Bioinformatic detection of genetic changes in the fungal endophyte *Epichloë festucae* AR37 during adaptation to a new perennial ryegrass host

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

Doctor of Philosophy (PhD)

in

Microbial Genetics

at Massey University, Palmerston North

New Zealand

Asad Razzaq

2020

ABSTRACT

Mutualistic association with the fungus *Epichloë festucae var lolii* improves the resistance to abiotic stress and herbivory of perennial ryegrass (*Lolium perenne*). Breeders are interested in moving select *E. festucae* strains between ryegrass cultivars. In one such attempt *E. festucae* strain AR37 was transferred from its original ryegrass host to two new ryegrass cultivars. Performance of the resulting novel associations was improved over several years in a breeding program. We wanted to determine if genetic changes in AR37 contributed to this enhanced performance and, if so, identify the nature of these changes. The *Epichloë* endophyte indeed changed during adaptation to both new host cultivars. We demonstrated this by comparing the genome sequence of AR37 in its original host with pooled "AR37 population genomes" from the two novel associations at the end of the breeding program. These comparisons revealed mutations associated with ~ 150 genes. Frequency of mutations in endophytes increased with the number of seed cycles their new host has gone through. A wide variety of genes including those encoding for certain binding proteins e.g. acting binding, zinc ion binding, DNA-binding, and calcium binding as well as genes encoding for proteins that form signal recognition particles and involved in intracellular signal transduction were amongst those affected by mutations. These genes and their products can play an important role in establishing symbiotic association with the host cultivar. These results indicate that an array of endophyte genes may be involved in establishing a successful association with the new host cultivar. I conclude that (i) the *Epichloë* genome undergoes functionally relevant alterations as the endophyte adapts to new cultivars and (ii) monitoring the genes encoding the proteins involved, may facilitate breeding programs aimed at improving the performance of new endophyte ryegrass associations.

ACKNOWLEDGEMENTS

Praise be to Allah, Lord of the worlds, the most Beneficent and the most Merciful, Who is the entire source of knowledge and wisdom endowed to mankind; Who gave me the courage and potential to pursue this goal and Who never spoils any effort of good deeds. Blessings of Allah be upon His Prophet Muhammad (PBUH), the city of knowledge and blessing for entire creations, who has guided His Ummah (nation) to seek knowledge from cradle to grave.

I will like to express my deepest appreciation to my supervisor Dr Jan Schmid for his guidance and assistance throughout these years. Without his intellectual suggestions and guidance, it would have been impossible for me to complete this thesis. He went out of his way to assist and guide me. Even after his retirement, he put a lot of time and effort in helping me to put this thesis together.

I would also like to express my deepest appreciation to my co-supervisor Dr David Wheeler, who affectionately guided me through all the steps related to bioinformatic analyses. His encouragement and support helped me a lot especially while learning basic bioinformatics. I would like to extend my thanks to Dr Paul Dijkwel, who initially helped me as a co-supervisor and after the retirement of Dr Jan Schmid, took over the role of my main supervisor. During last few months, despite his busy schedule, he left no stone unturned to help and guide me to complete this thesis. His valuable feedback helped me a lot to improve this thesis.

Many thanks to everyone at AgResearch who helped me during this project, especially Richard Johnson, who provided me with the AR37-infected ryegrass seeds for this study. Thanks to Anouck de Bonth for providing me with the AR37 culture and for developing my many immunoblots during this study. Thanks to Stuart Card and Wayne Simpson for their informative and motivating talks during my visits to AgResearch, Palmerston North.

I would like to thank Dr Ningxin Zhang for providing me all the assistance during my research. Thanks to everyone at SFS who provided technical help during these years.

I would like to offer my special gratitude to those who have funded this project; Higher Education Commission (HEC) of Pakistan, for funding my studies; Islamia College (Public Sector University) Peshawar, Pakistan for granting me with a paid study leave for

4 years; and Massey University Grants for providing me 5000$ to support last few months during this study.

My special thanks to my family and friends who are not directly involved in this project but without their support it would have been impossible for me to achieve this goal. To my father, for supporting me through every situation in whole my life. His support gave me courage to keep moving in situations when nothing seemed to work right. A special thanks must go to my wife, who had to sacrifice a lot so that I can achieve my goal. Despite the traumatic death of my younger daughter at the tender age of 3 months, my wife supported my decision to return to New Zealand to complete my PhD while she stayed back in Pakistan and had to cope with the trauma without me. A big thanks to my lovely daughter, Umama, who brings me to smile during the tough times.

Asad Razzaq

ABBREVIATIONS

A	Adenine
AR37-Orig	*Epichloe festucae* strain AR37 isolated from original host
AR37-SAM	*Epichloe festucae* strain AR37 isolated from Samson cultivar
AR37-KLP	*Epichloe festucae* strain AR37 isolated from KLP1102 cultivar
BLAST	Basic local alignment search tool
BLASTn	Nucleotide database search using a nucleotide query
BLASTp	protein database search using a protein query
BLASTx	protein database search using a translated nucleotide query
bp	Base pair(s)
BS	Blocking solution
BUSCO	Benchmarking Universal Single-Copy Ortholog assessment tool
BWA	Burrows Wheeler alignment
C	Cytosine
^{o}C	Degree Celsius
CDS	Coding sequence
Chr	Chromosome
Cl	Chlorine
cm	Centimetre
CNV(s)	Copy number variation(s)
CTAB	Cetyl trimethyl ammonium bromide
cv	Cultivar
DNA	Deoxyribonucleic acid
dNTP(s)	Deoxyribonucleotide triphosphate(s)
DP	Read depth
dsDNA	Double stranded deoxyribose nucleic acid
EAS	Ergot alkaloids
EDTA	Ethylene diamine tetra-acetic acid
ELISA	Enzyme linked immune sorbent assay
EtBr	Ethidium bromide
G	Guanine
g	Gram(s)
x g	Gravitational force used in centrifuge

GAT	Genome alignment toolkit
Gb	Gigabase (s), billion bases
g/L	Grams per litre
h	Hour(s)
HCl	Hydrochloric acid
HiSeq	A sequencing platform developed by Illumina
IDT	Indole diterpene alkaloid
Idt	Indole diterpene gene
Ig	Immunoglobulins
Indel	Insertion and / or deletions
jan	Janthitrem gene
kb	Kilo-bases
kbp	Kilo-basepairs
KLP	KLP1102 cultivar of perennial ryegrass
LOL	Loline alkaloid
ltm	Lolitrem gene
M	Molar
MAPQ	Mapping quality
Mb/MB	Megabase(s)
ME	Mecaptoethanol
mg	Milligram
min	Minute
MiSeq	A sequencing platform developed by Illumina
ml	Millilitres
mm	Millimeters
mM	Millimolar
MQ	Deionised water
Na	Sodium
NCBI	National center for biotechnology information
NCM	Nitrocellulose membrane
NGS	Next generation sequencing
ng	Nanograms
nt	Nucleotide
NZ	New Zealand

ORFs	Open reading frame(s)
PCR	Polymerase chain reaction
PDA	Potato dextrose agar
PDB	Potato dextrose broth
PDBG	Paired De Bruijn graph
PER	Peramine alkaloid
perA	Peramine gene
RAM	Random access memory
RNase	Ribonuclease
RO	Reverse osmosis
RPL	Reads placed left
rpm	Revolutions per minute
RPR	Reads placed right
SAM	Grassland Samson cultivar of perennial ryegrass
SAF	Number of alternate observations on forward strand
SAR	Number of alternate observations on reverse strand
SDS	Sodium dodecyl sulfate
Sec	Seconds
SNP(s)	Single nucleotide polymorphism(s)
SNV(s)	Single nucleotide variation(s)
SV(s)	Structural variation(s)
T	Thymine
TAE	Tris-acetic acid-EDTA
TBE	Tris-boric acid-EDTA
tBLASTn	Translated nucleotide database search using a protein query
TE	Tris EDTA
TNE	Tris-NaCl-EDTA
Tris	Tris (hydroxymethyl)aminomethane
UV	Ultra-violet
µg	Micrograms
µl	Micro-liter
µM	Micro-molar
V	Volts
VCF	Variant call format

w/v	Weight / volume ratio
2n	Diploid
%	Percent

TABLE OF CONTENTS

LIST FO FIGURES

LIST OF TABLES

List of Appendices

1. INTRODUCTION

This thesis describes the results of a bioinformatics analysis of changes in the genomes of populations of a fungal endophyte as it adapts to new grass hosts. I will therefore in the following first introduce the fungus and its hosts and then provide background on bioinformatic approaches applicable to detecting genomic changes in populations.

1.1. *Epichloë*-endophyte symbiosis

Seven genera of tribe Balanseiae of family Clavicipitaceae of fungi are known to establish systemic and long-term associations with some agriculturally important grasses (White & Reddy, 1998). One of these 7 genera is *Epichloë* (formerly *Acremonium* / *Neotyphodium*) which includes the best known endophytes forming such associations with around 80 genera of pooid grasses (Leuchtmann *et al.*, 1994, Clay, 1988, Leuchtmann, 1992, Saikkonen *et al.*, 2000). Before 1993, *Epichloë* spp. in symbioses with grasses were all considered as *E. typhina* but since then advancements in molecular phylogenetic analysis has led to recognition of nine separate *Epichloë* species. Imperfect or anamorphic stage of *Epichloë* spp. that lack sexual reproduction and only reproduce via asexual means had previously been accommodated in separate genus i.e. *Neotyphodium* (Glenn *et al.*, 1996) but they are now all classified as *Epichloë* (Leuchtmann *et al.*, 2014). The nature of relationship between *Epichloë* spp. and their hosts is a subject of ongoing debate (reviewed in (Saikkonen *et al.*, 2010a, Saikkonen *et al.*, 1998)). The effects of these associations on host vary with the stage of the host development, ecological circumstances and the environment (Schardl, 2001). It has been proposed that asexual *Epichloë* endophytes and their grass hosts are in a defensive mutualistic relationship where fungi produce alkaloid to defend their hosts against insects and herbivory in order to protect their own source (Clay, 1988) but it is debated by others (Ahlholm *et al.*, 2002, Faeth & Sullivan, 2003, Faeth, 2002).

During vegetative growth of the plant, the fungus grows systemically inside the plant host, and the host provides the endophyte with shelter, food (Saikkonen *et al.*, 2004) and means of dissemination. In return endophyte infection provides the host with a number of fitness enhancements such as resistance to insect and mammalian herbivores (Schardl,

2001, Tanaka *et al.*, 2005, Clay *et al.*, 1993, Shiba & Sugawara, 2005, Wilkinson *et al.*, 2000), disease (Niones & Takemoto, 2014, Christensen, 1996) and tolerance to drought (Hahn *et al.*, 2008, Malinowski & Belesky, 2000) and poor soil conditions (Kuldau & Bacon, 2008). Also, the presence of *Epichloë* endophytes has been associated with increased persistence against adverse conditions such as mineral imbalance (Lyons *et al.*, 1990) and soil acidity (Belesky & Fedders, 1995), and increased plant vigour and growth (Malinowski & Belesky, 2000). Plants infected with *Epichloë* endophytes have been reported to perform better than uninfected plants when planted together and shown have better root growth and higher rates of photosynthesis at high temperature and water stress (Prestidge & Gallagher, 1988, Crawford *et al.*, 2010, Saikkonen *et al.*, 2013, Bacon, 1993).

Some of these benefits have been attributed to the presence of various bioprotective secondary metabolites, commonly known as alkaloids, produced by the *Epichloë* endophytes. Four most widely studied secondary metabolites produced by *Epichloë* endophytes are indole-diterpenes (IDT), loline alkaloids (LOL), ergot alkaloids (EAS) and peramine (PER) (Fleetwood *et al.*, 2007, Tanaka *et al.*, 2005, Spiering *et al.*, 2005b, Schardl *et al.*, 2013a, Schardl *et al.*, 2013c, Panaccione, 2005). Peramine provides resistance to feeding by some invertebrate herbivores like Argentine stem weevil (Rowan *et al.*, 1994). Ergovalines are thought to provide resistance to ryegrass hosts from adult black beetle (Ball & Prestidge, 1993) and nematodes (Bacetty *et al.*, 2009). Lolines deter feeding by insects (Schardl *et al.*, 2007) and indole-diterpenes provide resistant to mammalian herbivory (Gallagher *et al.*, 1981) and reduce the development of Argentine stem weevil larvae (Prestidge & Gallagher, 1988). A large number of different genes are required to synthesize each of these products (Fleetwood *et al.*, 2007, Schardl *et al.*, 2013b, Schardl *et al.*, 2013a, Spiering *et al.*, 2005b, Saikia *et al.*, 2012) except for peramine which is synthesized by only one gene (Tanaka *et al.*, 2005).

Benefits to the host make *Epichloë* endophytes important from an agricultural perspective, as they increase the persistence of pasture grasses by increasing drought and herbivore resistance, crucial for the pastoral industry in New Zealand, USA and Australia (Young *et al.*, 2013). Their importance is likely to increase globally as droughts continue to increase in severity and frequency throughout the world due to climate change (Strzepek *et al.*, 2010, Kundzewicz *et al.*, 2008). However these desirable characteristics

of endophyte infection come at a cost as some of the secondary metabolites produced by *Epichloë* endophytes are toxic to cattle and sheep. Lolitrem B, an indole-diterpene, is a potent tremorgen and can cause ryegrass staggers in vertebrates (Fletcher and Sutherland, 2009). Ryegrass stagger affects nervous system of the livestock grazing on ryegrass pastures infected with toxic endophyte strains (Gallagher *et al.*, 1981, Gallagher *et al.*, 1982). Affected animals have muscular spasms accompanied by poor coordination of movement. It may result in poor weight gain and reduced milk production in dairy cows (Fletcher, 1982, Fletcher, 1983). Another alkaloid, ergovaline, although beneficial in terms of providing resistance to adult black beetle, causes heat stress in grazing livestock (Fletcher, 1993), significantly reduces milk production by dairy cows (Lean, 2001) and may be associated with impaired immune function, lameness and reduced reproduction rates in livestock (Stuedemann & Hoveland, 1988, Mostrom & Jacobsen, 2011).

Figure 1.1. Life cycles for *Epichloë festucae*. In the asexual cycle (left) the fungus is vertically transmitted, whereas in the sexual cycle (right) it is horizontally transmitted (Copied from Clay & Schardl, 2002).

Some *Epichloë* spp. are asexual and completely dependent on their host plants for dissemination through seeds or vegetative structures produced by the host (vertical transmission) (Moon *et al.*, 2004, Johnson *et al.*, 2003). Sexually reproducing *Epichloë* spp. in their vegetative state behave like asexual *Epichloë* spp. but they produce sexual spores, that can infect new hosts, inside structures called stromata during the sexual stage of their life cycle. These stromata are produced on and around host inflorescences, thus destroying them, a phenomenon known as choke (Bryant *et al.*, 2009) (Fig. 1.1).

1.1.1. Functional symbioses depend on compatibility between host and endophyte genotypes

Growth of *Epichloë* endophytes is tightly regulated within their natural host grasses. When new tiller buds originate from the meristematic zone located at the base of the tiller, they are colonized by the endophyte (Christensen & Voisey, 2007, Christensen *et al.*, 2000b). Growth of endophytes is synchronous with the growth of host grass leaves (Christensen *et al.*, 2002, Tan *et al.*, 2001). Hyphae grow in parallel to the longitudinal axis of the leaf and are mostly unbranched (Christensen *et al.*, 2000b). The average number of hyphae in a mature leaf remains constant (Christensen *et al.*, 2000b, Tan *et al.*, 2001, Christensen *et al.*, 2002). Hyphae within mature leaves remain metabolically active (Schmid & Christensen, 1999). Vascular tissues are usually free of endophytes (Christensen & Voisey, 2007, Christensen *et al.*, 1997). When the leaf starts to die, the symbiotic endophytes do not become saprotrophic to and do not degrade host tissue (Christensen & Voisey, 2007). This synchronized growth between the host and the fungal endophyte indicates signalling pathways between the two (Tan *et al.*, 2001, Tanaka *et al.*, 2012). Infection does not trigger a pronounced host defence response (Torres *et al.*, 2012, Zhang *et al.*, 2011a, Christensen *et al.*, 2002). The restriction of fungal growth to the intercellular spaces of the host grass may be one of the reasons why pronounced host defences are not observed (Christensen *et al.*, 2002).

The plant host may be playing a major role in controlling the fungal growth (Deitsch *et al.*, 2009, Torres *et al.*, 2012). Tissue age and tissue type all influence endophyte levels, and individual plant genotype determines the level of biomass of a given endophyte strains of the same host cultivar (Spiering *et al.*, 2005a, Herd *et al.*, 1997). Likewise the same strain will reach different levels of biomass in different cultivars of the same host

species (Christensen & Voisey, 2007, Christensen *et al.*, 1997) or in different host species (Christensen *et al.*, 1997). Concentrations of fungal alkaloids are also reported to vary when the same endophyte is present in different hosts, (Spiering *et al.*, 2005a, Agee & Hill, 1994) and the surrounding tissue type has a strong influence on endophyte gene expression (Schmid *et al.*, 2017).

Thus, genetic compatibility between host and endophyte seems to be a must for establishing a successful interaction. Further supporting the idea that complex cross-talk occurs between the host and its *Epichloë* endophyte comes from observations that these endophytes show host specificity in that they can usually only inhabit a particular host or closely related genera of host grasses, (Bryant *et al.*, 2009, Leuchtmann & Clay, 1993). The latter indicates that such cross-talk may have evolved over long periods of time, also suggested by *Epichloë* / grass co-cladogenesis; i.e. there are particular endophyte clades that are always associated with particular hosts.

1.1.2. Endophyte-grass combinations with mismatched genotypes fail to establish a functional symbiosis

Although the vast majority of *Epichloë* spp. show host specificity, it has been possible to generate novel associations of *Epichloë* spp. with closely related host genera by artificial inoculation (Leuchtmann & Clay, 1993, Siegel *et al.*, 1990, Christensen, 1995). In most cases, however, novel endophyte-host associations show incompatibility. This may manifest itself in the form of premature hyphal death (Koga *et al.*, 1993), or necrosis of host tissue (Christensen, 1995). In a novel associations between *N. coenophialum* and *Lolium perenne*, an electron-dense area between host cell wall and hyphae appeared which may be representing host defence response (Koga *et al.*, 1993). This response may reduce vigour, vacuolation, and even death of the fungal hyphae in artificial associations (Koga *et al.*, 1993). In some synthetic associations incompatibility may not be obvious at a cellular level but host growth may be affected (Christensen *et al.*, 1997). In other cases e.g. in an association of an asexual interspecific hybrid i.e. *Neotyphodium* spp. LpTG-2, with a tall fescue, apical meristem seems to die away slowly (Christensen, 1995). Unregulated endophyte growth can be another consequence of incompatibility (Tanaka *et al.*, 2007, Scott *et al.*, 2007).

Work on novel associations has demonstrated that incompatibility and compatibility are, at least in part, heritable *Epichloë* traits. This was shown by Christensen and co-workers. They demonstrated that compatibility of an *Epichloë* strain can predict the compatibility of its progeny (Christensen *et al.*, 1997). The concept of a genetic basis of compatibility, and that it is determined by multiple loci, is also supported by experiments in which two *Epichloë* spp., adapted to different host species, were mated: The progeny that was less compatible than the parents with either of the original hosts (Chung *et al.*, 1997).

1.1.3. *Epichloë* endophytes are constantly confronted with new host genotypes

According to red queen hypothesis, organisms need to continuously evolve in order to adapt to their changing environment (Brockhurst *et al.*, 2014). Evolution of sex may be a mechanism to provide organisms with a variety of alleles at all times in order to compete in a changing environment. Perennial ryegrass, the *Epichloë* spp. host used in my research, only reproduces through outbreeding (Cunningham *et al.*, 1994). Due to this, each ryegrass progeny is genetically different from either of its parents. According to Red Queen hypothesis, sexual recombinations may result in host progeny with an unusual genotype that makes it resistant to infection by parasites adapted to a common host genotype (Clay & Kover, 1996). Sexual reproduction in hosts thus makes them a moving target, that is difficult to hit for parasites (Clay & Kover, 1996). Same may apply to *Epichloë*- ryegrass associations. Sex could thus result in a grass host genotype that is no longer compatible with its native endophyte. This asymmetric interaction may result in loss of endophytes from out-breeding host populations, unless counterbalanced by selective advantages of infected over uninfected plants.

This continuous change in genotype of the grasses may result in an arms race between *Epichloë* endophyte and ryegrass hosts (Saikkonen *et al.*, 2004). There are studies demonstrating that endophytes do indeed get lost from the host plant populations over time, and loss of compatibility brought about by sexual reproduction of the host might be the cause (Saikkonen *et al.*, 2010b). Some systemic fungal parasites are shown to castrate their host plants to stop outbreeding and reduce the genetic variability in the host progeny (Clay & Kover, 1996). It has also been suggested that, as a "countermeasure" *Epichloë* endophytes may try to keep the genotype of their host constant by increasing their clonal propagation (Pan & Clay, 2002) or the rates of self-pollination (Berry *et al.*, 2015, Meijer

& Leuchtmann, 2001). Plant genotypes with faster clonal growth may have a selective advantage over other competing genotypes resulting in their dominance in a population (McLellan, 1997). However there is no strong evidence in support of this (Saikkonen *et al.*, 2004).

In some ways, sexually reproducing endophytes may face an even greater challenge than asexual endophytes in maintaining compatibility with the host. Not only are their spores faced with outbreeding host progeny whose genetic material is somewhat different than either of its parents. The spores themselves, resulting from mating between strains growing on different host plants, also represent genotypes different than either of the mating strains.

Thus there is a dynamic relationship between the endophytes and their host because new genotypes of both are regularly formed (Saikkonen *et al.*, 1998). What might be the mechanism(s) to bring necessary changes in endophytes genotype to remain adapted to their ever changing grass hosts is not fully known (Saikkonen *et al.*, 2010b). However that the endophytes can adapt is demonstrated by the stability of *Epichloë* spp./ host associations over time: the co-cladogenesis of *Epichloë* spp. and pooid grasses (Schardl *et al.*, 1997) over a period of approximately 40 million years (Schardl *et al.*, 2004). It must be added that the constant need for maintaining coadaptation also has positive consequences, and that coadaptation is not necessarily a one-way street, in that the host specie is also under selective pressure to maintain or improve a beneficial association with its *Epichloë* spp.. For example the diversification of loline alkaloids in endophyte-grass associations is a "collaborative effort" involving enzymes from both symbionts (Pan *et al.*, 2014).

1.1.3.1. Selection of host-adaptive mutations in asexual *Epichloë* spp.

Mutations arise in endophyte hyphae as they extend while colonizing the plant tissue. Mutations are a suitable mechanism for long-term adaptation as they build gradual variations in genome upon which natural selection may operate over long period of time. Sexual endophytes can combine beneficial mutations by recombination to generate genotypes highly compatible with a host genotype, in some case even allowing an *Epichloë* spp. to exist on different host species. Some sexually reproducing *Epichloë* spp.

are capable of forming associations with hosts from different plant families; *E. typhina* for example is capable of infecting a variety of distantly related grasses from different host families including the rush, Juncus effuses (Kilpatrick, 1961).

The object of my thesis, however, is an asexual *Epichloë* spp. and the genetic changes associated with its adaptation (in an outcrossing host species). Asexual *Epichloë* spp. are usually (see below) unable to combine beneficial mutations due to lack of recombination, but are nevertheless well-adapted to their particular host species. In the following I will discuss some of the processes likely to be involved in the selection of adaptive mutations in asexual *Epichloë* spp..

It should be noted that being asexual is not necessarily a disadvantage for an *Epichloë* endophyte. While asexual species are short-lived by evolutionary standards, they can be very successful in the short run (short run again by evolutionary time standards), as evidenced by the frequency with which such species arise, and not only among the *Epichloë* spp. (Rice, 2002). Indeed lack of sex can assist in generating and maintaining genotypes that allow an organism to be broadly adapted to a multitude of environments (Massicotte & Angers, 2012), such as, in the case of *Epichloë* spp., the constantly changing genotypes of an outcrossing host grass.

A mutation will initially arise in only one nucleus of the mycelium of an asexual *Epichloë* endophyte. Given that its mycelium is multinucleate, there are millions of nuclei in the mycelium colonizing a tiller (Tan *et al.*, 2001) and the mycelium will indeed contain a number of different mutations. As a mutation-carrying nucleus divides the mutation will be passed on to all of its daughter nuclei in the hypha in which it has arisen. Nevertheless, given the large number of hyphae in a tiller, the mutation, unless it has arisen very early in the development of the tiller when the mycelium was still small, is unlikely to significantly increase the fitness of the mycelium overall or of the symbiosis; overall fitness will be determined by the sum of all the mutations across all the nuclei (Christensen *et al.*, 2000b).

Mutations are more likely to be selected for or against by clonal segregation in new host tillers. Any new emerging plant host tiller seems to be colonized by hyphae that originated from a single nucleus in the maternal tiller (Schmid & Christensen, 1999). Likewise

although seed is colonized by large number of hyphae but only few hyphae seem to penetrate and infect the embryo sac and thus participate in infecting and colonizing new seedlings emerging from the infected seeds (Majewska-Sawka & Nakashima, 2004). All hyphae present between the aleurone layer and the seed coat and between the scutellum and the endosperm degrade progressively and do not infect new seedlings (Philipson & Christey, 1986). If a mutation-carrying hypha thus colonizes the meristem forming a new tiller or a seed, this will lead to the formation of tillers in which all or the majority of the endophyte mycelium contains the mutation. Selection can then further increase the frequency of the mutation if it increases the fitness of these tillers.

Aside from selection on the basis of increasing host fitness, the small number of hyphae in seed would also allow for selection of mutations that only enhance the fitness of the endophyte hyphae carrying them, such as mutations that would allow a hypha to more effectively colonize daughter tiller-forming meristems or seed, or increase its survival rate in seed during storage (Hume *et al.*, 2011).

Some advantageous mutations may be generally advantageous, others only in the context of a specific host genetic background. Both types of mutations could increase in frequency by these mechanisms - the latter however only if the benefits in some host genotypes are not outweighed by negative fitness effects in other genotypes. Without sexual recombination, these beneficial mutations can however not be combined and different clones carrying different mutations will compete with each other for selection, a phenomenon known as clonal interference (Gerrish & Lenski, 1998). This can result in a diverse array of endophyte lineages in both natural and artificial associations (Schmid & Christensen, 1999).

Hybridization between asexual endophytes, colonizing the same host, may be one mechanism that allows combining beneficial mutations present in different clones. Hyphae from two different endophytes strains have been shown to colonize the same host (Christensen *et al.*, 2000a), and on rare occasions the same tiller. In vitro studies suggest that while the meeting of such hyphae can generate genotypes combining their genetic markers, the process involved is not sexual recombination (Schardl & Leuchtmann, 1999, Leuchtmann *et al.*, 1994, Leuchtmann & Schardl, 1998). Rather a heterokaryon (an individual with two or more different nuclei (or genotypes) is formed, followed by

karyogamy resulting in polyploid nuclei. It has been proposed that hybridization between two asexual endophytes may occur immediately after they coinfect an "unusual' or novel host. The hybrid may be at selective advantage to either of maladapted parental strains and exclude them by competition (Selosse & Schardl, 2007). The genomes of numerous asexual endophytes show signs of past hybridization events (Schardl *et al.*, 2009, Moon *et al.*, 2000, Moon *et al.*, 2002, Tsai *et al.*, 1994). Analyses of beta-tubulin (*tub2*), actin (*act1*) and transcription elongation factor 1-alpha (*tef1*) genes has revealed that multiple copies of these genes with different phylogenetic origins were present in most asexual *Epichloë* endophytes (Craven *et al.*, 2001, Moon *et al.*, 2004, Gentile *et al.*, 2005). Two of these studies found that out of 59 endophytes isolated, 44 were interspecific hybrids (Gentile *et al.*, 2005, Moon *et al.*, 2004).

A potential additional advantage of hybridization is that it temporarily increases the number of alleles which can be utilized for adaptation, since for many genes multiple copies persist long after the hybridization event (Johnson *et al.*, 2003, Clay & Schardl, 2002) and duplicated copies of genes are free to mutate and undergo process of natural selection to evolve new functions while original copy of the gene may keep on performing normal function (Ohno, 1970). For example, in *Epichloë festucae* var. *lolii* four copies of cytochrome P450 monooxygenase and multiple copies of genes involved in indole-diterpene biosynthesis are found and believed to give it a selective advantage (Saikia *et al.*, 2008). Advantages of multiple copies of the same gene for adaptation to new hosts is also suggested by the finding that some hybrid endophytes like *Epichloë tembladerae* and *Epichloë occultans* occur in an unusually broad range of pooid grasses, latter forming associations even with many distantly related grasses (Moon *et al.*, 2004).

1.1.4. *Epichloë festucae*, a model for studying mechanism of *Epichloë* adaptation to the host

My study was aimed at determining how *Epichloë* spp. changes as they adapt to new hosts. From among the multitude of *Epichloë* / grass associations an association between an asexual *E. festucae* strain and *Lolium perenne* association was chosen for this study for a number of reasons. The first is that *E. festucae* has been used extensively as a model of endophyte-host interactions (Clay & Schardl, 2002, Eaton *et al.*, 2010, Schardl, 2001, Schardl *et al.*, 2004). It is the most intensely studied *Epichloë* spp. and is mostly

transmitted vertically. An important outcome of these efforts is the availability of genome sequences of two *E. festucae* strains, the best curated among the 15 different *Epichloë* spp. genomes (Schardl *et al.*, 2013b). Also during the course of this study, an ungapped chromosome level assembly of *E. festucae* Fl1 was made available (Winter *et al.*, 2018). This resource would facilitate interpretation of my data.

The second reason is the availability of novel, synthetic associations of *E. festucae* strains with different hosts, and the agronomic importance to New Zealand of such novel associations, to overcome the toxicity of natural associations that were originally introduced into New Zealand (NZ) (Milne, 2007). The predominant pasture grass in NZ is *Lolium perenne* (perennial ryegrass), a cool season grass belonging to family Poaceae. It originated in North Africa, Europe and temperate Asia (Jensen *et al.*, 2001). English settlers brought it from Europe in 1880 into New Zealand where it is used as an important pasture grass since 1930s, sometimes in combination with other pasture grasses or white clover (Charlton & Stewart, 1999, Minnee, 2011, Thom *et al.*, 1998). It establishes quite quickly, is tolerant to grazing, yields herbage of good quality and can grow in soils with a wide range of fertilities (Stewart & Charlton, 2006, Young *et al.*, 2013, Wilkins, 1991). It is naturally diploid but a number of different cultivars with different ploidy levels have now been made available to farmers in New Zealand. In some of the cultivars flowering is delayed than usual to avoid decrease in forage during peak lactation (Woodfield & Easton, 2004).

E. festucae endophytes appear beneficial to survival of ryegrass pastures in New Zealand. When British settlers brought perennial ryegrass seeds to New Zealand (Stewart & Charlton, 2006), the prolonged storage of ryegrass seeds may have caused the loss of endophytes from majority of the seeds, while only a small number of ancestral *E. festucae* strains may have remained viable in ryegrass seeds (Simpson *et al.*, 2012). All current *E. festucae* strains present in NZ pasture grasses may have descended from these small number of ancestral strains. These ancestral *E. festucae* strains may have provided selective advantage to their host due to toxic alkaloids produced by these endophytes. Additional support for this idea, came from the observation that *E. festucae* genotypes predominant in NZ were indistinguishable in a genotype test using eight microsatellite markers (Simpson *et al.*, 2012). Because of negative effects of the endophyte-derived alkaloids on livestock, attempts were made to remove *E. festucae* endophytes from

ryegrass. However, the endophyte-free pastures were found to be more susceptible to Argentine stem weevil (ASW) (Mortimer & Di Menna, 1983, Prestidge, 1982) and later it was shown that wild type *E. festucae* endophyte provided the ryegrass host with resistance to pasture mealybug (Pennell *et al.*, 2005), adult black beetle (Ball & Prestidge, 1993) and root aphids (Popay *et al.*, 2004, Popay & Gerard, 2007). Therefore, to reduce the toxic effects of *E. festucae* without compromising the added benefits provided by these endophytes, new strains were sought that may still produce anti-insect/pest alkaloids but no lolitrem B and ergovaline that are toxic to livestock.

Synthetic novel associations have been shown to mitigate the problem of livestock toxicity without compromising the additional benefits, but only fairly recently (Milne, 2007). What determines the success of attempts to generate new *E. festucae* / *L. perenne* associations of agronomic and economic interest is still largely unknown, as is what determines the long-term stability of the associations and their desirable traits. Thus, using such associations as an object of study can not only answer fundamental questions but can also assist NZ breeders in making, improving and maintaining agriculturally superior associations. Knowing how the endophyte changes in novel *E. festucae* / *L. perenne* associations may be important in addressing the issues related to intellectual property rights. Some of the novel associations have been commercialized and protected by patent. But in associations involving asexual *Epichloë* spp. as symbiont, it is possible that many divergent clones of the same lineage with multiple different mutations may be simultaneously present in the population and competing with each other for selection, a phenomenon known as clonal interference. If so then it is possible to mistake the multiple clones of the same ancestral lineage as separate strains. It may lead to commercialization of the same endophyte strain as a separate brand based on presence of some mutations. As a result it is challenging to find ways to protect intellectual properties for endophytes in new associations, namely to demonstrate that these divergent lineages are descendants of a protected strain, rather than representing other strains.

1.1.5. Characteristics of *E. festucae* AR37

Among *E. festucae* strains used in attempts to generate novel associations with improved agricultural properties, strain AR37 stands out as it represents the greatest success of New Zealand breeders in terms of generating novel associations (Hume *et al.*, 2007, Hume *et*

al., 2004, Johnson *et al.*, 2013). It was thus a natural choice for this work. In the following I will introduce its key properties.

AR37, also called Lp14, originated in France (Tian *et al.*, 2013a, Christensen *et al.*, 1993). Agronomic characteristics were compared for a perennial ryegrass cultivar infected with natural "wild type" endophyte and the same cultivar infected with 7 selected endophytes, including AR37, and endophyte-free cultivar in 11 trials in four different regions of New Zealand over a period of 3-4 years (Hume *et al.*, 2004). AR37 was selected as it apparently did not produce any of the then known bioactive alkaloids that are toxic to cattle (Johnson *et al.*, 2013, Hume *et al.*, 2007, Popay & Thom, 2009). AR37-ryegrass combinations outperformed all other combinations (Hume *et al.*, 2004) (https://www.grasslanz.com/understanding-the-science/15-ar37-endophyte). AR37 was commercialized in 2007 in New Zealand and by 2013 it had been introduced into 11 different ryegrass cultivars (Johnson *et al.*, 2013). There are many similar reports confirming that AR37 performs better than or similar to other existing endophytes in terms of resistance to insect pests and improvement of the agronomic characteristics of the host grass (Tian *et al.*, 2013b, Popay & Thom, 2009, Jensen & Popay, 2004, Popay & Cox, 2016, Popay & Hume, 2011, Thom *et al.*, 2014), leading to increased yields of perennial ryegrass in field experiments (Hume *et al.*, 2007). Specifically, AR37 endophyte provides resistance to a broad range of insect pests (5 out of total 6 insects in New Zealand) e.g. Argentine stem weevil (Popay & Wyatt, 1995), root aphid (Popay *et al.*, 2004, Popay & Gerard, 2007) pasture mealybug (Pennell *et al.*, 2005) and black beetle (Popay & Thom, 2009). AR37 also provides resistance against porina larvae (Johnson *et al.*, 2013). Porina consist of 7 related moth species found in New Zealand (Dugdale, Barlow *et al.*, 1986). AR37 has been shown to reduce growth, development and survival of the porina larvae which preferred ryegrass without endophyte in a choice experiment (Jensen & Popay, 2004). It is not certain whether AR37 also has an effect on root aphids and pasture mealy bugs that feed on ryegrass roots: epoxy-janthitrem, the anti-insect alkaloid characteristic to AR37 (see below) is not found or found in very low concentration in roots (Popay & Gerard, 2007).

A key characteristic of AR37 is that it does not produce any of the four alkaloid groups i.e. ergot alkaloids, lolines, peramines and indole diterpenes, that are commonly found in other *Epichloë* spp. (Tian *et al.*, 2013a). Some of these alkaloids (lolitrem and ergovaline)

are toxic to livestock (Fletcher, 1993, Fletcher, 1982, Gallagher *et al.*, 1981, Gallagher *et al.*, 1982) explaining why AR37 infection has no toxic effect on livestock. The only known alkaloids associated with AR37 are epoxy-janthitrems, and these alkaloids may underlie the bioprotective properties of AR37. However, epoxy-janithrems are unstable and it has been challenging to directly verify if they play a role in protecting against insect pests and herbivores. Purified epoxy-janthitrem did indeed negatively affected the growth of porina larvae (Finch *et al.*, 2010). The observation that the density of predation by porina larvae differed between different AR37-infected ryegrass cultivars (Popay *et al.*, 2012) suggested that plant genotype may have an effect on concentration of epoxy-janthitrems, as is the case for other *Epichloë* spp. alkaloids (Pańka *et al.*, 2013, Spiering *et al.*, 2005a).

Epoxy-janthitrems, like lolitrem B, are indole-diterpene compounds and the pathways leading to their biosynthesis may overlap (Figure 1.2). Epoxy-janthitrems, and lolitrem B are both lipophilic and are of similar structure (Rasmussen *et al.*, 2008, Rasmussen *et al.*, 2007, Rasmussen *et al.*, 2009). Lolitrem B translocation within the plant is reported to be difficult (Munday-Finch & Garthwaite, 1999, Spiering *et al.*, 2005a), and this may thus also apply to epoxy-janthitrems. Seasonal variations in the concentrations of epoxy-janthitrems have been reported, with concentrations peaking between December and April (Moate *et al.*, 2012). Similarly, the anti-feedant or insect-deterrent property of AR37 is affected by temperature, with higher temperatures (20 ^{o}C) increasing the deterrent activity and lower temperatures (7 ^{o}C) decreasing the deterrent activity (Hennessy, 2015). AR37 ryegrass pastures at cooler temperatures (7 ^{o}C) may thus be more prone to insect attack. AR37-containing new cultivars that produce more epoxy-janthitrem at lower temperatures could improve pasture persistence in cooler regions (Hennessy, 2015).

Table 1.1. AR37-perennial ryegrass combination is more effective than other endophytes-perennial ryegrass combinations against most prevalent insect pests in New Zealand. (modified from (https://www.farmlands.co.nz/Documents/Guides/Farmlands_Seed_Buyers_Guide_2018.pdf)).

Diploid perennial ryegrass							
Insect	**Argentine stem weevil**	**Pasture mealy bug**	**Black beetle adult**	**Root aphid**	**Porina**	**Grass grub**	**Field cricket**
AR1	♦♦♦♦	♦♦♦♦	♦	-	-	-	NT
NEA2	♦♦♦	(♦♦♦♦)	♦♦♦	♦♦	NT	-	NT
AR37	♦♦♦♦	♦♦♦♦	♦♦♦	♦♦♦♦	♦♦♦	♦	NT
SE	♦♦♦♦	♦♦♦♦	♦♦♦	♦♦	♦	-	NT
WE	-	-	-	-	-	-	NT
Tetraploid perennial ryegrass							
AR1	(♦♦♦)	(♦♦♦♦)	♦	-	-	-	NT
NEA2	♦♦	(♦♦♦♦)	♦♦♦	♦♦	NT	-	NT
AR37	(♦♦♦)	(♦♦♦♦)	♦♦♦	♦♦♦♦	(♦♦♦)	♦	NT
WE	-	-	-	-	-	-	NT

Key to tables

AR1, NEA2 and AR37 are novel endophytes

SE is "standard endophyte" that is naturally present in most New Zealand pastures

WE is ryegrass without endophyte.

NT is not tested

– No control

♦ Low level control: Endophyte may provide a measurable effect but is unlikely to give any practical control.

♦♦ Moderate control: Endophyte may provide some practical protection, with low to moderate reduction in insect population.

♦♦♦ Good control: Endophyte markedly reduces insect damage under low to moderate insect pressure. Damage may still occur when insect pressure is high.

♦♦♦♦ Very good control: Endophyte consistently reduces insect populations and keeps pasture damage to low levels, even under high insect pressure.

( ) Provisional result: Further results needed to support the rating. Testing is ongoing.

