007) and could be utilized in a follow-up study especially because it includes ABC transporter genes. In the presence of stronger evidence from further studies, a more advanced study using mutants could also be conducted to elucidate the mechanism of ABCG transporters in the grass-endophyte interaction.

Aside from ABC transporters, genes encoding pathogenesis-related (PR) proteins were also implicated in the grass-endophyte interaction especially PR class 10 protein (PR-10). N. Zhang et al. (2011) reported that high transcript level of PR-10 is correlated with the presence of *E. festucae* var. *lolii* in perennial ryegrass. Faville et al. (2015) suggested PR-10 as a candidate gene for the endophyte biomass QTL they detected in perennial ryegrass chromosome 4. In the present study, among the top SNPs commonly detected by several methods, only S4_118265340 is located in chromosome 4. Nevertheless, some interesting SNPs were identified with individual detection methods in this chromosome especially with the PCA-based method considering the first three components (Fig. 3.18 A). It would be interesting therefore to determine whether these SNPs of secondary interest were linked with the QTL detected by Faville et al. (2015). PR-10 was also previously determined by L. J. Johnson et al. (2003) as differentially expressed in endophyte-infected and non-infected tall fescue, alongside the ABC transporter mentioned above. The molecular basis of grass-endophyte interaction in the host is possibly mediated by several genes, especially those related to plant-microbe interaction such as PR-10. This study puts forward the case for ABCG transporters as genes of interest among the many genes in the host grass that interact with fungal endophytes.

3.6. Conclusion

Recurrent selection for AR501 transmission and viability resulted in changes in population structure and genetic diversity in the breeding population PGG04. Genotyping-by-sequencing (GBS) proved useful in this population genomics study of a breeding population in active development. As expected, most of the genetic variability occurred within generation than between them, based on AMOVA results. By comparing GBS data of PGG04-C2 and PGG04-C6, results show that there was an excess of rare alleles and reduction of genetic diversity (H_e) from the early to the late generation. These are characteristics of a population under selection. Since selection was targeted towards the trait of interest, the genetic changes were not uniform and thus allowed for the identification of selection signatures in the genome. Multiple analytical procedures identified a core group of SNPs significantly associated both with the observed changes in population structure and trait values. Specifically, twelve SNPs had relatively high F_{ST} values and were more correlated to principal components describing PCA-based population structure than other SNPs which are potentially neutral. In contrast to SNPs under selection, neutral SNPs were affected mainly by random genetic drift. Since the population was developed for breeding for endophyte compatibility, SNPs under selection have been confirmed by association analysis with infection data using logistic regression. Among genomic regions under selection, five regions tagged by the nine SNPs were of interest as they were associated with the infection data. The SNPs tagging these regions can increase the odds of the beneficial infection by more than five times (maximum), relative to the number of favourable alleles. Several important SNPs were located in chromosome 7 which is suggested to be rich in defence related (DR) genes. Hence, this study partly supports the hypothesis that DR genes may be involved in the grass-endophyte interaction. One of the significant SNPs was found to have annotations that suggested a role for ABCG-type transporters genes. These are identified elsewhere as having an ABCG proteins that are known to have important roles in symbioses and plant defence. The results of this study, therefore, provide additional evidence on the importance of ABCG gene in plant-microbe interaction, and may prove to be important in the perennial ryegrass-endophyte interaction. The nine most significant SNPs identified in this study, especially the one tagging an ABCG gene, may be useful for developing marker-assisted selection schemes.

3.7. References

Adcock, R., Hill, N., Bouton, J., Boerma, H., & Ware, G. (1997). Symbiont regulation and reducing ergot alkaloid concentration by breeding endophyte-infected tall fescue. *J Chem Ecol, 23*(3), 691-704.

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.

Antao, T., Lopes, A., Lopes, R. J., Beja-Pereira, A., & Luikart, G. (2008). LOSITAN: a workbench to detect molecular adaptation based on a FST-outlier method. BMC Bioinformatics, 9(1), 323.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics, 11*(3), 375-386.

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 Biol, 16*(1), 160. doi: 10.1186/s12870-016-0844-y

Auer, P., & Gervini, D. (2008). Choosing principal components: a new graphical method based on Bayesian model selection. Communications in Statistics—Simulation and Computation®, 37(5), 962-977.

Badri, D. V., Quintana, N., El Kassis, E. G., Kim, H. K., Choi, Y. H., Sugiyama, A., . . . Vivanco, J. M. (2009). An ABC transporter mutation alters root exudation of phytochemicals that provoke an overhaul of natural soil microbiota. *Plant Physiol, 151*(4), 2006-2017.

Barth, S., McGrath, S. K., Arojju, S. K., & Hodkinson, T. R. (2015). An Irish perennial ryegrass genetic resource collection clearly divides into two major gene pools. Plant Genetic Resources, 15(3), 269-278. doi: 10.1017/S1479262115000611

Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London. Series B: Biological Sciences, 263(1377), 1619-1626.

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.

Blackmore, T., Thomas, I., McMahon, R., Powell, W., & Hegarty, M. (2015). Genetic–geographic correlation revealed across a broad European ecotypic sample of perennial ryegrass (Lolium perenne) using array-based SNP genotyping. *Theor Appl Genet, 128*(10), 1917-1932. doi: 10.1007/s00122-015-2556-3

Blackmore, T., Thorogood, D., Skøt, L., McMahon, R., Powell, W., & Hegarty, M. (2016). Germplasm dynamics: the role of ecotypic diversity in shaping the patterns of genetic variation in Lolium perenne. *Scientific Reports, 6*, 22603. doi: 10.1038/srep22603

Bolaric, S., Barth, S., Melchinger, A. E., & Posselt, U. K. (2005). Genetic diversity in European perennial ryegrass cultivars investigated with RAPD markers. Plant Breeding, 124(2), 161-166. doi: 10.1111/j.1439-0523.2004.01032.x

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

Brazauskas, G., Pašakinskienė, I., & Lübberstedt, T. (2013). Estimation of Temporal Allele Frequency Changes in Ryegrass Populations Selected for Axillary Tiller Development *Breeding strategies for sustainable forage and turf grass improvement* (pp. 81-87): Springer.

Brazauskas, G., Xing, Y., Studer, B., Schejbel, B., Frei, U., Berg, P., & Lübberstedt, T. (2013). Identification of genomic loci associated with crown rust resistance in perennial ryegrass (Lolium perenne L.) divergently selected populations. *Plant Sci, 208*, 34-41.

Byrne, S., Conaghan, P., Barth, S., Arojju, S., Casler, M., Michel, T., . . . Milbourne, D. (2017). Using variable importance measures to identify a small set of SNPs to predict heading date in perennial ryegrass. *Scientific Reports, 7*(1), 3566. doi: 10.1038/s41598-017-03232-8

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., & Asp, T. (2013). Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. *PLoS One, 8*(3), e57438.

Byrne, S., Nagy, I., Pfeifer, M., Armstead, I., Swain, S., Studer, B., . . . Hentrup, S. (2015). A synteny-based draft genome sequence of the forage grass Lolium perenne. *The Plant Journal, 84*(4), 816-826.

Casler, M., Jung, H., & Coblentz, W. (2008). Clonal selection for lignin and etherified ferulates in three perennial grasses. *Crop Sci, 48*(2), 424-433.

Chapman, D., Bryant, J., Olayemi, M., Edwards, G., Thorrold, B., McMillan, W., . . . Moorhead, A. (2017). An economically based evaluation index for perennial and short-term ryegrasses in N ew Z ealand dairy farm systems. *Grass Forage Sci, 72*(1), 1-21.

Chapman, D., Muir, P., & Faville, M. (2015). Persistence of dry matter yield among New Zealand perennial ryegrass (Lolium perenne L.) cultivars: insights from a long-term data set. *Journal of New Zealand Grasslands, 77*, 177-184.

Cochran, W. G. (1954). Some methods for strengthening the common χ 2 tests. *Biometrics, 10*(4), 417-451.

Dinkins, R. D., Barnes, A., & Waters, W. (2010). Microarray analysis of endophyte-infected and endophyte-free tall fescue. *J Plant Physiol, 167*(14), 1197-1203. doi: https://doi.org/10.1016/j.jplph.2010.04.002

Dodds, K., Symonds, J., Brauning, R., McEwan, J., & Clarke, S. (2018). *A depth-adjusted FST calculation for low-depth sequencing data*.

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), e19379.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol, 14*(8), 2611-2620.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics, 164*(4), 1567-1587.

Faville, M., Briggs, L., Cao, M., Koulman, A., Jahufer, M., Koolaard, J., & Hume, D. (2015). A QTL analysis of host plant effects on fungal endophyte biomass and alkaloid expression in perennial ryegrass. *Mol Breed, 35*(8), 161.

Faville, M., Ganesh, S., Cao, M., Jahufer, M., Bilton, T., Easton, H., . . . Barrett, B. (2018). Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theor Appl Genet*. doi: 10.1007/s00122-017-3030-1

Fè, D., Ashraf, B., Pedersen, M., Janss, L., Byrne, S., Roulund, N., . . . Jensen, J. (2016). Accuracy of Genomic Prediction in a Commercial Perennial Ryegrass Breeding Program. *The Plant Genome, 9*(3). doi: 10.3835/plantgenome2015.11.0110

Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B., Pedersen, M., . . . Jensen, C. (2015). Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics, 16*(1), 921.

Flanagan, S. P., & Jones, A. G. (2017). Constraints on the F ST–Heterozygosity Outlier Approach. *J Hered, 108*(5), 561-573.

Frichot, E., & François, O. (2015). LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution, 6*(8), 925-929.

Gagic, M., Faville, M. J., Zhang, W., Forester, N. T., Rolston, M. P., Johnson, R. D., . . . Hudson, D. (2018). Seed transmission of Epichloë endophytes in Lolium perenne is heavily influenced by host genetics. *Frontiers in Plant Science, 9*, 1580.

Gallagher, J. A., Turner, L. B., Cairns, A. J., Farrell, M., Lovatt, J. A., Skøt, K., . . . Roldan-Ruiz, I. (2015). Genetic differentiation in response to selection for water-soluble carbohydrate content in perennial ryegrass (Lolium perenne L.). BioEnergy Research, 8(1), 77-90.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol Ecol Notes, 5*(1), 184-186.

Gutjahr, C., Radovanovic, D., Geoffroy, J., Zhang, Q., Siegler, H., Chiapello, M., . . . Guiderdoni, E. (2012). The half-size ABC transporters STR1 and STR2 are indispensable for mycorrhizal arbuscule formation in rice. *The Plant Journal, 69*(5), 906-920.

Hao, D., Zhang, Z., Cheng, Y., Chen, G., Lu, H., Mao, Y., . . . Xue, L. (2015). Identification of Genetic Differentiation between Waxy and Common Maize by SNP Genotyping. *PLOS ONE, 10*(11), e0142585. doi: 10.1371/journal.pone.0142585

Hothorn, T., Hornik, K., Van De Wiel, M. A., & Zeileis, A. (2008). Implementing a class of permutation pests: the coin package.

Hume, D., Cooper, B., & Panckhurst, K. (2009). *The role of endophyte in determining the persistence and productivity of ryegrass, tall fescue and meadow fescue in Northland.* Paper presented at the Proceedings of the New Zealand Grassland Association.

Hume, D., Ryan, D., Cooper, B., & Popay, A. (2007). *Agronomic performance of AR37-infected ryegrass in northern New Zealand.* Paper presented at the Proceedings of the conference-New Zealand Grassland association.

Johnson, L., de Bonth, A., Briggs, L., Caradus, J., Finch, S., Fleetwood, D., . . . Card, S. (2013). The exploitation of epichloae endophytes for agricultural benefit. *Fungal Diversity, 60*(1), 171-188. doi: 10.1007/s13225-013-0239-4

Johnson, L. J., Johnson, R. D., Schardl, C. L., & Panaccione, D. G. (2003). Identification of differentially expressed genes in the mutualistic association of tall fescue with Neotyphodium coenophialum. *Physiol Mol Plant Pathol, 63*(6), 305-317. doi: https://doi.org/10.1016/j.pmpp.2004.04.001

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ, 2*, e281.

Kang, J., Park, J., Choi, H., Burla, B., Kretzschmar, T., Lee, Y., & Martinoia, E. (2011). Plant ABC Transporters. *The arabidopsis book, 9*, e0153-e0153. doi: 10.1199/tab.0153

Khan, A., Bassett, S., Voisey, C., Gaborit, C., Johnson, L., Christensen, M., . . . Johnson, R. (2010). Gene expression profiling of the endophytic fungus Neotyphodium lolii in association with its host plant perennial ryegrass. *Australasian Plant Pathology, 39*(5), 467-476.

Kovi, M. R., Fjellheim, S., Sandve, S. R., Larsen, A., Rudi, H., Asp, T., . . . Rognli, O. A. (2015). Population Structure, Genetic Variation, and Linkage Disequilibrium in Perennial Ryegrass Populations Divergently Selected for Freezing Tolerance. Frontiers in Plant Science, 6(929). doi: 10.3389/fpls.2015.00929

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.

Krattinger, S. G., Lagudah, E. S., Wicker, T., Risk, J. M., Ashton, A. R., Selter, L. L., . . . Keller, B. (2011). Lr34 multi-pathogen resistance ABC transporter: molecular analysis of homoeologous and orthologous genes in hexaploid wheat and other grass species. *The Plant Journal, 65*(3), 392-403. doi: 10.1111/j.1365-313X.2010.04430.x

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 Sci, 41, 1565-1572. doi: 10.2135/cropsci2001.4151565x

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357.

Latch, G., Hunt, W., & Musgrave, D. (1985). Endophytic fungi affect growth of perennial ryegrass. *N Z J Agric Res, 28*(1), 165-168.

Latchs, G., & Christensen, M. (1985). Artificial infection of grasses with endophytes. *Ann Appl Biol, 107*(1), 17-24.

Lee, J. M., Matthew, C., Thom, E. R., & Chapman, D. F. (2012). Perennial ryegrass breeding in New Zealand: a dairy industry perspective. *Crop and Pasture Science, 63*(2), 107-127. doi: https://doi.org/10.1071/CP11282

Lewontin, R., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics, 74*(1), 175-195.

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol Ecol, 23*(9), 2178-2192.

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Molecular Ecology Resources, 19(3), 639-647. doi: 10.1111/1755-0998.12995

Luu, K., Bazin, E., & Blum, M. G. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular ecology resources, 17*(1), 67-77.

Martinoia, E., Grill, E., Tommasini, R., Kreuz, K., & Amrhein, N. (1993). ATP-dependent glutathione S-conjugate'export'pump in the vacuolar membrane of plants. Nature, 364(6434), 247.

Martins, H., Caye, K., Luu, K., Blum, M. G., & Francois, O. (2016). Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Mol Ecol, 25*(20), 5029-5042.

Mayer, K. F. X., Martis, M., Hedley, P. E., Šimková, H., Liu, H., Morris, J. A., . . . Stein, N. (2011). Unlocking the Barley Genome by Chromosomal and Comparative Genomics. The Plant Cell, 23(4), 1249-1263. doi: 10.1105/tpc.110.082537

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 Forage Sci, 56*(4), 383-394. doi: 10.1046/j.1365-2494.2001.00288.x

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics, 20*(2), 289-290.

Parsons, A., Edwards, G., Newton, P., Chapman, D., Caradus, J., Rasmussen, S., & Rowarth, J. (2011). Past lessons and future prospects: plant breeding for yield and persistence in cool-temperate pastures. *Grass Forage Sci, 66*(2), 153-172.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet, 2*(12), e190.

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS one, 7*(2), e32253.

Popay, A., & Hume, D. (2011). Endophytes improve ryegrass persistence by controlling insects. *Pasture persistence*.

Prevosti, A., Ocana, J., & Alonso, G. (1975). Distances between populations of Drosophila subobscura, based on chromosome arrangement frequencies. *Theor Appl Genet, 45*(6), 231-241.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved 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. *Anal Biochem, 191*(1), 31.

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., & He, Z. (2017). Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. Molecular Plant, 10(8), 1047-1064. doi: https://doi.org/10.1016/j.molp.2017.06.008

Rea, P. A. (2007). Plant ATP-binding cassette transporters. *Annu Rev Plant Biol, 58*, 347-375.

Shinozuka, H., Cogan, N. O. I., Spangenberg, G. C., & Forster, J. W. (2011). Comparative Genomics in Perennial Ryegrass (Lolium perenne L.): Identification and Characterisation of an Orthologue for the Rice Plant Architecture-Controlling Gene OsABCG5. *International Journal of Plant Genomics, 2011*, 12. doi: 10.1155/2011/291563

Skøt, L., Humphreys, M. O., Armstead, I., Heywood, S., Skøt, K. P., Sanderson, R., . . . Hamilton, N. R. S. (2005). An association mapping approach to identify flowering time genes in natural populations of Lolium perenne (L.). *Mol Breed, 15*(3), 233-245. doi: 10.1007/s11032-004-4824-9

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences, 100*(16), 9440-9445.

Stukkens, Y., Bultreys, A., Grec, S., Trombik, T., Vanham, D., & Boutry, M. (2005). NpPDR1, a pleiotropic drug resistance-type ATP-binding cassette transporter from Nicotiana plumbaginifolia, plays a major role in plant pathogen defense. *Plant Physiol, 139*(1), 341-352.

Sugiyama, A., Shitan, N., & Yazaki, K. (2007). Involvement of a soybean ATP-binding cassette-type transporter in the secretion of genistein, a signal flavonoid in legume-Rhizobium symbiosis. *Plant Physiol, 144*(4), 2000-2008.

Sun, X., Jia, Q., Guo, Y., Zheng, X., & Liang, K. (2015). Whole-Genome Analysis Revealed the Positively Selected Genes during the Differentiation of indica and Temperate japonica Rice. *PLOS ONE, 10*(3), e0119239. doi: 10.1371/journal.pone.0119239

The International Barley Genome Sequencing, C., Mayer, K. F. X., Waugh, R., Langridge, P., Close, T. J., Wise, R. P., . . . Stein, N. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature, 491*, 711. doi: 10.1038/nature11543

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res, 22*(22), 4673-4680. doi: 10.1093/nar/22.22.4673

Voisey, C., Khan, A., Park, Z., Johnson, L., Johnson, R., Ramakrishna, M., . . . McCulloch, A. (2007). *Development of an Affymetrix dual species (Noetyphodium lolli/Lolium perenne) symbiosis GeneChip.* Paper presented at the Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses, New Zealand Grass Association, Christchurch, New Zealand.

Wang, M., Kornblau, S. M., & Coombes, K. R. (2018). Decomposing the apoptosis pathway into biologically interpretable principal components. Cancer Informatics, 17, 1176935118771082.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38*(6), 1358-1370.

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 FST. *The American Naturalist, 186*(S1), S24-S36.

Wisser, R. J., Murray, S. C., Kolkman, J. M., Ceballos, H., & Nelson, R. J. (2008). Selection Mapping of Loci for Quantitative Disease Resistance in a Diverse Maize Population. *Genetics, 180*(1), 583-599. doi: 10.1534/genetics.108.090118

Zhang, N., Zhang, S., Borchert, S., Richardson, K., & Schmid, J. (2011). High levels of a fungal superoxide dismutase and increased concentration of a PR-10 plant protein in associations between the endophytic fungus Neotyphodium lolii and ryegrass. *Mol Plant-Microbe Interact, 24*(8), 984-992.

Zhang, Q., Blaylock, L. A., & Harrison, M. J. (2010). Two Medicago truncatula half-ABC transporters are essential for arbuscule development in arbuscular mycorrhizal symbiosis. *The Plant Cell, 22*(5), 1483-1497.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics, 28*(24), 3326-3328.

Chapter 4

General Discussion

The objective of this study is to investigate a perennial ryegrass breeding population under recurrent selection (RS) for compatibility (specifically, inter-generational transmission) with an endophyte sourced from tall fescue. In New Zealand, grass persistence is known to be partly mediated by the interaction of their mutualistic symbionts *Epichloë* endophytes, which increase defence against pest herbivores (Popay & Hume, 2011). Not all *Epichloë* endophytes are beneficial in the context of pastoral agriculture because some are not only toxic to insect pests but also to livestock. Research and discovery of novel endophytes is therefore important to the forage industry because these endophytes produce alkaloids that are non-toxic to mammals, and offer protection against insect pests. Perennial ryegrass breeding for improved compatibility with novel endophytes is a promising approach for improving persistence. The endophyte used in this study, AR501, represents a class of novel endophytes that is not naturally associated with perennial ryegrass. While novel associations of endophytes with forage grasses have been exploited, it is generally limited to associations occurring within the normal host species range, such as perennial ryegrass with *E. festucae* var. *lolii* strains or tall fescue with *E. coenophiala* strains (Johnson et al., 2013; Young et al., 2014). Nevertheless, exploitation of novel association of AR501 has been reported in a few studies (Easton et al., 2007; Fletcher, 2012; Freitas, 2017). AR501 is known to produce lolines (Easton et al., 2007; Faville et al., 2015), a class of alkaloids not produced in perennial ryegrass infected with its natural endophyte. Lolines are reported to have wide insecticidal properties but are safe for livestock (Schardl et al., 2007). In the present study, the investigation of AR501 transmission in a non-native host will contribute to the advancement of perennial breeding for endophyte compatibility. It is especially important if breeders are to successfully explore and exploit cross-species associations. Moreover, it will contribute to the broader understanding of the grass-endophyte interaction. The following sections discuss the summary of results of the present study, and possible future follow up work.

4.1. Summary of outcomes and discussion

The plant-fungal interaction is partially under genetic control from the host, and this is exploited in perennial ryegrass breeding. The genetic background of the host is known to influence endophyte biomass and alkaloid expression *in planta* (Easton et al., 2002 and Faville et al., 2015) as well as vertical transmission of the endophyte between generations (Gagic et al., 2018). Gagic et al. (2018) observed a higher percentage of viable endophyte in perennial ryegrass populations that had undergone selection for improved AR37 endophyte infection than those that had not. Similarly, in the present study, we found that four cycles of selection for endophyte compatibility in PGG04 may have resulted in an increased proportion of viable AR501 in the population. In Chapter 2, we investigated the effect of positive selection on infection rate by growing early and late generations of PGG04, and testing for endophyte infection using three methods. Tissue-print immunoblotting detected a higher percentage of viable endophytes in the late than in the early generation of PGG04. Using the microsatellite B11, endophyte genotyping confirmed the results of immunoblotting. Genotyping also identified AR501 as the endophyte strain present in the population. It also determined that there was no contamination by the common toxic endophyte or any other known novel endophyte strain. Endophyte detection in the seed showed that AR501 is present in the seeds at a very high level, both in the early and late generations of PGG04. The difference in the infection rate in the seed (i.e. seed squash) and seedling (i.e. immunoblotting) of PGG04 could be due to seed storage conditions and genetics. Seeds from the early generation (PGG04-C2) were under storage for nearly six years before the conduct of this study, while the late generation (PGG04-C6) was only stored a few months after harvest. Therefore, it is possible that the storage conditions of PGG04-C2 contributed to the decline in endophyte viability. The early generation seeds were stored under recommended conditions of low temperature and in moisture-resistant packaging. However, since the effects of storage conditions were not formally tested in the present study, their effects cannot be discounted. Endophytes detected in the seed include both viable and non-viable endophytes. It is also likely that endophyte loss from seed to seedlings was due to host incompatibility since endophytes are known to exhibit a high degree of host specificity. This is particularly relevant in the present study because it deals with cross-species association of a tall fescue endophyte infected in perennial ryegrass. Overall, under ideal conditions, host genetics largely influence endophyte viability. Based on the results of the current study, it can be concluded that recurrent selection in PGG04 could improve AR501 compatibility in terms of transmission, which was reflected by an increase in the proportion of viable endophyte from early to the late generation.

Breeding exploits the genetic variation in plant populations to improve a trait of interest. It is therefore important to study the available genetic variation in the breeding population. Several studies have characterized genetic diversity in perennial ryegrass populations, both natural and synthetic. However, in most diversity studies on perennial ryegrass populations, low molecular marker density systems were used (Barth et al., 2015; Brazauskas et al., 2011; Bolaric et al., 2005; Ghesquiere et al., 2003; Guthridge et al., 2001 and Kubik et al., 2001). Current molecular marker technologies such as genotyping-by-sequencing (GBS) are stimulating the re-examination of genetic variation in perennial ryegrass including advanced breeding populations which can contribute to the progress of breeding (Faville et al., 2018; Fè et al., 2015). In the present study, we investigated genetic diversity and population structure using at least 31K SNPs from GBS. Recurrent selection (RS), as practiced in the population under investigation, aims to increase the frequency of favourable alleles for the trait of interest. In the current case, selection should improve endophyte compatibility – specifically an improvement in AR501 transmission via seed, across generations. As mentioned above, this work provided evidence for the improvement of AR501 compatibility in terms of an increased proportion of the viable endophyte in PGG04 following multiple cycles of recurrent selection. Genomic regions associated with selected traits can also be studied in breeding populations. Selection mapping (SM) is a method described as "a range of approaches that identifies alleles, loci, and epistatic interactions using populations that have been subjected to iterative cycles of recombination and selection" (Wisser et al., 2008). SM has been successfully carried out in perennial ryegrass breeding populations to identify genes/QTLs associated with tiller development (Brazauskas, Pašakinskienė, et al., 2013) and crown rust resistance (Brazauskas, Xing, et al., 2013). In Chapter 3 of the current study, we determined selection signatures in the genome that indicate a response to selection for endophyte transmission. Selection results in allele frequency changes in genes responsible for trait variation, as well as in loci physically linked to such genes. Allele frequencies in the breeding population will be highly variable but loci for which one allele is favoured (i.e. selected) would be more genetically differentiated than neutral loci when comparing different generations of RS (Whitlock & Lotterhos, 2015). GBS proved useful in this population genomics study of a breeding population in active development. Analysis of data from GBS showed that RS for AR501 transmission and viability resulted in changes in population structure and genetic diversity in the breeding population PGG04. As expected, more genetic variability was within generation than between generations, based on AMOVA results. Comparing GBS data from the PGG04-C2 (early) and PGG04-C6 (late) generations, results show that there is an excess of rare alleles in the later generation and a small but significant reduction of genetic diversity (H_e). Both of these are characteristics of a population under selection.

In the current study, selection was targeted towards improving endophyte compatibility. The genetic changes across generations were therefore not uniform which allowed for the identification of selection signatures in the genome. Multiple analytical procedures identified a core group of SNPs significantly associated both with the observed changes in population structure and with trait values. Specifically, twelve SNPs had relatively high F_{ST} values and were more correlated to principal components describing PCA-based population structure than other SNPs which are potentially neutral. Since the population was developed for breeding for endophyte compatibility, SNPs under selection were confirmed by association analysis with infection data using logistic regression. Among genomic regions under selection, five regions tagged by nine SNPs were of particular interest as they were also associated with the infection data. In Chapter 2, it was mentioned that storage factors and genetics influenced variability in infection rate among PGG04 seeds. The discovery of a few SNPs that show signs of selection and are associated with infection data, points to the effects of genetics in influencing endophyte compatibility, that is, selection improves endophyte transmission.

The five interesting genomic regions were found to increase the odds of the beneficial infection by more than five times (maximum), relative to the number of favourable alleles. Several important SNPs were located in chromosome 7 which is suggested to be rich in defence-related (DR) genes based on high collinearity of grass genomes (Mayer et al., 2011). Hence, this study provides indirect evidence in support of the hypothesis that DR genes may be involved in the grass-endophyte interaction. One of the significant SNPs was found to have annotations that suggested a role for ABCG-type transporters. ABCG proteins are known to have important roles in symbioses and plant defence. The results of this study, therefore, provide additional evidence on the importance of ABCG genes in plant-microbe interactions and may prove to be important in the perennial ryegrass-endophyte interaction. The nine most significant SNPs identified in the current study, especially the one tagging an ABCG gene, may be useful for developing marker-assisted selection schemes.