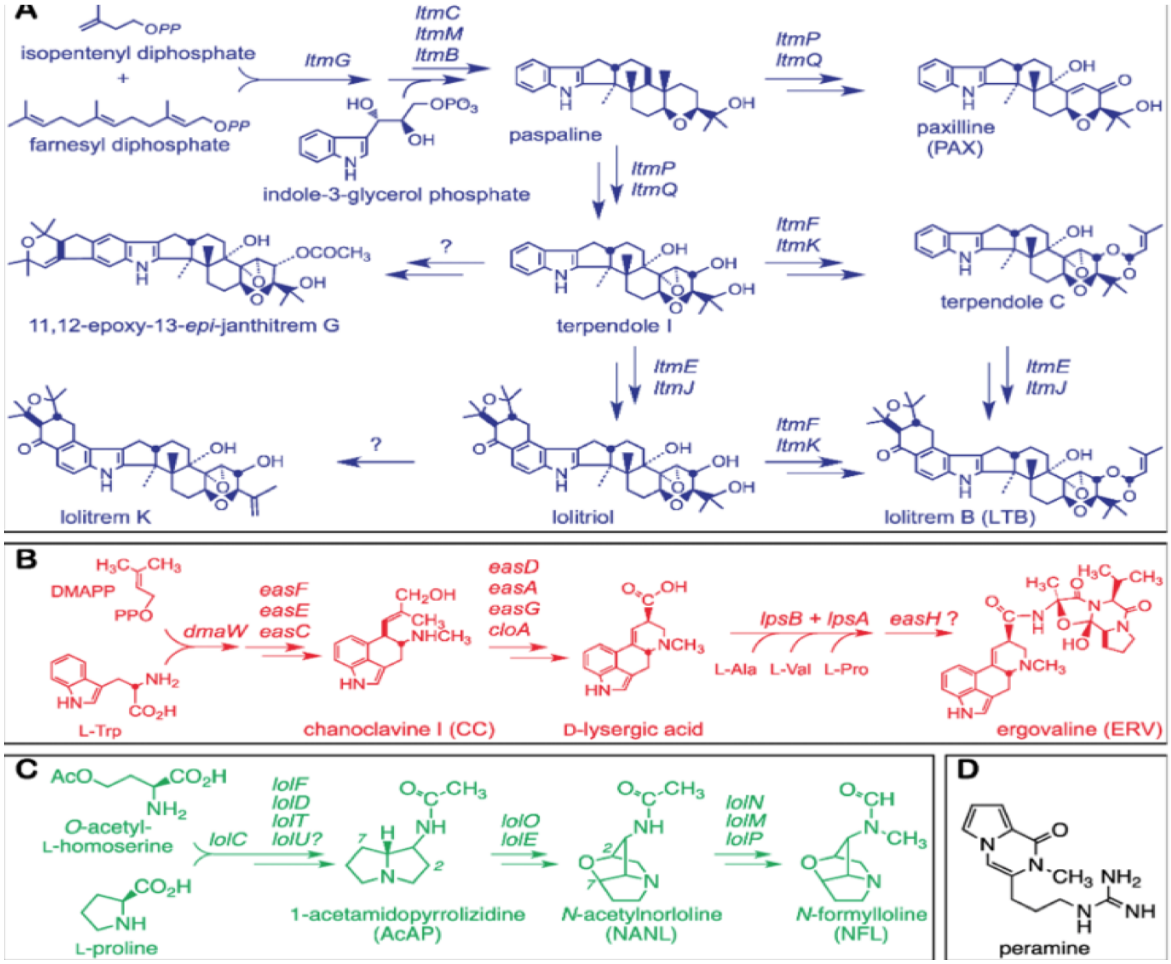


Figure 1.2. Pathways for the synthesis of four bio-protective alkaloids produced by endophytes. (A) Lolitrem biosynthesis pathway. Epoxy-janthitrem and lolitrems share the same core structure and early pathway steps (B) Ergot alkaloid biosynthesis pathway (C) Loline biosynthesis pathway (D) Peramine, produced by only one enzyme (Copied from Schardl *et al.*, 2013c)

1.1.6. Characteristics of novel AR37-containing cultivars and breeding towards improved compatibility

There is a long history of breeding programs aimed at improving the pastoral performance of perennial ryegrass in New Zealand. In these programs, selection of perennial ryegrass with desirable characteristics is usually done through sexual crosses (Humphreys *et al.*, 2006). Each progeny is assessed for desirable traits and the best one is chosen (Wilkins, 1991). The focus has always been the improvement of key factors i.e. forage quality, dry matter yield and persistence, in grasses (Humphreys *et al.*, 2006). Once the importance

of endophytes for grass performance had been discovered (in 1981; (Easton *et al.*, 2001)), it was realized that introduction of performance-enhancing endophytes would be another powerful means of creating better cultivars (Funk & White, 1997).

The commercial success of AR37 is attributable to such breeding programs. A trait of prime importance in these programs was AR37-host compatibility, in particular the improvement of AR37 seed transmission: only with high rates of AR37 seed transmission is the commercially viable production of AR37- infected seed possible (Hume, 2005). I used novel associations of AR37 with two new ryegrass cultivars i.e. SAMSON and KLP1102 in which compatibility / seed transmission improved with breeding. Both these associations initially showed signs of mild incompatibility but the associations survived and compatibility appeared to improve over the next few years. One key parameter indicating the improvement in compatibility was the rate of endophyte transmission in host seeds. It improved from 76% in first progeny to around 98% in fourth progeny. High rates of fungal transmission in host seeds are important for commercial exploitation of the beneficial characteristics of the association.

The AR37-SAMSON breeding program started in 1996 with artificial inoculation of AR37 into the 353 seedlings of the newly developed ryegrass cultivar SAMSON. Only 38 AR37-infected plants (T0 plants) survived the attempt. In 1997, seeds from these 38 plants (T0) were sown in Manawatu to generate a second generation (T1) of SAMSON plants. T1 plants showed an overall improvement over T0 plants especially in terms of endophyte infection rate of seedlings which was 76% in T1. Two years later, in 1999, seeds from 1200 endophyte infected T1 plants were sown in Lincoln area to generate the T2 generation. In T2 infection rate was 89% which was an improvement over T1. Between 2001 and 2006, these plants were tested in the field for their agronomic performance and were compared with many other endophyte-host combinations. Using randomly chosen seeds from only endophyte infected plants, T3 and T4 were generated in 2007 and 2008 respectively. The endophyte-infection rate for T3 and T4 was 96% and 98% respectively. In 2007 the AR37-SAM cultivar was commercialized. Seeds from T4 were stored for four years before they were again used in 2012 to generate T5. Seeds from T5 were then used to generate T6 plants in 2013. Hundred randomly picked seeds from T6 were used in this study. The percentage of endophyte infection dropped from 98% for T4 to 90% for T5 before again rising to ~95% for T6. The reason for this drop in

endophyte-infection percentage in T5 could have been loss of endophyte during seed storage: endophyte survival in seed is climate dependent (Hume *et al.*, 2011, Hume *et al.*, 2013) and thus can change between years. However, another potential reason for the drop in endophyte-infection percentage in T5 can be that no endophyte selection was done in T4 and it may have resulted in fewer endophyte-infected T4 seeds.

It is important to note that the seeds used to generate the next progeny did not come from one plant only. Instead the process started with seeds from 38 different plants and then randomly picked seeds from each generation were used to produce the next generation. In such a scenario it is less likely that all the plants in the next progeny have endophytes with the same mutations. Instead the chances are that endophytes from different host plants may have different mutations to adapt to that particular host plant.

Table 1.2. An overview of history of perennial ryegrass cultivar SAMSON. Seeds from T6 generation in 2013 were used in current study.

Year	Cultivar-breeding stage
1996	Inoculation of 353 seedlings (T0) – 38 survivors
1997	Seed harvesting from 38 plants (T1)
1999	Pre-nucleus crop from 1200 plants (T2)- No endophyte selection
2001-2006	Agronomic field performance tests (African black beetle, Root aphids etc)
2007	Spaced Plants (T3) - Endophyte Selection
2008	Nucleus production (T4) – No endophyte selection
2012	Spaced plants (T5)- endophyte selection
2013	Spaced plants (T6)- endophyte selection

Another, more complex, breeding program, for KLP1102, started around same time as for SAMSON. KLP1102 is based on the tetraploid cultivars ‘Grasslands Impact’, cv ‘Bealey’ (NW Spain) and cv Dutch (winter hardy N Europe). ‘Grasslands Impact’ itself is a result of crossing between NZ ecotype and NW Spain. Thus KLP1102 is a combination of 4 pools containing material from some or all 4 of the above cultivars. All these 4 pools had different endophytes including AR37 for different number of generations during the development of KLP1102. Once a KLP1102 cultivar was

established from these 4 pools, there have only been 2 cycles of AR37-endophyte seed transmission. For KLP1102, only three plants (T0) survived the initial inoculation attempt and seeds from these three plants were used to generate T1 plants. Infected T1 plants generated from T0 seed were used to generate T2 and the seeds from these T2 plants were used in this work. As for SAMSON, random seeds from each generation of KLP1102 were used to produce next generation.

Given that selection in both programs was for endophyte-infected plants, both plant and endophyte were under selective pressure to improve compatibility. Such improved compatibility should have manifested itself not only in higher percentages of infected seed and better survival of endophyte in infected seed. In addition the chance of an AR37-infected plant to contribute progeny to the next generation also depended on how well it could compete with other AR37-infected plants in the field, since this would determine how much seed it would contribute. At least some of the changes in the AR37 genome during these programs are thus likely to contribute to enhanced compatibility, and the genes affected by such mutations are determinants of compatibility and thus important parts of the cross talks that guides the interaction between the symbionts. In other words, identifying these changes can serve as a forward genetic screen for unravelling the still largely unknown genetic basis of *Epichloë*-host interaction. While such a forward screen has great potential, the necessary comparisons of whole AR37 genomes from different generations of infected plants for identification of such loci is not an easy task. Following is a brief introduction of the necessary steps and tools available for this complex task.

1.2. Finding genetic changes in populations using bioinformatics

Comparison of the *Epichloë* genomes mentioned above can identify loci in the genomes that may be involved in adaptation to the host. Such variant loci are mutations that have occurred during adaptation. DNA mutations provide raw material for evolution and have been a focus of research in nearly all biological fields. With the advent of Next Generation Sequencing (NGS) technologies, it is now possible to generate huge amounts of sequence data at a low cost and in a relatively short period of time. As a result there has been a tremendous increase in research aiming to find out mutations not only in haploid and diploid individuals but also in large populations, using pooled samples. Sequencing of

pooled samples is becoming more popular to study population genetics as it provides a better overview of population structure at a relatively low cost. There are challenges associated with each type of these samples and different frequency thresholds may be set to reliably detect true variants. Calling mutations from individual germline samples is relatively easy because in such cases a variant will have an allele frequency of either 50% or 100% (Xu *et al.*, 2012, DePristo *et al.*, 2011, Li, 2011). Sequencing errors usually occur at a much lower frequency. Error rates depend on multiple parameters e.g. sequencing platform, library preparation method etc but are highly unlikely to reach the 50% threshold that would cause problems in variant calling (Xu, 2018). However, calling variants from pooled samples is much more difficult because real variants may be present in the pool at frequencies comparable to sequencing errors; as a results it is challenging to distinguish between a real variant and an artefact caused by sequencing errors (Wei *et al.*, 2011). In non-diploid organisms, including prokaryotes, viruses and haploid fungal genomes, detecting variants in pooled samples can be a special challenge because low frequencies can occur not only due to an individual carrying an individual mutation but also due to the fact that microbes can have > 2 alleles per locus, only one of which may carry the mutation. Allele frequencies in such cases can range from close to zero to 100% (Xu, 2018, Wei *et al.*, 2011). Thus distinguishing between low-frequency true variants and sequencing errors is particularly challenging in such systems.

1.3. General workflow of variant calling

Finding variations from the huge data produced by current NGS technologies is a complex process involving multiple steps with each step requiring a specialized algorithm. Due to increasing popularity of NGS, for each step an array of such tools exists. The best choice of tools is crucial for obtaining meaningful results and this may depend on the type of sample, amount of data and type of sequencing platform used as well as the purpose of the study. There are many reviews of these tools available, such as (Pabinger *et al.*, 2014, Li & Homer, 2010, Bao *et al.*, 2011, Nielsen *et al.*, 2011, Koboldt *et al.*, 2012).

In the following I will introduce techniques and their challenges in the various steps of variant calling, namely:

1. DNA sample preparation
2. Sequencing
3. Quality assessment of raw sequencing data
4. *De novo* assembly
5. Mapping reads to *de novo* assembly
6. Variant calling
7. Variant filtration

1.3.1. DNA sample preparation

Most sequencing platform require a specific amount of high quality, nondegraded and intact DNA at a certain minimum concentration (Wong *et al.*, 2012). Obtaining DNA of such high quality and good concentration may be a challenge for certain species. DNA is checked on high resolution gel before sending for sequencing. Library preparation mostly involves PCR steps which may introduce artefacts in the downstream steps. There are PCR-free library preparation protocols offered by some of the platforms e.g. Illumina MiSeq.

1.3.1.1. Special considerations for preparing samples from pools of individuals

Although sequencing cost has dropped tremendously during the last decade, sequencing of large number of individuals from any population is still expensive and beyond the budget of many laboratories (Futschik & Schlötterer, 2010). One way to reduce the sequencing cost is to pool DNA from multiple individuals together before sequencing, a strategy called pool-seq (Arnheim *et al.*, 1985, Futschik & Schlötterer, 2010). DNA from multiple individuals can be pooled together and sequenced in two different ways: (i) to pool DNA together and sequence it without adding any tags or index sequences (barcodes) so that all the DNA is sequenced as one sample. (ii) to add tags or identifying sequences (barcodes) to individual DNA before pooling and sequencing so that DNA is sequenced as one pool but each DNA can be identified inside the pool, also called multiplexing (Guo *et al.*, 2013, Craig *et al.*, 2008). There are advantages and disadvantages of each technique (Zhu *et al.*, 2012). One big advantage of (i) is that it is less expensive than (ii) as there is no extra labour and cost for adding tags. However the

biggest disadvantage of (i) is that there is no way to trace any discovered variant back to the DNA (and thus the individual) containing that variant (Guo *et al.*, 2013). Regardless of which of the two approaches is used, if the number of individual DNAs pooled is too large then there is a higher probability that rare variants may not be identified correctly. The reason is that a variant present in DNA from a single individual will be typically represented by 1/2n reads for diploid species and 1/n reads for haploid species (as mentioned above in microbes the frequencies can be even lower). With a large number of individuals in the pool, a variant present in only one of the individuals may thus have a frequency lower than the sequencing error rate. Variants at an allele frequency of ≥ 1% have been reliably detected from pooled data generated using Illumina sequencing (Out *et al.*, 2009, Margraf *et al.*, 2011). Nevertheless, if the coverage of the variant-containing region is low a variant unique to one individual may not be sequenced at all.

1.3.2. Next Generation Sequencing (NGS)

The original Sanger method, also called first generation method, is still considered a gold standard in many situations that require accurate validation of small DNA regions (McCourt *et al.*, 2013). For example, Sanger sequencing is still regularly used to validate transgenic integration sites, plasmid insert sequences, and coding mutations. However, due to the low throughput nature of the Sanger method newer high throughput methods have now replaced this traditional approach for most applications (Martinez & Nelson, 2010). These newer methods are generally described as next generation sequencing (NGS) (Schuster *et al.*, 2008, Liu *et al.*, 2012), and include second generation sequencing, in which short DNA segments are sequenced at high throughput (Dolled-Filhart *et al.*, 2013) and more recently third generation sequencing, in which much longer DNA fragments are sequenced (Schadt *et al.*, 2010). All NGS methods make use of advances in nanochemistry and nanotechnology to dramatically increase the throughput of sequencing relative to the Sanger approach.

The first NGS machine developed was Roche 454 (Margulies *et al.*, 2005) system, despite the huge advance in throughput over Sanger sequencing it yielded relatively short (< 400 bp) error prone reads (Table 1.3). The ABI SOLiD sequencing system, commercialized in 2007, produced shorter reads (35 bases) but at a lower error rate than Roche 454 and

higher throughput. Another high throughput sequencing machine i.e. Solexa platform was commercialized in 2006. It was acquired by Illumina in 2007, and the development of Illumina HiSeq and MiSeq, whilst still generating relatively short reads (< 300 bp), dramatically improved read accuracy and increased data volumes by several magnitudes. The so called third generation sequencing machines, such as Pacific Biosciences (Eid *et al.*, 2009) and Oxford Nanopore (Jain *et al.*, 2015, Deamer *et al.*, 2016), take advantage of error prone single molecule sequencing to generate very long reads (100s of kb) that are then subsequently error corrected using read coverage. These latter technologies are ideal for genome polishing and genome assembly applications, and the long reads can dramatically improve on genome assemblies performed using high coverage short read technologies such as Illumina. Nanopore is also suitable for detecting any modifications on individual nucleotides which second generation technologies are incapable of detecting (Schreiber *et al.*, 2013, Wescoe *et al.*, 2014).

Illumina technology is the most popular platform for NGS sequencing based on its high accuracy and low cost per base-pair (Table 1.3). Illumina chemistry uses fluorescently labelled reversible terminator nucleotides to synthesize the second strand of template DNA (fragment), which has been immobilized on the surface of the glass flowcell and amplified by PCR to form a cluster (Bentley *et al.*, 2008). As each reversible terminator nucleotide is added to the growing second strand the fluorescence is recorded by a very high-resolution camera. At the end of each sequencing cycle the terminal nucleotide is then chemically reversed allowing the addition of another labelled nucleotide, and the process continues until the desired read length is obtained (Bentley *et al.*, 2008). If the machine is run in paired-end mode once the first read is complete, the cluster is regenerated and the other strand of the fragment is used as template for the sequencing process.

The two most common Illumina instruments are the 8 lane HiSeq2500 and the 1 lane MiSeq machine. The HiSeq machine uses read lengths of between 50 bp to 150 bp to generate up to 500-1800 Gb of sequencing data per run (5-6 billion reads). In contrast the MiSeq machine yields only 15 Gb on its single lane (~25 million reads), however the maximum read length is double that of the HiSeq at 2 x 300 bp (https://sapac.illumina.com/systems/sequencing-platforms.html). Due to the large size of most vertebrate genome, sequencing projects make use of the high yield of the HiSeq

machine, while the MiSeq capacity is ideal for sequencing genomes of microbes. For example, the 12 Mb genome of Saccharomyces cerevisiae can be sequenced to very high coverage (~500x) on just a single run of a MiSeq machine in paired-end mode. Multiple samples can also be included in a single run using adaptor tags in a process called multiplexing. Once the sequencing run is complete the sequenced reads are demultiplexed bioinformatically using the sample specific tags added during the library preparation.

The main drawback of Illumina HiSeq and MiSeq machines is the relative short length of the sequence reads. In *de novo* genome assembly applications long repetitive regions and large structural variants are almost impossible to accurately assemble using the short reads from these Illumina machines or any other sequencing platform (McCoy *et al.*, 2014). Reliable mapping of short reads to a genome is also difficult, especially in low complexity or duplicated regions of genomes (Salzberg *et al.*, 2012). For this reason Illumina developed mate-pair library preparation protocols, which allows sequencing libraries to be generated from DNA fragment up to 10 kb. The mate paired library is then sequenced as usual in pair-end mode, with up to 150 bp of DNA sequence being obtained for each end of the large genomic fragment. The large insert sizes of these mate-paired reads significantly increase the scaffolding rate of the *de novo* assemblies (Gnerre *et al.*, 2011).

Sequencing errors are common with all NGS techniques (Liu *et al.*, 2012, Bragg *et al.*, 2013, Hoffmann *et al.*, 2009) (Table 1.3). In many cases the existence of sequencing errors can be detected in high coverage regions (Rieber *et al.*, 2013), however, coverage comes at increased cost to the researcher. Even when relatively high coverage is obtained there will always be regions of low coverage due to the uneven distribution of sequenced reads across large genome. In the case of Illumina technologies (and others) each sequenced nucleotide is given an off machine phred quality score that describes the probability that the nucleotide identification was called correctly (Ewing & Green, 1998, Brockman *et al.*, 2008). Base quality scores tend to deteriorate with the number of cycles especially at 3` end of the reverse strand (Brockman *et al.*, 2008). Bioinformatic tools can then use the phred quality base scores to remove low quality base pairs from the raw data before they create issues in downstream applications (Abnizova *et al.*, 2017)(see below). Despite these precautions due to the large volumes of data generated by NGS machines,

even small error rates can have significant impacts on our ability to accurately characterise genome variation using this type of data.

Sequencing errors are particularly problematic for Illumina machines in GC rich or GC poor DNA regions (Rieber *et al.*, 2013, Chen *et al.*, 2013). Also lagging strand gets out of phase with time causing deterioration of sequence quality towards the end of the read. Accurate identification of genomic insertions and deletions (indels) is also a challenge for NGS technologies as these regions interfere with accurate mapping of short read data (Li *et al.*, 2013). Structural information carried by pair-end reads can help with correct mapping at regions containing indels (Albers *et al.*, 2011, Li *et al.*, 2013), but this is highly dependent on the size of the indel itself. Illumina also offers PCR-free library preparation protocol which is reported to decrease coverage bias across sequenced genome (Quail *et al.*, 2012, Kozarewa *et al.*, 2009).

The best platform to sequence a genome for variant calling may thus be the one with good read length and accuracy. Looking at table 1.3, Illumina platform with a read length of 300 bp (MiSeq) with an error rate of 0.1% seems ideal for small genomes.

Table 1.3. Comparison of read lengths and error rates of DNA sequencing platforms

Sequencing Platform	**Max Read length (bp)**	**Cost/Mbp (Gb)USD**	**Error rate**	**Error bias**
Illumina HiSeq	150	(41)	0.1%	Single base substitutions
Illumina MiSeq	300	0.07(502)	0.1%	Single base substitutions
Ion Torrent	200-500	(1000)	1.0%	Short deletions
Roche 454	400-1000	10	1.0%	Deletions/insertions
SOLiD	50-75 bp	0.13	0.1%	A-T bias
PacBio	150,000	(2000)		CG deletions
Nanopore	200,000			

1.3.3. Quality control of NGS data

As discussed above all NGS technologies generate a non-significant amount of sequencing errors, (Table 1.3) therefore the first step after sequencing is to analyse the quality of the off-machine (raw) sequencing reads in order to identify any potential error biases in the raw reads (Abnizova *et al.*, 2017). Popular software such as FastQC provides graphical summaries of important quality control metrics, such as read error distribution, GC bias, and read duplication levels and may detect problems in the library preparation or in sequencer (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Next read trimming is performed to remove low quality calls based on the off-machine phred quality score assigned to each nucleotide. For Illumina sequencing, sequence quality gets worse towards the ends of reads, especially at 3` end of reverse reads (Brockman *et al.*, 2008) and these are the bases that are usually removed by trimming applications (Kelley *et al.*, 2010, Del Fabbro *et al.*, 2013). High coverage data is often only lightly trimmed using a phred score of 10 (10% error probability), with downstream applications making use of the coverage to filter out any residual errors. For low coverage or complex populations a high stringy trimming cut-off of > 20 (1% error probability) is typically adopted, although there is always a trade-off between removing "all" bad calls and throwing away too much "good data". At this stage adapters, linkers and sequencing primer contamination is also removed from data. Typical tools for performing the trimming and adaptor filtering include BBMAP, SolexaQA (Cox *et al.*, 2010), FASTx-Toolkit (Gordon & Hannon, 2010), FastQC (Andrews, 2010) and Trimmomatic (Bolger & Giorgi, 2014). Illumina machines also add a small amount of a viral (PhiX) genome to each sequencing run to help with base-call calibration; these reads are removed by discarding any reads that successfully map to the PhiX genome.

1.3.4. *De novo* genome assembly

A *de novo* genome assembly is the *in silico* construction of a longer chromosome / DNA sequence from shorter read sequences produced by sequencing platforms (Abnizova *et al.*, 2017). A complete genome sequence is a crucial resource in genomics studies (Meader *et al.*, 2010). A complete record of genomic variations i.e. single nucleotide polymorphisms (SNPs), insertion-deletions (indels), copy number variations (CNVs) and structural variations (SVs) can be obtained by comparing whole genome resequencing

data to a complete genome assembly (Ellegren *et al.*, 2012, Heliconius Genome, 2012). Studies focussing on DNA-protein interactions (Auerbach *et al.*, 2013), gene expression (Vijay *et al.*, 2013) and epigenetic modifications (Herrera & Bazaga, 2011) also depend largely on availability of an accurate complete reference genome. An annotated genome is a great resource to find variations that may represent signatures of selection (Hohenlohe *et al.*, 2010).

Assembling a genome from short NGS reads is a considerable challenge, especially in presence of repetitive regions (Brenchley *et al.*, 2012, Nagarajan & Pop, 2013). Genes that are member of a large family are exceptionally difficult to assemble when the levels of sequence conservation across the family are high (Ekblom & Wolf, 2014). In such situations, genome assemblies can 'skip' genes because they collapse reads from very similar genes into a single assembled contig. Long read technologies can help resolve these regions and better scaffold the genome (Huddleston *et al.*, 2014). Library preparation protocols such as mate-paired libraries can help assemble these regions by providing large scale structural information to the assembler. In diploids, polyploids or pooled samples, variable regions are represented in the resulting genome assembly using a consensus sequence of chromosomes sets that contain the alternative sequences (Consortium, 2004). Some regions of the genome are extremely difficult to sequence e.g. areas around telomeres and centromeres as they contain very high levels of repetitive DNA (Ellegren *et al.*, 2012, McCoy *et al.*, 2014, Alkan *et al.*, 2010, Ye *et al.*, 2011).

The *de novo* assembly process begins with the generation of overlapping reads that are termed contigs. Higher read depth and longer reads typically result in a more reliable assembly due to larger regions of overlapping between reads. Larger reads can also be assembled using less computational resources, especially computer memory which is often limiting. Finally, structural information from paired-end reads or genetic maps is used to order the contigs into larger assembled regions called scaffolds. Although the goal of most assembly projects is to obtain a single scaffold per chromosome, this is most often not possible for all but the simplest of genomes.

There are many tools available for *de novo* assembly of sequencing data. All differ in their performance depending on the features of the genome, type and quality of data, available read lengths, and error profile of the sequencing platform used. Many older

assemblers developed to use longer Sanger or 454 reads e.g. Arachne (Batzoglou *et al.*, 2002), Celera assembler (Denisov *et al.*, 2008) and PCAP (Huang *et al.*, 2003) and some short read sequence assemblers e.g. FERMI (Li, 2012) and Edena (Hernandez *et al.*, 2008) use the overlap-layout-consensus (OLC) approach. However, these methods are often not capable of efficiently assembling the very large volumes of short read data produced by the popular Illumina sequencing platform. For this reason new assemblers designed to work with short read Illumina data were developed based on the De Bruijn graph. The De Bruijn or Eulerian graph approach uses defined windows of overlapping read sequence of length n (called k-mers) as nodes in the graph. A graph is then constructed in which k-mers nodes are connected whenever n-1 k-mer sequence is shared between them. Some of the most popular assemblers e.g. ABySS (Simpson *et al.*, 2009), Velvet (Zerbino & Birney, 2008), ALLPATHS-LG (Gnerre *et al.*, 2011), SOAPdenovo (Luo *et al.*, 2012), and SPAdes (Bankevich *et al.*, 2012) are based on this De Bruijn graph approach. The length of the k-mer window is often critically important, too small and the assembly graph becomes overly complexed (tangled), too long and the graph is unable to be extended due to the lack of sequence overlaps. Some assemblers allow to use a range of different k-mers and choose the best assembly.

Another popular approach for *de novo* assembly of microbial genomes relies on modified De Bruijn graph. Instead of using standard De Bruijn graph, it relies on paired De Bruijn graph (PDBG), an approach used to make good use of paired-reads (bi-reads). Instead of k-mers, this approach relies on a set of pair of k-mers (k-bimers) located at a specific distance (= d) in a read. It also introduced k-bimer adjustment to fix the distances of most of the adjusted k-bimers and improve the quality of the assembly (Bankevich *et al.*, 2012). SPAdes makes use of this approach.

Assembly algorithms have been reviewed several times (Miller *et al.*, 2010, Nagarajan & Pop, 2013, Chin *et al.*, 2014, Zhang *et al.*, 2011b, Desai *et al.*, 2013, Wajid & Serpedin, 2012, Schatz *et al.*, 2010, Bradnam *et al.*, 2013). No single assembler is good enough for all kind of data. It is better to use several different assemblers and then assess which of the assemblers produced the best assembly (Ekblom & Wolf, 2014).

1.3.5. Mapping of sequencing reads

One of the crucial step in variant calling is matching of sequencing reads with the corresponding region of a reference genome in a process referred to as mapping. Given the volumes of sequence data and the short read nature of many NGS technologies, correctly aligning a read to its corresponding region of the genome is a non-trivial process, and often represents the most computationally intensive step of most variant calling workflows. For this reason most aligners make use of the highly efficient Burrows-Wheeler algorithm (BWA) to align short reads to reference (Burrows & Wheeler, 1994). These BWA methods gain huge efficiencies by using hash tables to dramatically reduce the size and complexity of the data that needs to be processed.

There are different reports comparing performance of different mappers (Li and Homer 2010; Nagarajan and Pop 2013; Otto, Stadler, and Hoffmann 2014; Pightling, Petronella, and Pagotto 2014; Shang et al. 2014). Generally the final choice of mapper depends on the characteristics of the specific reference genome and the types of short read data that need to be aligned. For practical purposes the available computational resources can also limit the choice of read mapper, as there are often large differences in run time and memory (RAM) requirements for different software packages.

Some of the popular aligners are SOAP (Li *et al.*, 2009), Novoalign (http://novocraft.com), Bowtie / Bowtie2 (Langmead *et al.*, 2009, Langmead & Salzberg, 2012), BWA (Li & Durbin, 2010, Li & Durbin, 2009), MAQ (Li *et al.*, 2008), SSAHA (Ning *et al.*, 2001) and SOAP2 (Li *et al.*, 2009) etc. Updated list of aligners is available online (https://www.ebi.ac.uk/~nf/hts_mappers/). These algorithms vary in their sensitivity and methodology. So different results may be reported by different algorithms for the same sequence data, depending on sequencing platform, read length and complexity of the reference genome. Some mismatches are allowed between the reads and the reference genome using different approaches. SOAP2 splits read into 3 fragments and allows mismatches in maximum of 2 of the fragments. Bowtie and BWA allow user to pre-set criterion and the search stops when the pre-set criterion is met. Novoalign does not allow user to pre-set number of mismatches in a read but instead a threshold for alignment score can be set.

There are different reports comparing the performance of different mappers (Li & Homer, 2010, Nagarajan & Pop, 2013, Otto *et al.*, 2014, Pightling *et al.*, 2014, Shang *et al.*, 2014). An excellent review also compares different mapping algorithms (Fonseca *et al.*, 2012). Most mappers show good sensitivity towards long reads. Starting with accurate reads, it is observed that many of these alignment tools perform similarly in non-repetitive, simple genomes (Yu *et al.*, 2012). Novoalign is reported to perform better even with relatively shorter reads and complex reference genomes (Thankaswamy-Kosalai *et al.*, 2017).

1.3.6. Variant calling

Variant calling uses read alignments to identify differences between the sequenced sample and the reference genome (Zhang *et al.*, 2015, Liu *et al.*, 2012). Although it seems a straight-forward process of picking the sites that are different from the reference genome but in practice it is a complicated process due to presence of many different sources of sequencing errors. The three main types of variants are: (i) Single nucleotide variants / polymorphisms (SNVs / SNPs) are usually the most common type of variants and include transitions and transversions (ii) insertions and deletions (indels) and (iii) Structural variants (SV) that include copy number variants, translocation, duplication etc. Most variant callers make use of a probabilistic model that uses the phred quality of the nucleotide call, the number of reads that support that call, and other features such as the ploidy of the organism, to assign a probability that a variant is real. Correct calling for long indels and SVs is difficult, especially when any one read is not long enough to cover the whole variant region (Narzisi *et al.*, 2014, Kojima *et al.*, 2013). Problems associated with SV calling are described by Medvedev (Medvedev *et al.*, 2009). Some popular callers, such as GATK and FreeBayes make use of local re-alignment to improve the probability that an indel will be called correctly (Van der Auwera *et al.*, 2013). However, even these more advanced methods are subject to considerable error. At the population level the correct identification of rare variants is particularly challenging, as these may occur at a frequency at or only slightly above the underlying error rate of the sequencing technology (Chin *et al.*, 2013).

Performance of variant callers can be compared by measures of their sensitivity and specificity based on a data set with known variants (standard or truth data set). However, there is no one best caller, and ultimately the properties of the reference genome,

underlying short read data, and the mapped population will to some extent influence the performance of each caller. Some variant callers have also been calibrated to work best with model system: GATK is a good example of this because it utilizes a set of known variants to the variant calling calibration. In such cases these callers calibrated based on known data may perform poorly on non-model species.

Variants can be called by (i) directly aligning the reads to a reference genome followed by comparing read sequences to the reference (Pabinger *et al.*, 2014, Olson *et al.*, 2015) or (ii) by assembling reads into a *de novo* assembly and comparing this assembly to a reference assembly (Olson *et al.*, 2015). The former approach is used to detect SNPs and small structural variants while the latter approach is more suitable for individual genes but not for SNPs in whole genome because raw reads cannot be used to differentiate true variants and erroneous calls (Olson *et al.*, 2015).

There are numerous tools available to call variants in NGS data. Different tools use different types of algorithms to call variants and so they may report different outputs from the same data. Two of the most popular variant callers, namely FreeBayes and GATK use a Bayesian approach to find variants. Other tools e.g. VarScan uses heuristic/statistical approach to detect variants and some like SNVer use frequentist approach (Sandmann *et al.*, 2017). Performance of variant callers can be compared by measure of their sensitivity and specificity, based on a data set with known variants (standard or truth dataset). However, there is no one best caller, and ultimately the properties of the reference genome, underlying short read data, and the mapped population will to some extent influence the performance of each caller. Variant callers calibrated to work best with model systems e.g. GATK may perform poorly on non-model species, for which known data sets are not available.

1.3.7. Variant filtration

False positive variants called by variant callers need to be filtered out based on certain parameters like, base call frequency, base quality score, mapping quality score, minimum depth of coverage, and masking of repetitive and homopolymeric regions, etc (Zook *et al.*, 2014, Olson *et al.*, 2015). A deep and even sequencing coverage is directly related with more accurate variant calling. If an annotated reference genome is available along

with some known variants then recalibration of variant quality scores can outperform above mentioned hard filters in increasing the accuracy of variant calls (Jun *et al.*, 2015). Detailed filtration strategies are reviewed by (Guo *et al.*, 2014, Wyllie, 2013).

1.4. Aims and objectives of the project

A comparison of genomes of AR37 isolated from natural host grass and from 2 artificial host grass cultivars using bioinformatic tools will be discussed in this thesis. This comparison of genomes may provide clues about the genetic determinants of compatibility in AR37 endophytes. Both artificial host grass cultivars have gone through different number of seed cycles and this information can help to identify the effect of seed cycles on AR37 endophyte adaptation. The main aim for this study is to find if asexual *Epichloe festucae* endophytes undergo genetic changes when introduced into a new host cultivar in order to adapt to the new host?

The following objectives were set for this study.

1.4.1. Extract good quality AR37 endophyte DNA from multiple individuals of two artificial host cultivars i.e. SAMSON and KLP1102 as well as from original host and sequence them.

1.4.2. Make a *de novo* AR37 genome assembly of a comparable quality to already sequenced related *E. festucae* strains i.e. E2368 and Fl1.

1.4.3. Compare genomes of *E. festucae* strain AR37 after they were challenged to adapt to two new host grass cultivars and had gone through different number of seed cycles to find:

i. What types of genes are important for interaction in general? Are these genes more important than others?

ii. How much does the endophyte change over time in new associations. Are new associations stable? What good or bad changes may one expect in the long run when releasing a new association and in AR37 specifically?

iii. How diverse does an endophyte strain become when introduced into a new host cultivar after it has gone through multiple seed cycles?

1.4.4. Compare AR37 genome with the genomes of closely related strain i.e. Fl1 and AR37 from different host species to identify the impact of long term adaptation (evolution).

2. Materials and Methods

2.1. Fungal strains

Fungal strains used in this study are listed in Table 2.1. Table 1.2 lists the history of the perennial rye-grass (PRG) host that has been extensively selected through breeding.

Table 2.1. Fungal strains and host plants used in this study

Fungal Strain	Current host	Source
Epichloë festucae (AR37)	*Lolium perenne* (SAMSON)	AgResearch
Epichloë festucae (AR37)	*Lolium perenne* (KLP1102)	AgResearch
Epichloë festucae (AR37)	*Lolium perenne* (Original)	AgResearch

2.2. Media

All media were prepared using MilliQ water and sterilized at 121 ^{o}C for 15 min prior to use. Solid media were prepared by adding 1.5% agar to the liquid media. Both liquid and solid media were cooled to 50 ^{o}C before antibiotics were added to them. Media plates were stored at 4 ^{o}C before inoculating them with fungal cultures.

2.2.1. Potato Dextrose broth (PDB) and Potato Dextrose Agar (PDA)

12 g of potato dextrose broth (PD; Difco Laboratories, Detroit, Mich) was added to 500 ml of water to get Potato dextrose broth. 7.5 g of agar was added to PDB to get PDA medium. Medium was agitated and boiled for around 1 min before autoclaving to dissolve the medium in water. Melted media was cooled to approximately 50 ^{o}C and then filter sterilized chloramphenicol was added to the media to give a final concentration of 25 µg/ml.

2.2.2. Water agar (4%) (Latch & Christensen, 1985)

24 g of standard agar was put in 600 mL of RO water to yield 4% water agar.

2.3. Buffers and solutions

IUPAC standard buffers (Radiometer Analytical manufactured by HACH LANGE GmbH, Berlin) were used to calibrate pH meter before adjusting pH of solutions. MilliQ water was used to prepare all solutions and buffers used in DNA extraction protocols and gel electrophoresis. The pH of solutions was measured with a pH meter (Model PHM210, Radiometer Copenhagen).

2.3.1. Tris-EDTA buffer (TE buffer, pH 8)

10 mM Tris-HCl
0.1 mM EDTA

2.3.2. Lysis buffer

40 mM Tris-acetate, pH 7.8
20 mM Na-acetate, pH
1 mM EDTA
1% SDS
Prepared by mixing 10x concentrated stock solutions of i) Tris-acetate ii) sodium acetate iii) EDTA and iv) SDS. All solutions except SDS were autoclaved.

2.3.3. CTAB extraction solution

2% (w/v) CTAB
100 mM Tris.Cl, pH 8.0
20 mM EDTA, pH 8.0
1.4 M NaCl
Store at room temperature (stable several years)

2.3.4. CTAB precipitation solution

1% (w/v) CTAB
50 mM Tris.Cl, pH 8.0
10 mM EDTA, pH 8.0
Store at room temperature (stable several years)

2.3.5. High salt TE buffer

10 mM Tris-HCl, pH 8.0
0.1 mM EDTA, pH 8.0
0.2 M NaCl
Store at room temperature (stable several years)

2.3.6. CTAB / NaCl solution

10% CTAB in 0.7 M NaCl
Dissolved 4.1g NaCl in 80 ml water and slowly added 10 g CTAB while heating and stirring. Final volume was then adjusted to 100 ml.

2.3.7. TBE buffer

At 1x concentration the TBE buffer contained
89 mM Tris
89 mM Boric acid
2.5 mM Na_2 EDTA, pH 8.2
10x TBE stock solution was prepared and stored at room temperature. This stock solution was diluted ten times in RO water to get 1x TBE buffer.

2.3.8. TAE buffer

At 1x concentration the TAE buffer contained
40 mM Tris
20 mM Acetic acid

1 mM EDTA, pH 8.0

50x TAE stock solution was prepared and stored at room temperature. The stock solution was diluted with RO water to get 1x TAE freshly before use.

2.3.9. Chlorine bleach

Commercially available bleach (Janola) was diluted ten fold in milliQ water.

2.3.10. Aniline Blue

Contained 0.1% (w/v) aniline blue (Michrome) in bleaching solution (2.5 g chloral hydrate per ml water)

2.3.11. 10x Loading buffer

20% Ficoll 400

0.1 M Disodium EDTA, pH 8

1.0% Sodium dodecyl sulfate

0.25% Bromophenol blue

0.25% Xylene cyanol (was used only for monitoring very long runs)

2.3.12. Ethidium bromide solution

Stock solution:

10 mg/ml in water(Sigma)

Working solution:

Diluted stock solution 1:1000 for gels or stain solutions

2.3.13. Blocking solution

2.42 g Tris (hydroxymethyl) methylamine

2.92 g Nacl

5 g Non-fat milk powder

10 ml 1M HCl, pH 7.5

2.3.14. Chromogen

Two different solutions were combined to form the chromogen.

Solution A:
20 mg Fast Red TR
12.5 ml Tris buffer, pH 8.2

Solution B:
12.5 mg Naphthol as-MX phosphate
12.5 ml Tris buffer per 10 cm^2 of nitrocellulose membrane

2.4. Growth and maintenance of plant and fungal cultures

2.4.1. Obtaining *Epichloë* endophytes

All endophyte strains used in this work are derived from AR37. This included AR37 re-isolated from its original host plant (referred to in this thesis as AR37-Orig) maintained at AgResearch, Palmerston North, and isolated 3 weeks prior to use in this study. AR37-Orig culture was provided by Anouck de Bonth from AgResearch Grasslands, Palmerston North. AR37 infected seeds from perennial ryegrass cultivars Grasslands SAMSON and KLP1102 (referred to as AR37-SAM and AR37-KLP, respectively here in this thesis) were kindly provided by Dr. Richard Johnson from AgResearch, Palmerston North. These were most recent seeds after 2-6 seed cycles for KLP1102 and SAMSON respectively as per AgResearch records. AR37-SAM cultivar has been intensely selected for transmission and storage, so any genetic changes favourable to these traits may have been selected in endophyte also (Table 1.2).

Fifty AR37 infected seeds of each cultivar were grown on 4% water agar in Lab to avoid contamination by any other fungus. They were incubated in dark for 7-10 days at 22 ^oC and seedlings were then incubated in light for next ten days. Seedlings were then potted in root trainers in potting mixture inside a green house.

2.4.2. Plant culture and maintenance of symbioses

Plants were grown in ¾ polythene bags (PB). Potting mix was obtained from AgResearch (AgResearch Grasslands, Palmerston North) and Osmocote slow-release fertilizer was added to it. Pots were watered every alternate day. Plants were re-potted every six months to keep a steady supply of actively growing plants. In order to re-pot plants whole plant was removed from the pot and 6 healthy tillers were selected for re-potting. Old brown leaves and sheaths from these selected tillers were removed and remaining leaves and roots were trimmed to 2-4 cm length. The tillers were then inserted into the fresh potting mix to a depth of 2 cm and immediately watered to saturation and left in the green house. Plants were sprayed once with Vydate (DuPont) for control of insects/aphids.

2.4.3. Isolation of fungus from plant tissues (Christensen *et al.*, 2002)

Endophytes were isolated using procedure described by Christensen et al. (Christensen et al., 1998). All procedures were carried out using aseptic technique. Two tillers were cut from the base of each of the plants, about 0.5 cm above root level. Cut ends were trimmed and any necrotic tissue was removed. Remaining tillers (around 3 cm or so) of each plant were placed in a separate clean McCartney bottle and labelled. Tissues were surface sterilized with 95% ethanol prior to incubation with 10% domestic bleach (Janola) for 3-5 minutes. The tillers were then rinsed twice in sterile water and allowed to dry on filter paper. Both ends of the tillers were cut and discarded. Transverse segments (~2 x 2 mm) were cut along the length of each of remaining tiller. Cut segments were carefully manipulated to separate different layers (3-4) depending on the age and size of tiller. Separated sections were placed onto PDA plates (9-16 sections/plate) containing 25 µg/mL of chloramphenicol. The plates were incubated at 22 ^{o}C in the dark for 8-10 weeks. Once fungal colonies had grown around 20-25 mm in diameter, then were then sub-cultured. Initially 50 infected plants from each AR37-SAM and AR37-KLP (referred to as AR37-SAM1….AR37-SAM50 and AR37-KLP1…AR37-KLP50 respectively in this thesis) were chosen for isolation of the AR37 endophytes.

2.4.4. Maintenance of endophyte cultures (Simpson *et al.*, 2012)

Due to difficulties in extracting pure high molecular weight DNA from these endophytes, cultures (original stock) were continuously maintained to provide fresh material for new

extractions. Each AR37 isolate was at least sub-cultured twice to make sure that it was free of any impurity. In order to sub-culture, small pieces of mycelium were taken from the periphery of growing colony and transferred to fresh PDA plate using aseptic technique. These plates were again incubated in dark at 22 ºC. Small colonies developed within 3-5 weeks after incubation. The plates were constantly checked and if any bacterial or fungal contaminations were observed the plates were discarded and fresh plates were prepared for the discarded sample from the original stock.

2.4.5. Inoculation of endophyte into liquid media

All steps were carried out inside Laminar flow safely cabinet. A single colony of each isolate, growing on PDA plate for 8-10 weeks at 22 ºC and about 15-20 mm in diameter was scrapped from the surface of the agar using a sterile scalpel and shifted to a sterile petri dish. Sterile spatula and scalpel were used to remove agar from each colony. Mycelium was then cut into very fine pieces using a sharp sterile scalpel. These small pieces were then put into a 100 ml flask containing 50 ml sterilized PDB media. The flasks were labelled and placed at 22 ºC under both shaking (200 rpm) and non-shaking conditions. The growth was daily monitored and any flask showing signs of contamination was discarded and replaced by a fresh culture flask.

2.4.6. Harvesting from liquid media

Once enough growth had been observed in a flask, the cultures were harvested for further processing. Whole liquid media along with fungal cultures were passed through a sterile muslin cloth used as a filter paper. The media passed through the cloth leaving fungal mycelia at the top. Mycelia were collected in a sterile polypropylene tube and placed at -80 ºC for up to 12 hours. Holes were made in lids of all these tubes and samples were subjected to lyophilization for 20-24 hours. Lyophilized samples were put in a container containing silica gel and stored in -80 ºC freezer.

2.5. Endophyte detection

2.5.1. Immunoblotting (Hiatt *et al.*, 1997, Simpson *et al.*, 2012)

At 3-4 tiller stage, one of the tillers of planted seedlings was excised from the base (~5 mm above the soil level) with the help of a scalpel. The cut end of the tiller was pressed against a nitrocellulose membrane (NCM) so that circular outline of the cut end was formed on the NCM. Known endophyte infected and uninfected plants were also blotted on the same NCM sheet and acted as positive and negative controls respectively. Blotted NCM was either processed immediately or kept in refrigerator at 4 ^{o}C in dark and processed next day. The membrane was immersed for 2 hours in milk protein blocking solution (BS) (2.42 g Tris (hydroxymethyl) methylamine, 2.92 g NaCl, 5 g Non-fat milk powder, 10 ml 1 M HCl, pH 7.5) and shaken in an orbital shaker at room temperature. NCM were then rinsed twice with fresh BS. 25 µl of Primary antibody (Produced at AgResearch in collaboration with Massey University) was diluted in 25 ml BS (1:1000) and NCM sheets were then transferred to this solution and shaken for 15 minutes at room temperature and then incubated overnight at 4 ^{o}C. The sheets were rinsed twice with fresh BS to remove excess primary antibody and transferred to another solution containing 6.25 µl of secondary antibody (goat anti-rabbit IgG-AP, sc-2034, Santa Cruz Biotechnology, USA) diluted to 1:4000 in 25 ml BS, and shaken at room temperature for 15 min and then incubating at 4 ^{o}C for 5 h. NCM sheets were rinsed twice with BS to remove excess secondary antibody. Two different solutions were combined together to form the chromogen. The first solution was made by dissolving 20 mg Fast Red TR (Sigma F-2768) in 12.5 ml Tris buffer adjusted to pH 8.2. The second solution was made by dissolving 12.5 mg of naphthol As-MX phosphate (Sigma N4875) in 12.5 ml Tris buffer per 10 cm^2 of NCM). Both solutions were then combined and NCM sheets were immersed in it and shaken for 15 min at room temperature. The membranes were rinsed thrice with water to stop the reaction. Dark pink colour spots indicated endophyte positive tillers.

2.5.2. Aniline blue staining (Latch & Christensen, 1985)

Epidermises from the outer sheaths of host grass plants were peeled off, mounted on microscopic slides in aniline blue (0.05% aniline blue in lactic acid glycerol/water 1:2:1) and covered with a cover slip. Air bubbles were removed by briefly heating the slide and

slide was then observed under Zeiss compound microscope at 100x and 400x magnifications.

2.6. DNA extraction and quantification

Different approaches were used to get DNA out of endophytes.

2.6.1. Extraction using DNeasy® Plant Mini Kit (QIAgen)

QIAgen DNA extraction kit (DNeasy® Plant Mini Kit) was first used to extract fungal DNA. 100 mg of fresh fungal mycelium and 30 mg of lyophilized mycelium, isolated from each association and grown over a period of 8-10 weeks on PDA plates was used to extract DNA as per manufacturer's instructions. Fresh (100 mg) and lyophilized (50 mg) materials from liquid cultures grown over a period of 4-5 weeks in liquid culture were also used with QIAgen kits. In order to use as control, DNA from 2 grass samples was also extracted along with AR37 endophytes using QIAgen kit. Samples were stored in TE buffer at 4 ^{o}C and quantified using Qubit fluorometer.

2.6.2. Al-Samarrai method (Al-Samarrai & Schmid, 2000)

Fresh (100 mg) or lyophilized (50 mg) material grown in either PDA or PDB were ground in liquid nitrogen and lysed with 500 µl of freshly prepared lysis buffer. 165 µl of 5 M NaCl was mixed with the samples till the solutions became viscous. The samples were centrifuged at 13000 rpm (15493 x g, Sigma, rotor: 12024-H) for 15 min. The supernatant was mixed with equal volumes of phenol and chloroform and the samples were again centrifuged and clear top layer was transferred to a fresh tube. Additional lysis buffer was added and centrifugation steps repeated once more. Finally 2 volumes of cold 95% ethanol were added to the clear top layer and mixed gently with hand. Samples were centrifuged for 2 min and rinsed thrice with cold 70% ethanol, dried and suspended in 50 µl sterile Tris-EDTA (TE, pH 8) buffer.

2.6.3. Cetyl Trimethyl Ammonium Bromide (CTAB) method

2- Mercaptoethanol (2-ME) was added to CTAB extraction solution to give a final concentration of 2% (v/v) and this 2-ME/CTAB extraction solution was heated to 65 °C. Lyophilized (50 mg) mycelium of each isolate was ground in liquid nitrogen and 500 µl of heated 2-ME/CTAB extraction solution was then added to it and mixed thoroughly. Samples were then extracted with equal volume of 24:1 Chloroform/Isoamyl alcohol and the supernatant was transferred to a new tube and 1/10 volume 65 °C CTAB/NaCl solution was added followed once again by extraction with 24:1 Chloroform/Isoamyl alcohol. The supernatant was again transferred to fresh tube and was mixed with exactly 1 volume CTAB precipitation solution by inversion followed by centrifugation at 500 g (Sigma, rotor: 12024-H) for 5 min. The pellet was suspended in high salt TE buffer and DNA was precipitated using 0.6 volume isopropanol followed by centrifugation for 15 min at 7500 g (Sigma, rotor: 12024-H). Pellet was washed with 80% ethanol, dried and re-suspended in 50 µl Tris-HCl buffer (pH 8).

2.6.4. RNAse treatment

In order to make sure that DNA samples were not contaminated by RNA, DNAse-free RNAse was used. For DNA extractions using QIAgen kits, RNase provided with the kit was used as per the manufacturer's instructions. For Al-Samarrai and CTAB methods, RNase A manufactured by Roche Diagnostics Gmbh (25 mg form bovine pancreas) was used. RNase free of DNase was made by dissolving the RNase (1 mg/ml) in TE buffer and boiling for 30 min. Aliquots were stored at -20 °C. To get rid of RNA, RNase A was added in DNA preparations at a final concentration of either 10 mg/ml or 100 mg/ml and mixture incubated at 37 °C or 65 °C for 5-60 min. RNase was removed by extraction with buffered phenol (Invitrogen Ultra-pure™ buffer-saturated phenol) and chloroform / isoamyl alcohol.

2.6.5. Purification and concentration of DNA

To obtain a concentration of DNA high enough for the sample to be sequenced using high throughput sequencing (minimum 20 ng/µl, and 1.5 µg total), different DNA samples of the same isolate obtained from multiple extractions were pooled together and

concentrated using ethyl alcohol precipitation technique as given in current protocols. The DNA solution is first extracted with Phenol / Chloroform / Isoamyl alcohol mixture and then precipitated with 100% ethyl alcohol. DNA, which was now in pellet form, was washed thrice with 70% ethyl alcohol, dried and re-suspended in 50 µl of Tris-HCl buffer.

2.6.6. Fluorometric quantification of DNA concentration

A fluorometer (Qubit, Invitrogen) was used to quantify DNA using Quant-iT dsDNA BR assay kit (Invitrogen), as per manufacturer's instructions. Standards provided with the kit (Qubit® dsDNA HS Standard 1 and 2) were used.

2.7. Polymerase Chain Reaction (PCR) amplification (Davis *et al.*, 2012)

PCR was used to confirm the identity of AR37 endophyte.

2.7.1. Primers

Primers specific to rearranged region of perA locus that amplify AR37 were designed by Yanfei Zhou (Rosie Bradshaw; unpublished data) (Table 2.2). Primers were manufactured by Invitrogen and reconstituted to 100 µM final concentration in sterilized MilliQ water. Primers were diluted to a working concentration of 10 µM and stored at -20 ^{o}C.

Table 2.2. Sequence of primer pair used to identify AR37. The primers are designed for a rearranged region of *perA* genes.

Name of ORF	Primer Name	Primer sequence (5'-3')
perA	*perA*-AR37-forward-RT	CAGACTGAATGTGGAGATAAG
perA	*perA*-AR37-reverse-RT	CATAAGATCACTACCGACAAG

2.7.2. PCR reactions (Davis *et al.*, 2012)

Standard PCR amplification of genomic DNA was carried out in 10 µl reaction volumes containing 200 µM dNTPs, 10 µM of each primer, 0.5 U Taq DNA polymerase (Roche) in 1x PCR buffer (Roche) and 0.5 µl of template DNA using BioradiCycler and following conditions: One cycle at 94 ºC for 2 min followed by 30 cycles at 94 ºC for 30 sec, 57 ºC for 45 sec and 72 ºC for 1 min, with final extension at 72 ºC for 7 min. A pure sample of already identified AR37 genome was used as a positive control.

2.8. Agarose gel electrophoresis (Johansson, 1972)

The PCR products were tested on 0.7-2% agarose gel by electrophoresis. Tris Boric acid EDTA (TBE) buffer was used for electrophoresis, however for all experiments where attempts were made to extract DNA from gel, Tris Acetic acid EDTA (TAE) was chosen as electrophoresis buffer. For minigels 1 g agarose was dissolved in 50 ml of 1 x TBE / TAE by heating in microwave. For larger gels 2 g agarose was dissolved in 100 ml of TBE / TAE buffer. Molten agarose was cooled to 50 ºC before pouring it. Gels were run in TBE / TAE electrophoresis buffer and 1 µl of Bromophenol Blue loading buffer was added to all samples. In gels where high molecular weight DNA was to be visualized, a mixture of Bromophenol blue and Xylene Cyanol was used. Gels were stained with ethidium bromide (0.1 µg/ml) for half an hour and then de-stained in water. Gels were then visualized in trans-illuminator gel documentation system (Bio-Rad) and photographed. 1 kb plus DNA ladder was used as a standard to estimate the size of PCR products.

2.9. Pooling of samples

Concentrations of all DNA samples was equalized by comparing the DNA band of intensity following electrophoresis. For pooling, initially 22 individual AR37 DNA samples from each cultivar, containing approximately equal amounts of high molecular weight DNA, were pooled together. However, most of the pooled DNA did not move out of the wells and hence no high molecular weight DNA band was observed. Finally, only

eleven best AR37 DNA samples of each cultivar were pooled together. This time pooled DNA also showed a high molecular weight band in the gel.

2.10. Gel extraction

To get enough amount of high molecular weight DNA, DNA from multiple extractions of each isolate were pooled together and run on 0.7% gel. Multiple gels were run with different amounts of pooled DNA loaded ranging from 2 µg to 10 µg per well. In order to load this amount in one well, pooled samples had to be concentrated using phenol-ethanol precipitation protocol. The high molecular weight bands were then cut from the gel using a sharp sterile scalpel under UV light and DNA re-extracted using Roche high pure PCR purification kit, as per manufacturer's instructions.

2.11. High throughput sequencing of fungal genomes

DNA extracted from 11 AR37 endophytes isolated from AR37-SAM and AR37-KLP each was pooled together to make one sample each (total 2 samples). This pooling of DNA from many individuals of each strain is to ensure the full representation of variations within populations. DNA isolated from multiple extractions of original AR37 was also pooled together and served as standard. 2.5 µg DNA with a minimum concentration of 150 ng/µl of all three pooled samples were sent to New Zealand Genomics Limited (NZGL) for sequencing. Sequencing was done on an Illumina MiSeq platform using TruSeq PCR-free libraries.

2.12. Removing adapter sequences and correcting and trimming reads

Fastq-mcf tool from the ea-utils was used to remove adapter sequences from these reads. Resulting sequences were also trimmed to their longest contiguous segment with a quality score less than 0.01 using SolexaQA++ software. If whole read had to be trimmed away then only one base was left in place of that read to maintain the order of the reads.

2.13. Making a *de novo* genome assembly

The following assemblers were used in this project: Reads from all the libraries were used together to make a reference assembly. Assemblers included MIRA, SPAdes, SOAPdenovo, Velvet, ABySS and A5-MiSeq. Following commands were used to assemble genome from trimmed reads:

2.13.1. SPAdes

SPAdes-3.9.0 was used to make reference assembly at kmer values of 21, 33, 55, 77, 99 and 127. Following command type was used to assemble the genome from all the sequencing reads.

Spades.py –k 21,33,55,77,99,127 –careful –pe1-1 AR37-Orig-lib1-read1.fasta –pe1-2 AR37-Orig-lib1-read2.fasta –pe2-1 AR37-Orig-lib2-read1.fasta –pe2-2 AR37-Orig-lib2-read2.fasta –pe3-1 AR37-KLP-read1.fasta –pe3-2 AR37-KLP-read2.fasta –pe4-1 AR37-SAM-read1.fasta –pe4-2 AR37-SAM-read2.fasta –o AR37-reference-assembly

2.13.2. Velvet

Velvet 1.2.10 was used for assembly. Velvet is used in two steps. In first step velveth is used to process reads at different kmer values and then velvetg makes an assembly of each kmer value. I used velvet at kmer values of 33, 55, 77, 99, and 121. The following commands were used for velveth and velvetg steps:

Velveth AR37-assembly/ 33,127,22 –fastq –shortPaired All-Reads-1.gz All-Reads-2.fastq.gz
Velvetg AR37-assembly/_33 (to make assembly at kmer value of 33)

2.13.3. ABySS

ABySS version GNU MAKE 3.82 was used for assembly. Kmer values of 96 and 128 were used for ABySS. Following command was used.

Abyss-pe name=AR37-128 k=128 in= ‘All-reads-1.fastq.gz All-reads-2.fastq.gz’ contigs 2>&1 | tee ABySS.log

2.13.4. MIRA

Manifest file was set as follows:

project = AR37BothReads
job = genome,denovo,accurate
parameters = COMMON_SETTINGS -GE: -AS: nop=5 -NW:cnfs=warn -NW:cac=warn SOLEXA_SETTINGS -AL:mrs=90 -AS:mrl=40
readgroup = SolexaLib1
data = AR37_all_R1_001.fastq AR37_all_R2_001.fastq
technology = solexa
template_size = 50 1000 autorefine
segment_placement = FR
rename_prefix = M00933:80:000000000-AENKN: AR37-1
segment_naming = solexa

2.13.5. A5-Miseq

A5-miseq version 20160825 was used as part of an assembly pipeline that incorporates other third party tools including sspace and idba-ud for assembly and is specially designed for MiSeq data from haploid organisms. Following command was used for A5-miseq:

a5-pipeline.pl library.file AR37.out

Where library file contains path to all AR37 libraries and AR37.out is the base name for all output files by the assembler.

2.13.6. SOAPdenovo

A configuration file (config_file) was needed to assemble reads into a genome using SOAPdenovo. All values in configuration file were set to default and following command was used to make assembly.

all -s config_file -K 63 -R -o graph_prefix 1>ass.log 2>ass.err

Assembly was generated using two different values of -K i.e. 63 and 128

2.14. Assessing quality of the *de novo* genome assemblies

Quast version 4.6.3 was used to analyse the quality of the assembly based on assembly statistics Software was downloaded from http://quast.sourceforge.net/quast and installed as per instructions. An example command is shown below:

Quast.py –l AR37-velvet,AR37-ABySS,AQR37-SPAdes,AR37-MIRA,AR37-A5-miseq Velvet-assembly.fasta ABySS-assembly.fasta SPAdes-assembly.fasta MIRA-assembly.fasta A5-miseq-assembly.fasta –o assembly-comparisons

Another software tool called Benchmarking Universal Single-Copy Ortholog assessment tool (BUSCO) version 3.0.1 was also used to assess the quality of the assemblies. I used BUSCO v3 using Ubuntu virtual machine which had all dependencies included and is easier to install and operate. The software is freely available at https://busco.ezlab.org. A script i.e. “run_BUSCO.py” was run with following options to calculate the completeness of the assembly.

Python scripts/run_BUSCO.py -i sequence-file -o output-name -l Lineage -m geno

For option -l, sordariomycetes was selected as lineage.

2.15. Aligning reads to reference assembly

Three softwares were used to align reads to the reference assembly i.e. Bowtie2, Bwa-mem and Novoalign

2.15.1. Bowtie2

Bowtie2 version 2.3.4.1 was used to map reads against the reference. The software is freely available from https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.3.4.1. Bowtie2 needed to build an index before mapping reads. The following commands were used to build index and then map AR37-Orig reads:

Bowtie2-build –f AR37-assembly.fasta AR37-assembly
Bowtie2 –p 16 –X 1000 –end-to-end –very-sensitive –x AR37-assembly -1 AR37-Reads-1.fasta -2 AR37-Reads-2.fasta –S AR37-bowtie2.sam

2.15.2. Bwa-mem

Bwa version 0.7.17.r1188 was used. It also needs indexing the genome before mapping reads against it. Commands used were:

Bwa index AR37-assembly.fasta
Bwa mem AR37-assembly.fast AR37-reads-1.fasta AR37-reads-2.fasta > AR37-bwa.sam

2.15.3. Novoalign

Novoalign version 3.08.02 was used for mapping reads against reference. As for bowtie2 and bwa, Novoalign also needs reference genome to be indexed. Following commands were used for indexing and mapping:

Novoindex AR37.nix AR37-assembly.fasta

Novoalign –d AR37.nix –f AR37-reads-1.fasta AR37-reads-2.fasta –I PE 500,200 –o SAM > AR37-novo.sam

2.16. Converting .sam to .bam files and sorting and indexing them

Samtools version 1.9 was used to convert .sam files to a sorted .bam files using following command:
samtools view –bS AR37-bowtie2.sam > AR37-bowtie2.bam

.bam files were then sorted using following command:
samtools sort AR37-bowtie2.bam > AR37-bowtie2-sorted

To generate an index for the sorted bam files I used following command:
Samtools index AR37-bowtie2-sorted-rg.bam

2.17. Adding read groups

bamaddrg was used to add read groups to the sorted.bam files using the following command:
bamaddrg –s AR37vsRef-AR37 –b AR37-bowtie2-sorted.bam > AR37-bowtie2-sorted.rg.bam

2.18. Variant calling

FreeBayes version v1.0.2-15-g357f175 and CRISP were used to call variants. Command lines used are as follows.

FreeBayes –p 11 –q 20 –m 15 – C 1 –F 0.05 –min-coverage 10 –pooled-continuous –f AR37-assembly.fasta –b AR37-bowtie2-sorted-rg.bam –b KLP1102-bowtie2-sorted.rg.bam –b SAMSON-bowtie2-sorted.rg.bam > FreeBayes-bowtie2.vcf

CRISP --poolsize 11 --mbq 20 --mmq 15 --minc 1 --bam AR37-bowtie2-sorted-rg.bam --bam KLP1102-bowtie2-sorted.rg.bam --bam SAMSON-bowtie2-sorted.rg.bam --ref AR37-assembly.fasta --VCF Crisp-bowtie2.vcf > Crisp-bowtie2.log

2.19. Variant filtration

Multiple softwares were used to filter variants. Some of the tools are listed below

SnpSift

Vcflib

Bcftools

Bedtools

GATK

3. RESULTS

3.1. Overview of experimental strategy

Achieving the aims of this study (outlined in Aims and Objectives), required me to determine to what degree and how the endophyte AR37 had changed, genetically, after its introduction into two new cultivars, SAMSON and KLP1102 and its subsequent seed propagation for 6 and 2 generations, respectively, in these cultivars. Both the cultivars were artificially inoculated with the AR37 strain for the same amount of time. Any differences in the AR37 isolates from both these cultivars after same amount of time but different host seed cycles may point towards the effect of host seed cycles (sexual reproduction) on the endophyte strain.

To do so it was necessary to (i) obtain a good quality assembly of the original (ancestral) AR37 genome, (ii) obtain and sequence DNA from a number of serially seed-propagated AR37 endophyte clones and to (iii) detect how the clones had changed by comparing their sequences with that of their ancestor; to do so in a cost-effective manner I aimed to sequence pools of clones rather than a number of individual clones, an approach that has been successfully used for a similar type of analysis, namely quantitative trait mapping in rice (Takagi *et al.*, 2013).

As a reference for detecting genetic changes during serial propagation of AR37 in the new cultivars, sequencing DNA from the ancestral AR37 (isolated from the original host at the time of introduction of AR37 into the new cultivars) would have been ideal. This DNA was not available, but what was available was a recent AR37 culture isolated by Anouck de Bonth (AgResearch Grasslands, Palmerston North) from vegetatively propagated plants of the same original European *L. perenne* / AR37 association. This original host had not gone through any seed cycle and only endophyte positive tillers from the host were used in vegetative propagation. The AR37 in these original host plants is nearly in the same environment and should at least be largely identical to the ancestral AR37. There may have been a few random mutations in this AR37 during vegetative propagation in all these years but it should be the closest to the ancestral AR37 as compared to the AR37 isolated recently from the AR37-SAM and AR37-KLP

associations. This AR37 was termed as AR37-Orig. An assembly of AR37-Orig was generated using SPAdes, however, the assembly was quite fragmented and was of lesser quality than other *E. festucae* genomes available online. Therefore I used sequencing reads from all three samples i.e. AR37-Orig, AR37-SAM and AR37-KLP to generate another *de novo* assembly (AR37-pool) to use as a reference. This assembly would largely be the same

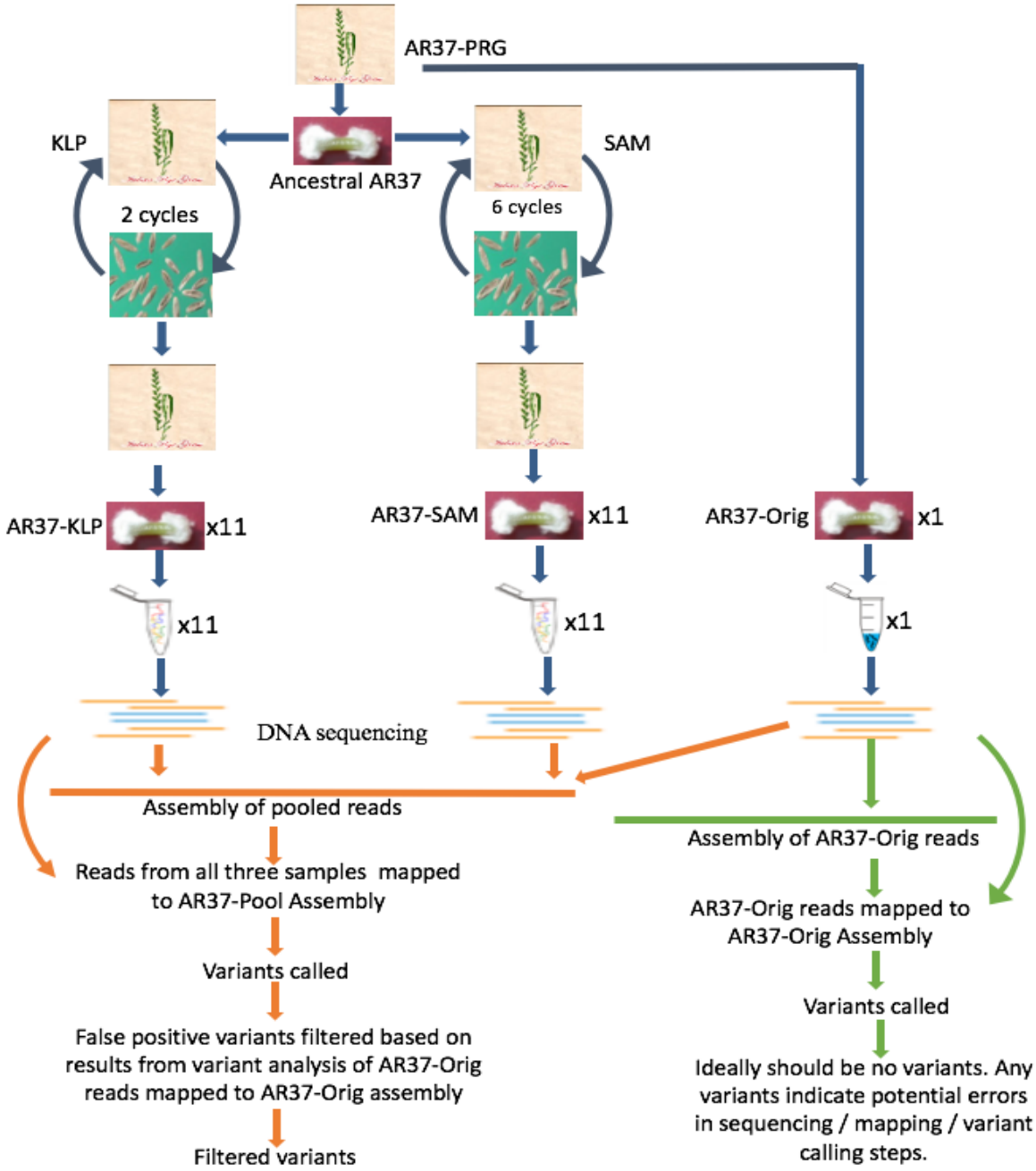


Figure 3.1. An overview of the experimental strategy. Blue arrows indicate the steps taken to obtain sequencing reads. Orange arrows indicate the steps to call variants from the pooled samples. Steps indicated by green arrows were taken to do "sanity check" of the variants.

as AR37-Orig because of the higher coverage of the AR37-Orig sample than either of the pools. The logic for using all the reads from all the samples for constructing the *de novo* AR37-pool assembly is discussed in section 3.4.1. This assembly was of better quality and comparable to the already sequenced *E. festucae* assemblies available online and thus was used as a reference assembly. Reads of all samples were then separately mapped against this assembly and variants called. Although the AR37-Orig assembly was not used as a reference assembly, mapping only AR37-Orig reads against this particular assembly and calling variants using the same tools as for the AR37-pool assembly provided useful information about the potential sequencing and mapping errors in our study (Figure 3.1 green arrows) thus helping to perform a "sanity check" of the final data set of variants.

3.1.1. Expectations from the study design

These particular samples were chosen for this study because they initially showed signs of mild incompatibility which improved over time and by the time the samples were sequenced, the associations appeared to behave normally. A comparison of these improved samples with that of ancestral AR37 would have been a better choice to underpin the genetic changes that may be associated with the compatibility but ancestral association or AR37 DNA from this association was not available. In absence of ancestral AR37, the closest sample to ancestral one was chosen i.e. AR37-Orig which was obtained from ancestral association propagated vegetatively through all these years. Following was expected from this study design.

3.1.1.1. If endophyte-host compatibility depends on one or few genes then such genes may be detected

A comparison of pools that have improved in compatibility to the AR37-Orig (or ancestral AR37) should reveal genetic changes that may be involved in determining compatibility to the host. However, such changes may only be easily detected if only one or a few genes are involved in compatibility. In such case, it is expected that all the clones that have survived and shown improvement in compatibility will have changes in those particular genes and the variations will be easy to detect due to their high frequencies (all or most clones will have same variation). The more likely scenario however is that

compatibility may depend on a host of genes and multiple different pathways as evidenced by multiple studies in which associations were disrupted by manipulating an array of genes. If so then different clones may have undergone different variations to adapt to the same host. In such case variations within each sequenced pool may be confined to single clone (individual) and hence only supported by few sequence reads (< 9%). Such low frequency variants are hard to detect.

3.1.1.2. AR37-Orig served as a reference

Since AR37-Orig did not have to adapt to a new host as it remained within its natural host that was only clonally (vegetatively) propagated so it served more like a reference standard. Any variations that are present in both the pools as well as AR37-Orig are likely to be sequencing / mapping errors. Only variations that are present in one or both of the pools but absent from AR37-Orig may be important from compatibility point of view.

3.1.1.3. AR37-Orig should have least number of variations

AR37-Orig sample was obtained from its natural perennial ryegrass host that has been maintained through vegetative propagation at AgResearch Palmerston North. Since this ryegrass host has not undergone any sexual cycle and has only been clonally propagated during all these years it was expected that its symbiont i.e. AR37-Orig should have least number of variations because it was well adapted to its natural host and did not have to adapt to a new host or somewhat changed host progeny. Both the pooled samples on the other hand came from artificial hosts to which they were not adapted. Also both pools had 11 clones each and all of these clones were isolated after their hosts has gone through multiple seed cycles (sexual cycles). So more variations were expected in each of the pools as compared to AR37-Orig because AR37 in these associations has to cope with a relatively different host after each sexual cycle.

3.1.1.4. Variants in AR37-Orig were expected to be of high frequency

AR37-Orig sample was isolated from a single clone obtained from its natural host and it was expected that any variations present in AR37-Orig sample should be indicated by all (or most of) the reads. On the other hand, any variation in a single clone of the pool should

be represented by 1/11 or 9% of the reads (given the equal coverage of each clone). For the pools the only possibility for all the reads to support the same variant would be when all the 11 clones within the pool have same variant at the same locus, which may be highly unlikely. It meant that it should be relatively easier to detect variants in AR37-Orig sample (~100% frequency) than to detect in AR37-pools (~9-100%).

3.1.1.5. Low frequency variants were expected for both the pooled samples

The 11 clones in each pool were randomly picked from a large population to better reflect the genome wide variations in the populations. This random picking meant that each clone can have a variation of its own irrespective of the variations in other clones. As AR37 only reproduces asexually, so beneficial mutations in different clones cannot recombine and due to clonal interference, different clones may compete with each other for selection. Over long period of time, the clone with best combination of mutations may out-compete other clones to establish itself. However, samples used in this study originated only around 15 years before they were sequenced which is quite a short time for any clone to dominate. The expectation was that different clones with variations of their own in each pool may be present and it was highly unlikely for a particular variation to be present in all the clones.

3.1.1.6. Effect of host sexual reproduction on the endophyte can be monitored

The study design also allowed to monitor the changes in AR37 symbiont in response to sexual reproduction in host. AR37-Orig, isolated from its natural host that has only been vegetatively propagated should serve as a reference again. AR37-KLP and AR37-SAM clones have been isolated from artificial (novel) hosts that have undergone two and six sexual cycles respectively. If sexual cycle in host plant brings any changes that force AR37 endophyte to adapt to relatively newer progeny of the host then it is expected that AR37-SAM should have more mutations than AR37-KLP and AR37-Orig should have least number of mutations. If sexual cycles in host do not bring any changes in AR37-symbiont then expectation is to find similar number of mutations in both pooled samples. These mutations will represent the changes needed to adapt to the new cultivars. AR37-Orig should still show no (or least) mutations as it was isolated from its original host and did not have to adapt to a new host.

3.2. Generation of endophyte DNA samples for sequencing

3.2.1. Obtaining biomass of AR37 colonizing the original European plant host (AR37-Orig) for extracting DNA

The AR37-Orig culture had been obtained by placing plant tissues on potato dextrose agar (PDA) plates (Anouck de Bonth, personal communication) and it is not impossible that the mycelium growing out of these samples could have harboured additional contaminant fungi or bacteria. In order to assure the purity of this culture, it was sub-cultured twice on PDA plates (see section 2.4.4) and screened for the appearance of morphologically distinct colony sectors or colonies that would indicate presence of such contaminants. None were observed. AR37 grew very slowly on PDA plates. Much of the colony growth was vertical and over time the growing colony appeared to lose contact with the PDA plate and thus the nutrients it contained. To achieve faster growth and sufficient biomass, one colony from PDA plate was cut into very small pieces and these were transferred to liquid cultures, potato dextrose broth (PDB), and allowed to grow for ~8 weeks (see Materials and Methods for details). Fungal biomass from these cultures was harvested by centrifugation, lyophilized and stored at -80 ^{o}C for subsequent use in DNA extraction.

3.2.2. Obtaining biomass of AR37 from AR37-infected SAMSON and KLP1102 plants for extracting DNA

My original plan was to obtain biomass from approximately 30 AR37 clones each (referred to in the following as AR37-SAM and AR37-KLP clones) from 30 different AR37-infected SAMSON (SAM) and 30 AR37–infected KLP1102 (KLP) plants, to extract DNA from each and to generate one AR37-SAM and one AR37-KLP pool containing equimolar amounts of high molecular weight DNA from each of these 30 clones for sequencing .

AR37-infected seeds from perennial ryegrass cultivars Grasslands SAMSON and KLP1102 were kindly provided by Dr. Richard Johnson (AgResearch Grasslands, Palmerston North). To mitigate instances of fungal or bacterial contamination or the spontaneous loss of the endophyte in the seed, 100 seeds of each cultivar were grown on

4% water agar plates as described in Materials and Methods). Seedlings showing any signs of epiphytic fungal growth or bacterial growth were discarded. The remaining 52 SAM and 48 KLP seedlings were then placed in potting mix in a green house. All but six seedlings of AR37-SAM and nine seedlings of AR37-KLP grew into multi-tillered plants. After 4 weeks in the green house, one tiller of each of these 46 SAM and 39 KLP plants was ELISA tested for the presence of endophyte as described in Materials and Methods. Thirty-eight tillers from the 46 SAM plants and all KLP tillers assessed tested endophyte-positive. The reliability of the ELISA results was confirmed by verifying, microscopically in aniline blue-stained epidermis sections, the presence of fungal hyphae in tillers from five randomly chosen SAM plants and five randomly chosen KLP plants that had tested positive for endophyte infection in the ELISA assay.

In order to obtain pure cultures of the endophytes for DNA isolation, two tillers from each endophyte-positive plant were surface sterilized and transverse segments (~2 x 2 mm) were cut along the length of each tiller. Multiple sheaths / layers of each segment were separated and 9-12 sections of such sheaths from each tiller were placed onto one PDA plate (original plate) containing 25 µg/ml of chloramphenicol. These plates were incubated at 22 ^{o}C in the dark and monitored daily to observe fungal growth and any other signs of contamination. Contaminated plates were discarded and replaced with plates containing freshly processed tiller sections of the same plant. Initially only two plates (each containing sections of one tiller) for each of the endophyte-positive plants were used. However, because of frequent loss of plates due to contamination this was subsequently increased to 3-5 plates (each inoculated from one tiller) per plant.

As already observed for AR37-Orig, these AR37 clones proved again to be extremely slow growing. Hyphae only emerged from the cut tillers after 10-14 days. Endophyte colonies had a viscous appearance and much of their growth was upwards rather than horizontal (Fig 3.2). In an attempt to accelerate growth of these colonies I transferred them to fresh PDA plates every 3-5 weeks. Nevertheless colonies took 12-14 weeks to reach a size deemed sufficiently large (20-25 mm diameter) for obtaining sufficient quantities of DNA. By the time the colonies had reached this size, their centres appeared less dense and darker in colour than their margins, possibly a sign of senescence-induced degradation of older parts of the mycelium. As the slow growth of the endophyte increased the risk of contamination, multiple peripheral sections of one of the colonies

growing on each original plate were cut out and transferred to new PDA plates (round-1 plates) after ~6 weeks. The process was repeated again to get "round-2" plates. Biomass from round-2 plates was harvested and either used directly for DNA extraction or lyophilized and stored in a desiccator at -80 ºC. Using these procedures I obtained, after 7 months, biomass of 30 AR37-SAM and 26 AR37-KLP clones.

As will be described in section 3.2.3 below, multiple attempts to extract enough high molecular weight DNA from fresh or lyophilized material harvested from the solid PDA plates using a variety techniques failed. Virtually no DNA was obtained - presumably largely due to the senescence of much of the mycelium (see above). As an alternative I therefore attempted to grow endophyte biomass in liquid cultures. One colony from each round-2 plate was cut into very small pieces and these were used to inoculate 50 mL of liquid PDB medium. Two flasks for each sample were prepared and placed at 22 ºC for up to 8 weeks. One flask was incubated in a shaker at 200 rpm another without shaking. The rationale was that shaking could be beneficial in terms of providing more oxygen and could shear the mycelium; the latter might be beneficial, since resulting small mycelial fragments can develop into new small mycelial pellets in which hyphae have better access to nutrients. On the other hand the slow-growing mycelium might have low oxygen requirements not requiring shaking and shearing the hyphae will have initial negative effects (Posch *et al.*, 2013). Flasks were regularly monitored, and if any sign of contamination were observed in a flask, it was discarded and replaced with a new flask inoculated with the same endophyte clone. Many of the flasks had to be discarded due to contamination.