4.2. Future research

In this study, improvement of compatibility of a tall fescue endophyte in a perennial ryegrass breeding population, PGG04, was investigated. Evidence of improved compatibility was based mainly on an increase in the proportion of viable AR501 from early to late generations of PGG04. However, as mentioned previously, endophyte compatibility in the context of PGG04 development was in terms of transmission and viability. In the current study, we were not able to separate the effects of each of these traits. Nevertheless, we provided indirect evidence that recurrent selection improved endophyte transmission based on the infection rate. That is, the increase in the proportion of viable endophyte is a direct consequence of improved endophyte transmission. The next step, therefore, is to look at transmission per se. This means growing progenies and measuring their infection rate to determine their maternal parent’s performance in terms of its transmission efficiency. The other trait component, viability, is also worth investigating. In the current study, the early generation showed a decline in AR501 viability when comparing endophytes in the seeds and the growing tillers. The early generation seeds were sourced from cold storage, hence the decline in viability could be due to storage conditions. Had we investigated viability formally, we could have ruled out environmental factors. More importantly, we would contribute to the understanding of the genetic basis of this specific trait in the host.

We also analysed the consequence of recurrent selection to the genetic variation in PGG04. In Chapter 3, we forwarded recommendations for the further improvement and validation of our analyses. The next step, therefore, is to implement these improvements starting with more accurate F_{ST} estimates by considering the read depth of the SNPs. This could be a proposed improvement in the R package OutFlank; that is, to include options specific for the analysis of GBS data (and related genotyping platforms). In addition to the adjustment of F_{ST} estimates, read depth can also be considered in the null distribution inference. In OutFlank, this is based on a trimmed distribution of F_{ST} calculated from quasi-independent SNPs. Read depth can be used as quality control criteria to ensure that inference of the null distribution is based only on high-quality SNPs (i.e. acceptable read depth). We also mentioned previously, that primer design can be pursued for eventual marker-assisted selection. This could be a practical way to validate the results of the present study and could provide plant breeders a tool for immediate application, if successful. It would also be interesting to compare the results of the current study and the QTL study of Faville et al. (2015). The QTL study investigated host genetic control of perennial ryegrass-endophyte interaction in terms of endophyte biomass and alkaloid expression. They also examined the interaction between perennial ryegrass and the tall fescue endophyte, AR501. The physical position of the loci under selection identified in the current study and QTLs in Faville et al. (2015) can be compared to determine genomic regions that were commonly detected. In addition, linkage maps can be constructed to include important loci detected in both studies, and primers designed for significant SNPs in the current study can be utilized. These follow up studies are only a few of the several future research areas opened up by the current study. While the results of the current study contribute to the advancement of research on grass-endophyte interaction and perennial ryegrass breeding for endophyte compatibility, more research work is still needed.

4.3. References

Barth, S., McGrath, S. K., Arojju, S. K., & Hodkinson, T. R. (2015). An Irish perennial ryegrass genetic resource collection clearly divides into two major gene pools. Plant Genetic Resources, 15(3), 269-278. doi: 10.1017/S1479262115000611

Bolaric, S., Barth, S., Melchinger, A. E., & Posselt, U. K. (2005). Genetic diversity in European perennial ryegrass cultivars investigated with RAPD markers. Plant Breeding, 124(2), 161-166. doi: 10.1111/j.1439-0523.2004.01032.x

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

Brazauskas, G., Pašakinskienė, I., & Lübberstedt, T. (2013). Estimation of Temporal Allele Frequency Changes in Ryegrass Populations Selected for Axillary Tiller Development *Breeding strategies for sustainable forage and turf grass improvement* (pp. 81-87): Springer.

Brazauskas, G., Xing, Y., Studer, B., Schejbel, B., Frei, U., Berg, P., & Lübberstedt, T. (2013). Identification of genomic loci associated with crown rust resistance in perennial ryegrass (Lolium perenne L.) divergently selected populations. *Plant Sci, 208*, 34-41.

Easton, H., Latch, G., Tapper, B., & Ball, O.-P. (2002). Ryegrass host genetic control of concentrations of endophyte-derived alkaloids. *Crop Sci, 42*(1), 51-57.

Easton, H., Lyons, T., Mace, W., Simpson, W., De Bonth, A., Cooper, B., & Panckhurst, K. (2007). *Differential expression of loline alkaloids in perennial ryegrass infected with endophyte isolated from tall fescue.* Paper presented at the Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses. New Zealand Grassland Association, Dunedin, New Zealand.

Faville, M., Briggs, L., Cao, M., Koulman, A., Jahufer, M., Koolaard, J., & Hume, D. (2015). A QTL analysis of host plant effects on fungal endophyte biomass and alkaloid expression in perennial ryegrass. *Mol Breed, 35*(8), 161.

Faville, M., Ganesh, S., Cao, M., Jahufer, M., Bilton, T., Easton, H., . . . Barrett, B. (2018). Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. *Theor Appl Genet*. doi: 10.1007/s00122-017-3030-1

Fè, D., Ashraf, B., Pedersen, M., Janss, L., Byrne, S., Roulund, N., . . . Jensen, J. (2016). Accuracy of Genomic Prediction in a Commercial Perennial Ryegrass Breeding Program. *The Plant Genome, 9*(3). doi: 10.3835/plantgenome2015.11.0110

Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B., Pedersen, M., . . . Jensen, C. (2015). Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics, 16*(1), 921.

Fletcher, L. (2012). *Novel endophytes in New Zealand grazing systems: The perfect solution or a compromise?* Paper presented at the Epichloae, endophytes of cool season grasses: implications, utilization and biology. Proceedings of the 7th International Symposium on Fungal Endophytes of Grasses, Lexington, Kentucky, USA, 28 June to 1 July 2010.

Freitas, P. (2017). Crossing the species barrier: investigating vertical transmission of a fungal endophyte from tall fescue within a novel ryegrass association (Doctoral dissertation, Lincoln University, Christchurch, New Zealand). Retrieved from https://researcharchive.lincoln.ac.nz/bitstream/handle/10182/8385/Freitas_PhD.pdf?sequence=3&isAllowed=y

Gagic, M., Faville, M. J., Zhang, W., Forester, N. T., Rolston, M. P., Johnson, R. D., . . . Hudson, D. (2018). Seed transmission of Epichloë endophytes in Lolium perenne is heavily influenced by host genetics. *Frontiers in Plant Science, 9*, 1580.

Ghesquiere, A., Calsyn, E., Baert, J., & Riek, J. (2003). Genetic diversity between and within ryegrass populations of the ECP/GR collection by means of AFLP markers. *Czech Journal of Genetics and Plant Breeding, 39*(Special issue), 333.

Guthridge, K. M., Dupal, M. P., Kölliker, R., Jones, E. S., Smith, K. F., & Forster, J. W. (2001). AFLP analysis of genetic diversity within and between populations of perennial ryegrass (Lolium perenne L.). *Euphytica, 122*(1), 191-201. doi: 10.1023/a:1012658315290

Johnson, L., de Bonth, A., Briggs, L., Caradus, J., Finch, S., Fleetwood, D., . . . Card, S. (2013). The exploitation of epichloae endophytes for agricultural benefit. *Fungal Diversity, 60*(1), 171-188. doi: 10.1007/s13225-013-0239-4

Kang, J., Park, J., Choi, H., Burla, B., Kretzschmar, T., Lee, Y., & Martinoia, E. (2011). Plant ABC Transporters. *The arabidopsis book, 9*, e0153-e0153. doi: 10.1199/tab.0153

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 Sci, 41, 1565-1572. doi: 10.2135/cropsci2001.4151565x

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol Ecol, 23*(9), 2178-2192.

Mayer, K. F. X., Martis, M., Hedley, P. E., Šimková, H., Liu, H., Morris, J. A., . . . Stein, N. (2011). Unlocking the Barley Genome by Chromosomal and Comparative Genomics. The Plant Cell, 23(4), 1249-1263. doi: 10.1105/tpc.110.082537

Popay, A., & Hume, D. (2011). Endophytes improve ryegrass persistence by controlling insects. *Pasture persistence*.

Schardl, C., Grossman, R., Nagabhyru, P., Faulkner, J., & Mallik, U. (2007). Loline alkaloids: Currencies of mutualism. *Phytochemistry, 68*(7), 980-996. doi: https://doi.org/10.1016/j.phytochem.2007.01.010

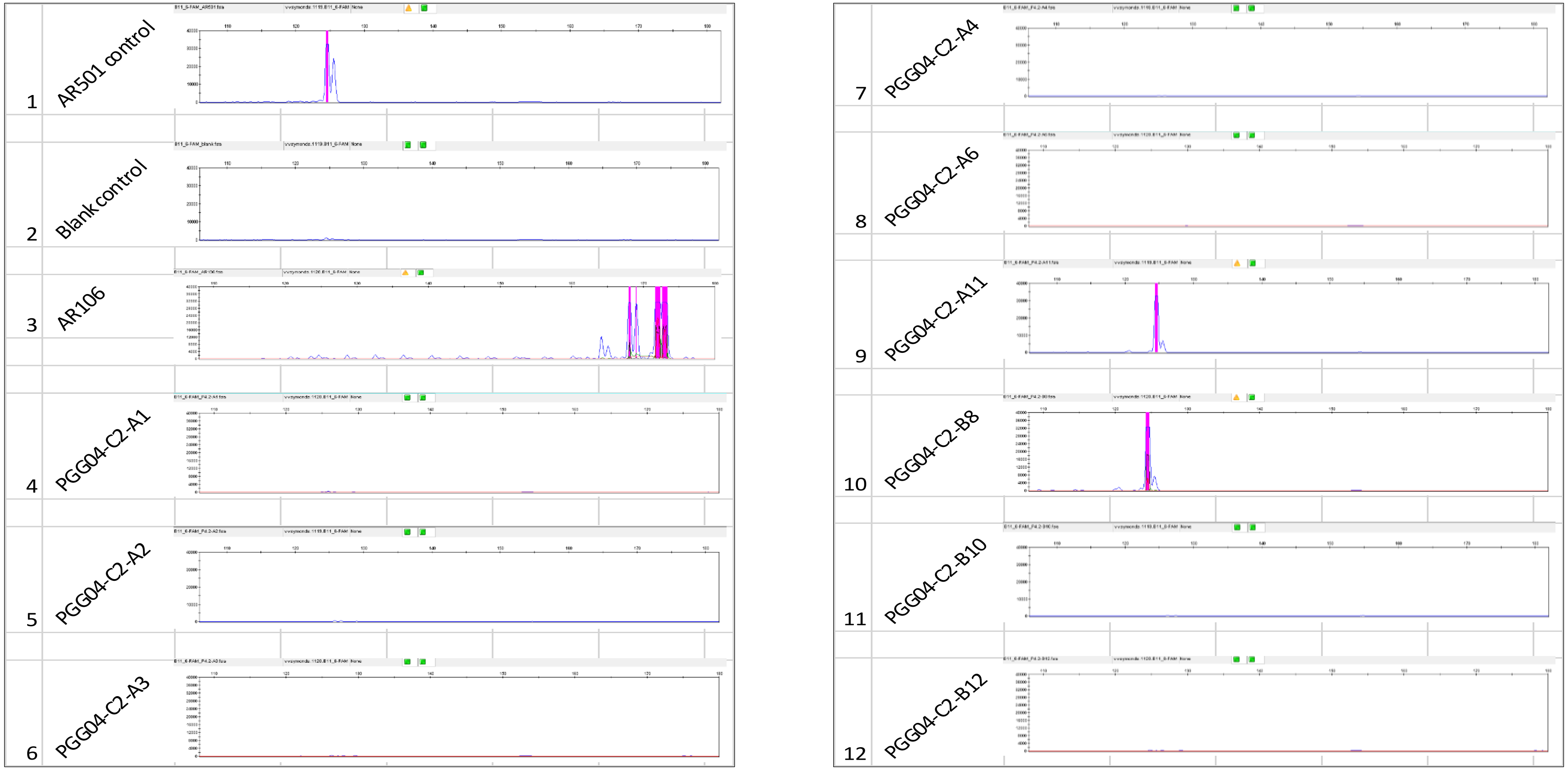
Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of FST. *The American Naturalist, 186*(S1), S24-S36.

Wisser, R. J., Murray, S. C., Kolkman, J. M., Ceballos, H., & Nelson, R. J. (2008). Selection Mapping of Loci for Quantitative Disease Resistance in a Diverse Maize Population. *Genetics, 180*(1), 583-599. doi: 10.1534/genetics.108.090118

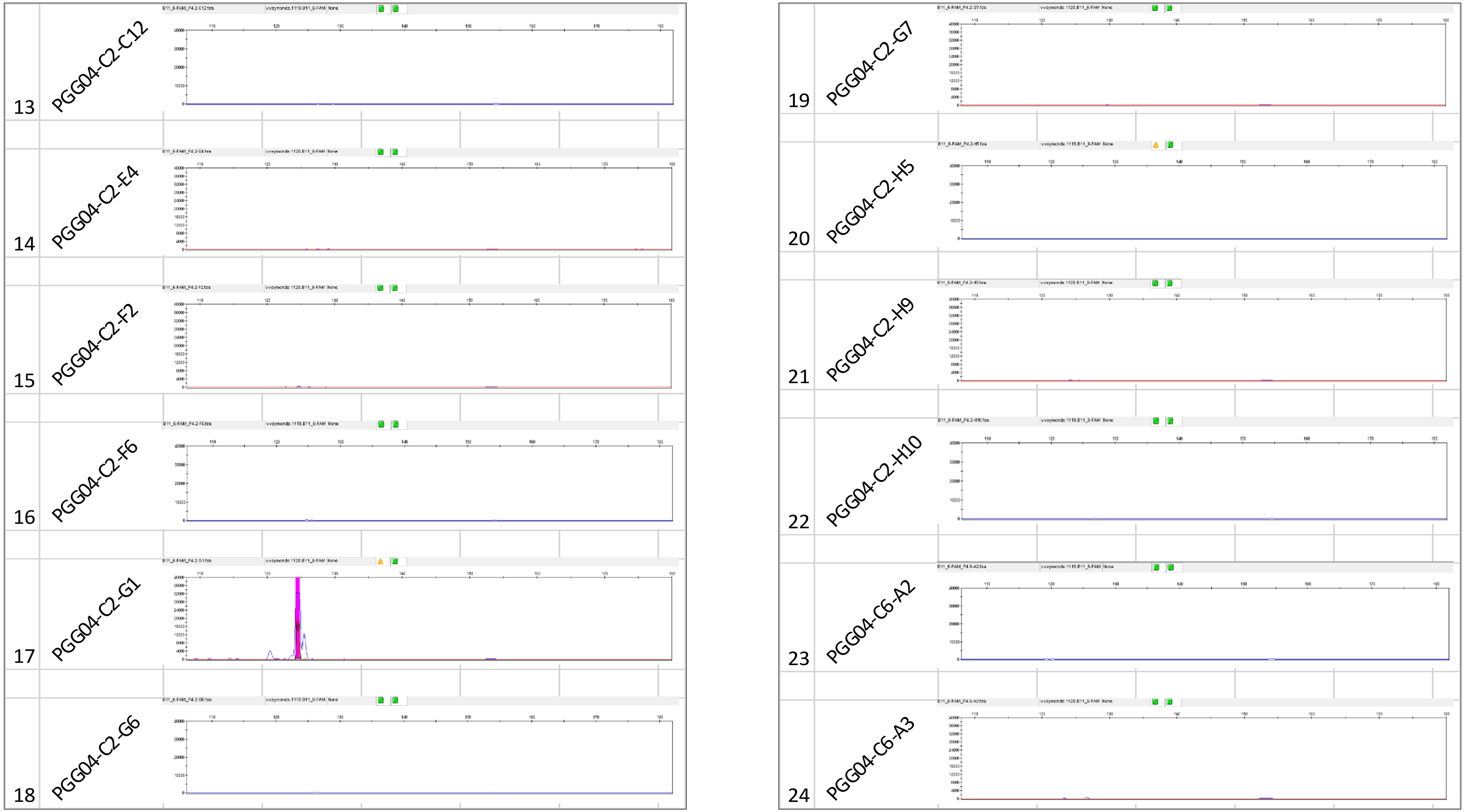
Young, C. A., Charlton, N. D., Takach, J. E., Swoboda, G. A., Trammell, M. A., Huhman, D. V., & Hopkins, A. A. (2014). Characterization of Epichloë coenophiala within the US: are all tall fescue endophytes created equal? *Frontiers in Chemistry, 2*(95). doi: 10.3389/fchem.2014.00095