However, eventually this approach yielded enough freeze-dried material from both shaken and unshaken flasks to extract DNA at least 3 times (> 150 mg; shaking had little or no impact on biomass increase or yield) from 22 AR37-SAM clones and 20 AR37-KLP clones, each from a different plant. This material from liquid cultures was stored at - 80 ºC for DNA extraction; the DNA extracted from it was what was used for pooling and subsequent sequencing.

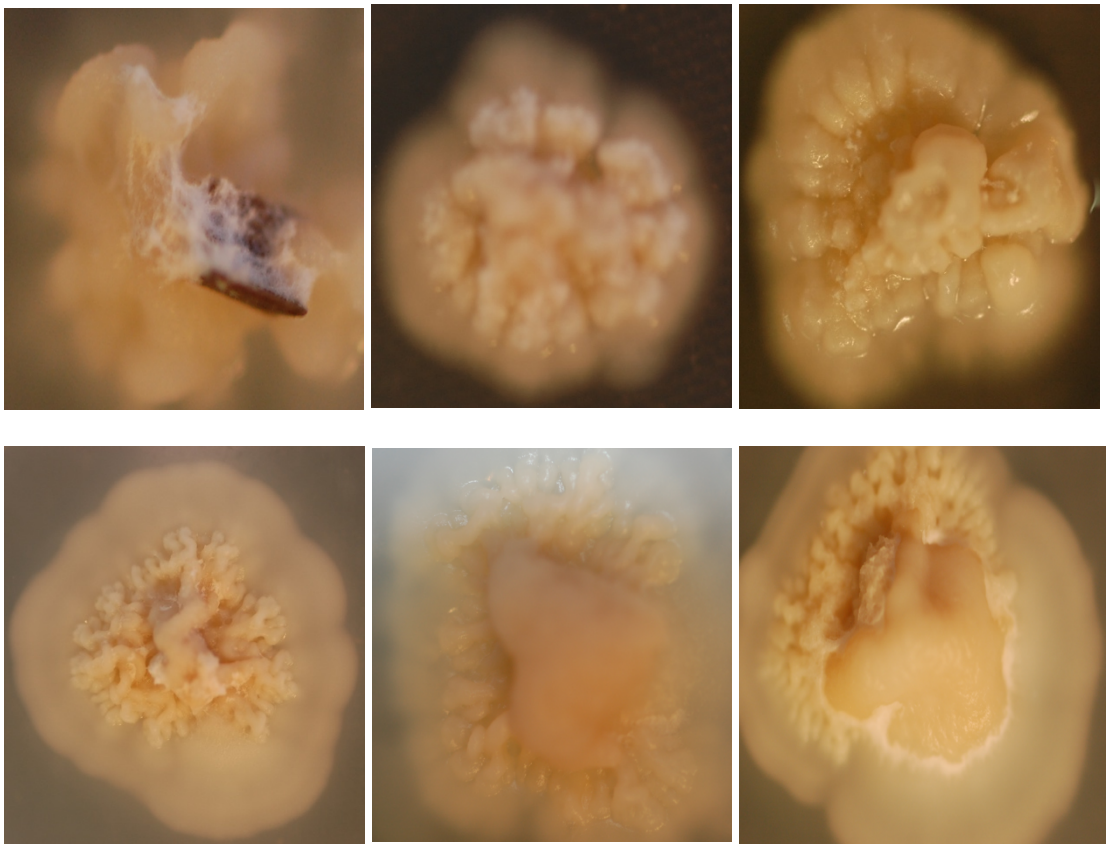


Figure 3.2. Six examples of AR37 colony morphologies on PDA medium with extensive vertical growth. Their slimy, glue-like nature is also evident from the figures.

3.2.3. DNA Extraction

Obtaining a sufficient quantity and quality DNA for sequencing from the limited amount of starting material, much of it containing old and thus presumably senescent and partially degraded biomass, proved a considerable challenge. Three different DNA extraction methods were trialled, namely (i) the DNeasy® plant mini kit (QIAGEN), (ii) the Al-Samarrai and Schmid (Al-Samarrai & Schmid, 2000) method and (iii) the Cetyltrimethylammonium bromide (CTAB) method (Webb & Knapp, 1990, Murray & Thompson, 1980).

3.2.3.1. DNA extractions using a QIAgen kit resulted in very low quantities of low-quality DNA

I first attempted to use the DNeasy® plant mini kit (QIAGEN) reported to be suitable for extraction of *Epichloë* DNA (Fleetwood *et al.*, 2011). This method failed to extract significant amounts of DNA from colonies from PDA plates, either freshly picked or lyophilized (20-25 mm diameter colonies; one colony corresponded to approximately 50 mg dry weight). It likewise failed for liquid culture-derived material. What little DNA was obtained proved to be of low molecular weight, unlike the DNA extracted from grass as a control, demonstrating that in my hands the procedure did allow the extraction of high molecular weight DNA (Fig. 3.3).

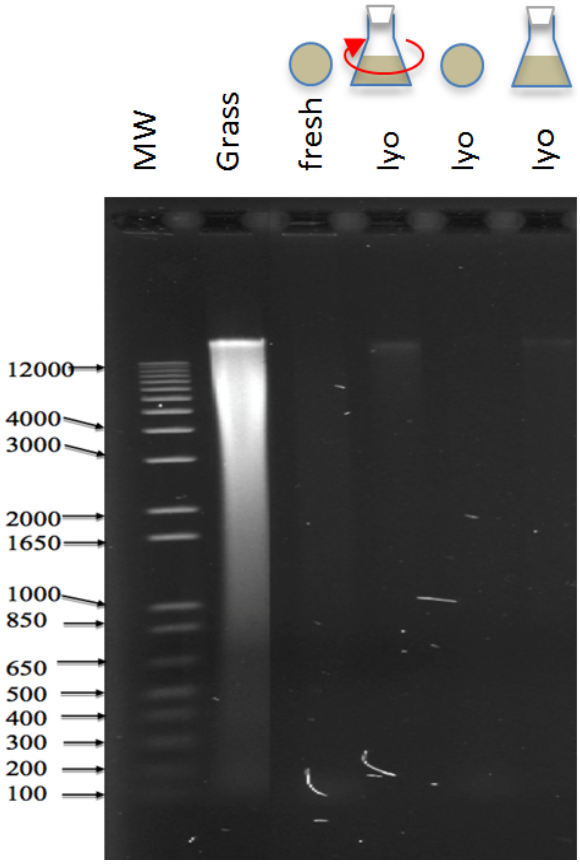


Figure 3.3. Gel image showing DNA extracted using DNeasy® Plant Mini kit. 20 µl of a 50 µl DNA extract were loaded onto 0.8% gel and run for 90 min at 70 V. 20 µl of a grass DNA extract (Grass) is loaded for comparison. Extracts were prepared from either 100 mg of fresh (fresh) or 30 mg of lyophilized (lyo) material from plates ( ), shaking ( ), or nonshaking ( ) liquid cultures of AR37-SAM clone 43 (third and fifth lane) or AR37-SAM clone 23 (fourth and sixth lane). MW: 1 kb plus molecular weight standard.

3.2.3.2. The Al-Samarrai and Schmid method yielded DNA of insufficient purity

Given the poor DNA yields of DNAeasy extractions, I next tried the Al-Samarrai method (Al-Samarrai & Schmid, 2000), which is frequently used for *Epichloë* endophytes (Schardl *et al.*, 2013b, Spiering *et al.*, 2005b, Moon *et al.*, 2002, Craven *et al.*, 2001). This method yielded significantly more DNA than the Qiagen kit (200-4000 ng per 50 mg sample), but much of it was degraded (Fig. 3.4). Additional purification steps using either ether precipitation or phenol treatment and ethanol precipitation also had no significant effect on the quality of the DNA.

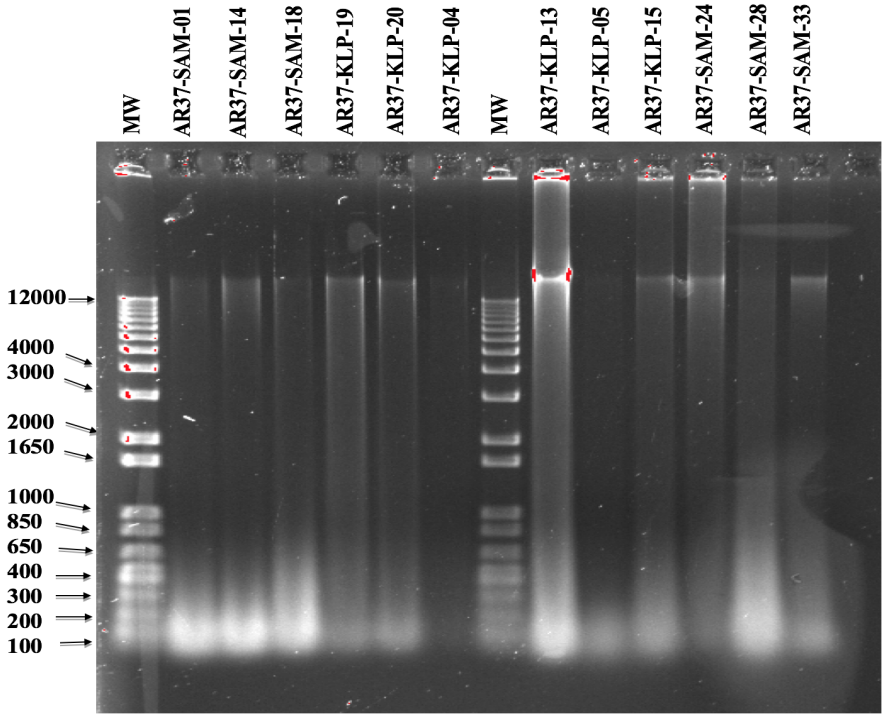


Figure 3.4. DNA extracts obtained by the Al-Samarrai and Schmid method, using different concentrations of RNAse. 13 µl of each sample was loaded onto 0.7% gel and run for 90 min at 50 V. Extracts loaded had been prepared from freeze-dried material harvested from shaken liquid cultures of AR37-SAM clones 1, 14, 18, 19, AR37-KLP clone 20, 04, 13, 5, 15, AR37-SAM clones 24, 28 and 33. Samples in lanes 2–5 had been treated with 2 µl, and samples in lanes 6–7 with 4 µl of 10 mg/ml RNAse. Samples in lanes 9–11 and 14 had been treated with 2 µl, and lanes 12 and 13 with 4 µl of 100 mg/ml RNAse. MW: 1 kb plus ladder. Numbers on the left indicate molecular weights in bp.

In addition, intense ethidium bromide staining at a molecular weight of around ~150 bp was visible on the agarose gels. Intense bands are usually associated with RNA and it suggested that the RNAse treatment may not have been fully effective. However, I was

unable to reduce the intense staining at ~150 bp by increasing in RNAse concentration, incubation temperature, and incubation length.

3.2.3.3. The CTAB method yielded DNA of sufficient quantity and quality

DNA extracted by Cetyl Trimethyl Ammonium Bromide (CTAB) method (Webb & Knapp, 1990, Murray & Thompson, 1980) also appeared to be smeared but with a prominent high molecular weight component (Fig. 3.5). Little DNA (~100 ng) was obtained from fresh material obtained from PDA plates (Fig. 3.5). Somewhat higher yields were obtained when freeze-dried material from PDA plates was used (23 – 4665 ng). Use of broth culture biomass produced the best yields (312 ng – 5500 ng) (Fig. 3.5). The CTAB method was therefore used for all DNA extractions.

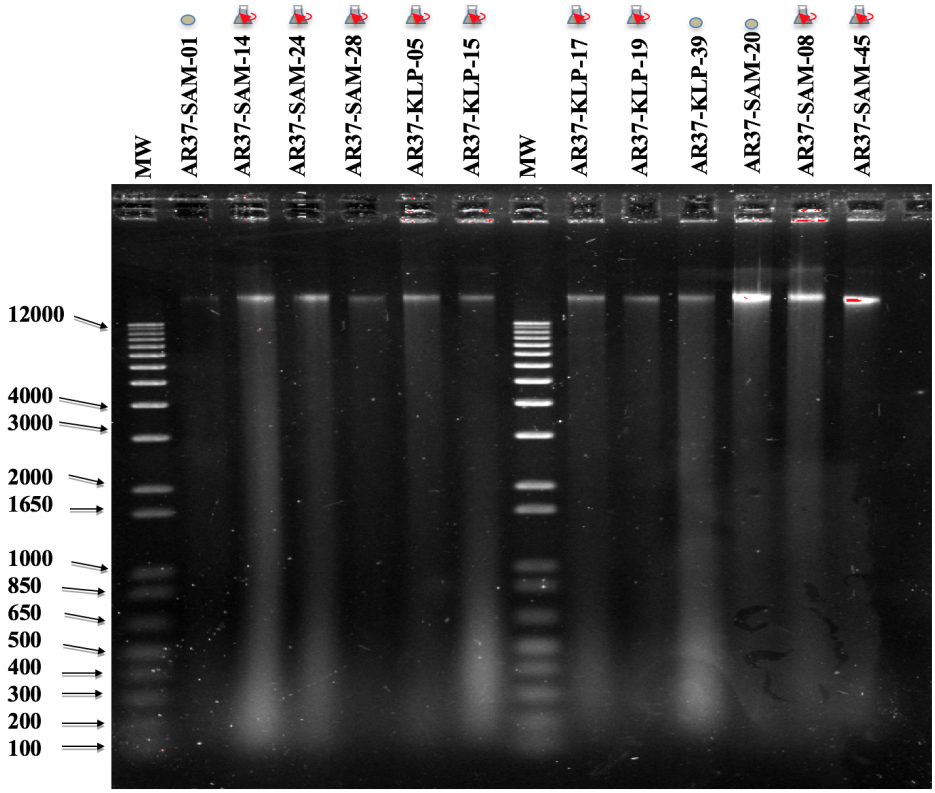


Figure 3.5. CTAB extracted DNA. 10 µl of each sample was loaded on 0.8% gel and run for 2 hours at 60 V. 2 µl of 10 mg/ml RNAse was used and samples were incubated at 65 ^{o}C. Extracts loaded were prepared from fresh material obtained from PDA plates of AR37-SAM clone 01 (lane 2) and freeze-dried material from either solid plates of AR37-KLP clone 39 and AR37-SAM clone 20 (lane 11 and 12 respectively) or from shaken liquid cultures of AR37-SAM clones 14, 24, 28 and AR37-KLP clones 05, 15, 17, 19, 39 and AR37-SAM clones 08 and 45. MW: 1 kb plus ladder. Numbers on the left indicate molecular weights in bp. High molecular weight band is very clear.

Even with the CTAB method multiple extractions had to be performed to obtain, when extracts were combined, an amount of DNA per clone deemed sufficient for further analysis (> 1 µg). Nanodrop readings indicated that some of the extracts to be combined had protein contamination or low DNA concentrations (< 20 ng/µl). Such samples were purified and concentrated using an additional phenol and ethanol extraction step. I succeeded in obtaining >1 µg DNA from 26 AR37-SAM clones and 21 AR37-KLP clones

3.2.4. Eleven samples of each AR37-SAM and AR37-SAM containing high molecular weight DNA were pooled

As mentioned above, my intention was to sequence pools of AR37-SAM clone DNA and AR37-SAM clone DNA to cost effectively find mutations in these populations. I therefore needed to combine the DNAs I had obtained so that all clones would contribute equally to the sequences obtained from a pool. I therefore needed to consider not just the total amount of DNA contributed to a pool by each clone DNA sample but rather the amount of high molecular weight DNA, since smaller DNA fragments would yield less sequence.

Clone DNA samples considered for pooling were therefore run on a gel alongside each other (see Fig. 3.6 for an example), and visually compared. On this basis I determine in what ratio they needed to be combined. When samples were pooled, however, no strong high molecular weight band was detectable and fluorescence in the well suggested the presence of a contaminant that interfered with the movement of DNA (Fig. 3.7). Some individual clone samples produced stronger fluorescence inside wells than others when loaded on a gel (Fig. 3.6), and I argued that omitting such samples from the pool might alleviate the problem. This was indeed the case. Combining only those samples that had low levels of fluorescence inside the well of a gel, I was able to generate pools that had a strong high molecular weight DNA band with little DNA retained in the well (Fig. 3.7).

The resulting two pools contained DNA from 11 AR37-SAM clones and 11 AR37-KLP clones, representing only 1/3 the number intended at the onset of the project. Nevertheless I decided I should continue to the next stage of the project with these pools. This appeared

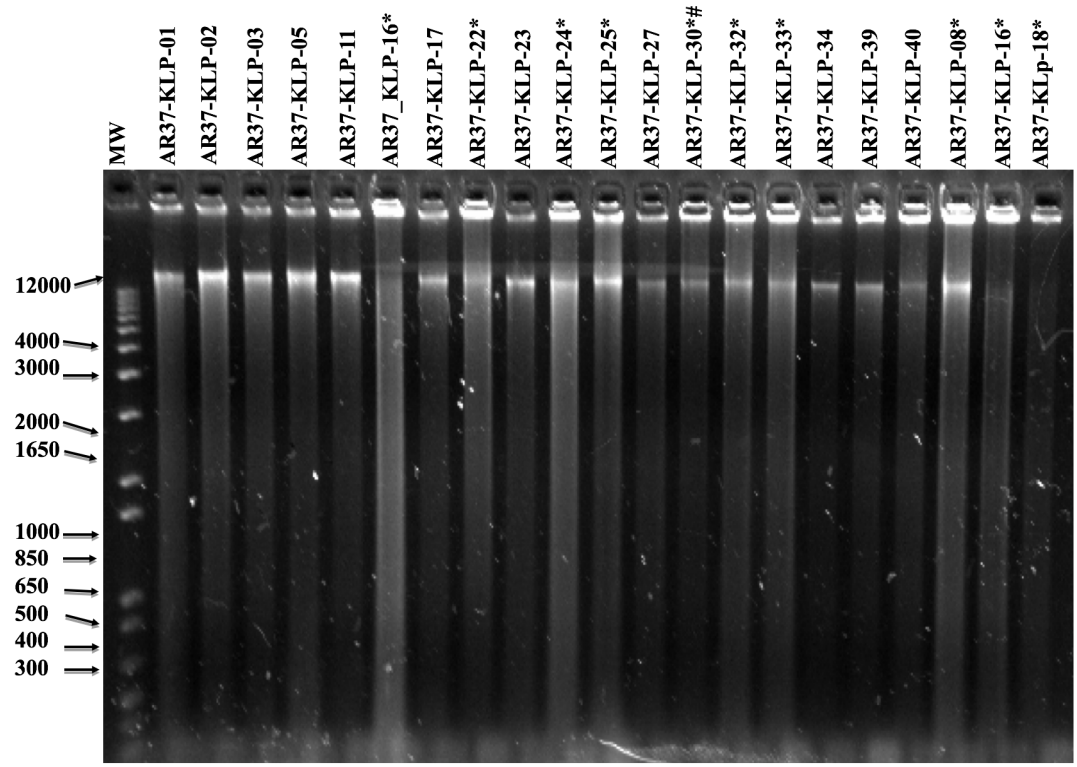


Figure 3.6. DNA from 21 AR37-KLP clones run side by side before pooling. The amount of total DNA loaded varied between 60–1360 ng, adjusted so as to produce high molecular weight bands of similar intensity. Nine samples, marked with * were omitted from the final 11 sample pool because of retention of material in the well and (probably as a result) smearing. One sample (labelled with #) did not have enough high molecular weight DNA to go into the pool. Lanes are labelled with the numbers of the AR37-KLP clone from which the DNA was derived. MW: 1 kb plus ladder. Numbers and arrows on the left indicate molecular weights in base pairs.

justifiable not only on the basis by time restraints: While pools of 11 samples would be expected to contain ~2/3 fewer variants, each variant would be present in ~3 times more sequencing reads. I argued that bioinformatic distinction between variant bases and sequencing errors would be difficult and would be more likely to eliminate a variant present in a single clone if it was represented by only 1/30th of reads compared to a variant represented by 1/11th of all reads in a smaller pool. Thus the number of variants detectable could potentially even be increased by reducing pool size. Figure 3.8 shows a gel with the two pools and the AR37-Orig DNA sample used for sequencing.

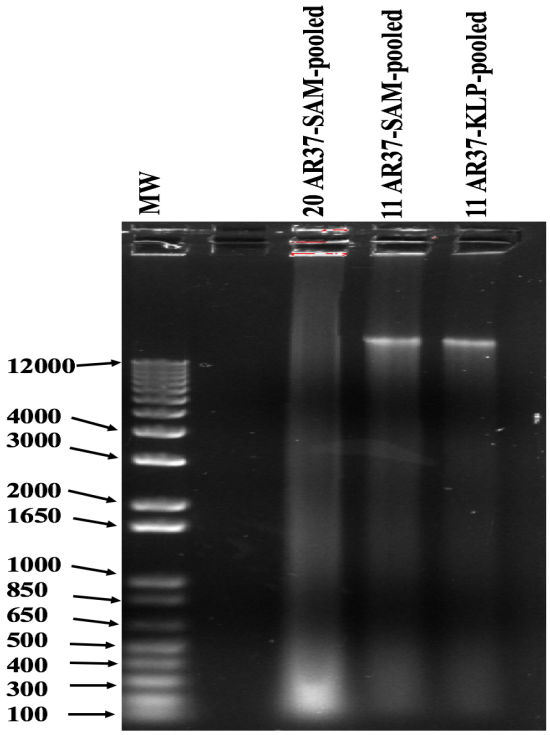


Figure 3.7. Comparison of the AR37-SAM 20 clone pool (20 AR37-SAM-pooled) with two pools made by combining only samples in which little material was retained in the well (11 AR37-SAM-pooled, 11 AR37-KLP-pooled). The lane labelled MW contains the 1 kb plus molecular weight marker. Numbers and arrows on the left indicate molecular weights in base pairs.

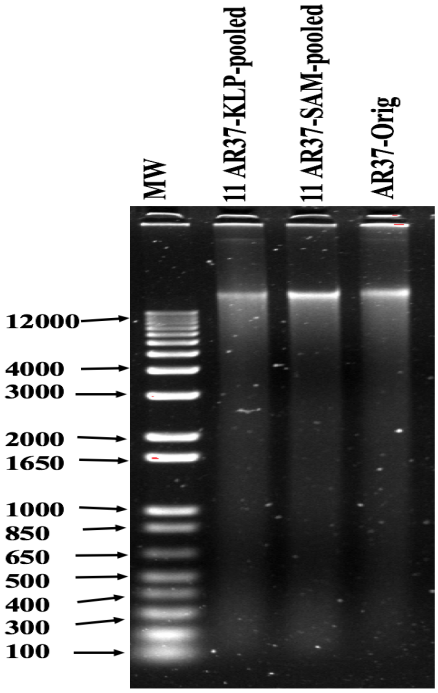
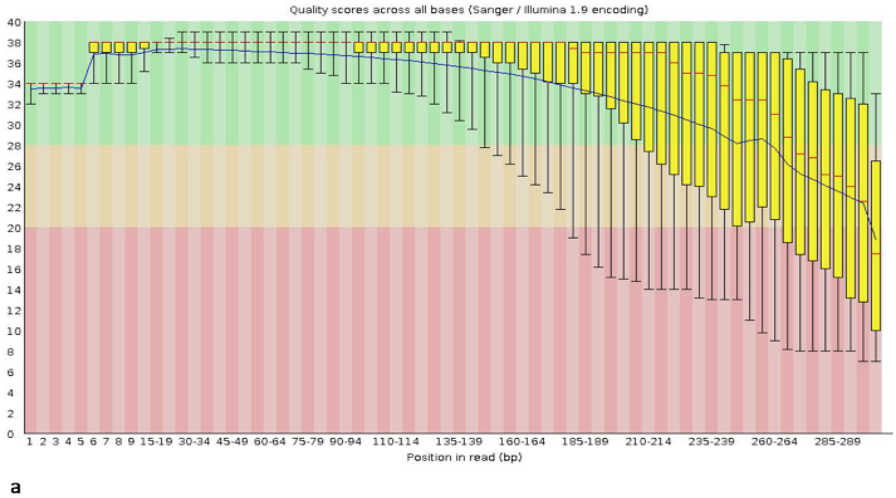


Figure 3.8. DNA samples used for sequencing. The lane MW contains the 1 kb plus molecular weight marker. Numbers and arrows on the left indicate molecular weights in base pairs.

3.3. Sequencing and processing and trimming of reads

AR37-Orig, isolated from original host plant maintained at AgResearch (section 2.4.1), and the two 11 clone pools were sequenced using the Illumina MiSeq platform which produces 2 x 300 bp paired-end reads. AR37-Orig was sequenced a second time using a 2 x 250 bp paired-end approach to increase coverage, as it was to be the main source for deducing the ancestral AR37 genome, which both the pools were to be compared against. The sequence reads were provided as "raw" sequence files in fastq format by the sequencing centre (Massey Genome Service). As per standard practice in Illumina sequencing protocols a small amount of PhiX DNA was also sequenced along with my samples as a quality control. Initially, sequence reads were mapped using BWA against the PhiX reference genome, with all reads that aligned to this genome being removed prior to further processing. The software Fastq-mcf from the ea-utils (Aronesty, 2011) was then used to remove adapter sequences from the remaining reads. FastQC quality score plots (Figure 3.9 a,b) indicated that the sequencing quality degraded towards the 3` end of the reads. Therefore quality trimming was performed to remove low quality bases. This trimming improved the overall quality of the reads so that now nearly all bases had a phred quality score of ≥ 30 throughout; i.e. a probability of any base being called incorrectly of ≤ 0.001 (Figure 3.9 c, d).



a

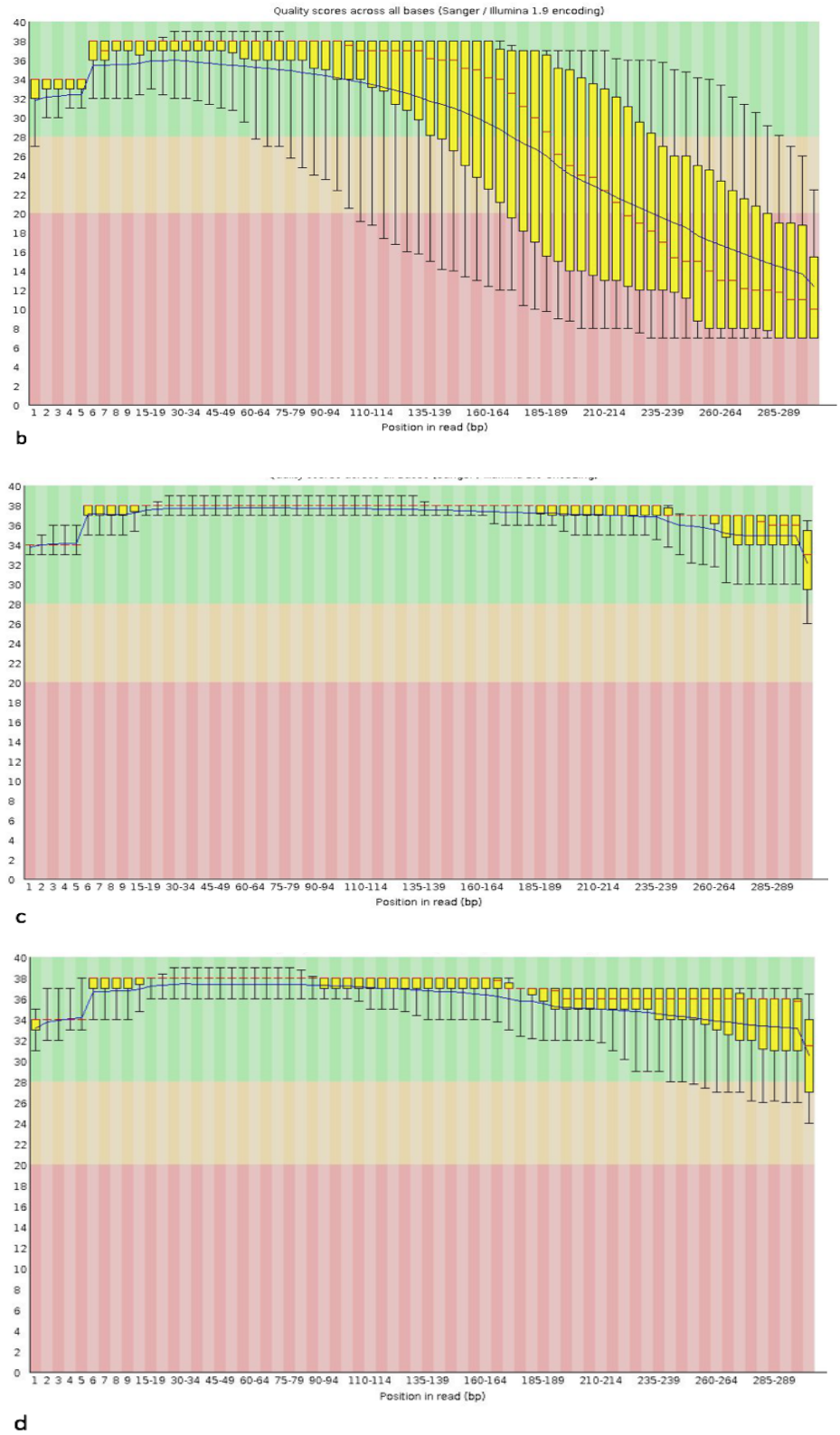


Figure 3.9. Quality scores across all bases of a) raw forward reads and b) raw reverse reads c) trimmed forward reads and d) trimmed reverse reads. Forward reads from all samples were combined in one file and so were reverse reads to generate these graphs.

Table 3.1 gives an overview of the number of reads before and after processing. Trimming only removed low quality bases from the ends of individual reads and not the whole reads, so it resulted in decrease in length of reads but not their total number. If whole read was a low quality one and had to be trimmed, even then only one base was kept at that place to maintain order of the paired reads. Only the trimmed reads were used for all subsequent analyses. Judging by previously reported *E. festucae* genome sizes (Schardl *et al.*, 2013b), even after trimming the combined sequences lengths of the reads in each sample were likely to be equivalent to > 60 times the size of the AR37 genome, i.e. the coverage of the genome was expected to be > 60x on average.

Table 3.1. No. of Illumina MiSeq reads before and after initial processing and trimming

Sample	Total reads	Processed reads*	Processed-trimmed Reads**	Reads Discarded
AR37-Orig	13,014,591 x 2	12,963,639 x 2	12,963,639 x 2	50,952 x 2
SAMSON	6,370,227 x 2	6,357,010 x 2	6,357,010 x 2	13,217 x 2
KLP1102	8,729,636 x2	8,718,058 x2	8,718,058 x 2	11,578 x 2

* Reads remaining after removing reads that mapped to PhiX genome.
**Reads remaining after removal of adapter sequences and trimming low quality bases

3.4. How much variation is there between AR37 clones?

3.4.1. Strategy for the search for genetic variation in AR37

As explained in the introduction the sequences obtained were now to be utilized to answer the following two questions, namely

1. How much and how has the genome of AR37 changed as its compatibility with the new ryegrass hosts SAM and KLP improved?
2. How much genetic variation is there between different clones of AR37?

The answer to the first questions required finding differences between each AR37 pool genome and the ancestral AR37 genome. One way of identifying these differences is to directly map sequencing reads from each pool against a reference genome assembly and then search for sites in mapped reads that differ from the reference (Pabinger *et al.*, 2014). Another approach is to make a *de novo* assembly for each pool and then compare this *de*

novo assembly against a reference assembly (Olson *et al.*, 2015). The latter approach seemed less advisable for the current project, because a pool genome would represent a consensus sequence and variations present only in one of the clones would be eliminated in the assembly. Given clonal interference I expected most variations to be present only in one clone and thus the second approach would likely prevent me from detecting the majority of variations.

Comparison of pool reads with a reference also seemed a better approach for answering the second question, again because only then would variations specific to one clone within a pool be apparent. Subsequently comparing the variations relative to the reference found in the AR37-SAM pool with the variations relative to the reference in the AR37- KLP pool would then allow me to also indirectly infer the differences between the two pools.

For AR37 an assembly existed (unpublished data obtained from Richard Johnson, AgResearch Ltd, Palmerston North) but several quality parameters indicated that it was of poor quality. The assembly was highly fragmented, containing 35,463 contigs, of which the largest was only 28,588 bp long, and the N50 was only 1836 bp. The N50 represents the size of the contig that contains 50% of the entire genome length - thus the AR37 assembly was largely made up of relatively small contigs < 2 kb in size. The genome was sequenced using the 454 platform which is known to have high error rate especially in homopolymer regions, and it contained 187.21 N's per 100 kbp, a further indication of its poor quality. The assembly also had a very low coverage (7x) and the overall length of assembly was ~16 Mb longer than closely related strains indicating potential assembly artefacts. Another shortcoming of the assembly was that it might contain mutations that were not present in the ancestral AR37 used to inoculate SAM and KLP.

It would thus be necessary to make a *de novo* assembly for use as a mapping reference during variant calling of our sequenced samples. A summary of my approach for variant identification is shown in Figure 3.1 (including variant filtering as a final step, necessary to distinguish true variants from sequencing errors; see sections 3.4.4.1.2 and 3.4.4.2).

3.4.2. Construction and characteristics of an ancestral AR37 reference genome assembly

The depth of sequencing for each sample in our study was at least > 60x, which is at least 8 times higher coverage than the previously available 454 assembly and thus using the AR37–Orig reads should produce a better assembly. However, an even greater depth was desirable, because Illumina reads are shorter in length than 454 reads and this could result in shorter contigs due to the lack of sufficient read overlaps, especially in the repetitive and low complexity regions that are abundant in the genomes of most eukaryotes. In addition, using only the AR37-Orig sample reads would not necessarily reflect the ancestor of the SAM and KLP clones, because this DNA was from an AR37 clone that had been propagated (albeit in the original host and only vegetatively) for several years after the AR37-SAM and AR37-KLP lines had been initiated. To overcome both of these pitfalls I decided to use not only the AR37-Orig reads but also the AR37-SAM and AR37-KLP reads when making a new assembly. Pooling the data gave a hypothetical genome coverage of ~150x based on a predicted genome size of ~40 Mb. Variants specific to one or both the pools would be too few and too low in frequency to affect the assembly algorithm. As a result of clonal interference (Gerrish & Lenski, 1998), most nucleotide variations in pools would probably be specific to one of the 11 clones in the pool and thus occur at a frequency of approximately 1/11 or 9%. Mixed with the reads from the other samples their frequency in the assembly would usually drop below 5%. All sequencing platforms are prone to produce errors and error rates vary depending on the sequencing platform ranging from < 1% (Illumina) to > 5% (Ion Torrent, Pacbio). Genome assemblers are therefore designed to ignore low frequency divergent base calls. Thus it should be possible to combine reads from all pools without a major negative impact on the overall assembly. Furthermore, variant base calls specific to AR37-Orig and absent from both pools would also not be reflected in the consensus assembly. Such base calls most likely reflect mutations that have occurred as AR37-Orig diverges from the common ancestor of AR37-Orig, AR37-SAM and AR37-KLP. Thus an assembly made from all AR37-Orig, AR37-SAM and AR37-KLP reads should largely reflect the ancestral AR37 genome.

3.4.2.1. Evaluation of six different assemblers for generating a *de novo* AR37 assembly

For producing a new AR37 assembly, six assemblers were initially evaluated; Velvet, SOAPdenovo, MIRA, ABySS, A5-miseq-pipeline and SPAdes. Each assembler was used to generate an assembly and the software QUAST was then used to compare the assemblies. A comparison of some of the key assembly statistics are shown in Table 3.2.

I initially evaluated the assemblies generated using four parameters: Total number of contigs in an assembly provided the first clue towards the quality of the assembly, because an assembly with a low number of contigs indicates a high level of contiguity. The N50 value provided an important metric for assessing the ability of an assembler to generate large contigs. It is defined as the minimum contig length representing 50% of the assembly. "It means, half of the genome sequence is in contigs larger than or equal the N50 contig size" (http://www.metagenomics.wiki/pdf/definition/assembly/n50) (Consortium, 2001). Therefore assemblies that have a large N50 were targeted as they are likely to be better than others (Gurevich *et al.*, 2013). The number of ambiguous bases i.e. N's per 100 kbp is important because it represents missing data. Finally, I would expect the AR37 *de novo* assembly to be within the size range of a other closely-related *Epichloë* strains (*Epichloë festucae* E2368 and *E. festucae* Fl1) (i.e. ~34 MB, http://www.endophyte.uky.edu)(Schardl *et al.*, 2013b). Assembly sizes that are much larger or smaller than expected could indicate problems with the assembly, such as low / uneven coverage or the presence of contaminating DNA from other organisms.

Judging by these criteria, the best assemblies were produced by SPAdes and A5-miseq (Table 3.2). These two assemblies had the least number of contigs and significantly higher N50 values than the remaining assemblies, and were close to the genome size of the closely related strains i.e. *Epichloë festucae* E2368 and *Epichloë festucae* Fl1. The MIRA assembly shown in Table 3.2 is only based on AR37-Orig clone sequences as attempts to generate a full assembly using pooled samples failed after 3 weeks of runtime on the Massey server.

It must be noted that A5-miseq assembly did not report contigs smaller than 500 bp in size while SPAdes assembly reported all the contigs. If only contigs that are greater than

1000 bp are considered, then the SPAdes assembly appeared as good as A5-miseq or even better. Importantly, the SPAdes assembly was largely free of ambiguities, an indication that when combining all reads from all samples, sample-specific polymorphisms did not significantly interfere with the assembly process. I also investigated if combining reads from all samples might cause other problems in assembling, for instance as a result of major genome rearrangements as the clones diverged. To do so I carried out an assembly with the SPAdes assembler using reads from only AR37-Orig sample (SPAdes-Orig) and compared it with the SPAdes assembly based on reads from all samples (SPAdes-pool) (Table 3.2). The assembly produced from reads of only AR37-Orig sample was inferior to the one produced using all samples (Table 3.2). The opposite would have been expected if combining samples had interfered with the assembly process. SPAdes-Orig was henceforth not used in any analysis and SPAdes assembly will refer to SPAdes-pool assembly from this point onwards.

The above assembly statistics do not take into consideration the completeness of the assembly in terms of presence of genes. Since the interpretation of the functional significance of polymorphisms would be largely based on which genes they affected, I also evaluated the assemblies using the benchmarking single copy orthologues (BUSCOs) approach (Simao et al., 2015). It uses a dataset of highly conserved single copy genes to predict the completeness of open reading frames found in the genome assemblies. For the entire phylum fungi 290 highly conserved genes, i.e. BUSCOs that are found in > 90% of the sampled species, have been identified. For phyla within fungi a larger set of BUSCOs have also been identified that provide a better resolution at each sub-clade e.g. 1315 BUSCOs for Ascomycota and 3725 BUSCOs for the Sordariomycetes. The presence of a high number of unfragmented BUSCOs is a measure of the completeness of an assembly, as is a low frequency of multiple copies; the latter would indicate misassembly because BUSCOs are supposed to be single-copy genes. The BUSCO software package uses group consensus sequences to searches for matches on genome loci of assemblies using tBLASTn and makes amino acid BUSCO group block-profiles, which are used to guide gene annotations by another software tool, Augustus. This approach provides a fair comparison of different assemblies (Simão *et al.*, 2015).

The SPAdes and the A5-miseq AR37 *de novo* assemblies, both identified as superior in terms of contig lengths (see above), also scored well in this analysis, containing > 95%

complete Sordariomycete BUSCOs with very few of these fragmented or duplicated (Figure 3.10). The BUSCO analysis of the previous 454 assembly revealed that more than half of the Sordariomycete BUSCOs (Fig. 3.10) were missing, further confirming the need for a better assembly for this project.

Considering all aspects of these evaluations, the SPAdes assembly was chosen as the basis of further analyses over the A5-miseq assembly. Both were superior to the other assemblies and comparable in quality overall. However, since my goal was to detect variants, the lack of ambiguities in the SPAdes assembly constituted a significant advantage of this assembly.