Appendices

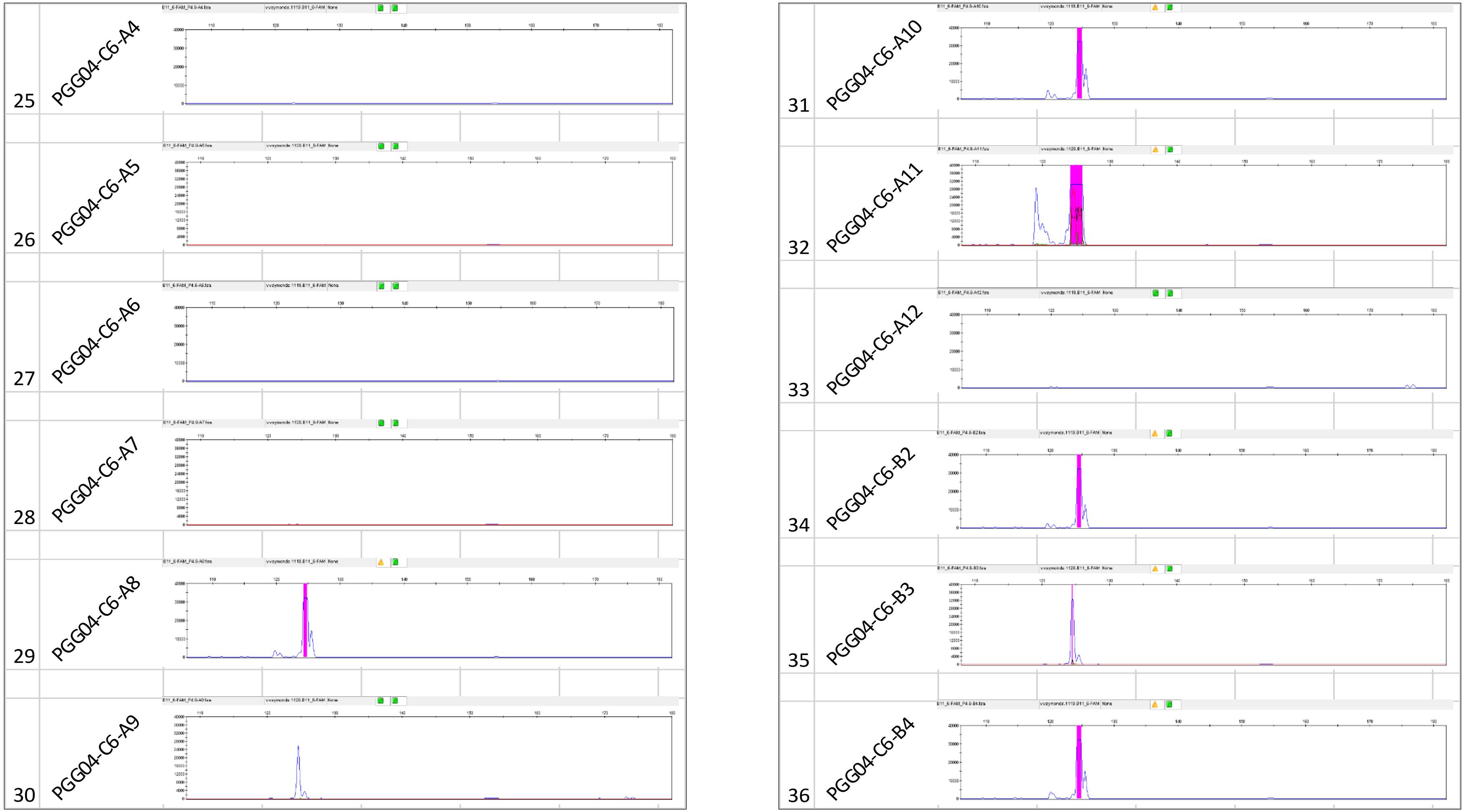
Chapter 2 Appendices



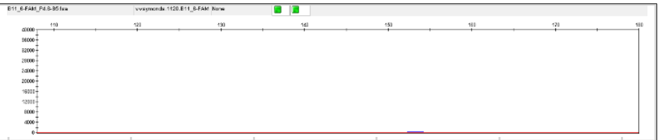
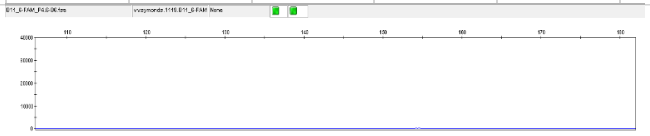
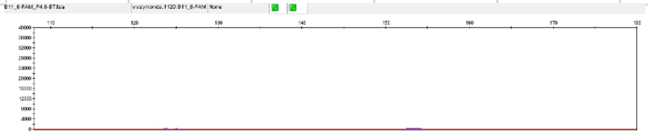
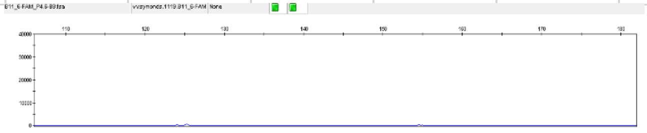
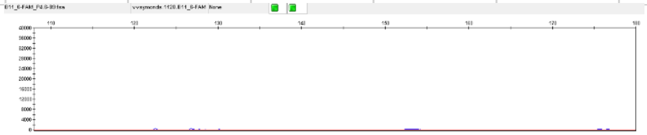
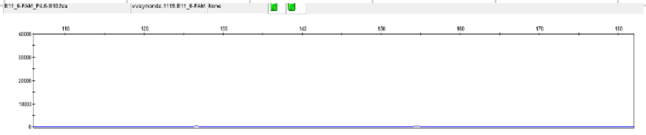
Appendix Figure 2.1. Microsatellite (B11) analysis of AR501 in the perennial ryegrass breeding population, PGG04.

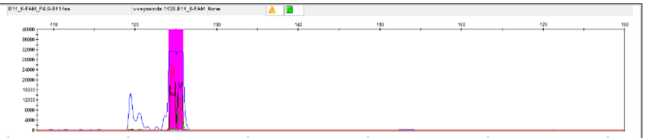
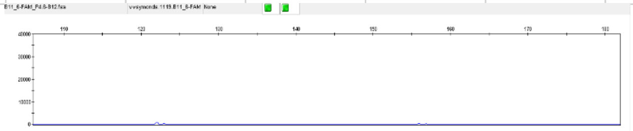
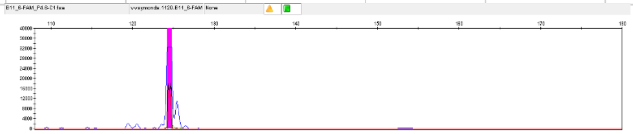
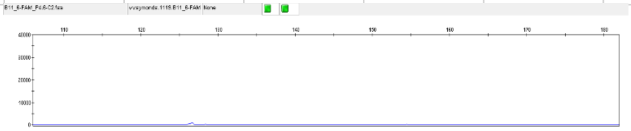
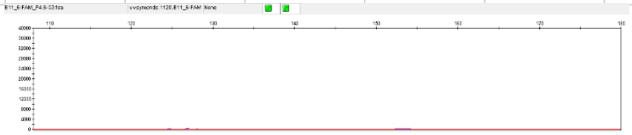
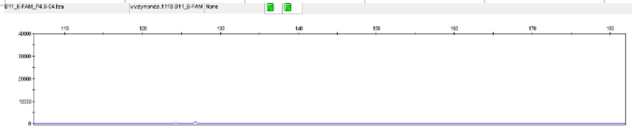


Appendix Fig. 2.1. continued

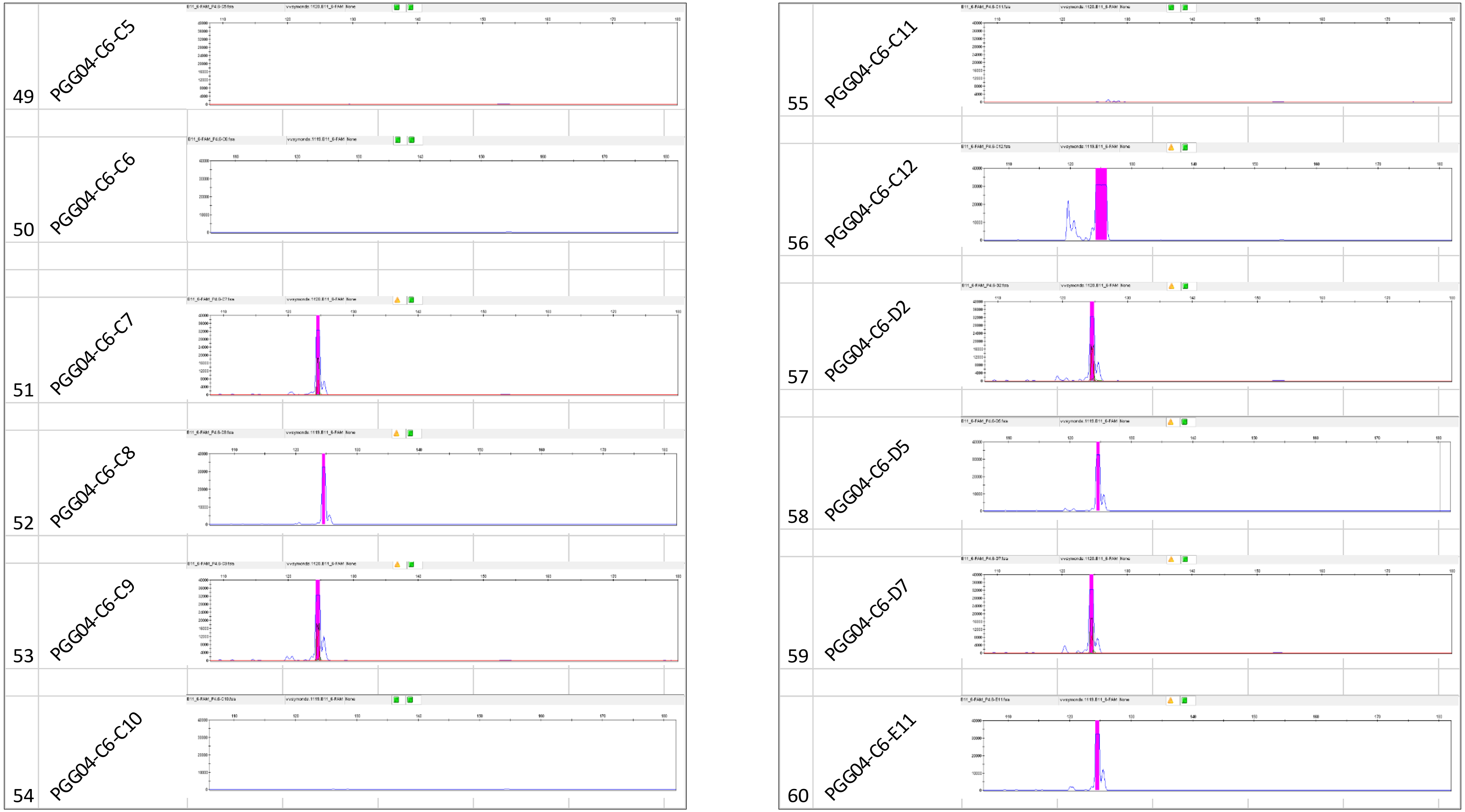


Appendix Fig. 2.1. continued

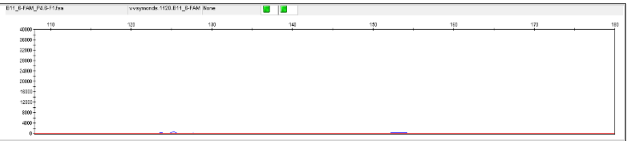
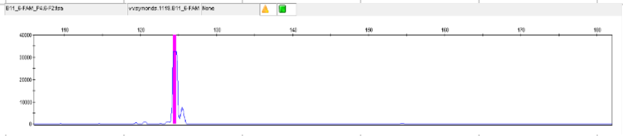
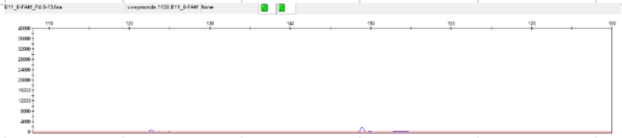
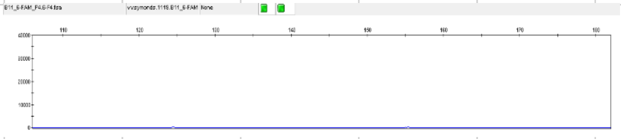
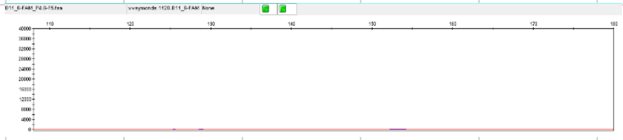
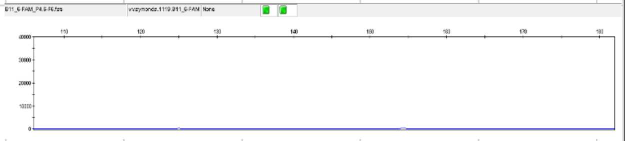
37	PGG04-C6-B5	
38	PGG04-C6-B6	
39	PGG04-C6-B7	
41	PGG04-C6-B8	
41	PGG04-C6-B9	
42	PGG04-C6-B10	