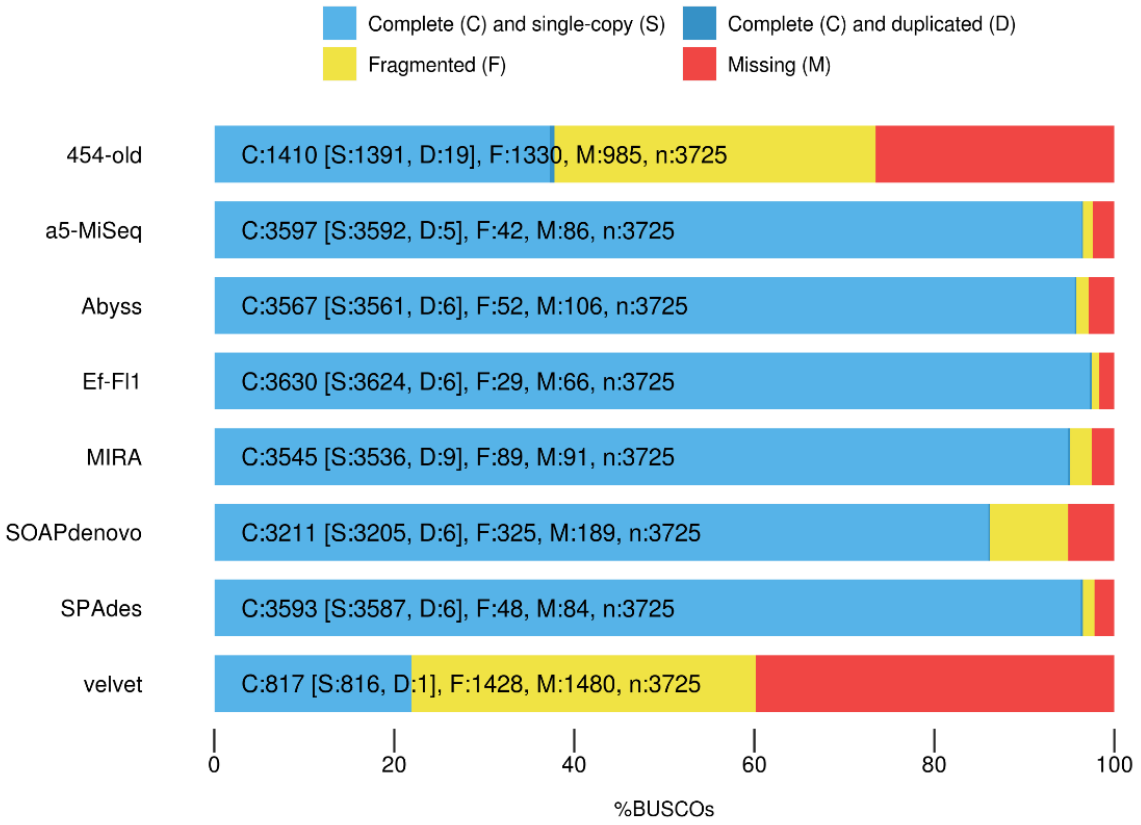


Figure 3.10. Assessment of completeness of assemblies using BUSCOs from Sordariomycetes. Numbers on the bars indicate no. of BUSCOs in each category i.e. C:Complete, S:Single, D:Duplicate, F:Fragmented and M:Missing.

The quality of the SPAdes assembly was also confirmed by a comparison with a recent ungapped assembly of a closely related strain i.e. *E. festucae* Fl1 (Ef-Fl1) produced by Winter et al. (Winter *et al.*, 2018). This assembly consists of 7 chromosomes and one contig representing mitochondrial DNA. As such this Fl1 assembly was of course vastly

superior to my AR37 assembly in terms of contig lengths. The N50 for the Fl1 assembly was 6,201,951 bp, and length of the largest contig (chromosome) was 7.9 Mbp. However in terms of BUSCO, my AR37 SPAdes assembly was a fairly close match to the Ef-Fl1 assembly. A search for Sordariomycete BUSCOs in this Fl1 assembly failed to find 66 genes (< 1%) compared to 84 genes (~2%) missing in AR37. The size of the Fl1 (~35 Mb) assembly was similar to my AR37 assembly (~33 Mb), another indication that my assembly was a reasonable representation of the AR37 genome.

In summary my work had yielded a good quality AR37 assembly. With bases at each position based on the prevalent base call across all three samples is should largely reflect the ancestral state of AR37 prior to the divergence of the clones from which I had isolated DNA for sequencing.

Table 3.2. Comparison of key characteristics of AR37 assemblies produced from Illumina MiSeq reads using different assemblers. Statistics for previously available assembly are also given. Bold numbers are the top two values for the given parameter.

	454-old[1]	Velvet[2]	SOAPdenovo[3]	MIRA[4]	ABySS[5]	A5-miseq[6]	SPAdes-pool[7]	SPAdes-Orig[8]
Total length	50,896,160	25,136,713	36,536,660	28,069,245	33,388,992	32,762,814	32,899,942	32,622,234
Total Contigs	35,463	51,558	28,058	6,865	7,298	**759**	**1,343**	2,137
Contigs>1000bp	18,114	6,209	6,644	2,147	1,549	**690**	**601**	1,341
Contigs>10,000bp	62	0	616	861	584	**390**	**429**	755
Largest contig	28,588	5,955	70,508	148,148	351,083	**638,446**	**513,618**	243,211
N50	1,836	1,118	6,431	23,652	63,305	**141,564**	**115,316**	51,988
L50	7,474	5,027	1,441	334	145	**70**	**84**	184
N's per 100 kbp	187.21	1.37	**0.00**	5.25	317.60	28.37	**0.00**	0.00

[1]previous assembly sequenced using 454 technology and assembled using MIRA.
[2]Velvet was used to assemble Illumina reads with kmer values of 33, 77 and 121. The statistics here are for the best assembly obtained at kmer 33
[3]SOAPdenovo was used with kmer values of 63 and 128 and statistics are for better of the two assemblies i.e. kmer 128
[4]Assembly is produced by using reads from only AR37-Orig sample. Reads from all combined samples took 3 weeks without finishing so job had to be terminated.
[5]ABySS was used to assemble genome at kmer values of 96 and 128. Statistics are for assembly obtained with kmer 128
[6]No option to choose any kmer values for A5-miseq
[7]SPAdes assembly with reads from all the samples combined and with kmer values of 33,55,77,99 and 127. Statistics here are for kmer 127
[8]SPAdes assembly with reads from only AR37-Orig sample and with kmer values of 33,55,77,99 and 127. Statistics here are for kmer 127

3.4.3. Mapping reads to reference genome

The next step was to map reads from the different samples to the ancestral AR37 references assembly, a prerequisite for finding variants, as positions in the genome at which base call frequencies in reads from a sample suggested presence of a base different from that in the reference assembly.

3.4.3.1. Three mappers, Bowtie2, Bwa-mem and Novoalign all appear suitable for aligning reads to the assembly

Alignment software packages differ in algorithms and a mapper's performance can be dramatically affected by the choice of alignment parameters and the underlying complexity of the reference genome. I therefore tried three different packages, Bowtie2, Bwa-mem and Novoalign to align reads to the SPAdes AR37 assembly. Summary mapping statistics from the three aligners, such as mean coverage, mean mapping quality and the percentage of genome at a particular coverage, were calculated using FASTQC. Tables 3.3 and 3.4 show some of the key parameters

Whilst all of three mapping packages allow reads to map to multiple positions, all provide a primary alignment position, which represents the most likely mapping position. I note that all secondary and supplementary alignments were removed before further analysis and data in table 3.3 and 3.4 include only these primary mapping positions.

Bowtie2, Novoalign and Bwa-mem successfully mapped ~90 – 95% of sequencing reads from all three samples against the AR37 assembly (Table 3.3). Coverage data were calculated, because they provide important indication of the number of the number of reads with divergent base calls expected at positions where pooled samples differ from the reference.

The mean coverage values reported for each of the samples by all three aligners were also similar (Table 3.4). The lowest mean coverage was ~60x for the AR37-SAM pool. Since pools contained DNA from 11 individuals it means that any variation present in only one individual of the pool should be present in ~9% of the reads (assuming even coverage across all samples). At a mean coverage of 60x, for half of the AR37-SAM assembly any

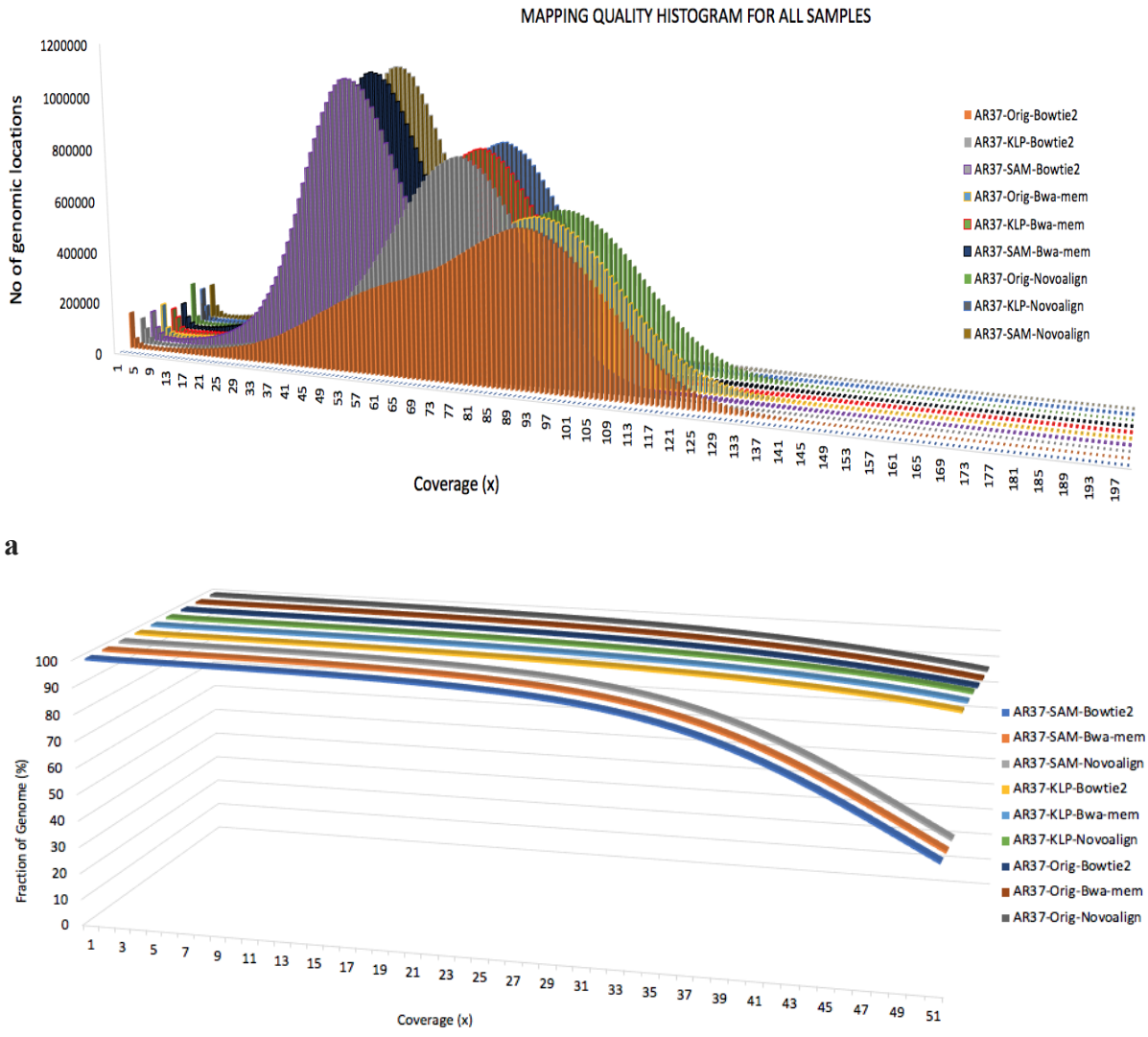


Figure 3.11. Coverage of all samples by all 3 aligners. a) Coverage histogram. Different aligners for a particular sample provided nearly identical coverage. b) Genome Fraction coverage.

polymorphism present in one clone should on average be represented by 5 reads. For AR37-KLP pool with a mean coverage of ~90x, any polymorphism present in one clone should on average be represented by 8 reads.

Equally important is the uniformity of coverage across the assembly, as it determines what percentage of the genome can be assessed for the presence of variants. Coverage was fairly uniform across the assembly (Fig.3.11b). For the AR37-KLP pool, coverage

exceeded 40x in > 90% of the assembly, indicating that a polymorphism present in 1/11 clones in a pool would on average be represented by 4 reads in > 90% of the assembly. For the AR37-SAM pool coverage of 30x was exceeded for ~95% of the AR37 assembly and thus polymorphisms in most of the genome would on average be represented by 3 reads.

In summary, all mappers provided suitable coverage to theoretically detect polymorphisms across the majority ~95% of the assembly (Fig.3.11b).

3.4.3.2. Three mappers, Bowtie2, Bwa-mem and Novoalign assess read alignment quality in different ways

For variant finding, two different values assigned by mappers are important. The first value is alignment score and represents how well individual reads align with the reference assembly. Ideally, unless a mutation is present, reads should align perfectly (i.e. without any mismatch) to the genome. To assess this, all aligners calculate alignment scores of reads based on matches and mismatches of read to the reference. The maximum possible score is zero, indicating that there is no mismatch between the read and the reference. Each mismatch or gap is given a penalty (minus score); however, the three aligners assign different penalties to mismatches / gaps. Therefore, the different aligners report different alignment scores, even when they align a read to the same position in the reference genome.

The second value is mapping quality and this value indicates the log-scaled probability that the read is incorrectly mapped. Sometimes, a read may have an equal alignment score at more than one position on the reference and aligners have no way to prefer one place over the other. Each aligner thus reports a parameter called mapping quality (MAPQ) which is the probability that a read is placed incorrectly. MAPQ is calculated using $-10 \log10 \times p$; where p is an estimate of the probability that read is wrongly mapped (Ruffalo *et al.*, 2012). The bigger the mapping quality is and the bigger the difference between the mapping quality of best alignment and the second best alignment is, the more unique the best alignment is. Although calculation of MAPQ seems straightforward, in practice it is not an easy task to put a p-value on the likelihood that any given read is incorrectly

mapped, especially when there might be many equally possible alignments. This latter situation occurs for reads that map in highly repetitive areas of the genome. Different aligners deal with this situation in different ways and thus their MAPQ scores differ from one another.

Bowtie2 does not use the number of times a read mapped to the reference in the calculation of MAPQ (http://biofinysics.blogspot.com/2014/05/how-does-bowtie2-assign-mapq-scores.html). Instead it compares alignment scores of the best and second best read and assigns a MAPQ value. Uniquely-mapped reads get a MAPQ value > 40 (with a max value of 42) and multi-mapped reads with equal alignment scores get a MAPQ value of 0 in bowtie2. Thus, to remove all reads with multiple alignments, filters must be set at removing any read with a MAPQ value < 41. Lower filter values will remove reads with varying degrees of difference between primary and secondary alignment.

Bwa-mem follows the above equation (−10 log10 x p) and reports MAPQ values as Phred scores (Max score is set to 60; Fig 3.12). It takes into account number of best alignments and number of sub-optimal alignments as well as the Phred scores of the bases which differ between the best alignment and the sub-optimal alignments (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/).

Novoalign also takes into account primary and secondary alignments while calculating MAPQ values but it also considers the likelihood that a read may come from a region of the genome that is absent from the reference assembly. The maximum MAPQ value for Novoalign is set to 70 (https://sequencing.qcfail.com/articles/mapq-values-are-really-useful-but-their-implementation-is-a-mess/).

These different ways to calculate MAPQ make it difficult to directly compare mapping qualities between different aligners. A plot of MAPQ values against the number of genomic locations for all three mappers showed that the MAPQ values differed considerably between three mappers. It was expected because of different mapping quality algorithms used by these three mappers (Figure 3.12). However, the MAPQ values were fairly uniform for all samples for any particular mapper. Novoalign mapped

most reads to the reference genome with near maximum MAPQ (large peaks on extreme right in Fig.3.12). It was followed by Bwa-mem (center peaks in Fig.3.12) and bowtie2 (short peaks on extreme left in Fig. 3.12). For bowtie2, although ~90% reads appeared to have MAPQ values very close to the maximum value (i.e. 42), a fair percentage (~10%) of reads had lower MAPQ values of 30-35 (Fig. 3.12). MAPQ scores are reported to mirror Phred scores (Ewing & Green, 1998) and so scores > 30 would be considered very good as score of 30 would indicate 1 in 1000 chance that the read is wrongly mapped (Ruffalo *et al.*, 2011).

In summary these results indicated that all three mapping methods performed reasonably well with the sequence data and reference genome used in this project. Bowtie2 mapped most reads (> 99%) to the reference genome for all samples, followed by Novoalign (95-98%) and Bwa-mem (89– 97%). Higher number of reads mapped by Bowtie2 came at a cost of relatively lower (but still acceptable) MAPQ scores for ~10% of the reads. Bwa-mem and Novoalign mapped > 95% of the reads with near maximum MAPQ score with Novoalign slightly outperforming Bwa-mem. Together these results suggested that all three mappers produced alignments with sufficiently high number of reads mapped to AR37 assembly at an acceptable MAPQ scores to be used in downstream analysis. Thus, short read alignment files were generated for all three in order to maximize the possibility of finding sample specific variants.

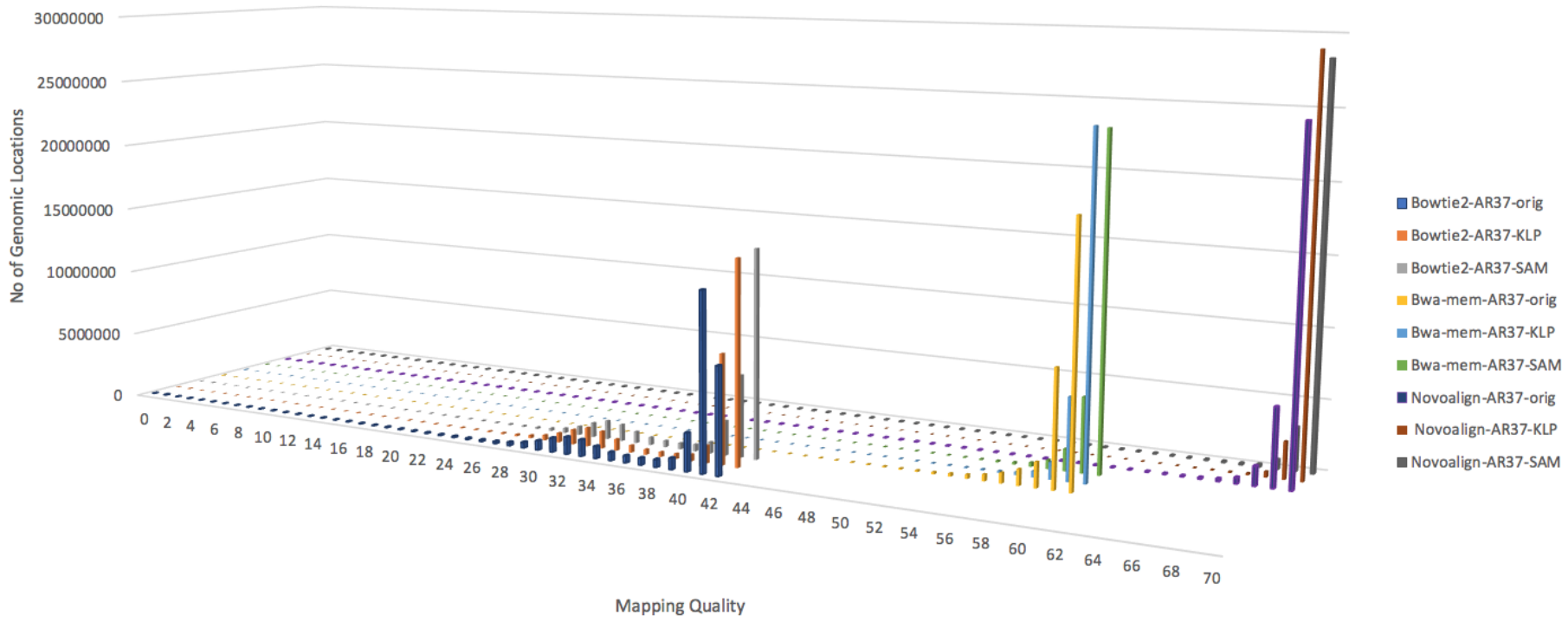


Figure 3.12. Mapping qualities of all 3 samples by all 3 aligners.

Table 3.3. Number of reads from all samples that mapped to the AR37 reference assembly using bowtie2, Bwa-mem and Novoalign. All duplicate, supplementary and secondary reads were filtered from the alignments and stats calculated from BAM files.

Samples	Processed-trimmed Reads	Bowtie2	Bwa-mem	Novoalign
AR37-Orig	12,963,639 x 2= 25,927,278	25,877,127 (99.81%)	23,077,990 (89.01%)	24,599,305 (94.88%)
SAMSON	6,357,010 x 2= 12,714,020	12,683,414 (99.76%)	12,132,574 (95.43%)	12,430,771 (97.77%)
KLP1102	8,718,058 x 2 =17,436,116	17,400,698 (99.80%)	16,945,629 (97.18%)	17,228,809 (98.81%)

Table 3.4. A comparison of Mean Mapping Quality (MQM) and Mean Coverage (Mean Cov) values for all three aligners as calculated from BAM files produced by aligning reads from AR37-Orig, SAMSON and KLP1102 against the AR37 reference genome.

Samples	Bowtie2		Bwa-mem		Novoalign	
	MQM (42)	Mean Cov	MQM(60)	Mean Cov	MQM(70)	Mean Cov
AR37-Orig	25.99	92.15	45.1	91.10	56.02	91.60
SAMSON	26.91	60.97	46.74	60.79	57.58	60.74
KLP1102	27.84	87.24	48.05	87.10	58.61	87.06

3.4.4. Choosing variant calling algorithms for the discovery of sequence variations in AR37

Having mapped reads to the *de novo* AR37 assembly, the next step was to choose software for variant calling, i.e. for detecting sites at which each sample differed from the ancestral AR37 assembly; since the assembly represents the consensus base calls between the three samples (see section 3.4.1 such sites would also be expected to differ from or both of the remaining two samples. FreeBayes and CRISP were chosen as variant callers for this study. The reason for the use of these packages is described below.

A good variant caller for my study needed to be able to call low frequency variants, given that in pools many variants may on average be represented by only $1/11^{th}$ (9%) of the reads. It would also need to perform well with output from the multiple aligners I used. Among variant callers with these properties, those requiring minimal processing of data files would be preferable. A third criterion would be that the caller has been designed to work with pooled data.

A survey of some popular variant callers was carried out to identify, based on these criteria, the most suitable variant caller(s) for this work. I evaluated assessments in the literature of the popular packages GATK, FreeBayes, LoFreq, VarDict, SNVer CRISP and VarScan, (Cornish & Guda, 2015, O'Rawe *et al.*, 2013, Sandmann *et al.*, 2017, Highnam *et al.*, 2015, Yu & Sun, 2013, Hwang *et al.*, 2015, Laurie *et al.*, 2016, Alioto *et al.*, 2015).

VarScan was eliminated as unsuitable as it is reported to consider divergent base calls as possible sequence variants only if their frequencies exceed 15% (Sandmann *et al.*, 2017) too low for detecting variants in the AR37 pools. The existing literature did not allow an unequivocal ranking of the remaining packages, GATK, FreeBayes, LoFreq, VarDict, SNVer and CRISP. They were all reported to perform better than other variant callers in specific situations used for testing (Laurie *et al.*, 2016, Highnam *et al.*, 2015, Cornish & Guda, 2015, Hwang *et al.*, 2015, Warden *et al.*, 2014), but in each of the above studies a different combination of aligner and variant caller performed best. GATK performed better in calling SNPs (Cornish & Guda, 2015, Highnam *et al.*, 2015, Liu *et al.*, 2013, Pirooznia *et al.*, 2014) and indels (Hwang *et al.*, 2015) from Illumina datasets but there

is a possibility that benchmarking standard used in these studies, may have some bias towards GATK as the variants in the dataset used for testing had also been identified by GATK (Hwang *et al.*, 2015, Zook *et al.*, 2014).

It also appeared from the literature that different variant callers may produce the best results depending on type of data, sequencing platform used, quality of data, sequencing depth and evenness, and type of variants. Also, the applicability of evaluations in the literature to my work may be limited because most of these evaluations used sequence data from model organisms (mostly human) and compared the performance of callers in a specific region of DNA for which variants had been previously identified. The complexity of the DNA varies between organisms and in different regions of the chromosomes (Laurie *et al.*, 2016) with non-coding regions, especially repetitive regions around telomeres and centromeres being extremely complex (McCoy *et al.*, 2014, Alkan *et al.*, 2010, Ye *et al.*, 2011). Results in one region of DNA from one organism for comparison of performance of a variant caller may thus not necessarily be an indication of its performance in general.

Processing requirements and other prerequisites indicated that GATK, LoFreq and VarDict were not well suited for my work. GATK, is extensively used for detecting variants in human genomes when trained with human data (DePristo *et al.*, 2011, Pirooznia *et al.*, 2014, Liu *et al.*, 2013, Nekrutenko & Taylor, 2012). However it is less suitable for analyzing data from non-model organisms where few if any verified sequence variants are known (Nekrutenko & Taylor, 2012). Similarly, LoFreq recommends pre-processing Illumina data files using GATK best practice protocol, which again need a high quality set of known variants (http://csb5.github.io/lofreq/commands/). VarDict requires substantial additional data processing to generate a so-called bed file defining regions in which to call variants (https://github.com/AstraZeneca-NGS/VarDict).

The remaining two packages, FreeBayes and CRISP seemed the most suitable for my work. FreeBayes is designed for use with pooled data, requires little pre-processing of data, can detect variants represented by less than 5% base calls, and do so simultaneously in multiple samples at a time. It has also been reported to work nearly equally well with different read aligners (Hwang *et al.*, 2015) and performed well in comparisons with other variant callers (Hwang *et al.*, 2015, Laurie *et al.*, 2016). FreeBayes is also one of the most

sensitive variant caller currently available; however this comes at a cost of a slight increase in the false positive error rate. The overall precision can be improved by removing the low quality variants as suggested by (Hwang *et al.*, 2015).

CRISP (Comprehensive Read analysis for Identification of SNVs from Pooled sequencing data) is designed specifically for pooled data and performed well on data sets with varying coverages and different numbers of samples per pool (Huang *et al.*, 2015). Therefore, FreeBayes and CRISP were used for the discovery of nucleotide variants.

3.4.5. Detection of AR37 sequence variations

I intended to use two complementary strategies for detecting true variants; i.e. genuine differences between the reference assembly, AR37-Orig, AR37-SAM and AR37-KLP, and to eliminate, as much as possible, instances in which sequencing errors and other artefacts falsely indicated the presence of a variant; i.e. false variants. One strategy was to independently use each of the two variant callers (FreeBayes and CRISP) on read mappings produced by each of the three different mappers (Bowtie2, Bwa-mem, and Novoalign). Any variant present in all of the six resulting sets of putative variants should be more likely to represent a true variant than variants present only in some of these sets. The second strategy was to filter the initially detected putative variants, based on criteria such as mapping quality, DNA region-specific differences in the probability of sequencing errors etc. This should (predominantly) eliminate false variants and lead to a set of variants enriched for true variants.

3.4.5.1. Identification of variants using FreeBayes

In an initial attempt at variant identification, I used FreeBayes. The cut-offs for a number of parameters, namely base call frequency threshold, mapping quality, and base quality needed to be considered beforehand.

3.4.5.1.1. Cut-off settings for detecting potential variants

Given the pooled nature of AR37-SAM and AR37-KLP, the minimum frequency of reads with a variant base call that FreeBayes should consider indicative of the presence of a

variant was a key parameter. Nominally the average frequency of such base calls should be $1/11^{th}$ or 9%. Although considerable effort was made to make sure that equimolar amount of DNA from every individual went into the final pool yet there is a possibility that some individuals may have contributed less DNA than others to the pool, due to technical errors in the sample concentration measurement and during sequencing. Also during size-selection step of the library preparation, different proportions of DNA from different samples may have been retained for further processing. Furthermore stochastic variation in read numbers from each clone is also to be expected. As a result variants may often be represented by less than 9% of reads. Thus it was desirable to set the cut-off for calling of a variant as low as possible. On the other hand, if the cut-off was set too low, the resulting putative variant set would consist largely of false variants, caused by sequencing and mapping errors. It was thus necessary to use a cut-off likely to maximize true variant detection while minimizing the number of false variants. To do so it was necessary to estimate the sequencing error rate in my data.

I derived an initial estimate of the sequencing error rate from a recent study cl in which 7 well-characterized *Candida albicans* loci were sequenced in multiple strains, using the MiSeq platform. The authors reported that, when assessing ~50,000 base calls, on average 99.9% of calls indicated the correct base. Only for one site an error rate of 5% was observed. In other words, based on these data only once every 50 kb would one expect sequencing errors to generate ≥ 5% divergent base calls (Zhang *et al.*, 2018).

I next carried out a similar analysis using my sequence data. For this I mapped the reads from the AR37-Orig single clone against a AR37-Orig assembly (section 3.4.1), using Bowtie2 aligner. The resulting alignment files were then used to calculate various statistics. The mapping rate was very high (99.8%) and the coverage was 98 x, with a mean mapping quality of around 29. The error rate was 0.11%. In other words, at a given site 99.89% of all reads indicated the presence of the same base. This error rate included both sequencing and mapping errors. Also, being a genome wide average, it included low complexity / repetitive / homopolymeric regions that are difficult to sequence and to map reads to - and thus more error-prone. When I assessed the error rates for low-complexity, repetitive and homopolymeric regions I found the average error rate to be 19%, and some homopolymeric regions had error rates as high as 70%. Conversely, based on manually scoring low frequency mismatches in three randomly selected 10 kb regions that

contained a high percentage of non-repetitive DNA, I arrived at an error rate of 0.06% for this DNA.

Next, I attempted to estimate the best cut-off for variant calling. Given the error rates in the difficult-to sequence regions of the genome, I would not be able to find variants in these and therefore the cut-off needed to be based only on the error rate for non-repetitive DNA. I decided on 5%, based on the following calculations: Assuming an error rate of 0.06% at a coverage of 60x (lowest coverage obtained from three samples) the binomial probability of getting ≥ 5%, i.e. 3 or more incorrect base calls out of 60 by chance is 7 x 10^{-6}. Thus, using a 5% cut-off would falsely suggest a possible SNP once every 140 kb. This number is something of an overestimate because most of the SNPs would only be called if the three diverging reads all contained the same alternate base. As there are 27 permutations of the three possible erroneous base calls in three reads and in only three of these all three reads have the same base call (for example G,G,G; C,C,C; or A,A,A, when the base at this position is a T). This would suggest that consistent divergent base calls may occur at a frequency roughly 10 times lower than the error rate, generating false impressions of a putative SNP only around once every 1 Mb. However it must be noted that Illumina platform is reported to generate more errors in GC-rich regions and errors are mostly preceded by 'G' (Dohm *et al.*, 2008) or if the read contains GC-rich motifs, especially the GGC motif, followed by another G/C (Quail *et al.*, 2012). Single base substitution errors are more common in Illumina data than insertions and deletions (Indels) (Hoffmann *et al.*, 2009) and transition and transversion substitutions are not symmetrical (Dohm *et al.*, 2008, Abnizova *et al.*, 2012), implying that sequencing error rates are not the same and thus the likely number of sequencing errors falsely indicating putative SNPs is somewhere between these two estimates. In summary using a 5% cut-off should produce a manageable number of false variants for non-repetitive DNA. Conversely, the binominal probability that a polymorphism present in 9% of the DNA sequenced is represented by < 3/60 reads is 8.5%; i.e. more than 90% of true variants in normal DNA should be identifiable using a 5% cut-off.

While base quality score is another important parameter that can affect variant calling, no cut-off was set prior to variant calling. The reason is that reads had already been trimmed (section 3.3) to their longest contiguous segment for which the probability of calling each base correctly was 99% (p value = 0.01 or Phred scale base quality ~= 20; Fig. 3.8), and

the above error calculations and the setting and evaluation of the 5% cut-off was based on error rates determined using these trimmed reads.

The third important parameter to be set was the minimum mapping quality of reads containing a possible variant above which FreeBayes would report this variant: This would reduce reporting of false variants caused by mapping errors, a mapping quality cut-off can be set. As discussed in section 3.4.2.2 different mappers calculate mapping quality in slightly different ways and thus mapping quality scale used by these three mappers is different. All three mappers generated alignment files with most reads having MAPQ values > 30 which is considered very good (Ruffalo *et al.*, 2011). Since FreeBayes can call variants from multiple alignment files together with better accuracy, I decided to set an initial cut-off value of 15 (for all three aligners), which will call nearly all the variants from all the alignments. Mapping quality is reported for each variant of each alignment in final variant call format (vcf) file and more stringent cut-offs were applied when the variant initially reported by FreeBayes were filtered further with the aim to increase the ratio of true to false variants.

3.4.5.1.2. Variants called by FreeBayes and enrichment for true variants

Using these settings (5% frequency, base quality score > 20 and mapping quality > 15) FreeBayes reported, approximately 30,000 variants, i.e. sites in which one or several samples differed from others or from the assembly, regardless of which of the three aligners had been used (Table 3.5 Column 2). It seemed unlikely that the number of true variants would be this high. As already discussed high error rates in parts of the genome that were problematic in terms of sequencing and mapping would, for instance, be likely to generate false variants, and other factors might further inflate the number of false variants. Since the number of potential variants reported was too high to allow assessment of the individual variants, I applied a number of filters to arrive at a smaller, manageable pool of variants likely to be enriched in true variants, even at the cost of possibly losing some true variants.

A closer inspection of a sample of variants revealed that many were present in all three samples at the same locus. Given that the sequence of the assembly used for comparison was the consensus of the three samples it was unlikely that any of these represented true variants. These variants were therefore removed from the analysis. This reduced the

number of bowtie2 alignment-based, Bwa-mem alignment-based and Novoalign alignment-based variants by ~14%, 22% and 28%, respectively, (Table 3.5, column 3).

I also noted that many of the remaining variants were called in regions that showed read mapping strand bias i.e., variants which were supported by reads on only one of the strands. In addition many showed placement bias i.e. the divergent base calls were positioned at the end of mapped reads. Strand-bias, and placement and position bias are indicative of sequencing errors (Garrison, 2015), and on this basis these were removed as unlikely true variants. To eliminate variants at bases with strand bias, I used two parameters provided by FreeBayes for each potential variant, namely the "number of alternate observations on forward strand" (SAF) and the "number of alternate observations on reverse strand" (SAR). I set threshold of > 0 for both these, meaning that variants were only retained if they were supported by at least one read supporting the alternate allele on each strand. This removed approximately half of the variants that had not been eliminated in the previous step (Table 3.7, Column 4). In order to remove variants that are close to the ends of contigs / reads, the FreeBayes parameters "Reads placed left" (RPL) and "Reads placed right" (RPR) were used which provided a way to remove such variants by counting the reads 'balanced' or 'centered' on both sides of the called variants. Both RPL and RPR were set to > 0 meaning that unless there was at least one read on each side of the variant, it would be removed. This step reduced the number of variants approximately by a further 30%, but the number of remaining variants was still ~10,000 regardless of which aligner had been used (Table 3.5 Column 5).

Many of these variants were in repetitive areas. Although repetitive regions are likely to be rich in variation due to their high mutability (Legendre *et al.*, 2007), they present substantial difficulties in assembling and subsequently mapping reads. I had established in last section (page 89), that this affected error rates to the point that detection of true variants in such regions was impossible in pools due to the error-induced prevalence of false variants. There was no easy way to directly identify in the FreeBayes output which putative variants were located in repetitive areas. However, since it was not repetitiveness itself, but the resulting sequencing and alignment errors that generated false variants, I could use assembly and mapping quality indicators to identify any problematic regions in the AR37 genome and remove variants mapping to such regions.

One such indicator is mapping quality (MAPQ). As discussed earlier in section 3.4.2.2 each aligner uses different scales e.g. max MAPQ value for bowtie2 alignments is 42, for Bwa-mem alignment it is 60 and for Novoalign it is 70. Also, it is evident from figure 3.12 that most of the reads had a MAPQ values of close to maximum value for Bwa-mem and Novoalign while Bowtie2 alignment had ~10% reads with a slightly lower than maximum MAPQ values. For these reasons different MAPQ cut-offs were set for all three alignments so as to remove reads with low MAPQ but still retain enough reads to call variants from most part of the genomes. I set a MAPQ threshold of 20 for bowtie2 alignments, 35 for Bwa-mem alignments, and 40 for Novoalign alignments. This removed a further 3% of variants (Table 3.5 Column 6).

Regarding MQM filtering, I should add that that while the presence of true polymorphisms does reduce mapping quality it does not do so significantly. Mapping quality is related to "uniqueness" and indicates how confidently the aligner can assign a read to the true origin of read in reference (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#mapping-quality-higher-more-unique). An alignment will be considered unique if it has a considerably higher alignment score than the rest of the alignments of that read. The bigger the gap between the alignment with the best alignment score and the second best alignment, the higher its mapping quality. Since mapping quality is assessed based on the unique mapping of the entire read, SNPs would have little impact. The same applies to indels (a) because a sliding window is used when assessing mapping quality and thus most of the read would usually still be a good match and (b) because uniqueness of the mapping is a key determinant of mapping quality-in non-repetitive regions at least, even a read representing an indel would be likely to only map to a unique position.

A large number of the remaining variants were in low coverage regions i.e. regions where total read depth (DP) was low (< 40). In these, stochastic variations are more likely to cause divergent base frequencies to exceed the 5% minimum frequency set for detecting polymorphism, especially since the low coverage tends to be associated with "unreliable" areas of the genomes, i.e. areas in which sequencing or assembly are problematic (Benjamini & Speed, 2012, Oyola *et al.*, 2012, Bentley *et al.*, 2008, Dohm *et al.*, 2008, Aird *et al.*, 2011, Laurie *et al.*, 2016). Thus low coverage does not only directly generate

problems in variant detection but also indicates areas in the genome that are "problematic". I therefore used low DP as an additional criterion for eliminating variants.

In choosing a DP cut off, the sample with the lowest coverage i.e. AR37-SAM (mean coverage ~60x; Table 3.4) was taken into consideration and a DP cut-off value of 40x was set. At this DP cut-off value, variants could still be called from ~80% of the AR37 assembly for AR37-SAM sample while for other two samples variants could be called from ~95% of the AR37 assembly; Figure 3.10b). Applying the 40x DP filter removed approximately 91% of the remaining variants (Table 3.5 column 7).

Visual inspection of some of ~1000 remaining variants for each aligner revealed that many were, or were in close proximity to insertions or deletions (indels). The accuracy of most variant callers including FreeBayes to call variants is reduced around indels (O'Rawe *et al.*, 2013, Fang *et al.*, 2014, Hasan *et al.*, 2015). Even Sanger sequencing struggles to correctly detect most indels (Bhangale *et al.*, 2004). On this basis, I decided to remove indels as well and thus only single nucleotide variations (SNVs) were selected for further analyses. After removing indels, 341 SNPs from bowtie2-aligned data, 366 SNPs from Bwa-mem aligned data and 332 SNPs from Novoalign aligned data remained (Table 3.5, Column 8).

Visual inspection of these remaining SNPs revealed that many of them were at the ends of homopolymers i.e. long stretches of the same nucleotide. Homopolymers are problematic for all sequencing platforms and short read aligners struggle to map reads correctly to such regions. Therefore SNPs within 3 bp of a homopolymeric region (defined as region in which the same base occurred more than 4 times in a row) of were also removed. A custom Python script (by Dave Wheeler) was used to identify such homopolymers, in the reference genome and then all variants within 3 bp of these regions were removed. This step eliminated a further ~15% variants from bowtie2 and bwa data and ~20% variants from Novoalign data, leaving 293 SNPs from bowtie2-aligned data, 308 SNPs from Bwa-mem aligned data and 266 SNPs from Novoalign aligned data (Table 3.5, Column 9).