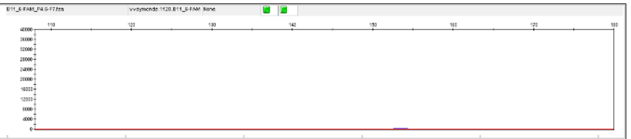
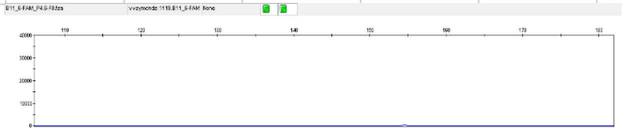
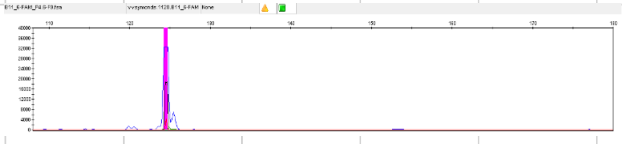
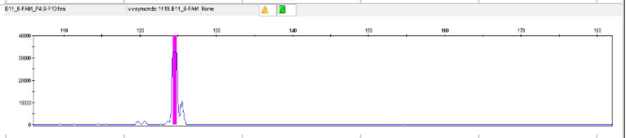
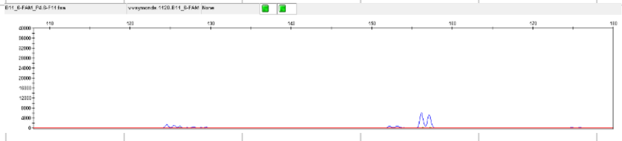
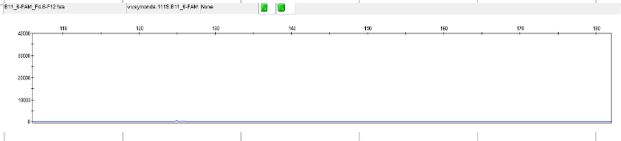
43	PGG04-C6-B11	
44	PGG04-C6-B12	
45	PGG04-C6-C1	
46	PGG04-C6-C2	
47	PGG04-C6-C3	
48	PGG04-C6-C4	

Appendix Fig. 2.1. continued

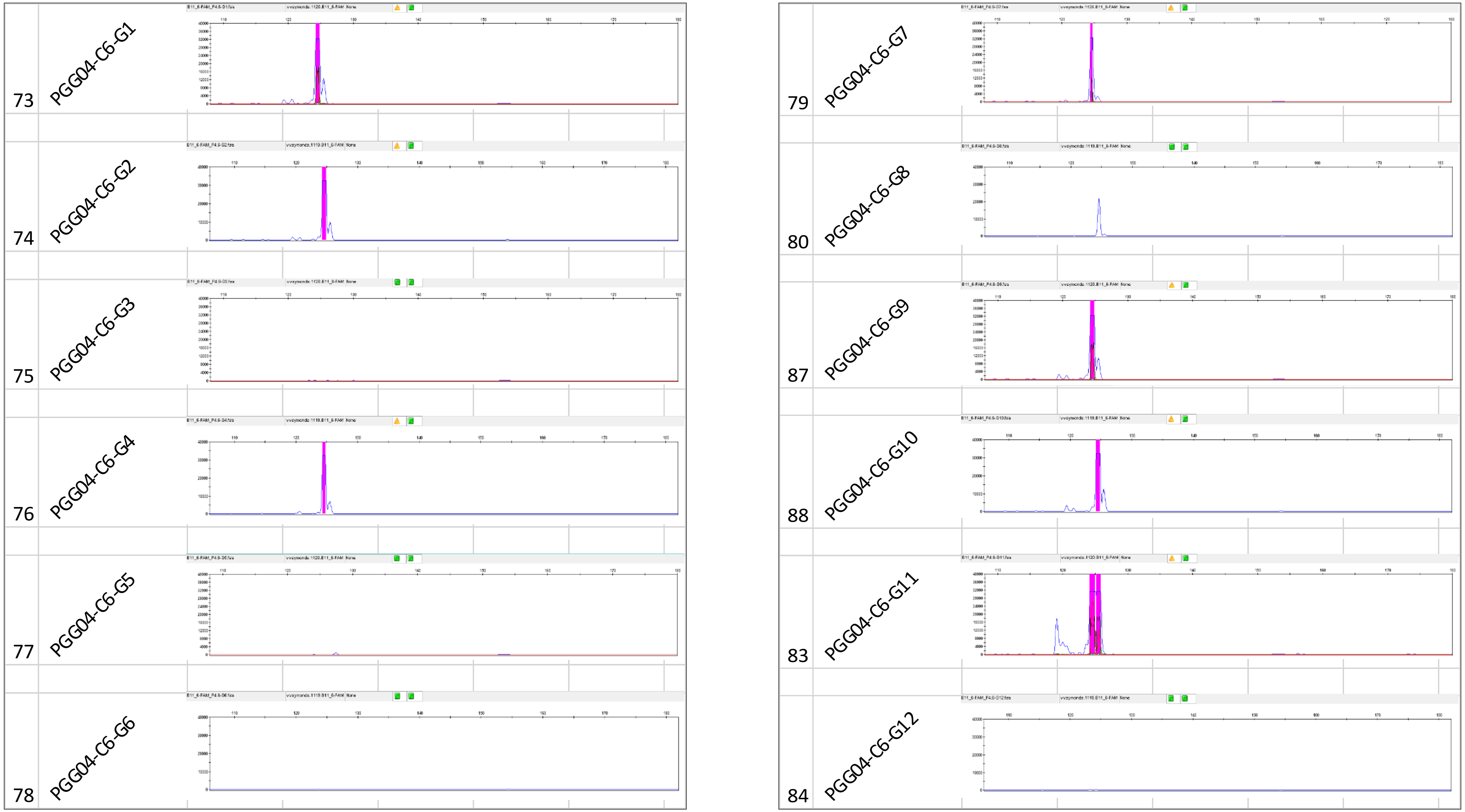


Appendix Fig. 2.1. continued

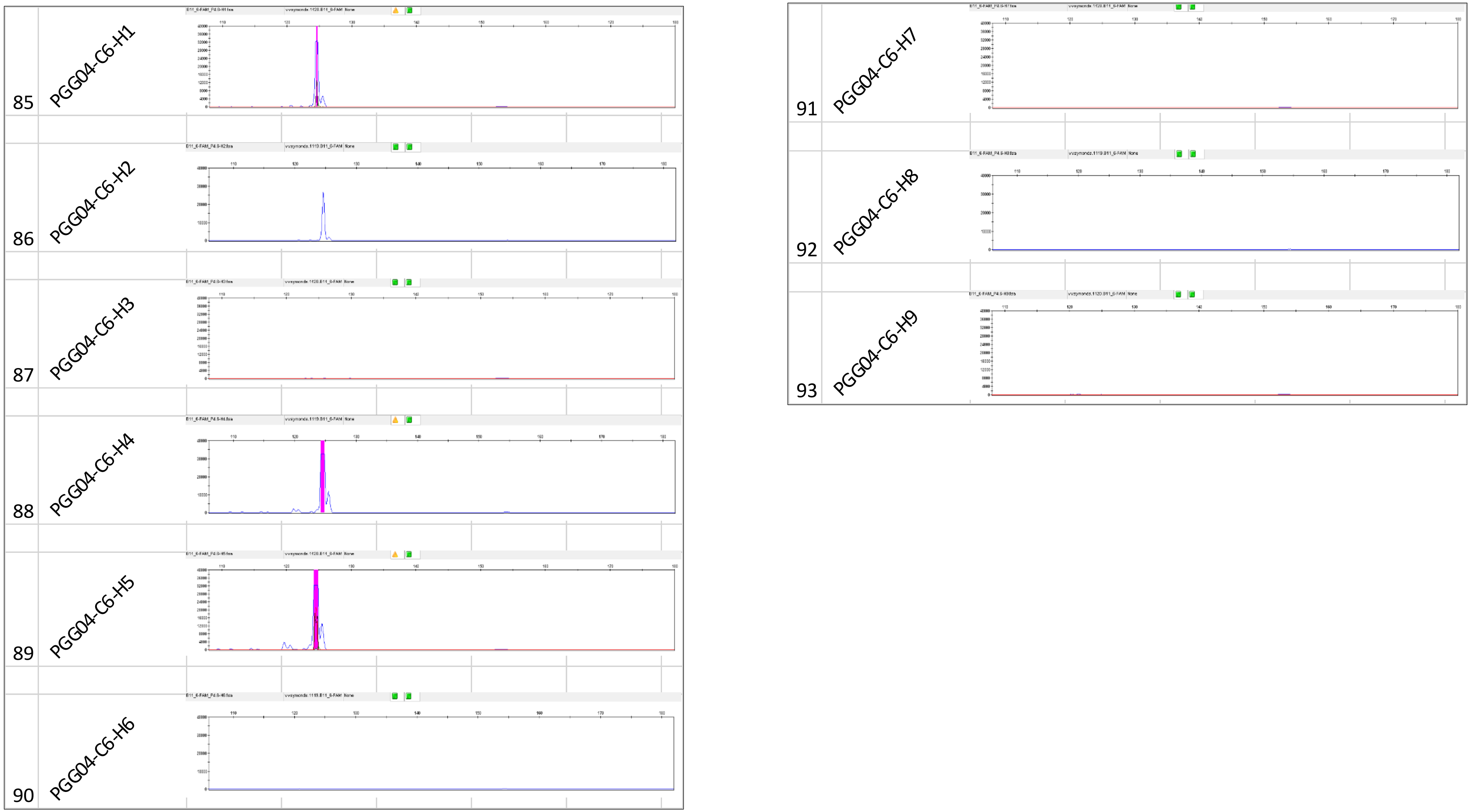
61	PGG04-C6-F1	
62	PGG04-C6-F2	
63	PGG04-C6-F3	
64	PGG04-C6-F4	
65	PGG04-C6-F5	
66	PGG04-C6-F6	

67	PGG04-C6-F7	
68	PGG04-C6-F8	
69	PGG04-C6-F9	
70	PGG04-C6-F10	
71	PGG04-C6-F11	
72	PGG04-C6-F12	

Appendix Fig. 2.1. continued



Appendix Fig. 2.1. continued



Appendix Fig. 2.1. continued

Chapter 3 Appendices

Pippin prep procedure

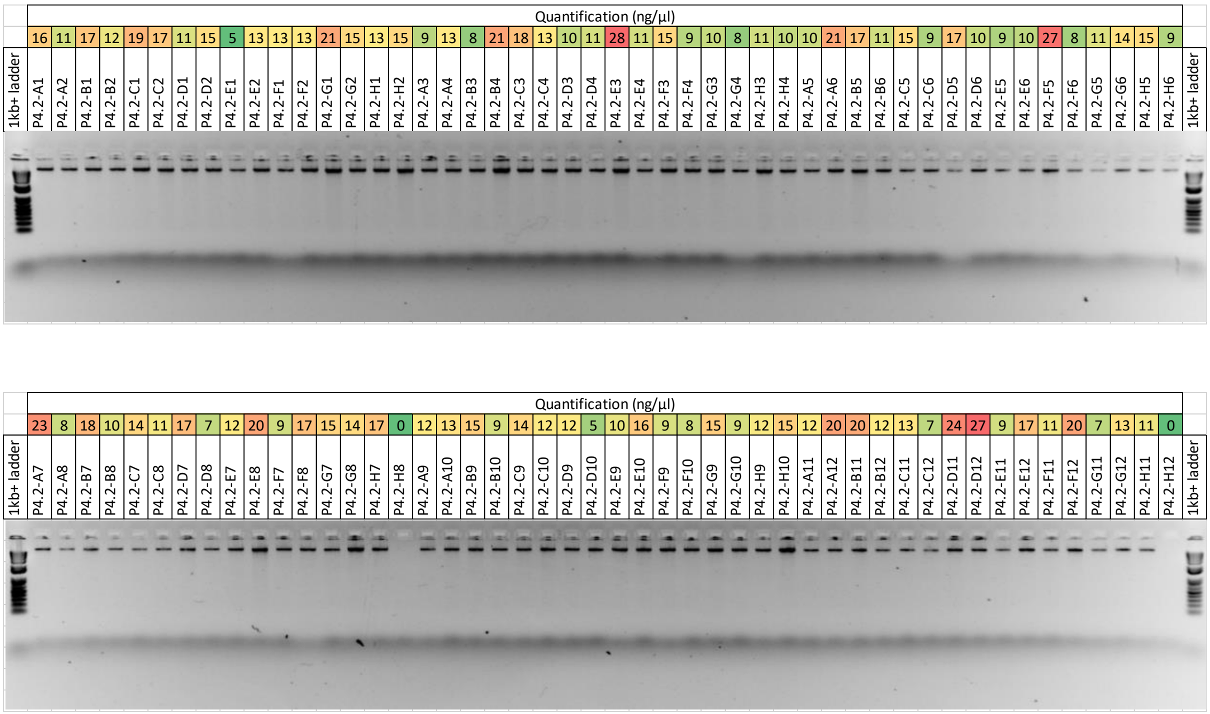
The Pippin prep software was first launched and a protocol for size selection of 193 to 313 base pairs (bp) was created. The following parameters were set in the protocol editor tab: 2% DF Marker L for Cassette type; 193 and 313 for BP Start and BP End (i.e. range of size to select) respectively; and using internal standards (which will automatically fill the Ref Lane values). Since the machine can run up to five lanes (i.e. samples) and only one was needed, unused lanes were turned off. The lane numbers were also checked to be matched. Second, an optical calibration was performed using the "CALIBRATE" control option of the software and with a calibration fixture. Third, the gel cassette was prepared. The gel was inspected to have the appropriate amount of buffer, that it had no breakage and that bubbles were not present in the optical path of the gel. The buffer in the elution well was also refreshed. Next, the gel was placed in the Pippin prep machine and a continuity test was done. This is a quality control procedure to check the current in the elution and separation channel in the gel. Then, 30 µL of the library was mixed with 10 µL of ladder L (internal standard) and spun quickly. Lastly, the library mixed with ladder L was loaded in the gel cassette and the Pippin prep protocol was run. After approximately 1.5 hours, elution of 40 µL of size-selected DNA was completed. The sample in the elution well was collected and this was the final library sent for sequencing. The library was validated by comparing 2 µL of pre- and post-Pippin samples using an Agilent 4200 TapeStation system.

Tape Station procedure

TapeStation was first prepared by launching the software and loading the machine with tips and the High Sensitivity D5000 ScreenTape device. Then, the ladder was prepared by mixing 2 µL of High Sensitivity D5000 sample buffer and 2 µL of High Sensitivity D5000 ladder in a strip tube. The library samples were prepared next. Two µL of the library, collected following Pippin prep size selection, was sampled and diluted in water five-fold while the pre-Pippin sample was diluted 10-fold. In the same manner, 2 µL of the diluted samples (pre and post-Pippin) were combined with 2 µL of the High Sensitivity D5000 sample buffer in separate tubes. The tubes of ladder and samples were spun down, vortexed using an IKA vortexer at 2000 rpm for 1 min and spun down again. The ladder and the samples were then loaded in the machine. The ladder needs to be placed at the A1 position and the samples were placed at B1 and C1. TapeStation was started and after the run, gel image and electropherograms were generated.

TASSEL-GBS pipeline

SNP calling using the TASSEL-GBS pipeline is described below. It started with creating a database of good quality barcoded reads, i.e. tags, connected to their corresponding samples based from the sequence files (FASTQ). The pooled DNA fragments can be mapped back to the samples based on their barcode sequences. The adapter and barcode sequences are also trimmed from sample sequences. These are based on a key file that identifies the samples based on their barcode and the restriction enzymes. In addition, a user-defined minimum quality score based on counts is used to determine whether to process a particular read or not. For each sequence FASTQ file, the set of unique sequences, i.e. a tag, is identified, their length (bp) and the number of times they were observed are determined. The output per file is then merged into a master list in the database. A user-defined minimum count of tags is used as a quality score. The minimum count controls for the compromise between the lowest minor allele frequency possible and the sequencing error allowed. The more frequent a tag has been observed, the lower the probability of a sequencing error. In the succeeding steps, subsequent filtering and downstream processing will further eliminate a lot of possible sequencing errors. The tags were then aligned to the constructed *L. perenne* reference genome, as described previously. First, the tags in the database are converted to FASTQ format, which is required by the aligner program. Alignment is basically matching the sequences (DNA bases) of the tags to the sequences of the reference genome to determine how and where they are similar. It determines the position of tags in the genome based on the match, allowing for a few mismatches. We have used the alignment program bowtie2. Its output, a SAM file that can be read by TASSEL, was used to update the database to include the position of tags in the genome. In TASSEL, this is called a TagsonPhysicalMap file. When the database has information on tags (with their count) and their position in the genome, polymorphism can be identified. Tags located in the same position in the genome have sequences that are largely identical. Polymorphism is called, as the name implies, by a single nucleotide difference or mismatch. SNP discovery was only performed with tags with starting position that perfectly matched a genomic position beginning with a cut site remnant and at the correct strand. This tag is called TagLocus in TASSEL. The number of times each tag occurred per sample was also determined to form a matrix of tags per taxa (i.e. samples). SNP discovery is then performed by conducting a multiple sequence alignment of all the tags using the CLUSTAL W algorithm (Thompson et al., 1994) to determine possible alleles (nucleotides). SNP genotype per sample was then determined based on a binomial likelihood ratio method of quantitative SNP calling with a specified expected sequencing error rate (i.e. 1%) using the tags per taxa matrix. The quality of SNPs detected was also determined by scoring the markers in coverage, depths and other genotypic statistics for a group of samples and this information were included in the database. Based on the quality scores, the final SNPs are filtered in the database. Finally, SNP production was conducted by utilizing information from the GBS database and the key file to process the raw sequence data (FASTQ files). Multiple SNP genotype was produced for every individual sample based on the results of the SNP discovery pipeline.



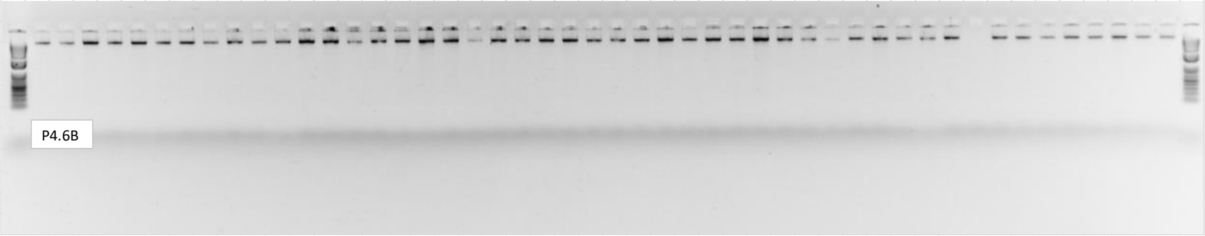
Appendix Figure 3.1. DNA quantity and quality of PGG004-C2 plant samples (i.e. P4.2). Quality was assessed based on the appearance of the bands (i.e. absence of smears, size, intensity, etc.).

	Quantification (ng/µl)											
	15	20	19	17	15	17	11	9	13	11	11	21
1kb+ ladder	P4.6-A7	P4.6-A8	P4.6-B7	P4.6-B8	P4.6-C7	P4.6-C8	P4.6-D7	P4.6-D8	P4.6-E7	P4.6-E8	P4.6-F7	P4.6-F8

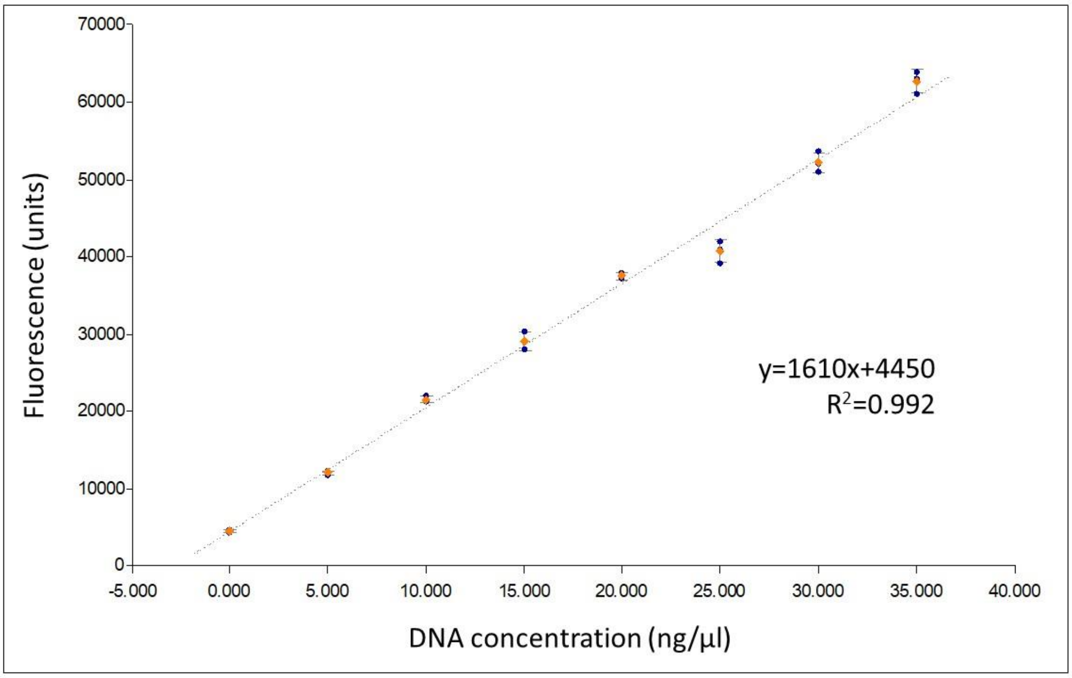
19	13	16	12	18	26	11	12	17	17	15	4
P4.6-G7	P4.6-G8	P4.6-H7	P4.6-H8	P4.6-A9	P4.6-A10	P4.6-B9	P4.6-B10	P4.6-C9	P4.6-C10	P4.6-D9	P4.6-D10

7	15	17	12	15	15	20	14	15	6	15	23
P4.6-E9	P4.6-E10	P4.6-F9	P4.6-F10	P4.6-G9	P4.6-G10	P4.6-H9	P4.6-H10	P4.6-A11	P4.6-A12	P4.6-B11	P4.6-B12

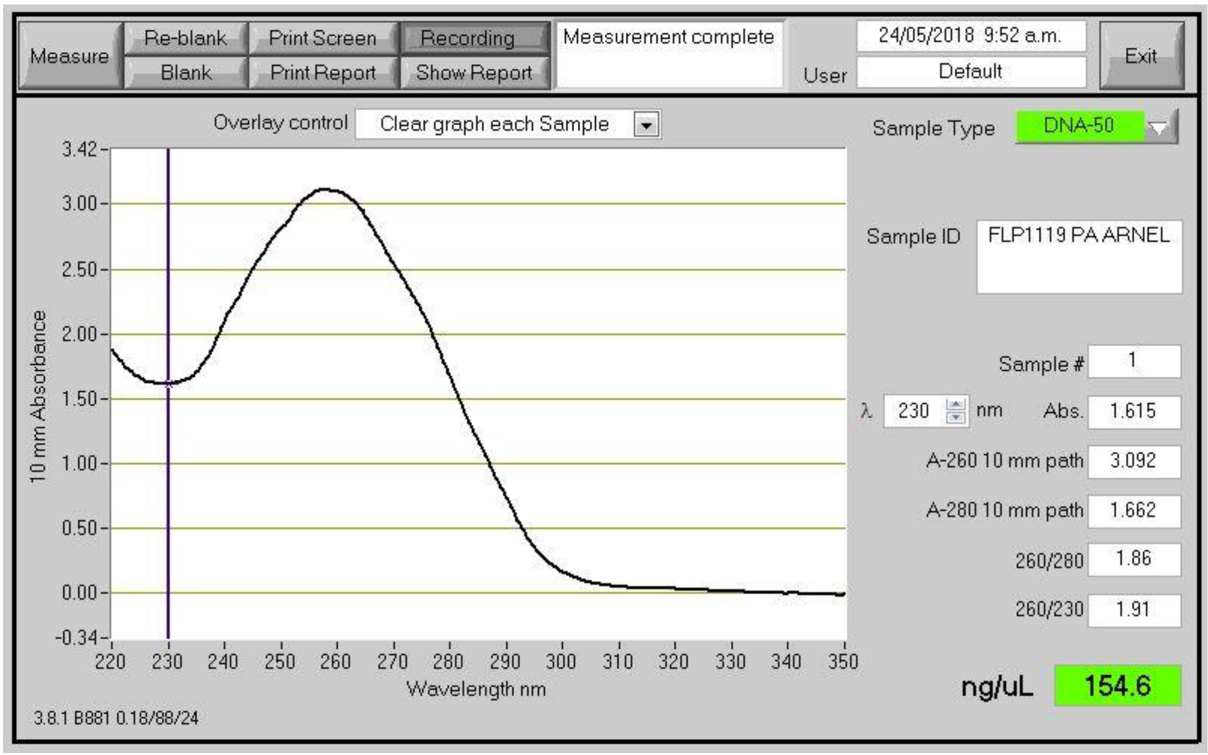
17	35	23	0	19	22	13	17	14	28	12	18	
P4.6-C11	P4.6-C12	P4.6-D11	P4.6-D12	P4.6-E11	P4.6-E12	P4.6-F11	P4.6-F12	P4.6-G11	P4.6-G12	P4.6-H11	P4.6-H12	1kb+ ladder



Appendix Figure 3.2. DNA quantity and quality of PGG004-C6 plant samples (i.e. P4.6). Quality was assessed based on the appearance of the bands (i.e. absence of smears, size, intensity, etc.).



Appendix Figure 3.3. Fluorescence values of the standards for DNA quantification of PGG004-C6 plant samples. There's a strong positive linear relationship between DNA concentration and fluorescence.



Measurement complete

24/05/2018 9:52 a.m.

User Default

Overlay control Clear graph each Sample

Sample Type DNA-50

Sample ID FLP1119 PA ARNEL

Sample # 1

λ 230 nm Abs. 1.615

A-260 10 mm path 3.092

A-280 10 mm path 1.662

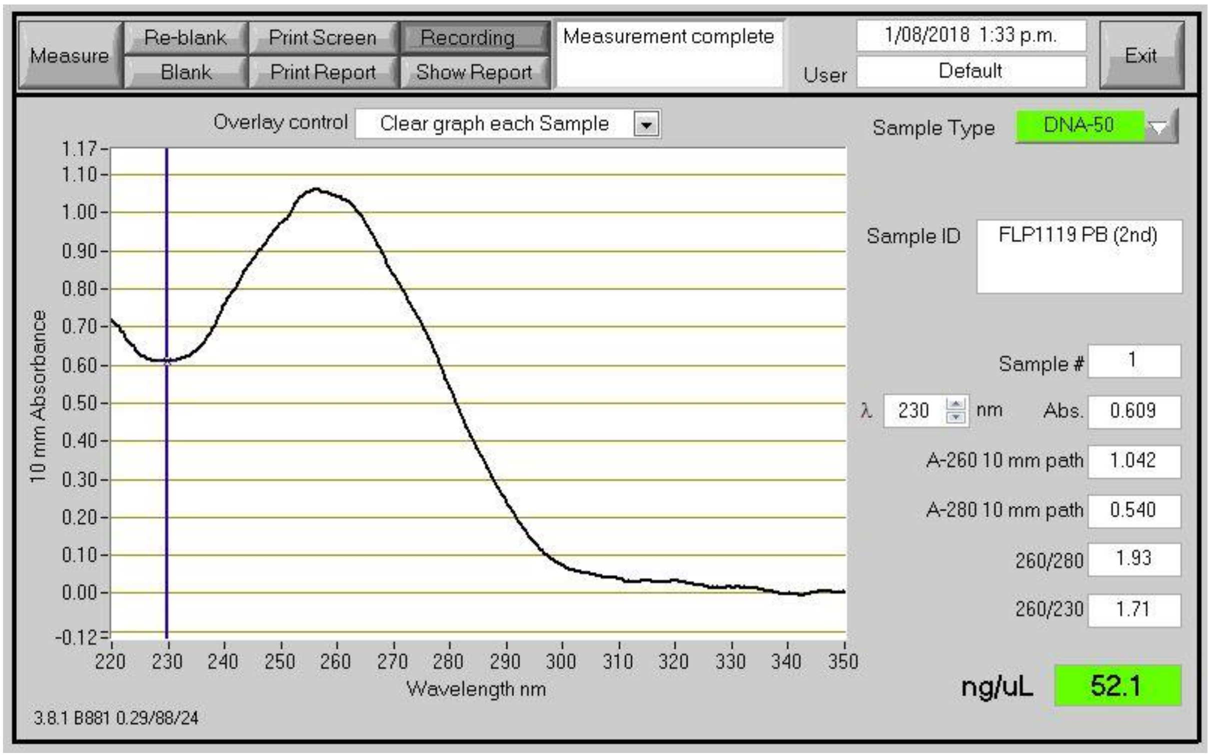
260/280 1.86

260/230 1.91

ng/uL 154.6

3.8.1 B881 0.18/88/24

Appendix Figure 3.4. Quantification and absorbance pattern of the first GBS library.