3.4.5.1.3. AR37-Orig assembly and reads were used to validate the above filters

As discussed above (section 3.1.1 and 3.4.1.1.) a *de novo* assembly was generated from only AR37-Orig reads. This particular assembly was not used as a reference for mapping reads from pooled samples and calling variants. Instead another assembly generated from combined reads from all the three samples was used as a reference standard. However, presence of AR37-Orig reads at a coverage of ~100x provided me with an alternate approach to reduce to number of erroneous variant calls. Since the assembly was generated only from AR37-Orig reads, mapping these reads back to this assembly and calling variants should ideally have resulted in no variants at all. However, this was not the case and more than > 10,000 variants were called at a 5% frequency cut-off, the same cut-off used to call variants from all three samples using "AR37-pool" assembly as reference. All these variants were a result of errors during sequencing, mapping or variant calling steps and a careful look at these variants helped to identify potentially miss-assembled areas of the assembly and potential sequencing and mapping errors. Most obvious false variants seemed to be located (i) at the ends of contigs ii) at the ends of homopolymers and simple sequence repeats iii) at the ends of sequencing reads iv) in low coverage areas v) in reads with low mapping quality vi) variants supported by only forward or reverse reads. Applying filters to remove variants from all such variants from AR37-Orig data reduced the total number of remaining variants to < 30 indicating that these filters actually removed most of the false variants. Ideally there should be no variants after the filters have been applied but no combination of the filters was able to bring the variants down to zero. One possible reason could be the presence of gene families. If there are very similar multiple genes within a gene family then chances are that such regions may not be assembled accurately using Illumina short reads and may merge together as one region in the final assembly. In such case reads that originated from these multiple genes would not be able to map accurately to the collapsed one region in the assembly giving rise to false variant calls despite all the parameters indicating it a true variant.

The knowledge obtained from mapping AR37-Orig reads against AR37-Orig assembly was used to identify the types of erroneous calls and remove them from our variant data set of pooled samples. It also supported all the filters used in section 3.3.3.1.2 to remove

false variants. These filters when applied to AR37-pool data reduced the number of variants from 30,000 to ~250. After applying each filter, a random sample of the remaining variants was screened visually using Integrative Genomic Viewer (IGV). This visual screening at each step allowed me to assess the effectiveness of the applied filter in removing false positive variants. Fifty randomly picked variants from the final dataset for each mapper were screened visually to ascertain that all of them were located in non-repetitive regions of good coverage away from the read and contig ends and were supported by reads with good mapping qualities and no strand bias. Such variants were good candidates to be considered as true positive (Figure 3.16). As discussed above, repetitive regions mutate more often than the non-repetitive areas of the genome and they may be an important source of variations in asexually reproducing endophytes. Exclusion of repetitive areas from our analyses means that there is a fair chance that a good number of true positive variants have also been filtered out. However, these regions are difficult to sequence and map accurately and retaining them may have resulted in too many false positive variants and compromised the analyses.

3.4.5.1.4. Of all SNPs 139 were detected in all three alignments

In an attempt to further enrich for true variant among the remaining combined 867 variants obtained by analysing data produced by the three different aligners, I investigated to what degree the three sets of variants overlapped and identified SNPs common to all sets. Files of filtered variants from each of the three alignments were intersected and venn diagrams generated showing the overlap in SNP calls from the different alignments (Figure 3.13). I found 139 SNPs (16% of all SNPs) common to the three sets, and this set was likely to contain the highest percentage of true variants.

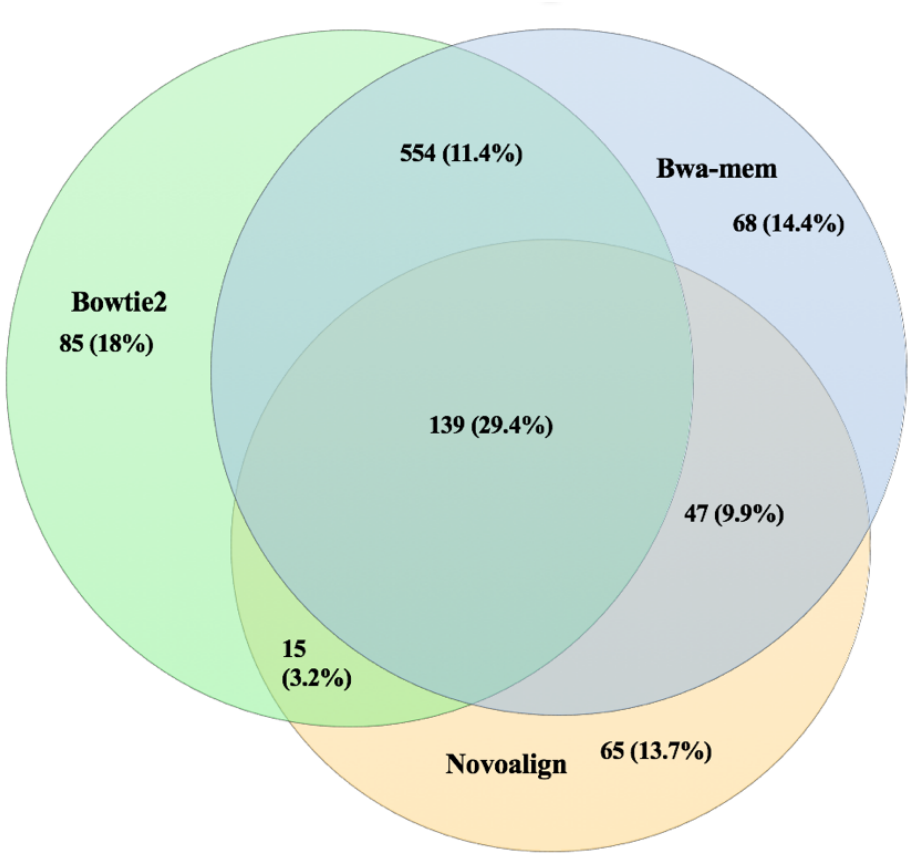


Figure 3.13. Venn diagram showing the intersection of variants called by FreeBayes after filtering, identified in reads aligned to the reference AR37 assembly by bowtie2, Bwa-mem and Novoalign.

Table 3.5. No of variants called by FreeBayes using data from three aligners i.e. Bowtie2, Bwa-mem and Novoalign. Also indicated are the number of remaining variants after 7 different filters were applied (column 3 to 9).

Aligners	Raw Entries	All 3 samples variants removed	SAF & SAR < 1 removed	RPL & RPR < 1 removed	Low MQM removed	DP < 40 removed	Indels removed	Homopolymers Removed
Bowtie2	31405	27016	12726	9282	9017 (20)	761	341	293
Bwa-mem	32844	25577	13911	10116	9804 (35)	829	366	308
Novoalign	27393	19644	12055	8657	8430 (40)	796	332	266

3.4.5.2. Identification of probable variants using CRISP

I next used 'Comprehensive Read analysis for Identification of single nucleotide variations (SNVs) from Pooled sequencing data' (CRISP) to find potential variants, i.e. sites in which AR37-Orig. AR37-SAM or AR37-KLP differed from the reference assembly, using as before the bowtie2, Bwa-mem and Novoalign alignments. CRISP compares DNA sequences from the multiple pools to find rare and common SNPs, an analysis not built into FreeBayes. This cross-pool comparison approach helps to identify rare variants from sequencing errors (https://github.com/vibansal/crisp). The parameters used for CRISP were the same as in the FreeBayes attempt: minimum base quality for variant call was set to 20 and minimum read mapping quality (mmq) cut-off was set to 20 for bowtie2, 35 for Bwa-mem and 40 for Novoalign data (as previously in section 3.4.4.1.2).

CRISP reported 10,409, 10,063 and 7,100 variants, respectively, in bowtie2, Bwa-mem aligned, and Novoalign alignment data (Table 3.6 column 2). The resulting variant call format (VCF) file was filtered largely as described for the FreeBayes analysis. Briefly, in the first step all variants present in all 3 samples at same locus were removed. This filter removed highest percentage of variants from Novoalign data. Then variants showing strand bias, low coverage, and low mapping quality were removed followed by removal of indels and variants from homopolymeric regions. All these filters had greater impact on bowtie2 and Bwa-mem data than on Novoalign data (Table 3.6). No placement bias filter could be applied as CRISP does not report "Reads Placed Left" (RPL) and "Reads Place Right" (RPR) or equivalent parameters.

As mentioned in the previous paragraph that the minimum read mapping quality (mmq) cut-offs were set to 20, 35 and 40 for Bowtie2, Bwa-mem and Novoalign data respectively for calling variants using CRISP. The same cut-offs were set for calling variants using FreeBayes. These cut-offs ensured that no variants were called from the reads having mapping quality lower than these defined thresholds. However, the final CRISP output file also reported mapping quality for reads that supported reference alleles at each variant site. The number of reference reads were split in four different mapping quality ranges i.e. (i) number of reads having mapping quality (MQ) between 0 and 9 (ii) number of reads having MQ between 10 and 19 (iii) number of reads having MQ between 20 and 39

and (iv) number of reads having MQ > 40. Since MQ cut-offs were set for reads supporting variants before variant calling and no variants was called from reads with MQ < 20, 35, and 40 for bowtie2, Bwa-mem and Novoalign data respectively. The reads with MQ < 20 were those supporting reference allele and may indicate potential misassembly. To remove the potentially misassembled areas out of analysis, all the sites where more than 10% of reads had a MQ less than 20 were also filtered out of analyses. No such values are reported by FreeBayes, which only reports one mean mapping quality score for reference reads at each position. The difference in output of both variant callers i.e. FreeBayes and CRISP makes it difficult to compare all the parameters as such. This additional filter used in CRISP analysis had least effect on Novoalign data as only ~35 of the variants were removed while ~50% of variants from bowtie2 and Bwa-mem aligned data were removed by this filter (Table 3.6 column 6). This was expected as Novoalign data has higher mean (and absolute) mapping quality and is expected to contain only few reads with < 20 mapping quality. On the other hand, mean mapping quality for bowtie2 is the lowest among the three aligners and it is expected to contain more reads with mapping quality < 20.

As before indels and variants associated with homopolymeric regions were also removed. After these filtration steps bowtie2-aligned data contained 56 variants, Bwa-mem 98 and Novoalign 243, significantly fewer than in the FreeBayes analysis.

3.4.5.2.1. Using CRISP a smaller percentage of CRISP SNPs was shared among variants detected by the three aligners

The intersection of filtered variants based on data produced from all three aligners using CRISP is depicted in Figure 3.14. Only 7% (29 SNPs) had been detected in all three alignments, less than ½ of the percentage among FreeBayes-called variants.

Table 3.6. Variants called by CRISP using three aligners and number of remaining variants after each filtration step.

Aligners	Total variants called	3-samples variants Removed	SAF/SAR<1 removed	DP<40 removed	LowMQ10 removed	Indels removed	HomoPolymers Removed
Bowtie2	10409	3043	929	249	128	109	56
Bwa-mem	10063	2647	966	363	188	174	98
Novoalign	7100	1430	910	354	342	327	243

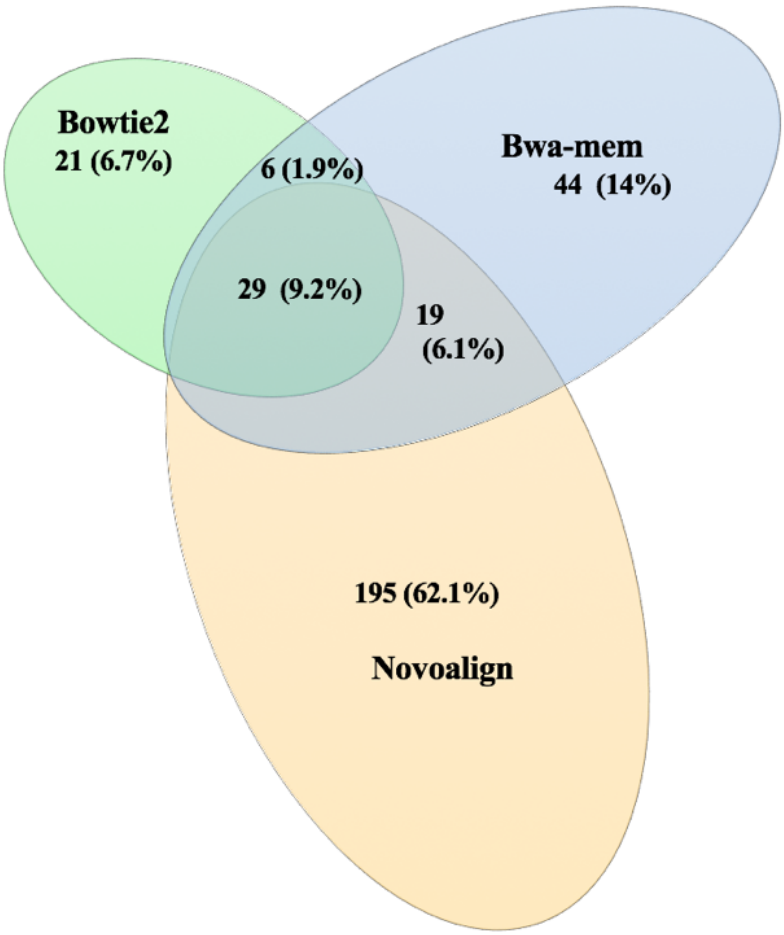


Figure 3.14. Venn diagram showing intersection of variants called by CRISP after filtering identified in reads aligned to the reference AR37 assembly by bowtie2, Bwa-mem and Novoalign data.

3.4.5.3. Nineteen out of twenty variants, identified by both FreeBayes and CRISP are specific to the seed-propagated lines, as expected for true variants

A total of 20 variants had been called by both FreeBayes and CRISP regardless of which aligner was used (Figure 3.15). This set would be expected to contain the highest percentage of true variants. Looking at all 20 variants using the Integrative Genomics Viewer (IGV) revealed that although the majority of them were located in somewhat repetitive regions (Figure 3.16) their good coverage and high mapping qualities made them the most promising candidates to be considered true variants. Eleven of these variants were unique to AR37-SAM and 8 were unique to AR37-KLP but only a single variant was unique to AR37-Orig. This distribution indicated that filtering had indeed generated a set of 20 variants enriched in true variants: Variants should be easiest detectable in AR37-Orig, given the high coverage and the fact that this represented a single clone and variants should, on average be represented by a high percentage of all reads. Indeed the AR37-Orig specific variant was supported by high percentage of reads i.e. 80%. On the other hand, true variants distinguishing the samples from the ancestral AR37, represented by the assembly would be expected to occur predominantly in AR37-SAM and AR37-KLP, given that, unlike AR37-Orig, these were separated from the ancestral AR37 by several generations of seed transfer in a new host, likely to selectively favour variants. Thus the distribution of variants was in accordance with biological expectations, as expected for true variants.

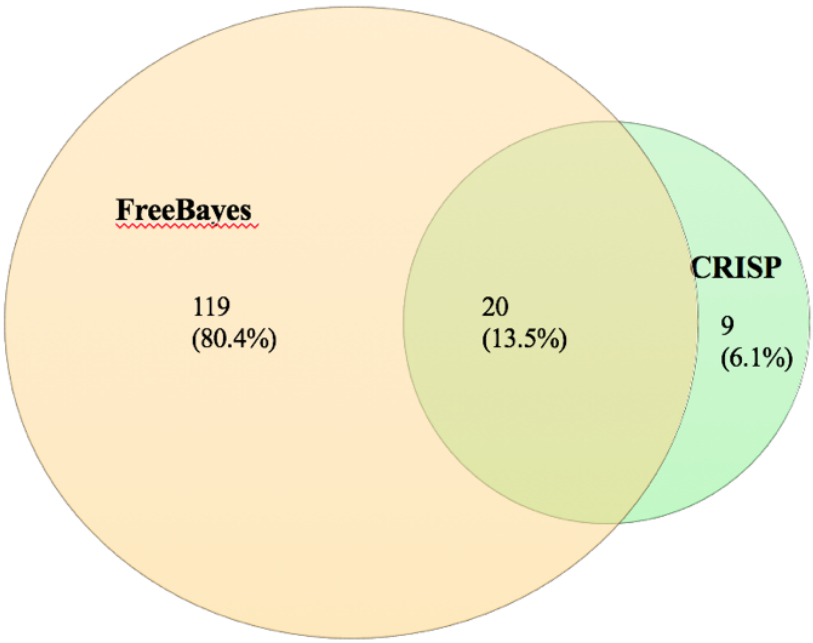
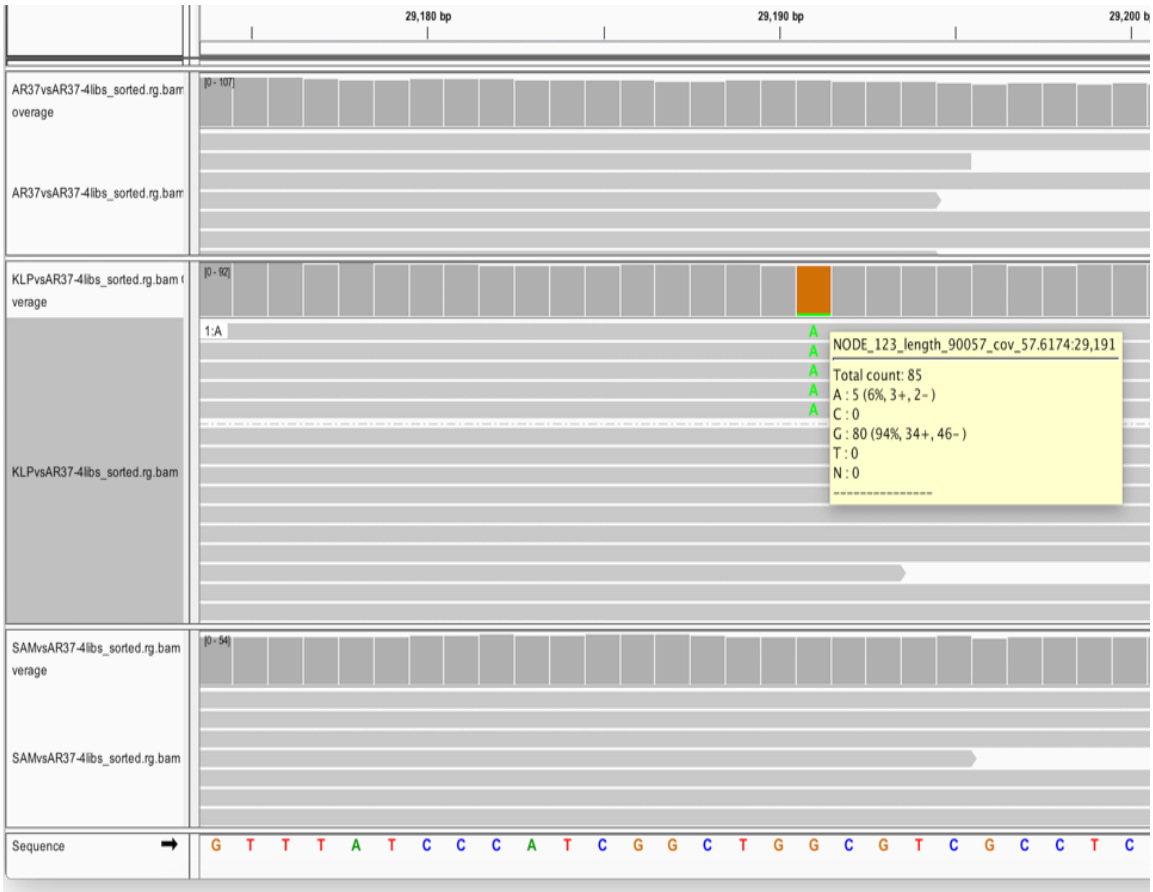
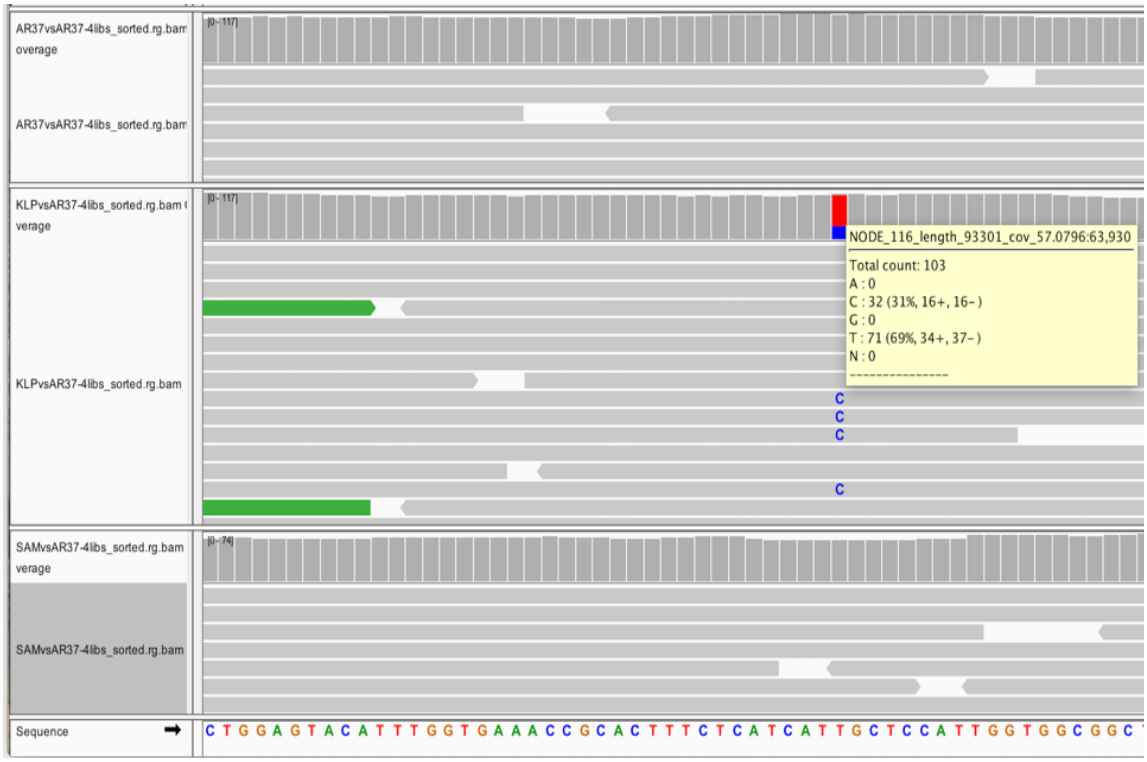


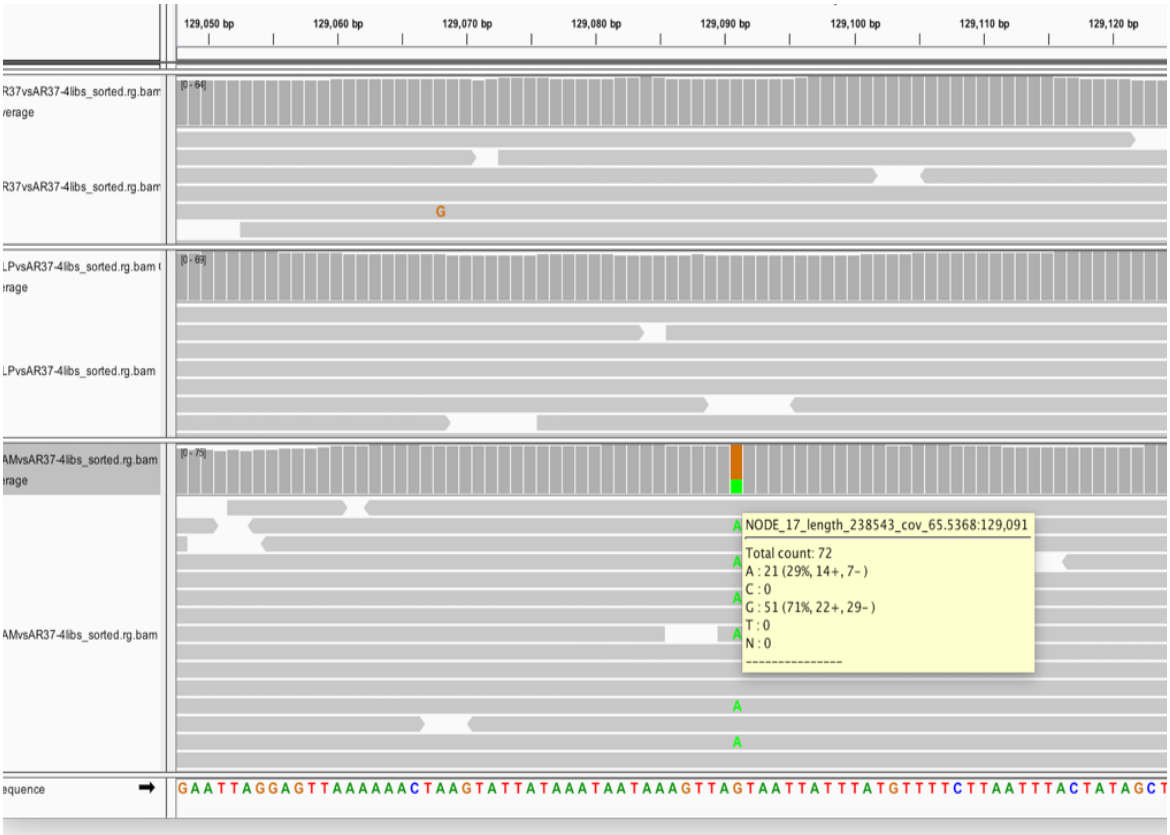
Figure 3.15. Venn diagram showing intersection of variants called by FreeBayes and CRISP.



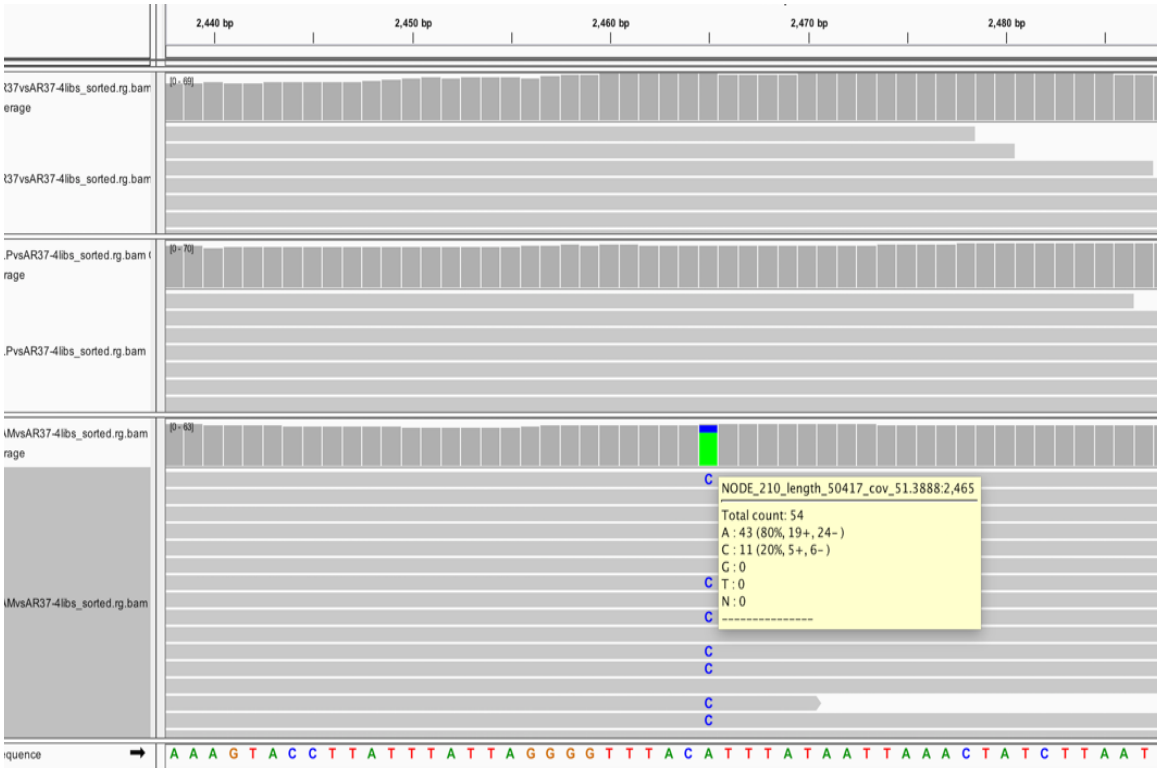
a


AR37vsAR37-4libs_sorted.rg.bam
overage
AR37vsAR37-4libs_sorted.rg.bam
KLPvsAR37-4libs_sorted.rg.bam
verage
KLPvsAR37-4libs_sorted.rg.bam
SAMvsAR37-4libs_sorted.rg.bam
verage
SAMvsAR37-4libs_sorted.rg.bam
Sequence
NODE_116_length_93301_cov_57.0796:63,930
Total count: 103
A : 0
C : 32 (31%, 16+, 16-)
G : 0
T : 71 (69%, 34+, 37-)
N : 0

b


129,050 bp
129,060 bp
129,070 bp
129,080 bp
129,090 bp
129,100 bp
129,110 bp
129,120 bp
R37vsAR37-4libs_sorted.rg.bam
verage
R37vsAR37-4libs_sorted.rg.bam
LPvsAR37-4libs_sorted.rg.bam
rage
LPvsAR37-4libs_sorted.rg.bam
AMvsAR37-4libs_sorted.rg.bam
rage
AMvsAR37-4libs_sorted.rg.bam
equence
NODE_17_length_238543_cov_65.5368:129,091
Total count: 72
A : 21 (29%, 14+, 7-)
C : 0
G : 51 (71%, 22+, 29-)
T : 0
N : 0

c



d

Figure 3.16. Screenshots of 4 potential variants as seen through IGV. The yellow box contains information on total read count, variant and reference bases, number and percentage of reads supporting variant and reference base and number of reads on forward and reverse strand supporting variant and reference base. The sequence at the bottom is that of reference assembly.

3.4.6. How much has AR37 altered during propagation?

While the above set of 20 SNPs probably contains the highest percentage of true variants, it is unlikely to contain all true variants, as it is the end-product of a very stringent filtering process. I next sought to determine if a somewhat less stringent filtering (considering all 139 variants called by FreeBayes using all three mappers (section 3.4.4.1.3), and all 29 variants called by CRISP using all three mappers (section 3.4.4.2.1) also generated a distribution of variants between the three samples in accordance with biological expectations and thus indicating that these sets were also enriched in true variants. Indeed this was the case, and thus, based on these larger numbers of variants some estimate of the degree of variation during seed propagation in new hosts can be arrived at (Fig. 3.17):

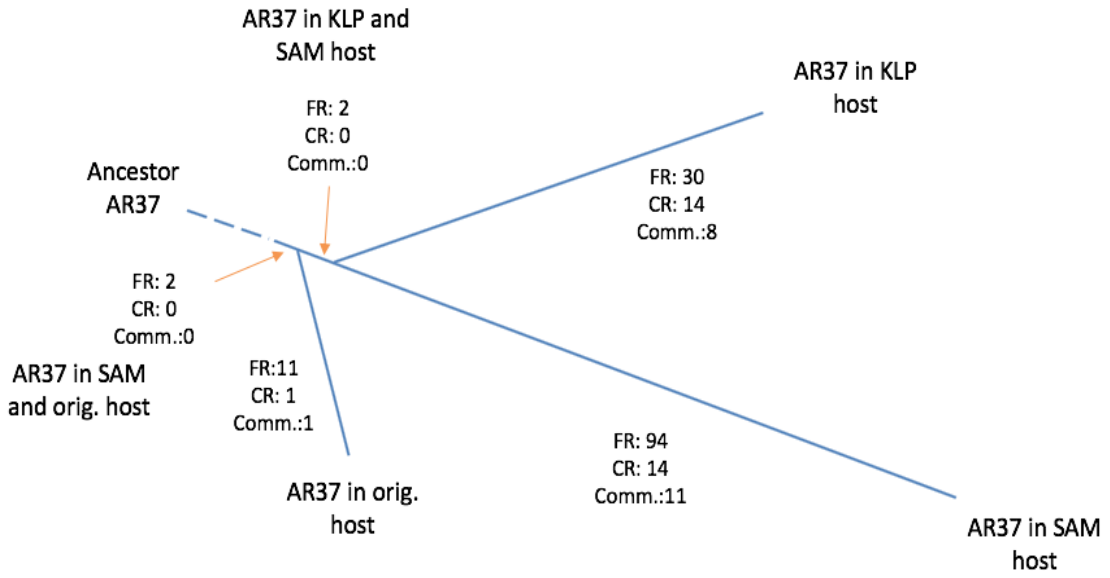


Figure 3.17. Tree representation of the SNPs identified by both variant callers and inference of when these mutations have arisen. FR: SNPs inferred using FreeBayes; CR: SNPs inferred using CRISP, Comm: SNPs inferred by both. The length of the branches indicates, semi-quantitatively, the number of mutations separating the different AR37 lineages analysed from the ancestor. The sequence of *de novo* AR37 genome assembly, generated by combining reads from all three samples and used as reference in this study, is assumed to represent the sequence of a common ancestor to all these three samples.

as was the case for the set of 20, the two larger sets of variants also indicated that significantly more variants had arisen during seed propagation in new hosts than during vegetative propagation in the original host and, in addition that the number of variants kept increasing with the number of seed transfers (7 for AR37-SAM and 2 for AR37-KLP). I note that the variants in AR37-SAM have been identified in a relative smaller portion of the genome than other two samples. While all other parameters may have equally affected all 3 samples, the coverage cut-off filter of 40 retained around 80% of AR37-SAM genome. Both other samples i.e. AR37-Orig and AR37-KLP retained around 95% of their genomes to be analysed at this cut-off value (Fig.3.11b).

Interestingly SNPs were found that were present in AR37-KLP and AR37-SAM but were absent in the AR37-Orig clone. These would most likely represent differences that have arisen before introduction into new SAM and KLP grass cultivars and could characterize a sub-population of AR37 more easily transferred into new hosts (Figure 3.17). Also there were two SNPs that were shared between AR37-Orig and AR37-SAM but not found in AR37-KLP. These may have arisen during serial propagation of AR37-Orig before it was

inoculated into AR37-SAM, alternatively this SNP could have been lost in the AR37-KLP line.

The frequencies of alternate base calls were high for variants unique to AR37-Orig (5-90%) Even higher numbers would have been expected, because it is believed that during vegetative propagation each new tiller is infected by a very small number of hyphae (Christensen *et al.*, 2000b). Because of these bottlenecks I expected AR37-Orig to be represented by a single clone, but it is conceivable that it does represent a collection of a small number of clones.

Similar high percentages of reads supporting alternate base calls may have been expected for variants unique to AR37-SAM or AR37-KLP, if one or a few exceptionally well-adapted clones had expanded through the host population in the seed propagation process. However most of the variants unique to either AR37-SAM or AR37-KLP- were represented at low frequencies (5-30% and 5-46%, respectively), in line with the existence of large numbers of competing variant clones in these populations, most represented by only one of the 11 clones contributing to the pooled DNA.

3.4.7. A significant portion of SNPs identified by each variant caller was located within or in close proximity of ORFs and could impact on phenotype

An important question was how much functional impact of the SNPs identified were likely to have. A functional impact is more likely, and its nature easier to deduce, if the SNP is located adjacent to or within an ORF. I therefore used the 1500 bp flanking sequences on either side of a SNP as a query in a BLASTn search against a local BLAST database created from the GO annotated well-curated M3 gene (exon) models of *E. festucae* E2368. This was carried out for (i) SNPs identified as common in all three alignments by FreeBayes (139 SNPs). (ii) SNPs identified as common in all three alignments by CRISP (29 SNPs) and (iii) the 20 SNPs that were shared between (i) and (ii) i.e. SNPs identified both by FreeBayes and CRISP.

The FreeBayes SNPs had 401 hits against 209 exons. The CRISP SNPs had 18 hits against 10 exons and the SNPs that were called by both CRISP and FreeBayes had 13 hits against 7 exons. A list of all ORFs and their annotations is given in Supplementary Table 1. Many

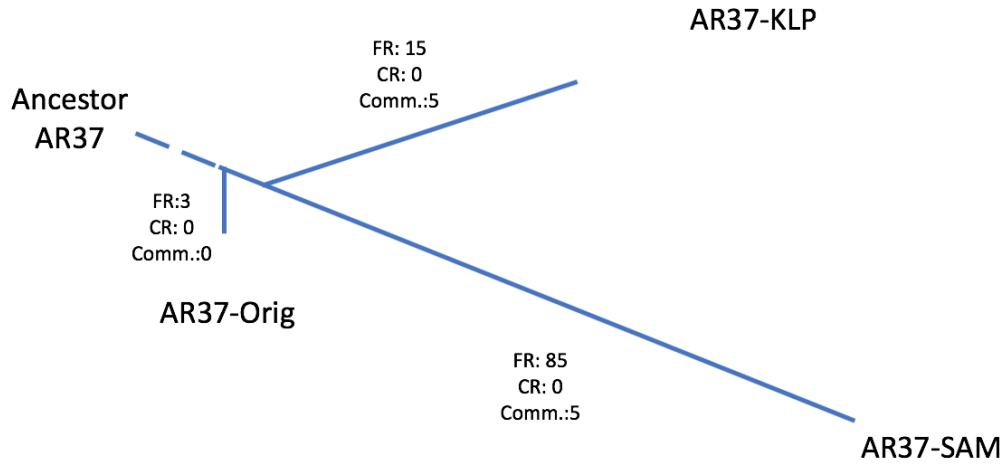
of these hits were located upstream, rather than within gene models. Such SNPs still can potentially have some impact on phenotype if they affect promoter regions. A preferred association of SNPs with particular types of genes may indicate genes that are particularly important in adaptation to a new host on the basis that mutations in such genes would enhance fitness, i.e. be more likely be passed on to the next generation. However a GO enrichment analysis, using the E2368 GO annotations, of the three sets of SNP-affected genes (with SNPs either upstream or within the ORF) showed no statistically significant over- or under-representation of any categories in these sets.

I next repeated this analysis considering only genes in which the SNP was located in the protein-coding region. To do so, ORFs were predicted in the AR37 assembly and for those matching E2368 ORFs the impact of identified SNPs on the amino acid sequence of proteins was predicted. FreeBayes-called SNPs were predicted to impact on 78 proteins, and CRISP-called SNPs on 3 proteins (all of the latter had also been called by FreeBayes; Supplementary Table 1).

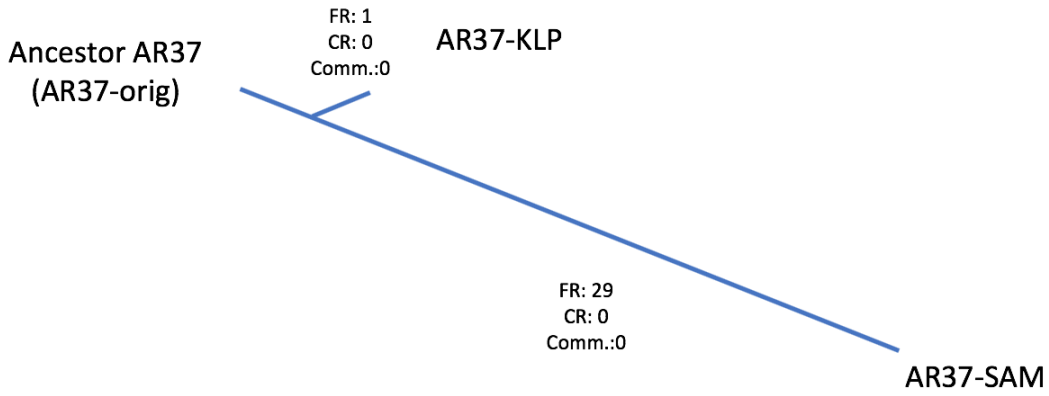
It was conceivable that different types of genes were important in (improving the) interaction with the SAM host, and the KLP host. I therefore charted separately the GO terms associated with coding regions affected by AR37-SAM specific SNPs and AR37-KLP- specific SNPs. A simplified version of the GO categories, achieved by removing the intermediate GO terms and only showing the broader terms, associated with each sample is shown in Figure 3.18. AR37-KLP SNP–affected genes were mostly predicted to function as nuclear proteins, involved in intracellular signal transduction, with affinity to substances such as zinc ions, calcium ions, DNA and actin. AR37-SAM SNP–affected genes were associated with signal recognition particle, ribosomes, endoplasmic reticulum and integral component of membranes, all of which may have role in communicating with the host and establishing a symbiotic relationship.