Measurement complete

1/08/2018 1:33 p.m.

User Default

Overlay control Clear graph each Sample

Sample Type DNA-50

Sample ID FLP1119 PB (2nd)

Sample # 1

λ 230 nm Abs. 0.609

A-260 10 mm path 1.042

A-280 10 mm path 0.540

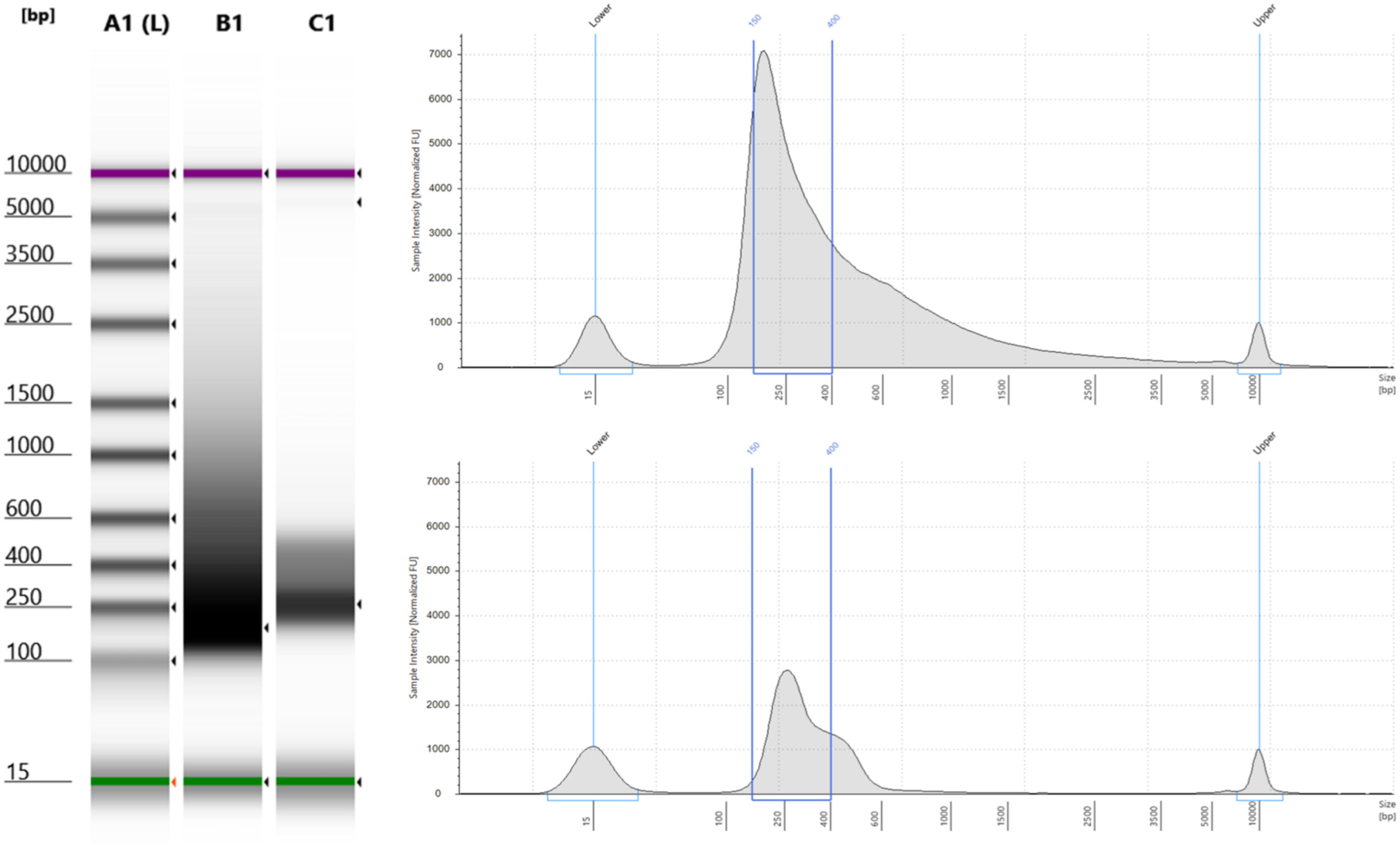
260/280 1.93

260/230 1.71

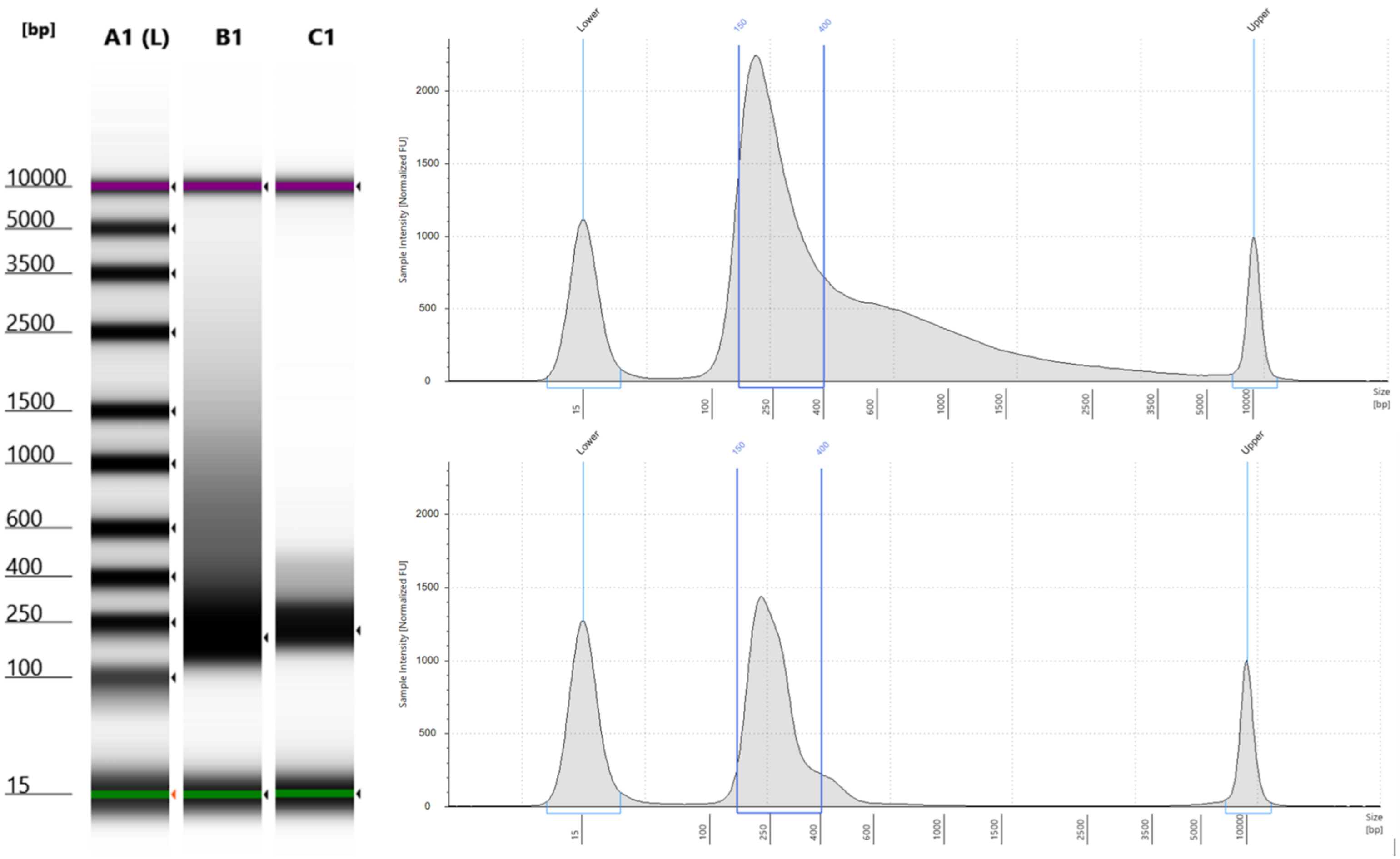
ng/uL 52.1

3.8.1 B881 0.29/88/24

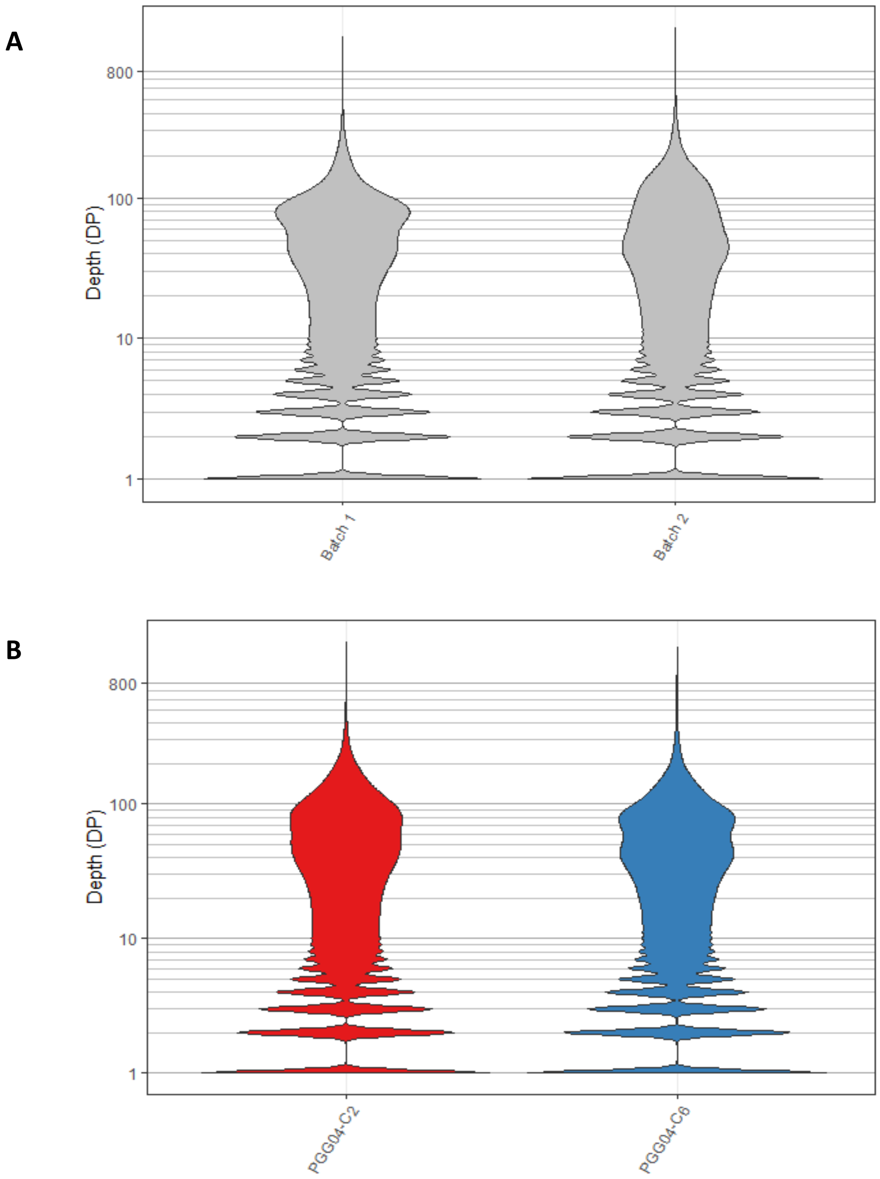
Appendix Figure 3.5. Quantification and absorbance pattern of the second GBS library.