However, again no enrichment of any categories was observed among genes with alterations in protein-coding regions affected by SNPs unique to AR37-SAM or AR37-KLP. As the number of proteins was fairly small, there was a possibility that 5% false discovery rate (FDR) cut-off may be too stringent (type 2 error). However, when I repeated the GO enrichment analysis for all the above categories with a 10% FDR cut-off value, I again saw no over- or under-representation for any of the GO category

Having identified SNPs that mapped closely to or within ORFs, I also analysed how the number of such SNPs increased during propagation. I did so for SNPs mapping close to and within ORFs (Figure 3.18a) and SNPs with predicted impact on proteins i.e. within coding sequences (Figure 3.18b). The accumulation of these SNPs over time (Fig. 3.18) was similar to those of all SNPs (Figure 3.17) with the possible exception that 6 generations of adaptation to SAM, compared to 3 generations of adaptation to KLP, was associated with up to 29 times more SNPs affecting proteins (29:1) compared to 3.1 times more SNPs overall (94:29 SNPs)

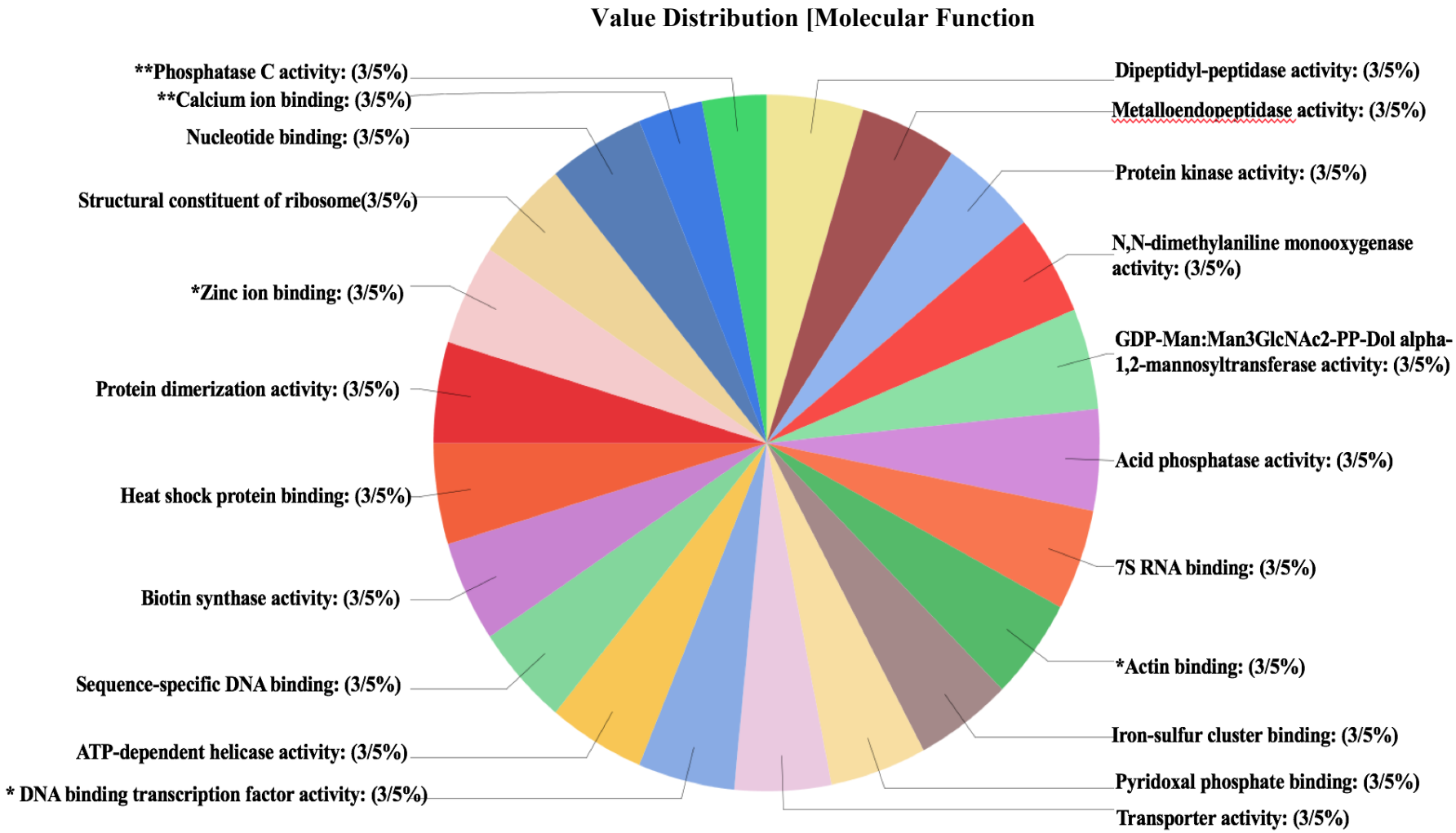


a

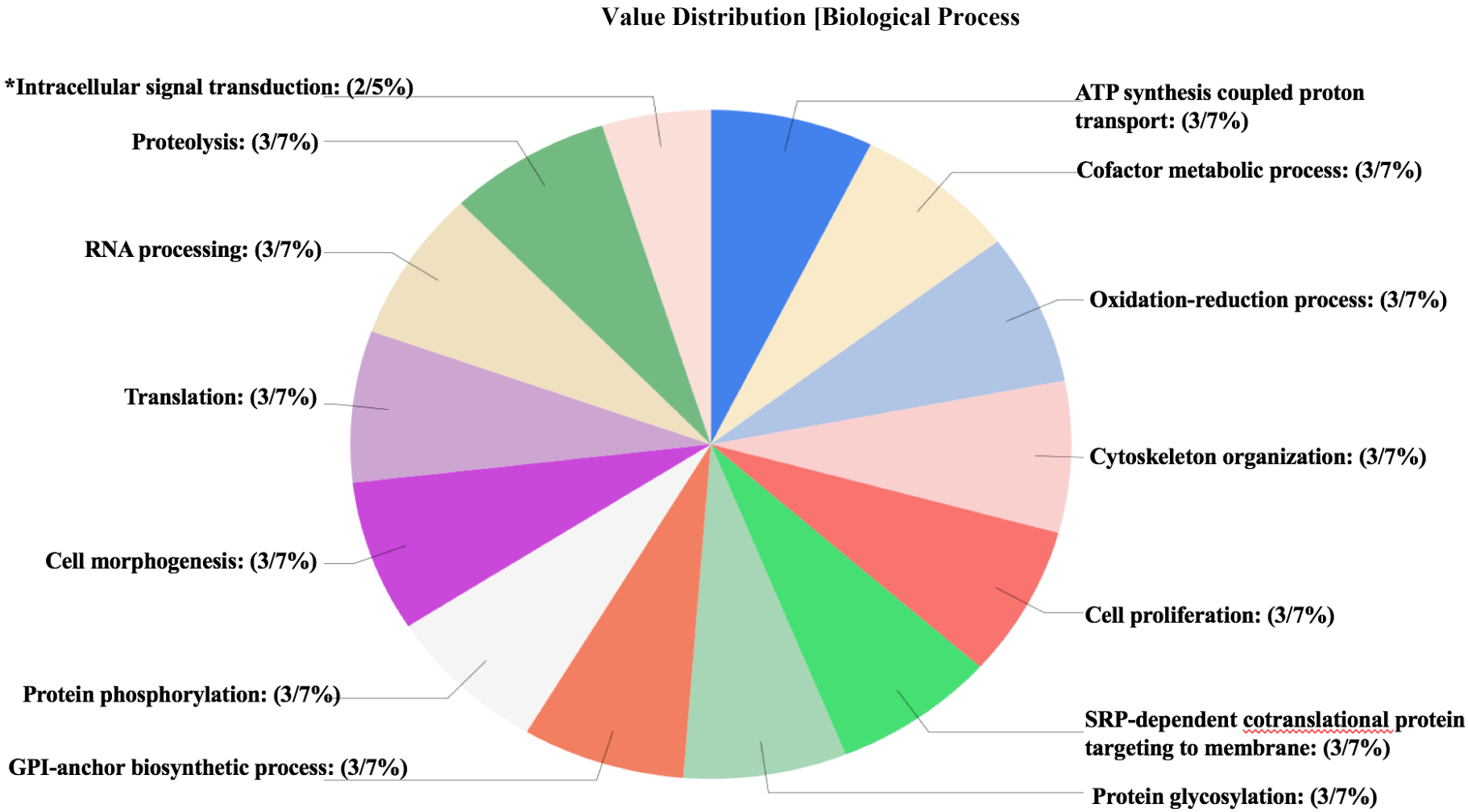


b

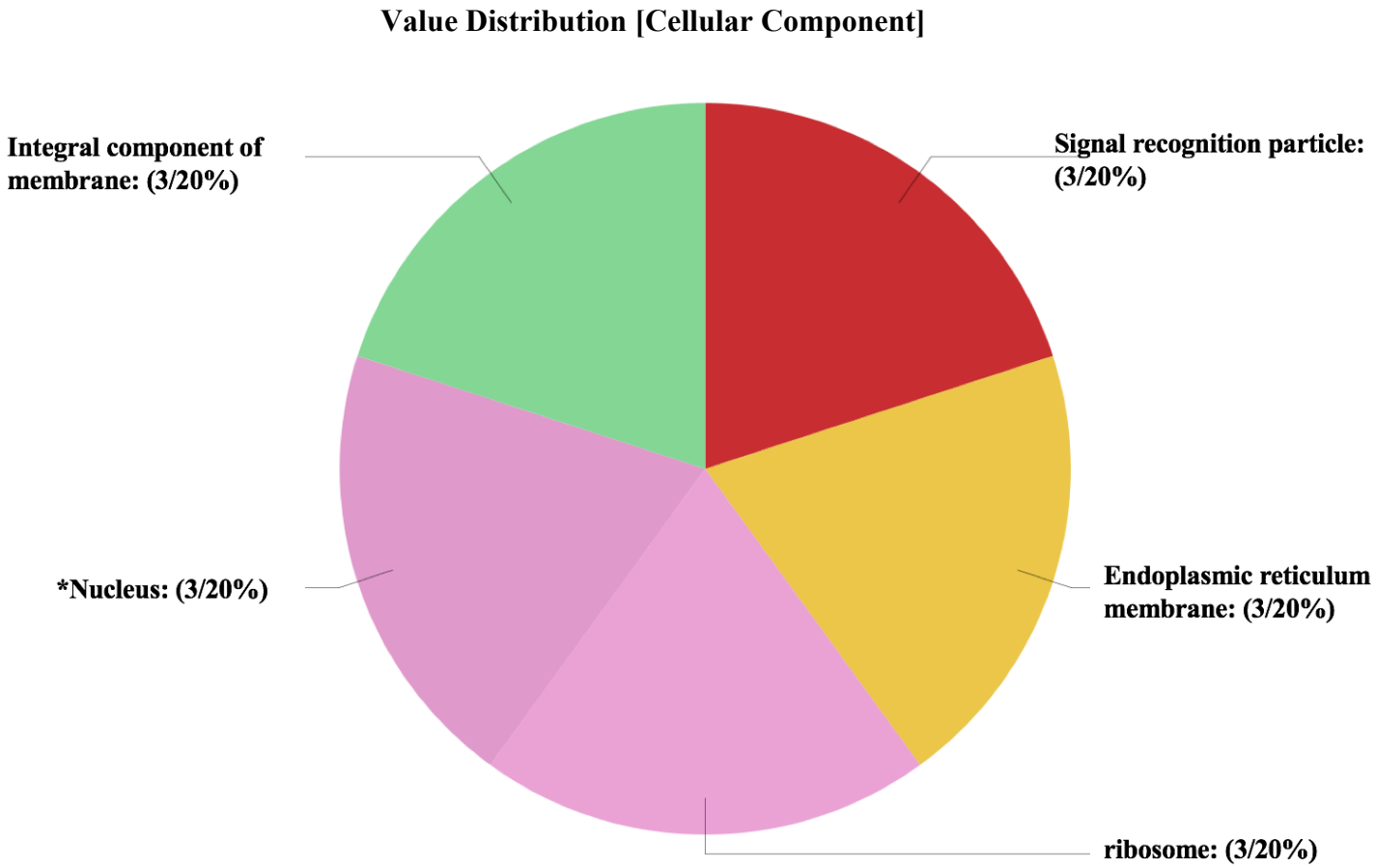
Figure 3.18. Tree representation of the SNPs identified and inference of when a) SNPs mapped close to or within ORFs and b) SNPs with predicted impacts on proteins. FR: SNPs inferred using FreeBayes; CR: SNPs inferred using CRISP, Comm: SNPs inferred by both. The length of the branches indicates, semi-quantitatively, the number of mutations separating the different AR37 lineages analysed from the ancestor.


Value Distribution [Molecular Function
**Phosphatase C activity: (3/5%)
**Calcium ion binding: (3/5%)
Nucleotide binding: (3/5%)
Structural constituent of ribosome(3/5%)
*Zinc ion binding: (3/5%)
Protein dimerization activity: (3/5%)
Heat shock protein binding: (3/5%)
Biotin synthase activity: (3/5%)
Sequence-specific DNA binding: (3/5%)
ATP-dependent helicase activity: (3/5%)
* DNA binding transcription factor activity: (3/5%)
Dipeptidyl-peptidase activity: (3/5%)
Metalloendopeptidase activity: (3/5%)
Protein kinase activity: (3/5%)
N,N-dimethylaniline monooxygenase activity: (3/5%)
GDP-Man:Man3GlcNAc2-PP-Dol alpha-1,2-mannosyltransferase activity: (3/5%)
Acid phosphatase activity: (3/5%)
7S RNA binding: (3/5%)
*Actin binding: (3/5%)
Iron-sulfur cluster binding: (3/5%)
Pyridoxal phosphate binding: (3/5%)
Transporter activity: (3/5%)

a


Value Distribution [Biological Process
*Intracellular signal transduction: (2/5%)
Proteolysis: (3/7%)
RNA processing: (3/7%)
Translation: (3/7%)
Cell morphogenesis: (3/7%)
Protein phosphorylation: (3/7%)
GPI-anchor biosynthetic process: (3/7%)
ATP synthesis coupled proton transport: (3/7%)
Cofactor metabolic process: (3/7%)
Oxidation-reduction process: (3/7%)
Cytoskeleton organization: (3/7%)
Cell proliferation: (3/7%)
SRP-dependent cotranslational protein targeting to membrane: (3/7%)
Protein glycosylation: (3/7%)

b



c

Figure 3.19 . Filtered (only showing broad terms) GO terms associated with SNPs in each sample in 3 major groups i.e. a) Molecular Function b) biological process and c) Cellular component. ** shows GO terms associated only with AR37-KLP sample. * shows GO terms shared by AR37-SAM and AR37-KLP . All GO terms without any * are associated with AR37-SAM

3.5. How does AR37 differ from other *E. festucae* Fl1

E. fectucae AR37 differs from many other *E. fectucae* strains in that it does not produce a number of key secondary metabolites i.e. indole-diterpenes, lolines, ergot alkaloids and peramine (Hume *et al.*, 2007) (http://www.ar37.co.nz/ar37-vs-other-endophytes/) that are normally associated with improved survival of the host (Schardl *et al.*, 2004). As the survival of seed-borne endophytes benefits from increasing the fitness of the hosts, it is likely that other secondary metabolites fulfil this role in AR37. One possible candidate is the alkaloid molecule epoxy-janthitrem, which is synthesized in AR37 but not in other *Epichloë* strains (http://www.ar37.co.nz/ar37-vs-other-endophytes/) (Tapper and Lane 2004). It is uncertain whether epoxy-janthitrem suffices to compensate for AR37's inability to synthesize other alkaloids known to protect the host or to increase its fitness by other means. The new high quality AR37 assembly I produced provided an opportunity to investigate AR37s potential metabolic capabilities, and how these differed from other *Epichloë* spp.. To do so I compared the AR37 genome with the well-characterized *E. festucae* genome of strain Fl1.

The Fl1 assembly is a complete and ungapped assembly (Winter *et al.*, 2018), containing 7 chromosomes and a mitochondrial genome, ideally suited for the detection of the presence and absence of AR37 homologues by mapping AR37 reads to it. I did so, using bowtie2, and visualized the results using the integrative genomics viewer (IGV).

3.5.1. A significant number of Fl1 genes were absent in AR37

No AR37 reads mapped to 164 Fl1 genes suggesting that these genes are absent from AR37. I used BLASTx search against NCBI non-redundant protein database to investigate putative functions for these 164 Fl1 specific genes. These searches identified a wide range of putative functions for the genes, including genes involved in the synthesis of lolium and ergot alkaloids (see below for a more detailed analysis of alkaloid gene complements in the two strains). To find out if these 164 genes are enriched in particular functional categories, an enrichment analysis was performed against a background of the full Fl1 set of genes using BLAST2GO. A Fisher Exact Test revealed that monosaccharide binding and L-ascorbic acid binding proteins were significantly over-represented in the test set. Conversely protein categories involved in the formation of

cellular components (cell part, cell membranes, organelles, intracellular etc) were underrepresented (Fig. 3.19)

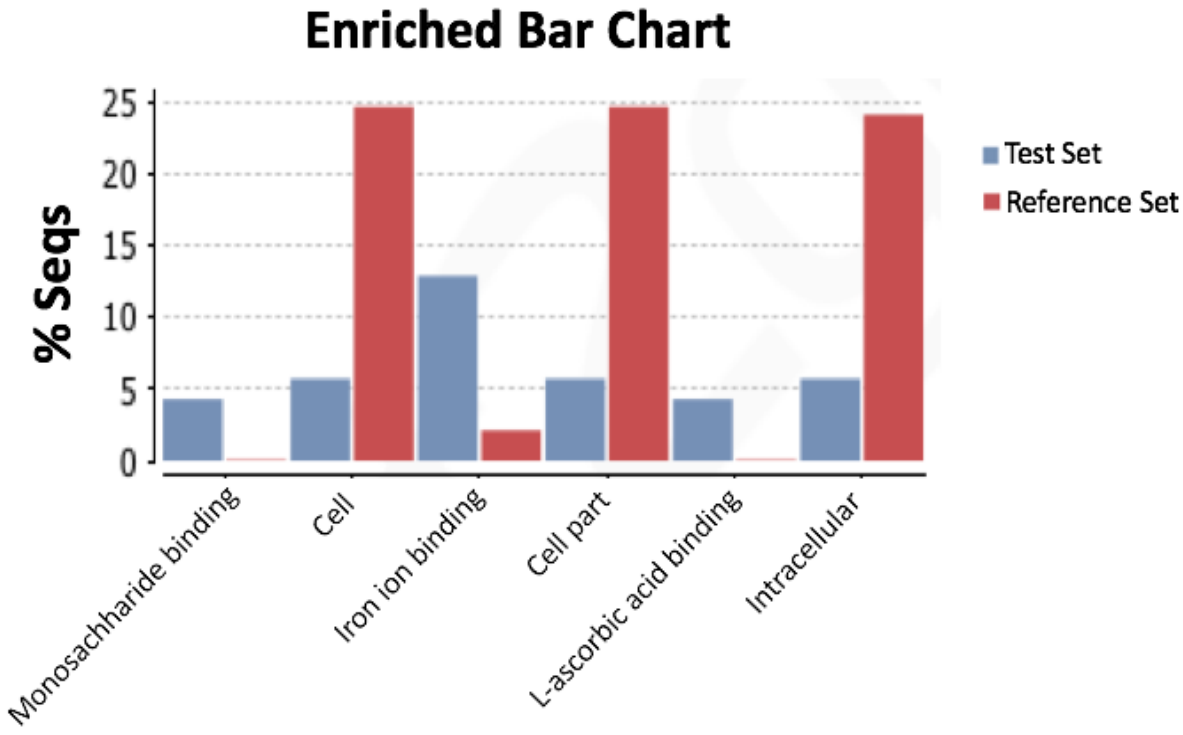


Figure 3.20. A screenshot of BLAST2GO analysis showing over- and under-representation of certain gene categories for Fl1 genes that are missing in AR37 assembly.

I also attempted to identify genes that are present only in AR37 but absent in Fl1, in particular as these may be candidates for novel epoxy-janthitrem synthesis pathways or so-far uncharacterized compounds which may be responsible for beneficial characteristics associated with AR37. I used an *ab-initio* gene prediction tool i.e. Glimmer to predict genes in the AR37 assembly. Glimmer predicted 15571 transcripts containing 19406 coding sequences (CDS). The CDS were compared to a local database of conceptually translated Fl1 coding regions using BLASTx. No BLAST hits were observed for 56 AR37 CDS regions. When I used these sequences in a BLAST search against the NCBI non-redundant nucleotide database and the translated nucleotide database, only four of them returned a hit (against a phosphatidylinositol N-acetylglucosaminyltransferase mRNA, a trypsin-like serine protease, a hypothetical protein and a transposase).

The remaining 52 CDS may represent additional proteins unique to AR37. However at this stage it is difficult to rule out the possibility that these loci represent gene prediction artefacts in AR37.

3.5.2. Alkaloid biosynthesis genes in AR37

AR37, unlike other *Epichloë festucae* strains, is not known to produce ergot alkaloid, indole diterpenes, peramine and lolines (Tapper *et al.*, 2011) (Tapper and Lane, 2004). My new assembly, in conjunction with the sequences of known alkaloid genes (Schardl *et al.*, 2013b), provided an opportunity to investigate if AR37 is genetically incapable of producing these compounds, and why.

3.5.2.1. Peramine gene in AR37

A single gene, *perA*, is required for peramine synthesis. The *perA* gene encodes three protein domains i.e. peptide synthetase, methyltransferase and reductase that together synthesize peramine (Schardl *et al.*, 2013b). An Fl1 *perA* homologue is present in AR37 but the encoded protein is no longer capable of synthesizing peramine. The AR37 gene has multiple SNPs, plus a 12 bp insertion, as well as a ~1320 nt deletion in the 3` region that encodes the reductase domain.

3.5.2.2. Lolitrem B biosynthesis genes in AR37

No homologues of genes involved in the biosynthesis of ergot and lolium alkaloids were identified in the AR37 assembly. Of the 11 known lolitrem B biosynthesis genes (Schardl *et al.*, 2013b, Schardl *et al.*, 2013c) only two, (*ltmE*) and lolitrem J (*ltmJ*) are apparently missing in AR37, explaining the absence of lolitrem B in AR37-infected plants - the genes encode enzymes that carry out the final two steps in lolitrem B biosynthesis. The coding sequences of six of the remaining genes, *ltmB*, *ltmC*, *ltmF*, *ltmP*, *ltmQ* and *ltmS* were completely identical to the coding sequences in the lolitrem-producing strain Fl1. Of the three remaining genes one, *ltmG* had 1 synonymous SNP and one 3 nt deletion, *ltmK* had 4 non-synonymous and 3 synonymous SNPs and *ltmM* had 2 non-synonymous and 3 synonymous SNPs in AR37. None of the SNPs or indel resulted in non-sense codon so genes still appeared to be potentially functional - indeed the paucity of mutations in the

9 genes would suggest that the truncated pathway could still be under selection, and thus has a biological function. One possibility is that these 9 genes present in AR37 may be synthesizing certain intermediate compounds of the lolitrem B biosynthesis pathway. The mixture of these intermediate compounds may play a role in protecting the host plant from herbivores and insects. The dn/ds analysis was not feasible due to nearly identical nature of coding sequences of the 9 lolitrem genes present in both AR37 and Fl1.

3.5.2.3. Epoxy-janthitrems and lolitrem B biosynthesis may share early pathway genes

The only class of alkaloids that has been associated with AR37 are epoxy-janthitrems. Five different epoxy-janthitrems have been reported so far to be produced by AR37 (Finch *et al.*, 2012, Finch *et al.*, 2013, Tapper *et al.*, 2011) but genes involved in the biosynthesis pathway have not been identified, because they are not produced by AR37 in culture and they are highly instable when isolated from AR37-ryegrass symbioses (Babu *et al.*, 2018).

In an attempt to identify genes that may be involved in the biosynthesis of epoxy-janthitrem in AR37, I used genes demonstrated as being involved in biosynthesis of more stable janthitrems, in the *Penicillium janthinellum* strain PN2408 (Nicholson *et al.*, 2015). Nine individual janthitrem genes from publicly available janthitremane gene cluster of *Penicillium janthinellum* strain PN2408 were searched against the AR37 genome using tBLASTn using default parameters (Table 3.7). The BLAST searches identified potential AR37 homologues (> 30% amino acid sequence identity) for seven of these genes, *JanD*, *JanG*, *JanM*, *JanB*, *JanC JanO* and *JanP*. Details of the blast hits are given in table 3.7. *JanQ* had a hit with good query coverage but < 30% identity, making it border-line hits (Pearson, 2013). Although two sequences with more than 30% identity over their entire length are almost always considered homologous, yet the 30% threshold may miss some valid hits in certain cases (Pearson, 2013). Two sequences with < 30% identity can be homologues if the evalue is < 1e-10 (Pearson, 2013). Given these criteria, *JanQ* with a sequence identity of 27% and evalue of 6e-72 may have a potential homologue in AR37 assembly. No significant hit was found for *JanA*.

Table 3.7. Best tblastn hits for 9 Janthitrem genes from *P. janthinellum* against AR37 and Fl1.

	AR37 assembly				Fl1 assembly			
Gene	Location	QCov	e-value	Ident.	Location	QCov	e-value	Ident.
JanD	Node_383	94%	2e-130	52%	Chr 1	96%	1e-36	25%
JanG	Node_304	79%	1e-102	52%	Chr 3	79%	2e-102	52%
JanM	Node_304	88%	2e-82	47%	Chr 3	88%	3e-82	47%
JanB	Node_235	91%	6e-54	45%	Chr 3	91%	6e-54	45%
JanC	Node_235	79%	2e-62	50%	Chr 3	83%	8e-62	49%
JanP	Node_235	94%	3e-75	32%	Chr 3	94%	2e-75	34%
JanQ	Node_235	94%	6e-72	27%	Chr 3	94%	6e-72	27%
JanO	Node_319	98%	2e-123	42%	Chr 3	79%	5e-09	46%
JanA	Node_180	26%	1.9	25%	Chr 3	26%	2.0	25%

The above tblastn results were in the same region where lolitrem genes were identified and suggested a possible homology between janthitrem genes and lolitrem genes. To confirm the homologous relationship between janthitrem genes and lolitrem genes, their protein sequences were compared against each other using blastp (Table 3.8). Blastp result further supported the notion that 6 of the janthitrem genes i.e. *JanG*, *JanM*, *JanB*, *JanC*, *JanP* and *JanQ* were homologous to the respective lolitrem genes. *JanD* may be a potential homologue to *ltmF* as both share around 25% identity but it has a better hit in another contig (NODE_383) . No homology for *JanA* and *JanO* was found within lolitrem gene cluster. Since late pathway genes for the lolitrem biosynthesis are missing in AR37, it is possible that the early pathway genes for lolitrem biosynthesis pathway are part of a janthitrem biosynthesis pathway. It must be noted that epoxy-janthitrems produced by AR37 are structurally similar to lolitrems and both belong to a large, diverse group of compounds i.e. indole-diterpenes. All indole-diterpenes share a basic core structure consisting of cyclic diterpene skeleton and an indole moiety (Saikia *et al.*, 2006). An early pathway compound i.e. paspaline is common to all indole diterpenes including epoxy-janthitrems (Saikia *et al.*, 2006). Out of 9 lolitrems genes present in AR37, only the presence of the first four genes in the pathway (*ltmB*, *ltmC*, *ltmG* and *ltmM*) is explicable solely on the basis of their involvement in janthitrem biosynthesis, as they are required for the biosynthesis of paspaline. Additional prenylation and ring substitutions steps may

then increase the complexity of the core structure to give rise to different types of indole diterpenes (Saikia *et al.*, 2006). Further elaboration of the core structure in AR37 may be done by some other genes including *JanD* and *JanO* which had significant hits in two different contigs i.e. contig_383 and contig_319.

Table 3.8 Blastp results for comparing janthitrems proteins against lolitrem proteins.

Janthitrem	Lolitrem	Qcov%	e-value	Identity%
JanG	LtmG	80%	2e-120	54.61%
JanM	LtmM	88%	2e-115	41.24%
JanB	LtmB	92%	5e-90	56.00%
JanC	LtmC	83%	4e-85	45.29%
JanP	LtmP	93%	1e-117	38.49%
JanQ	LtmQ	94%	1e-127	36.66%
JanD	LtmF	91%	9e-31	25.00%
JanA	Nil			
JanO	Nil			

The presence of 9 lolitrem genes in AR37 with 6 of them sharing homology to respective janthitrem pathway genes suggested that epoxy-janthitrem production may be dependent on the products of these genes. As most of the alkaloid producing genes are found in clusters, there is a possibility that some other AR37 genes involved in the biosynthesis of epoxy-janthitrems are positioned adjacent to these genes. To test this possibility, the AR37 contigs containing the nine lolitrem genes (contigs 235 and 304) were aligned against the Fl1 genome. Aligning each of these contigs against the Fl1 genome produced 100% identity over the entire length of contig 235, and 93% identity for contig 304 over 99% of its length. The first ~1500bp of contig 304 did not align to Fl1 genome. No ORF was detected in this unaligned ~1500 bp. Most of the differences between the contig 304 and Fl1 chromosome were in the non-genic regions.

To find any potential genes that may function to modulate the products of *ltm* genes in order to produce epoxy-janthitrems, ORFs were predicted in all 6 reading frames in the areas immediately upstream or downstream of the lolitrem gene clusters of both the

contigs i.e. 235 and 304 using ORFfinder. The minimum ORF length was set to 150 bp. In both contigs the areas immediately upstream or downstream of *ltm* gene clusters were largely devoid of any ORFs. No function could be attributed to any putative ORFs identified in these regions.

Although alkaloid genes are reported to occur in cluster, the sub-telomeric nature of these genes, presence of abundant repeats within and around these genes and fragmented nature of the assembly may be the reason that these genes got hits on different contigs as contigs did not assemble well. There is also a possibility that late pathway genes for synthesis of epoxy-janthitrems in AR37 may be different than used by *Penicillium janthinellum* as there are differences between the janthitrems produced by the two.

In summary these analyses confirmed that different complements of alkaloid genes form the basis of the differences between the alkaloid profiles in AR37- infected plants and plants infected with other *Epichloë* spp.. Some candidates for early part of the pathway for epoxy-janthitrem synthesis cluster could be identified. They do form a cluster but additional janthitrem biosynthesis genes seem to be located elsewhere. The presence of a truncated but possibly functional lolitrem pathway suggests that this pathway is at least partially explicable on the basis of its likely involvement in early steps of janthitrem biosynthesis pathway. It may in addition lead to the formation of yet uncharacterized alkaloids that may increase the fitness of the AR37 infected host.

4. DISCUSSION

4.1. *Epichloë* endophyte adaptation to a new host: challenges and opportunities

Epichloë endophytes have, and continue to, coevolve with their host grasses (Schardl *et al.*, 1997). As a result they are adapted specifically to their natural host. This apparently makes transition to another type of host difficult, as indicated both by their natural restricted host range, and by the difficulty of artificially transferring *Epichloë* spp. to new hosts (Leuchtmann, 1992, Christensen, 1995, Koga *et al.*, 1993). However such transitions are possible (Leuchtmann & Clay, 1993, Christensen, 1995), even in *Epichloë* spp. which lack a sexual cycle and have therefore little or no opportunity of attempting to expand their host range as part of their natural lifecycle (Christensen, 1995), and in which the ability to do so should not confer significant selective advantages. Thus even when an endophyte is moved to and able to initially survive in a new host, it is to be expected that the symbionts are initially only marginally compatible. There will be selective pressure on the symbionts to acquire mutations that improve their interaction –in particular on the asexual endophyte, whose very existence depends on its survival in the host and its vertical transmission.

Monitoring genetic alterations during adaptation of an *Epichloë* to a new host provides opportunities for identifying what genes are the most important determinants of symbiotic interaction between *Epichloë* and their hosts since mutations in these genes should have the greatest impact on compatibility. This is of fundamental scientific interest but could also improve our ability to establish and screen for improved compatibility novel *Epichloë* / grass associations. In addition it can tell us how much novel associations can change over time.

The latter can have applications in novel association IP protection, if the endophyte indeed rapidly acquires compatibility-enhancing mutations that distinguish it from the strain originally introduced. As a result, what is present in seed after a few generations would no longer exactly match the original strain. Indeed, because of clonal interference

(Gerrish & Lenski, 1998), in asexual endophytes, the original strain is likely to diverge into numerous clonal lineages, each seeking a different path to enhanced compatibility.

Perhaps more important, the properties of commercially distributed novel associations may change as a result over time. Some of these changes, such as improved seed transmission, may be beneficial, others, such as changes in endophyte / plant biomass ratio and in plant protective properties could lead to reduced growth of the grass and agronomically undesirable changes in the secondary metabolite profile – including metabolites that enhance survival by being detrimental to livestock.

I investigated changes in *Epichloë festucae* AR37, introduced, in the late 1990s, into two NZ commercial grass cultivars viz SAMSON (The endophyte associated with SAMSON is called AR37-SAM) and KLP1102 (endophyte associated with KLP1102 is called AR37-KLP). Due to its broad range resistance against pests and observed increase in the overall yield, it is estimated that AR37 endophyte may have contributed NZ$ 42 million to the farming sector (Caradus *et al.*, 2013). AR37 is well compatible with original European host grass (Christensen *et al.*, 1993) but it was initially mildly compatible with both new cultivars. However, compatibility was reported to improve over time in both the new cultivars (unpublished data, AgResearch).

4.2. A new AR37 assembly reveals that epoxy-janthitrem may not be the only reason for enhanced agronomic traits observed in host grasses

Most of the *Epichloë festucae* strains produce one or more of the four commonly found bioactive alkaloids namely lolines, peramines, ergot alkaloids and indole-diterpenes. Lolines and peramines are reported to protect specifically against the insect pests (Tanaka *et al.*, 2005, Schardl *et al.*, 2007) and maybe against some invertebrates (Bacetty *et al.*) while ergot alkaloids and indole-diterpenes are associated with protection against mammals and insects (Schardl *et al.*, 2013a, Knaus *et al.*, 1994). AR37 does not produce any of these four alkaloids, instead it produces indole diterpene-like compounds called epoxy-janthitrems. Epoxy-janthitrems comprise of 5 compounds: epoxy-janthitrem I, epoxy-janthitrem II, epoxy-janthitrem III, epoxy-janthitrem IV and epoxy-janthitriol

(Tapper & Lane, 2004). Epoxy-janthitrems produced by AR37 are considered to protect AR37-infected host against a broad range of insect pests such as Argentine stem weevil larvae (Thom *et al.*, 2014, Popay & Wyatt, 1995), pasture mealybug (Pennell *et al.*, 2005), African black beetle (Popay & Thom, 2009, Thom *et al.*, 2014), porina (Jensen & Popay, 2004) and root aphid (Popay & Thom, 2009, Popay & Cox, 2016, Thom *et al.*, 2014). AR37-infected ryegrass was reported to exhibit up to 36% increase in dry matter production as compared to ryegrass infected with other endophyte strains (Hume *et al.*, 2007). Since AR37 does not produce any other known bioactive alkaloids, epoxy-janthitrems are thought to be the main reason for the improved characteristics associated with AR37 (Johnson *et al.*, 2013). If agriculturally relevant differences between AR37 and other related *Epichloë* are indeed largely restricted to AR37's inability to synthesize ergot and alkaloid metabolites, and its ability to synthesize epoxy-janthitrem, then this should be reflected in the differences between the genomes of AR37 and other *Epichloë.*

The *de novo* assembly of the AR37 genome I produced as a foundation for finding changes that occurred during the adaptation of AR37 suggests that matters may be more complex, judging by a comparison with the chromosome-level assembly of *Epichloë festucae* Fl1. In particular, while it is true that AR37 lacks genes encoding for ergot alkaloids and loline alkaloids, it does however, have most genes (9 out of total 11) encoding the enzymes of the lolitrem B biosynthesis pathway. The missing two genes of the lolitrem B pathway, *ltmE* and *ltmJ*, encode enzymes catalyzing the final steps in the pathway and their absence likely explains why lolitrem B cannot be the end product in the AR37. Late pathway genes have also been reported to be missing in hybrid *Epichloë* (Schardl *et al.*, 2013c). The remainder of the pathway for lolitrem B biosynthesis seems however to be functional in AR37. One possible effect of the loss of only the final parts of the pathway may be the accumulation of mixture of intermediate alkaloids that are postulated to provide added benefit to the host (Schardl *et al.*, 2013c). This is consistent with the observation that a mixture of intermediate ergot alkaloid, loline and indole diterpene pathway compounds have been observed in plants with endophytes (Panaccione, 2005, TePaske *et al.*, 1993, Spiering *et al.*, 2008, Young *et al.*, 2009).

A single gene (*perA*) encodes for a multifunctional protein with 3 different domains that together may synthesize peramine in other *Epichloë*. A *perA* homologue was found in AR37 but with many SNPs and a large deletion at the 3`end in the reductase coding

region. A *perA* gene with an identical deletion (called *perA-ΔR*) has been found in other *Epichloë* as well, most notably in E2368 where *perA-ΔR* is reported to show expression (Schardl *et al.*, 2013b). Novel SNPs and / or indels within *perA* or its flanks have been reported to make it non-functional and it has been postulated that *perA-ΔR* may encode for another multifunctional protein, which may help to synthesize a compound similar to peramine if other appropriate domains / enzymes are available (Berry *et al.*, 2015). Nevertheless, comparison of amino acid sequence of *perA-ΔR* in AR37 and E2368 revealed a non-sense mutation halfway through the gene in AR37 making it unlikely, if not impossible, that compounds similar to peramine may play a role in insect protection of grasses by AR37.

As epoxy-janthitrems are the only known alkaloids found in AR37- infected grass, an attempt was made to identify genes encoding enzymes involved in the biosynthesis of epoxy-janthitrem in AR37. It must be noted that epoxy-janthitrems are a class of compounds with many similar, related structures, and five such compounds have been identified in AR37 in planta (Finch *et al.*, 2012, Finch *et al.*, 2013). Epoxy-janthitrems are highly unstable and AR37 does not produce any of these compounds in culture (Babu *et al.*, 2018). However, the relatively stable and closely related class of compounds, Janthitrems, are present in *Penicillium janthinellum* and biosynthesis pathway for one such compound viz janthitrem B is well characterized in *P. janthinellum* (Nicholson *et al.*, 2015).

There are nine genes forming a cluster (janthitremane cluster) that are involved in biosynthesis of janthitrem B in *Penicillium janthinellum*. A blast (tblastn) search of amino acid sequences of these nine genes against the translated nucleotide database of AR37 assembly identified that > 80% of the sequence from 6 of these proteins had at least 40% identity with a predicted protein from the AR37 database. Six of these shared homology with lolitrem genes both in AR37 and Fl1 (Table 3.8) suggesting that lolitrem and janthitrem biosynthesis may share the same early steps. Epoxy-janthitrems are structurally similar to lolitrem B (Rasmussen *et al.*, 2009). Both share a common core structure consisting of a cyclic geranylgeranyl diphosphate ring and an indole moiety. Both also share the biosynthetic pathway intermediates / precursors (Figure 4.1). It is possible that the product of these early pathway genes of lolitrem are modified by variant *ltmP* / *idtP* and *ltmQ* / *idtQ* genes or some other yet-to-be identified late pathway genes

that encode proteins involved in the synthesis of epoxy-janthitrems or some other undetected alkaloid that may provide beneficial characteristics associated with AR37. Since alkaloid genes are in clusters and located in sub-telomeric regions that are filled with repetitive areas and are hard to assemble, it is possible that janthitremane gene cluster has not been fully assembled in the AR37 assembly presented in this thesis, hindering the identification of late pathway genes. Also janthitrems, synthesized by *Penicillium janthinellum*, are structurally different from epoxy-janthitrems synthesized by AR37, raising the possibility that different pathways may be involved in synthesis of these alkaloids in AR37 and *Penicillium janthinellum*.

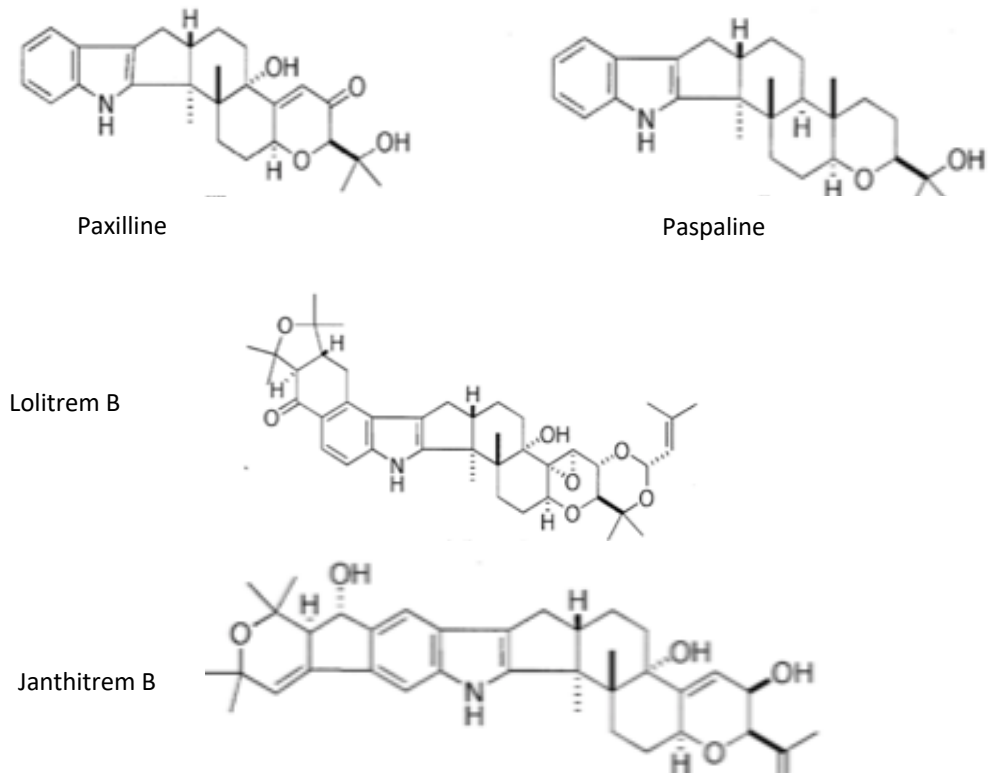


Figure 4.1. Structures of paxilline, paspaline, lolitrem B and janthitrem B. Both lolitrem B and janthitrem B share a common core structure.