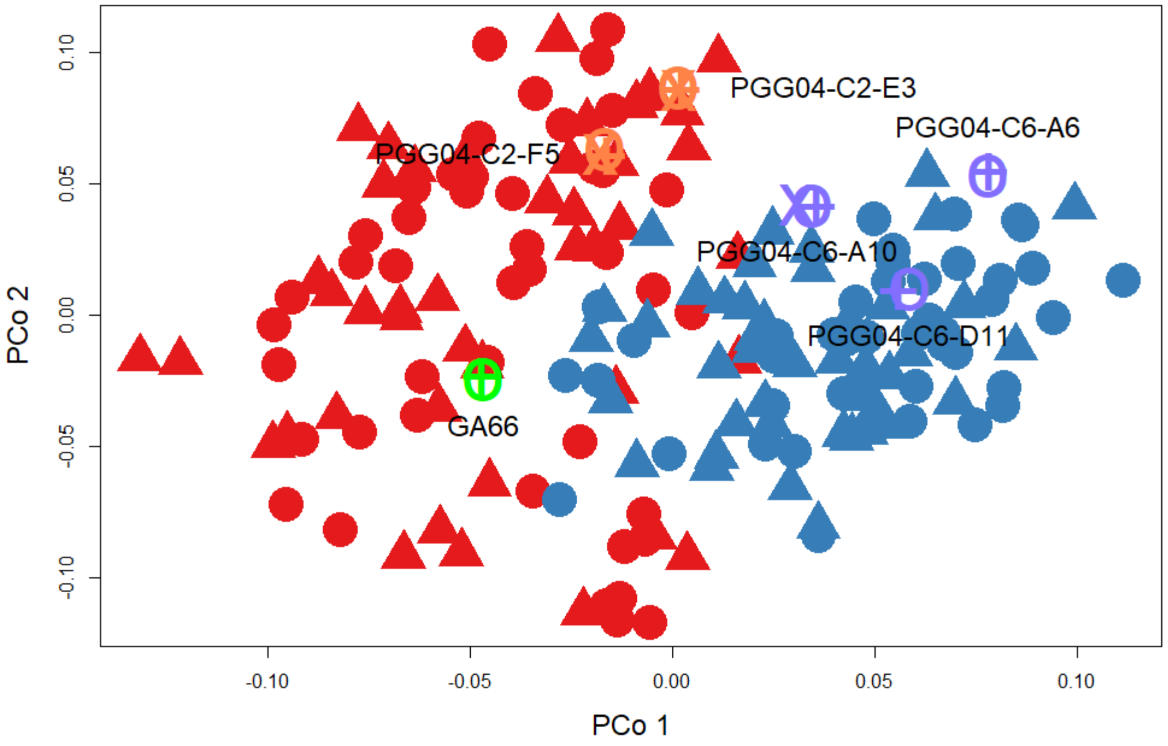
Appendix Figure 3.6. Gel (left) and electropherogram (right) of before (pre-Pippin prep) and after (post-Pippin prep) size selection of the first GBS library. In the gel, A1 is the size ladder, B1 is the pre-Pippin prep sample while C1 is the post-Pippin prep sample. The graph on the right also shows the fragment size distribution before size selection (top) and after (bottom) size selection, which resulted in mostly 100 – 400 bp fragments.



Appendix Figure 3.7. Gel (left) and electropherogram (right) of before (pre-Pippin prep) and after (post-Pippin prep) size selection of the second GBS library. In the gel, A1 is the size ladder, B1 is the pre-Pippin prep sample while C1 is the post-Pippin prep sample. The graph on the left also shows the fragment size distribution before size selection (top) and after (bottom) size selection, which resulted in mostly 100 – 400 bp fragments.

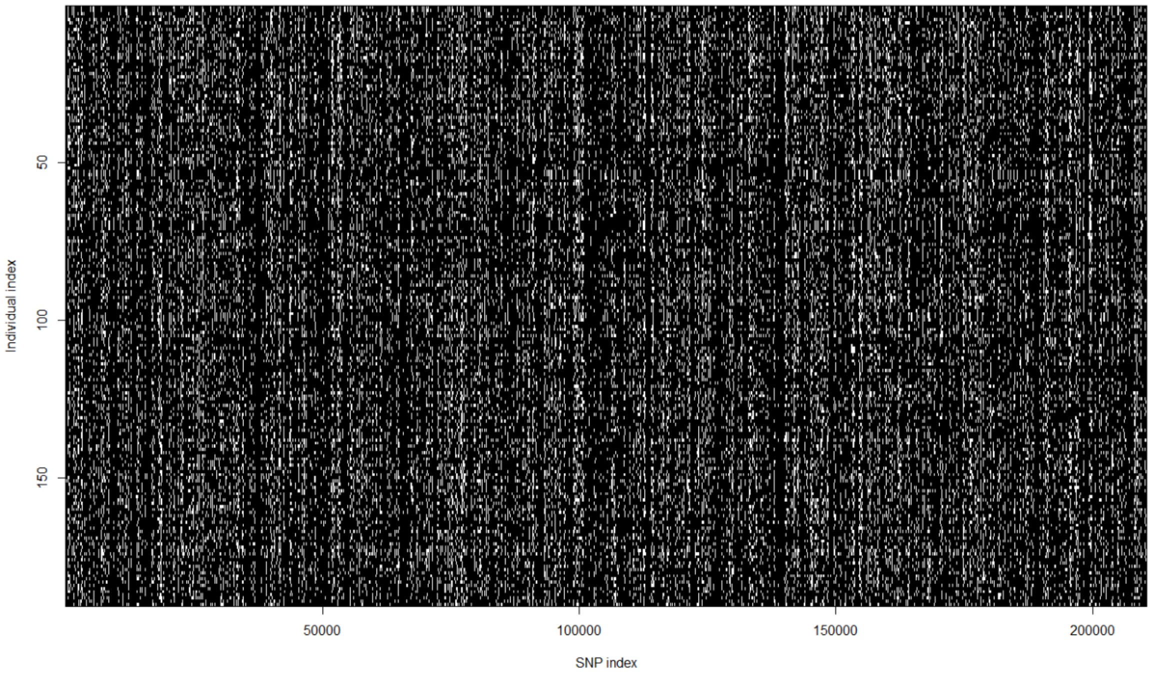


Appendix Figure 3.8. Distribution of read depth comparing two GBS libraries (A) and two generations (B) of PGG04. The shape of distribution differs between the batches of libraries but is more similar between the two generations. This supports the strategy of pooling across generations per library to minimize possible unintended batch effects.

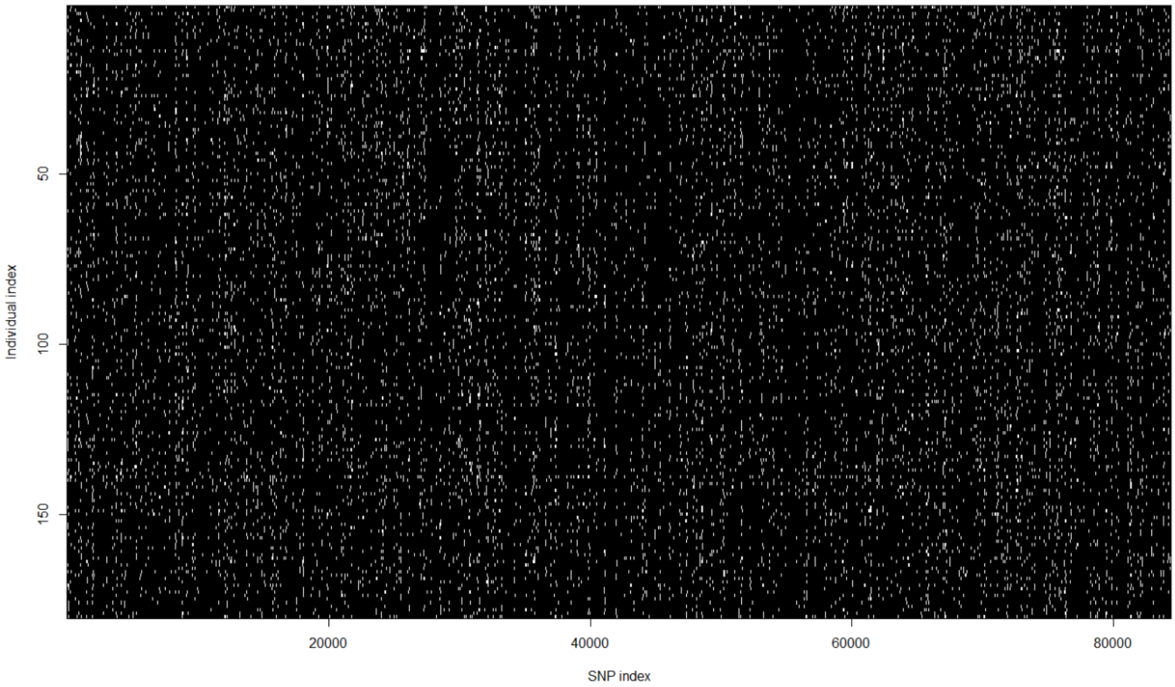


Appendix Figure 3.9. The relationship between early (PGG04-C2; red) and late (PGG04-C6; blue) generations of PGG04 in terms of genetic distance (Provesti's). PCo1 separates the two generations while the two library batches (circle or triangle) are spread randomly across the two axes. Biological replicates ("0", "+", "x") generally clustered very closely especially, the GA66 control.

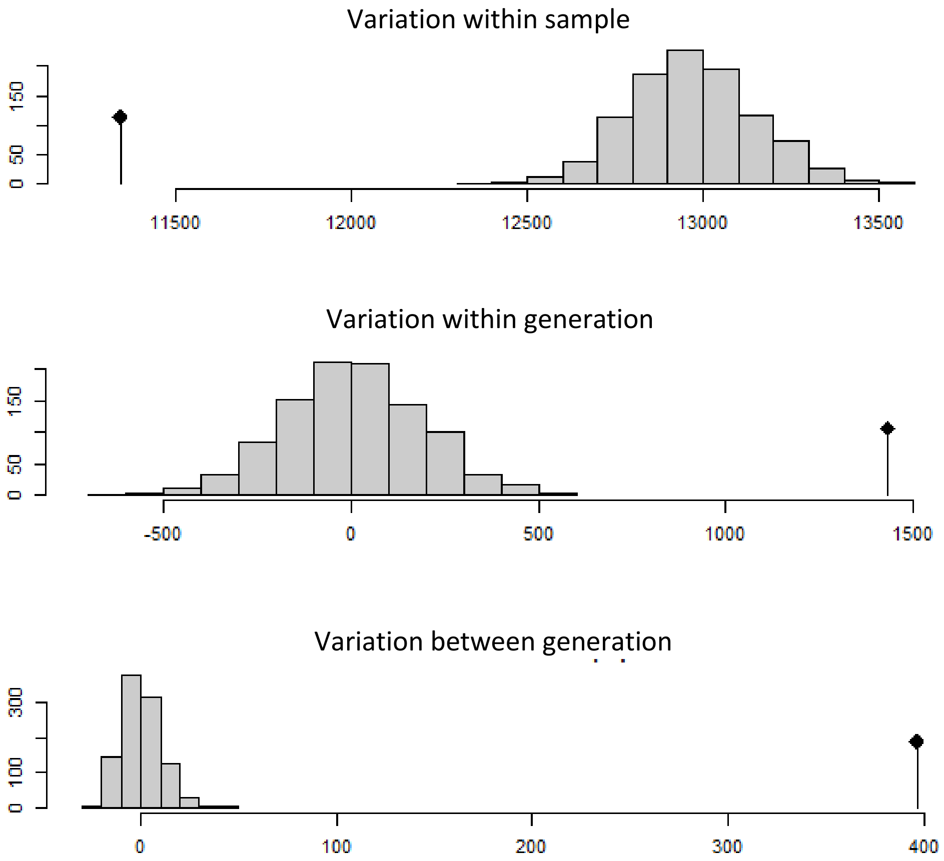
A



B



Appendix Figure 3.10. Plot of missing data (white) before (A) and after (B) filtering. In A, SNPs with a high amount of missing data show white strips running continuously from top to bottom while in B the missing data appears to be relatively random across all SNPs.

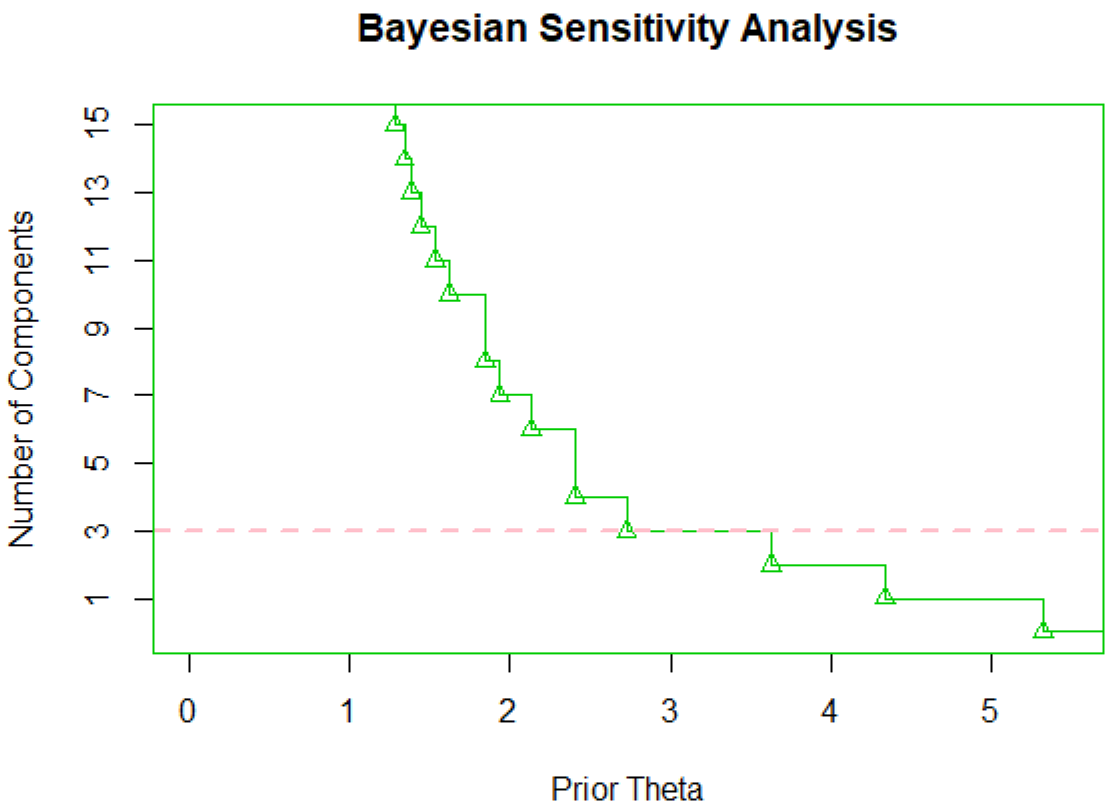


Variation within sample

Variation within generation

Variation between generation

Appendix Figure 3.11. Significance testing in AMOVA as implemented in R/poppr. Null distributions (histograms) are inferred by random permutation and compared to the variance components, that is, the observed data represented by a black line.



Appendix Figure 3.12. The number of important components was verified with Bayesian model selection approach. Posterior estimate of the number of components was plotted as a step function of the prior theta and a large step length corresponds to the optimum number of PCs given a number of prior model probabilities. The maximum estimate or the highest long-step is found at PC = 3, highlighted by red dashed line.