I also searched for additional differences between AR37 and Fl1. Coding sequences (CDS) were predicted in our *de novo* AR37 genome assembly using AUGUSTUS and their homologues searched for in *E. festucae* Fl1 assembly. There were 56 AR37 CDS

that were not found in *E. festucae* Fl1. Since Fl1 assembly is a complete chromosomal level assembly, it is less likely to miss these CDS due to fragmentation of the assembly. However given that even the Fl1 assembly lacked certain core Sordariomycete genes (see section 3.4.1.1) some of the genes in repetitive areas may not be detected due to assembly artefacts. BLAST search for these 56 CDS against NCBI non-redundant database did not come up with any valid hit for 52 of these CDS. These 52 CDS appear to be unique to AR37 and their functions are not yet known. Many of them have small size i.e. < 300 bp long, which hints that some of the coded peptides/proteins may be secreted proteins. However, as these CDS are *insilico* predictions, there is a possibility that some may be artefacts by the AUGUSTUS gene prediction tool. Nevertheless, some of the proteins may act as "chemical messenger" and play a role in establishing and/ or improving symbiotic relationship with new hosts. Indeed, as Fl1 has a different host than AR37, it is possible that these unique CDS may have role in maintaining AR37 in its specific host range. Also important to note is that AR37 assembly is fragmented and if there are 56 CDS from a fragmented assembly that have no homologs in *E. fesutcae* Fl1 then it is likely that there will be more unique CDS in a complete AR37 assembly. Given the number of CDS that are unique to AR37, it can be assumed that epoxy-janthitrem may not be the only source for improved characteristics of AR37 and some of these unique CDS may play a role in improved agronomic characteristics seen in AR37-host symbioses.

Only 4 out of 56 CDS had valid hits against NCBI non-redundant database. These CDS coded for phosphatidylinositol N-acetylglucosaminyltransferase mRNA, trypsin like serine protease, a hypothetical protein and a transposase. These proteins play roles that may help endophyte in establishing / maintaining symbiotic relationship with the host. Phosphatidylinositol N-acetylglucosaminyltransferase is necessary for the synthesis of N-acetylgluscosaminyl-phosphotidylinositol, which is in an intermediate in the biosynthesis of Glucosylphosphatidylinosital (GPI)-anchor. GPI-anchored proteins are mostly associated with membranes especially with outer surface of the cell membrane (Kinoshita, 2016). One of the first proteins to come into contact with a new host are cell surface proteins and secreted proteins. These proteins may play a key role in sensing the outer environment and sending signals to the cell to help adapt to the outer environment/host (Simons & Toomre, 2000, Hořejší *et al.*, 1999, Jones & Varela-Nieto, 1998). Endophytes are reported to produce secreted proteins, which may have a role in

modulating the interaction between the endophyte and the host. Trypsin-like serine proteases help insects in digesting the food from their host (Telleria *et al.*, 2010). Also modulation of serine protease activity has been linked with increase or decrease in parasite infection in insects. Trypsin-like serine proteases have been shown to help fungi digest food from their insect hosts (Lopez-Llorca & Robertson, 1992, Lopez-Llorca, 1990). It may help *Epichloë* endophyte digest food from the host apoplast as well. These proteases may breakdown some proteins from host apoplast and may help endophyte in evading host defense responses. Although proteins with similar functions may be present in Fl1, the absence of CDS for these 4 particular proteins from Fl1 may indicate that these proteins in AR37 provide an advantage in adapting to new hosts.

In summary, comparison of AR37 with Fl1 for the presence and absence of whole genes indicated that there are significant differences between AR37 and Fl1. Around 164 genes are exclusively present in Fl1, with monosaccharide binding and L-ascorbic acid binding proteins showing overrepresentation against the total Fl1 genetic background. On the other hand, 56 genes were exclusively present in AR37. Both strains have different natural hosts and these exclusive genes may play important roles in adaption of each strain to its respective host. There were considerable differences found in alkaloid coding genes as well between the two strains. Genes encoding secondary metabolites reflected the known alkaloid profile of each strain except that AR37 had 9 genes for early lolitrem pathway while no lolitrem had ever been reported in AR37-infected hosts. Also 6 of these lolitrem genes shared a significant homology with 6 janthitrem genes. Given the similarity between lolitrems and epoxy-janthitrems, there is a possibility that both share the same early biosynthesis pathway to form a core structure, which may be modified by some yet unknown genes to produce epoxy-janthitrems.

4.3. Bioinformatic analysis can detect adaptation signatures in AR37

There are multiple steps involved in variant calling (Fig. 3.7) and it is difficult to distinguish bioinformatically between true variants and sequencing artefacts in a genome-wide survey, in particular if these are inferred from pools of clones. The reason is that even rare events, such as an alternate base call in a large percentage of reads caused by sequencing or alignment errors, are sufficiently likely to occur on occasions when an

entire genome is analyzed. I applied numerous filtering strategies aimed at increasing the ratio between true variants and false variants. That this strategy was successful is indicated by the fact that the clone pools, isolated after multiple instances of seed propagation differ significantly more from the ancestral AR37 reference genome than the single AR37 that was isolated from its vegetatively propagated original host. Not only would one expect more mutations upon propagation in a new host. Given clonal interference, i.e. the inability of AR37 lineages with compatibility-enhancing mutations to combine these, the frequency of true variant base calls is expected to be lower in pools of clones than in the single AR37 clone from its native host. Indeed highest call frequencies indicating variants after filtering were observed for AR37 clone. If my enrichment for true variants had not worked, or if there were no true variant calls, the clones would be expected to have a lower rather than higher number of variants after filtering- the opposite of what I observed.

A necessary drawback of the filtering (which also involved elimination of parts of the genome from the analysis is that it will eliminate an unknown percentage of true variants, and likely a higher percentage in the clones. Thus the approximate mutation rates calculated from filtered variants (0.04 per generation per Mb) are an underestimate, in particular as far as the clones are concerned. What I can say is that AR37 altered as the novel associations were propagated, and probably considerably faster than when remaining in the original host (because more variants will likely have been missed when the clones were analyzed). Another tentative conclusion is that the rate at which mutations accumulate in AR37 during serial vertical transmission does not markedly diminish over the time of my observations, the above rates being similar for the two cultivars even though they were seed-propagated for a different number of generations. Lastly the frequency of the variant calls in the pools indicate that these are in different lineages and that at the time of analysis there was no indication that a single AR37 clone had reached prevalence.

As far as the identity of AR37 is concerned, it is clear that at this stage AR37 in commercially available seedlots is no longer a single strain, but, for the time being, a collection of increasingly diverging clonal lineages. This has several important implications. It takes considerable amount of time and resources to develop and market an artificial association with desired characteristics and breeders and farmers can only

avail maximum benefits of such association if it remains stable for longer period of time and performs as desired in the field. However due to the asexual nature of these endophytes, multiple clones with different mutations may exist and compete with each other. This could lead to unexpected changes in the performance of the associations and possible decrease in market value of the association. There is also a possibility that the mutations in some of the clones may even increase the performance of the association or one clone may overtake other clones to get established in a new host. In such cases, the possibility of marketing such a clone as separate strain by competing companies can lead to intellectual property rights conflicts. Monitoring the changes in endophytes in artificial associations for a long period of time over many generations may help document changes and to avoid such conflicts.

Perhaps the most significant negative effect of filtering is that I had to eliminate variants in tandem repeat regions, because of the challenges they present for sequencing and assembly. Tandem repeats are known to constitute a significant part of nearly all genomes and can mutate 10 to 100,000 times more frequently than other parts of the genome (Gemayel *et al.*, 2010, Fan & Chu, 2007). Mutation frequency in repetitive regions is directly related to the size and purity of repeat unit, with longer and more pure repeat tract showing higher frequencies (Legendre *et al.*, 2007). It has been known in many microorganisms that hypermutable repeat-containing genes are involved in rapid phenotypic changes in response to change in environment or host (Gemayel *et al.*, 2010, Moxon *et al.*, 2006, van Belkum, 1999, Verstrepen & Fink, 2009). In coding regions of microbial genomes, changes in number of repeat units are found to generate new proteins by changing number and type of amino acids, which may help microbes evade host immune response (Goto *et al.*, 2008, Stern *et al.*, 1986, Smith *et al.*, 2001, Snyder *et al.*, 2001). An infection of wheat plants by *Fusarium* resulted in contraction of CT repeats thus producing a mutant allele, suggesting that external environment can directly select for mutations that occur in repetitive areas (Schmidt & Mitter, 2004).

A significant number of variations in this study were also linked to repetitive DNA, especially homopolymeric regions and simple sequence repeats (SSRs). I also noted that nearly all indels were associated with repetitive regions. Previous studies have suggested that repetitive DNA does indeed change more frequently via addition and / or deletion of whole repeat units rather than SNPs. All homopolymeric regions were omitted because

such regions are not assembled and mapped as well as the "normal" regions (Kececioglu & Ju, 2001). Given that most of the agronomic benefits associated with presence of *Epichloë* endophytes are linked to secondary metabolites and nearly all of the genes encoding such metabolites are either in repetitive areas or in close proximity to such areas, it is likely that repetitive DNA may play a major role in adaptation to the new host.

In the absence of an ungapped reference genome, it is a challenging task to analyze all the repeat regions in a whole genome sequencing project to detect repeat signatures in tandem repeat regions. In such case, a better approach may be to target a few promising repetitive areas in further analyses.

4.4. The distribution of variants across gene categories suggests that many types of genes are involved in determining compatibility

Investigations aimed at determining key genes involved in *Epichloë* -host interactions have so far mainly relied on reverse genetics, disrupting individual endophyte genes and identifying the effect of the loss of these genes on compatibility. Subsequently incompatible associations are characterized by different symptoms in different associations including stunting of the host, browning of host tissue, increased branching of the fungal hyphae, hyphal colonization of host vascular bundles, reduced host vigour, death of hyphae, activation of plant defense response, and death of host plant. Many such genes have been identified e.g. *noxA* and *noxR* (Scott *et al.*, 2007), stress-activated mitogen-activated protein kinase (*sakA*) (Eaton *et al.*, 2010), polarity proteins Bem1 and Cdc24 (Takemoto *et al.*, 2011), *RacA* and NADPH oxidases (Tanaka *et al.*, 2008, Tanaka *et al.*, 2006), $p67^{phox}$-like regulator (Takemoto *et al.*, 2006), *soft* gene (Charlton *et al.*, 2012), *ProA* (Tanaka *et al.*, 2013), *acyA* (Voisey *et al.*, 2016), *MobC* (Green *et al.*, 2016), G-protein and cAMP/PKA signalling genes (Bisson, 2017), velvet-domain protein coding gene (*velA*), *sidN* (Johnson *et al.*, 2007), *cAMP* (Voisey *et al.*, 2016), and *MAPK* (Becker *et al.*, 2015). Following initial success with one gene researchers have often found evidence that other genes with related functions are also important for the interaction. However this does not necessarily mean that these selected pathways are more important than others. It has also been observed, in attempts to complement *Epichloë* mutants and

in other attempts to manipulate *Epichloë* that the symbiosis is very easily "accidentally" disrupted (Simpson *et al.*, 2012, Simpson *et al.*, 2007, Zhang *et al.*, 2006).

This could indicate that a large number of genes of numerous functions play roles in maintaining a delicate balance between the symbionts. The distribution of functions of variant-marked genes in my analysis supports this idea. Neither were GO categories significantly overrepresented among genes marked by SNPs in AR37-KLP or AR37-SAM clones, nor was there a significant overlap of GO categories when AR37-KLP and AR37-SAM variants were compared.

Interestingly, and further supporting the concept that many gene categories determine compatibility, none of the SNPs I found marked genes that had previously been identified by reverse genetics as important determinants of symbiotic interaction. This may have been expected if these previously identified genes were exceptionally important in symbiotic interaction. Nevertheless, such mutations may be present but may have been removed during the filtering process.

Finally, the overlap of GO categories between KLP and SAMSON lineage mutations was also not less than expected by chance, as might be expected if different types of genes were of different importance in different hosts.

Thus overall my results support the idea that endophyte- host compatibility is multifactorial and do not provide evidence that some types of gene are more important than others.

4.5. The discovered variants suggest new categories of genes that could play a role in *Epichloë* -grass symbiosis.

All of the many categories of genes marked by SNPs may play a role in *Epichloë* host interaction, and for some of these, this is also supported by research in other systems. Examples are actin binding, zinc ion binding, calcium ion binding, iron-sulphur cluster binding, cytoskeleton organization, cell proliferation, cell morphogenesis, oxidation-reduction process and DNA binding transcription factor activity. A significant number of

genes encoding binding proteins were differentially expressed in stress-activated mitogen-activated protein kinase (sakA) deletion mutants of *E. festucae* (Eaton *et al.*, 2010, Eaton *et al.*, 2011). Actin binding proteins may help fungal hyphae synchronize their growth with that of the surrounding plant cells. Intercalary growth shown by fungal hyphae in the expanding leaves of their host will need reorganization of the whole fungal cytoskeleton, including actin (Scott *et al.*, 2012). Actin binding proteins may help in this reorganization of the actin filaments. An iron siderophore is suggested to be involved in maintaining the symbiotic association between *E. festucae* and ryegrass host (Johnson *et al.*, 2007). Similarly oxidation-reduction processes, controlled by a multi-subunit NADPH oxidase complex have already been proven to play a key role in the symbiotic interaction between *E. festucae* endophyte and its ryegrass host (Tanaka *et al.*, 2006, Tanaka *et al.*, 2008, Scott *et al.*, 2007). DNA binding proteins may influence the transcription of genes involved in establishing the symbiotic relationship with the ryegrass host. Zinc and copper are reported to influence the activity of anti-microbial peptides in another fungus i.e. *Verticillium kibiense* (formerly *Epichloë kibiensis*) (Nishikawa & Ogawa, 2004). A protein that can bind Zn may help to maintain the concentration of Zn ions to a level at which functionality of other peptides is maximum. Zinc binding proteins may also help in alleviating the heavy metal toxicity to their host (Göhre & Paszkowski, 2006). Zinc may have a role in scavenging of toxic free radicals (Van Ho *et al.*, 2002).

4.6. Future Directions

Due to constraints on time and resources, SNPs identified *in silico* in this study could not be validated by PCR amplification followed by sanger sequencing. This remains one of the goals that should be done in future. Since two of our samples were pools of 11 clones each and none of the identified SNPs was present in all of the clones, it suggests that, in order to confirm the validity of any SNP, all clones will need to be screened individually. It would be a major task to confirm the validity of all the identified SNPs in each of the clones. A more efficient approach would be to randomly choose a few SNPs from this study and screen all the clones individually for the presence/absence of the chosen SNP. This approach will not confirm the validity of all SNPs, but provide a measure of the validity of any given SNP and its prevalence in the pools.

As AR37 has been recently introduced into SAMSON and KLP1102 and compatibility has been reported to improve over a period of decade or so, it will be interesting to monitor these symbioses in coming years. As mentioned earlier that each endophyte clone may have mutations of its own and in competition with other clones. Monitoring the agronomic performance of these symbioses and performing similar analyses as this study for other time points down the selection cycles in coming years may yield further clues about the genetic determinants of the endophyte adaptation. Monitoring changes in AR37 is also important because not all the mutations may be beneficial and there is a likelihood that some of the mutations may not have desirable effects on the association. Farmers and agronomist will be interested to know if AR37-grass association is performing as good as expected over the years.

This study has focused on SNPs only and so homopolymeric regions and indels associated with such regions were excluded from in this study due to difficulties associated with sequencing and mapping homopolymers with accuracy. SNPs may only represent the tip of the iceberg as mutation rates of such regions is known to be higher and changes in homopolymeric and repetitive regions may be more important in adaptation to the new host (Gemayel *et al.*, 2010, Fan & Chu, 2007). A significant portion of coding regions are known to contain repeats (Gemayel *et al.*, 2010, Verstrepen *et al.*, 2005, Legendre *et al.*, 2007). Eukaryotes are found to have higher internal protein repeats than prokaryotes and archaea. One reason can be that these repeats may provide an extra source of variations to eukaryotes thus compensating for their long generation time (Ekman *et al.*, 2005). Mutations in homopolymeric and repetitive regions associated with protein coding genes may provide a repertoire of proteins, some of which may be beneficial and selected for adaptation to the new host (Marcotte *et al.*, 1999). It will be important to also focus on selected homopolymeric regions (and indels) in order to ascertain their role.

Host grass genotype has not been considered in our study. Since endophyte grass associations are considered to be symbiotic, it is reasonable to assume that both host and endophyte genotypes play crucial role in establishing the mutualistic association. Host genotype is reported to have a crucial role in successfully transmitting the endophytes to the new progeny (Gagic *et al.*, 2018). Another recent study suggested that epigenetic mechanisms operating in the host may be involved in modulating the symbiotic and

pathogenic interactions between microbes and host plants (Zogli & Libault, 2017). Perennial ryegrass is obligatory outcrossing and its genome is far more complex than AR37 endophyte genome. Ryegrass populations are genetically heterogeneous with a high level of heterozygosity (Sweeney & Danneberger, 1994) and polyploidy populations exist. This makes it even more challenging to conduct a genome wide study for the identification of genetic determinants of compatibility / adaptation. A good quality reference genome for perennial ryegrass has not yet been established but the size of the genome is estimated to be around 2.6 Gb (Kopecký *et al.*, 2010). Pooling multiple individuals together for sequencing is not a viable option and sequencing multiple clones individually, to a coverage sufficiently high to identify variants, will be too costly using current technologies. With the advancement in technologies such as PacBio long read sequencing, and genotyping by synthesis (GBS) it may be feasible in near future to sequence and analyze multiple genomes of ryegrass for genetic determinants of compatibility.

Finally, the involvement of epigenetic processes cannot be ruled out in adaptation of an endophyte to the host grass. In one study, a frequent reversible change in the colony morphology of the same endophyte was observed and was related to (in)compatibility, suggesting that epigenetic mechanisms may be involved (Simpson *et al.*, 2012). Methylation-sensitive amplified polymorphism (MSAP) analyses can be used to explore the link between adaptation and epigenetic characteristics of the endophyte.

In summary, experimental validation of some of the SNPs and monitoring changes in AR37 clones in future selection cycles may provide valuable clues regarding the nature of SNPs and their impact on the pasture performance. Given that AR37 has only recently been introduced in these new host cultivars, it seems probable that epigenetic changes may have played an important role in adaptation of these endophytes to host grasses. Also important is to consider the role of host genotype in establishing such symbiotic interactions.

5. Bibliography

Abnizova, I., Leonard, S., Skelly, T., Brown, A., Jackson, D., Gourtovaia, M., Qi, G., Te Boekhorst, R., Faruque, N., and Lewis, K. (2012) Analysis of context-dependent errors for Illumina sequencing. *Journal of Bioinformatics and Computational Biology* **10**: 1241005.

Abnizova, I., te Boekhorst, R., and Orlov, Y. (2017) Computational errors and biases of short read next generation sequencing. *Journal of Proteomics and Bioinformatics* **10**: 1-17.

Agee, C., and Hill, N. (1994) Ergovaline variability in *Acremonium*-infected tall fescue due to environment and plant genotype. *Crop Science* **34**: 221-226.

Ahlholm, J.U., Helander, M., Lehtimäki, S., Wäli, P., and Saikkonen, K. (2002) Vertically transmitted fungal endophytes: different responses of host-parasite systems to environmental conditions. *Oikos* **99**: 173-183.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**: R18.

Al-Samarrai, T., and Schmid, J. (2000) A simple method for extraction of fungal genomic DNA. *Letters in Applied Microbiology* **30**: 53-56.

Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., and Durbin, R. (2011) Dindel: accurate indel calls from short-read data. *Genome research* **21**: 961-973.

Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck, T.A., Simpson, J.T., and Tonon, L. (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications* **6**: 10001.

Alkan, C., Sajjadian, S., and Eichler, E.E. (2010) Limitations of next-generation genome sequence assembly. *Nature Methods* **8**: 61-65.

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. *URL* http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Arnheim, N., Strange, C., and Erlich, H. (1985) Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. *Proceedings of the National Academy of Sciences* **82**: 6970-6974.

Aronesty, E. (2011) ea-utils: Command-line tools for processing biological sequencing data. *URL* http://code.google.com/p/ea-utils *https://expressionanalysis.github.io/ea-utils/*.

Auerbach, R.K., Chen, B., and Butte, A.J. (2013) Relating genes to function: identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool. *Bioinformatics* **29**: 1922-1924.

Babu, J.V., Popay, A.J., Miles, C.O., Wilkins, A.L., di Menna, M.E., and Finch, S.C. (2018) Identification and structure elucidation of janthitrems a and d from *Penicillium janthinellum* and determination of the tremorgenic and anti-insect activity of janthitrems A and B. *Journal of Agricultural and Food Chemistry* **66**: 13116-13125.

Bacetty, A., Snook, M., Glenn, A., Noe, J., Hill, N., Culbreath, A., Timper, P., Nagabhyru, P., and Bacon, C. (2009) Toxicity of endophyte-infected tall fescue alkaloids and grass metabolites on *Pratylenchus scribneri*. *Phytopathology* **99**: 1336-1345.

Bacon, C.W. (1993) Abiotic stress tolerances (moisture, nutrients) and photosynthesis in endophyte-infected tall fescue. *Agriculture, Ecosystems and Environment* **44**: 123-141.

Ball, O., and Prestidge, R. (1993) The effect of the endophytic fungus *Acremonium lolii* on adult black beetle (*Heteronychus arator*) feeding. *New Zealand Plant Protection* **45**: 201-204.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., and Pevzner, P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**: 455-477.

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., and Song, Y.-Q. (2011) Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics* **56**: 406-414.

Barlow, N.D., French, R.A., and Pearson, J.F. (1986) Population ecology of *Wiseana cervinata*, a pasture pest in New Zealand. *Journal of Applied Ecology* **23**: 415-431.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Research* **12**: 177-189.

Becker, Y., Eaton, C.J., Brasell, E., May, K.J., Becker, M., Hassing, B., Cartwright, G.M., Reinhold, L., and Scott, B. (2015) The fungal cell-wall integrity MAPK cascade is crucial for hyphal network formation and maintenance of restrictive growth of *Epichloe festucae* in symbiosis with *Lolium perenne*. *Molecular Plant-Microbe Interactions* **28**: 69-85.

Belesky, D.P., and Fedders, J.M. (1995) Tall Fescue development in response to *Acremonium coenophialum* and soil acidity. *Crop Science* **35**: 529-533.

Benjamini, Y., and Speed, T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* **40**: e72-e72.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., and Bignell, H.R. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.

Berry, D., Takach, J.E., Schardl, C.L., Charlton, N.D., Scott, B., and Young, C.A. (2015) Disparate independent genetic events disrupt the secondary metabolism gene perA in certain symbiotic *Epichloë* species. *Applied and Environmental Microbiology* **81**: 2797-2807.

Bhangale, T.R., Rieder, M.J., Livingston, R.J., and Nickerson, D.A. (2004) Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Human Molecular Genetics* **14**: 59-69.

Bisson, A. (2017) The role of the G protein and cAMP/PKA signalling pathway in establishment and maintenance of the mutualistic *Epichloë festucae*-ryegrass association. *Doctoral Thesis, Massey University, New Zealand* **PhD**.

Bolger, A., and Giorgi, F. (2014) Trimmomatic: a flexible read trimming tool for illumina NGS data. *http://www.usadellab.org/cms/index.php?page=trimmomatic*.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T.R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N.A., Ganapathy, G., Gibbs, R.A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J.B., Ho, I.Y., Howard, J., Hunt, M., Jackman, S.D., Jaffe, D.B., Jarvis, E.D., Jiang, H., Kazakov, S., Kersey, P.J., Kitzman, J.O., Knight, J.R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M.D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T.D., Paten, B., Paulo, O.S., Phillippy, A.M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F.J., Richards, S., Rokhsar, D.S., Ruby, J.G., Scalabrin, S., Schatz, M.C., Schwartz, D.C., Sergushichev, A., Sharpe, T., Shaw, T.I., Shendure, J., Shi, Y., Simpson, J.T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B.M., Wang, J., Worley, K.C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., and Korf, I.F. (2013) Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* **2**: 10.

Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P., and Tyson, G.W. (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLOS Computational Biology* **9**: e1003031-e1003031.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L., D'Amore, R., Allen, A.M., McKenzie, N., Kramer, M., Kerhornou, A., and Bolser, D. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**: 705-710.

Brockhurst, M.A., Chapman, T., King, K.C., Mank, J.E., Paterson, S., and Hurst, G.D. (2014) Running with the Red Queen: the role of biotic conflicts in evolution. *Proceedings of the Royal Society B: Biological Sciences* **281**: 20141382.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research* **18**: 763-770.

Bryant, M.K., Schardl, C.L., Hesse, U., and Scott, B. (2009) Evolution of a subtilisin-like protease gene family in the grass endophytic fungus *Epichloe festucae*. *BMC Evolutionary Biology* **9**.

Burrows, M., and Wheeler, D.J. (1994) A block-sorting lossless data compression algorithm. *System Research Center, California*.

Caradus, J., Lovatt, S., and Belgrave, B. (2013) Adoption of forage technologies by New Zealand farmers – case studies. *International Grasslands Congress* **22**.

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

Charlton, N.D., Shoji, J.-Y., Ghimire, S.R., Nakashima, J., and Craven, K.D. (2012) Deletion of the fungal gene soft disrupts mutualistic symbiosis between the grass endophyte Epichloë festucae and the host plant. *Eukaryotic Cell* **11**: 1463-1471.

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., and Hwang, C.-C. (2013) Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PloS One* **8**: e62856.

Chin, E.L., da Silva, C., and Hegde, M. (2013) Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. *BMC Genetics* **14**: 6.

Chin, F.Y., Leung, H.C., and Yiu, S. (2014) Sequence assembly using next generation sequencing data—challenges and solutions. *Science China Life Sciences* **57**: 1140-1148.

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. *Mycological Research* **99**: 466-470.

Christensen, M., Simpson, W., and Al Samarrai, T. (2000a) Infection of tall fescue and perennial ryegrass plants by combinations of different Neotyphodium endophytes. *Mycological Research* **104**: 974-978.

Christensen, M., and Voisey, C. (2007) The biology of the endophyte/grass partnership. *Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses* **6**: 123-133.

Christensen, M.J. (1996) Antifungal activity in grasses infected with *Acremonium* and *Epichloë* endophytes. *Australasian Plant Pathology* **25**: 186-191.

Christensen, M.J., Ball, O.J.P., Bennett, R.J., and Schardl, C.L. (1997) Fungal and host genotype effects on compatibility and vascular colonization by *Epichloë festucae*. *Mycological Research* **101**: 493-501.

Christensen, M.J., Bennett, R.J., and Schmid, J. (2002) Growth of *Epichloë / Neotyphodium a*nd p-endophytes in leaves of *Lolium a*nd *Festuca g*rasses. *Mycological Research* **106**: 93-106.

Christensen, M.J., Leuchtmann, A., Rowan, D.D., and Tapper, B.A. (1993) Taxonomy of *Acremonium* endophytes of tall fescue (*Festuca arundinacea*), meadow fescue (*F. pratensis*) and perennial ryegrass (*Lolium perenne*). *Mycological Research* **97**: 1083-1092.

Christensen, M.J., Spiering, M.J., and Schmid, J., (2000b) Metabolic activity, distribution, and propagation of grass endophytes in planta: investigations using the GUS reporter gene system. In: Microbial Endophytes. CRC Press, pp. 309-336.

Chung, K.-R., Hollin, W., Siegel, M.R., and Schardl, C.L. (1997) Genetics of host specificity in *Epichloë typhina*. *Phytopathology* **87**: 599-605.

Clay, K. (1988) Fungal endophytes of grasses: a defensive mutualism between plants and fungi. *Ecology* **69**: 10-16.

Clay, K., and Kover, P. (1996) Evolution and stasis in plant-pathogen associations. *Ecology* **77**: 997-1003.

Clay, K., Marks, S., and Cheplick, G.P. (1993) Effects of insect herbivory and fungal endophyte infection on competitive interactions among grasses. *Ecology* **74**: 1767-1777.

Clay, K., and Schardl, C.L. (2002) Evolutionary origins and ecological consequences of endophyte symbiosis with grasses. *The American Naturalist* **160**: S99-S127.

Consortium, I.H.G.S. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Consortium, I.H.G.S. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931.

Cornish, A., and Guda, C. (2015) A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed Research International* **2015**: 11.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**: 485.

Craig, D.W., Pearson, J.V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J.J., Pawlowski, T.L., Laub, T., Nunn, G., and Stephan, D.A. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods* **5**: 887-893.

Craven, K., Hsiau, P., Leuchtmann, A., Hollin, W., and Schardl, C. (2001) Multigene phylogeny of *Epichloë* species, fungal symbionts of grasses. *Annals of the Missouri Botanical Garden* **88**: 14-34.

Crawford, K.M., Land, J.M., and Rudgers, J.A. (2010) Fungal endophytes of native grasses decrease insect herbivore preference and performance. *Oecologia* **164**: 431-444.

Cunningham, P., Blumenthal, M., Anderson, M., Prakash, K., and Leonforte, A. (1994) Perennial ryegrass improvement in Australia. *New Zealand Journal of Agricultural Research* **37**: 295-310.

Davis, L., Dibner, M., and Battey, J. (2012) Basic Methods in Molecular Biology. *Elsevier*.

Deamer, D., Akeson, M., and Branton, D. (2016) Three decades of nanopore sequencing. *Nature Biotechnology* **34**: 518-524.

Deitsch, K.W., Lukehart, S.A., and Stringer, J.R. (2009) Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nature Reviews. Microbiology* **7**: 493-503.

Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F.M. (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS One* **8**: e85024.

Denisov, G., Walenz, B., Halpern, A.L., Miller, J., Axelrod, N., Levy, S., and Sutton, G. (2008) Consensus generation and variant detection by Celera Assembler. *Bioinformatics* **24**: 1035-1040.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., and Hanna, M. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**: 491-498.

Desai, A., Marwah, V.S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V., and Jere, A. (2013) Identification of optimum sequencing depth especially for *de novo* genome assembly of small genomes using next generation sequencing data. *PLoS One* **8**: e60204.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36**: e105.

Dolled-Filhart, M.P., Lee, M., Ou-yang, C.-w., Haraksingh, R.R., and Lin, J.C.-H. (2013) Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. *The Scientific World Journal* **2013**: 10.

Dugdale, J.S., (1994) Fauna of New Zealand 30: Hepialidae (Insecta: Lepidoptera). In: Fauna of New Zealand. pp.

Easton, H., Christensen, M., Eerens, J., Fletcher, L., Hume, D., Keogh, R., Lane, G., Latch, G., Pennell, C., and Popay, A. (2001) Ryegrass endophyte: a New Zealand Grassland success story. *Proceedings of the conference-New Zealand Grassland Association* **63**: 37-46.

Eaton, C.J., Cox, M.P., Ambrose, B., Becker, M., Hesse, U., Schardl, C.L., and Scott, B. (2010) Disruption of signaling in a fungal-grass symbiosis leads to pathogenesis. *Plant Physiology* **153**: 1780-1794.

Eaton, C.J., Cox, M.P., and Scott, B. (2011) What triggers grass endophytes to switch from mutualism to pathogenism? *Plant Science* **180**: 190-195.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., and Bettman, B. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133-138.

Ekblom, R., and Wolf, J.B.W. (2014) A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications* **7**: 1026-1042.

Ekman, D., Björklund, Å.K., Frey-Skött, J., and Elofsson, A. (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *Journal of Molecular Biology* **348**: 231-243.

Ellegren, H., Smeds, L., Burri, R., Olason, P.I., Backström, N., Kawakami, T., Künstner, A., Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., Uebbing, S., and Wolf, J.B.W. (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**: 756.

Ewing, B., and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**: 186-194.

Faeth, S.H. (2002) Are endophytic fungi defensive plant mutualists? *Oikos* **98**: 25-36.

Faeth, S.H., and Sullivan, T. (2003) Mutualistic asexual endophytes in a native grass are usually parasitic. *The American Naturalist* **161**: 310-325.

Fan, H., and Chu, J.-Y. (2007) A brief review of short tandem repeat mutation. *Genomics, Proteomics & Bioinformatics* **5**: 7-14.

Fang, H., Wu, Y., Narzisi, G., ORawe, J.A., Barrón, L.T.J., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M.C., and Lyon, G.J. (2014) Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Medicine* **6**: 89.

Finch, S., Fletcher, L., and Babu, J. (2012) The evaluation of endophyte toxin residues in sheep fat. *New Zealand Veterinary Journal* **60**: 56-60.

Finch, S., Wilkins, A., Popay, A., Babu, J., Tapper, B., and Lane, G., (2010) The isolation and bioactivity of epoxy-janthitrems from AR37 endophyte-infected perennial ryegrass. Poster 80. In: Proceedings of the 7th International Symposium of Fungal Endophytes of Grasses'.(Ed. C Schardl) Available at http://www.ars.usda.gov/research/publications/Publications.html. pp.

Finch, S.C., Thom, E., Babu, J., Hawkes, A., and Waugh, C. (2013) The evaluation of fungal endophyte toxin residues in milk. *New Zealand Veterinary Journal* **61**: 11-17.

Fleetwood, D.J., Khan, A.K., Johnson, R.D., Young, C.A., Mittal, S., Wrenn, R.E., Hesse, U., Foster, S.J., Schardl, C.L., and Scott, B. (2011) Abundant degenerate miniature inverted-repeat transposable elements in genomes of epichloid fungal endophytes of grasses. *Genome Biology and Evolution* **3**: 1253-1264.

Fleetwood, D.J., Scott, B., Lane, G.A., Tanaka, A., and Johnson, R.D. (2007) A complex ergovaline gene cluster in *Epichloe* endophytes of grasses. *Applied and Environmental Microbiology* **73**: 2571-2579.

Fletcher, L. (1982) Observations of ryegrass staggers in weaned lambs grazing different ryegrass pastures. *New Zealand Journal of Experimental Agriculture* **10**: 203-207.

Fletcher, L. (1983) Effects of presence of *Lolium* endophyte on growth rates of weaned lambs, growing on to hoggets, on various ryegrasses. *Proceedings of the New Zealand Grassland Association* **44**: 237-239.

Fletcher, L. (1993) Heat stress in lambs grazing ryegrass with different endophytes. *Proceedings of the 2nd International Symposium on Acremonium/Grass Interactions'.(Eds DE Hume, GCH Latch, HS Easton) pp*: 114-118.

Fonseca, N.A., Rung, J., Brazma, A., and Marioni, J.C. (2012) Tools for mapping high-throughput sequencing data. *Bioinformatics* **28**: 3169-3177.

Funk, C.R., and White, J.F. (1997) Use of natural and transformed endophytes for turf improvement. *Neotyphodium/Grass Interactions*: 229-239.

Futschik, A., and Schlötterer, C. (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* **186**: 207-218.

Gagic, M., Faville, M.J., Zhang, W., Forester, N.T., Rolston, M.P., Johnson, R.D., Ganesh, S., Koolaard, J.P., Easton, H.S., Hudson, D., Johnson, L.J., Moon, C.D., and Voisey, C.R. (2018) Seed transmission of *Epichloë* endophytes in *Lolium perenne* is heavily influenced by host genetics. *Frontiers in Plant Science* **9**.

Gallagher, R., Smith, G., Di Menna, M., and Young, P. (1982) Some observations on neurotoxin production in perennial ryegrass. *New Zealand Veterinary Journal* **30**: 203-204.

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

Garrison, E. (2015) FreeBayes in depth: model, filtering, and walk-through.

Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics* **44**: 445-477.

Gentile, A., Rossi, M.S., Cabral, D., Craven, K.D., and Schardl, C.L. (2005) Origin, divergence, and phylogeny of Epichloë endophytes of native Argentine grasses. *Molecular Phylogenetics and Evolution* **35**: 196-208.

Gerrish, P.J., and Lenski, R.E., (1998) The fate of competing beneficial mutations in an asexual population. In: Mutation and Evolution. R.C. Woodruff & J.N. Thompson (eds). Dordrecht: Springer Netherlands, pp. 127-144.

Glenn, A.E., Bacon, C.W., Price, R., and Hanlin, R.T. (1996) Molecular Phylogeny of Acremonium and Its Taxonomic Implications. *Mycologia* **88**: 369-383.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., and Sykes, S. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* **108**: 1513-1518.

Göhre, V., and Paszkowski, U. (2006) Contribution of the arbuscular mycorrhizal symbiosis to heavy metal phytoremediation. *Planta* **223**: 1115-1122.

Gordon, A., and Hannon, G. (2010) Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished)* http://hannonlab.cshl.edu/fastx_toolkit **5**.

Goto, Y., Carter, D., and Reed, S.G. (2008) Immunological dominance of Trypanosoma cruzi tandem repeat proteins. *Infection and Immunity* **76**: 3967-3974.

Green, K.A., Becker, Y., Fitzsimons, H.L., and Scott, B. (2016) An *Epichloë festucae* homologue of MOB3, a component of the STRIPAK complex, is required for the establishment of a mutualistic symbiotic interaction with *Lolium perenne*. *Molecular Plant Pathology* **17**: 1480-1492.

Guo, Y., Samuels, D.C., Li, J., Clark, T., Li, C.-I., and Shyr, Y. (2013) Evaluation of allele frequency estimation using pooled sequencing data simulation. *The Scientific World Journal* **2013**: 895496-895496.

Guo, Y., Zhao, S., Sheng, Q., Ye, F., Li, J., Lehmann, B., Pietenpol, J., Samuels, D.C., and Shyr, Y. (2014) Multi-perspective quality control of Illumina exome sequencing data using QC3. *Genomics* **103**: 323-328.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072-1075.

Hahn, H., McManus, M.T., Warnstorff, K., Monahan, B.J., Young, C.A., Davies, E., Tapper, B.A., and Scott, B. (2008) *Neotyphodium* fungal endophytes confer physiological protection to perennial ryegrass (*Lolium perenne* L.) subjected to a water deficit. *Environmental and Experimental Botany* **63**: 183-199.

Hasan, M.S., Wu, X., and Zhang, L. (2015) Performance evaluation of indel calling tools using real short-read data. *Human Genomics* **9**: 20.

Heliconius Genome, C. (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94-98.

Hennessy, L. (2015) Epoxy-janthitrems, effects of temperature on in planta expression and their bioactivity against porina larvae. *Doctoral Thesis, University of Waikato, New Zealand* **PhD**

Herd, S., Christensen, M.J., Saunders, K., Scott, D.B., and Schmid, J. (1997) Quantitative assessment of in planta distribution of metabolic activity and gene expression of an endophytic fungus. *Microbiology* **143**: 267-275.

Hernandez, D., François, P., Farinelli, L., Østerås, M., and Schrenzel, J. (2008) *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research*.

Herrera, C.M., and Bazaga, P. (2011) Untangling individual variation in natural populations: ecological, genetic and epigenetic correlates of long-term inequality in herbivory. *Molecular Ecology* **20**: 1675-1688.

Hiatt, E.E., Hill, N.S., and Bouton, J.H., (1997) Monoclonal antibody immunoblot procedure for detecting *Neotyphodium coenophialum* in seedling tall fescue. In: Neotyphodium/Grass Interactions. C.W. Bacon & N.S. Hill (eds). Boston, MA: Springer US, pp. 261-263.

Highnam, G., Wang, J.J., Kusler, D., Zook, J., Vijayan, V., Leibovich, N., and Mittelman, D. (2015) An analytical framework for optimizing variant discovery from personal genomes. *Nature Communications* **6**: 6275.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C.M., Khaitovich, P., Vogel, J., Stadler, P.F., and Hackermüller, J. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Computational Biology* **5**: e1000502.

Hohenlohe, P.A., Phillips, P.C., and Cresko, W.A. (2010) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International Journal of Plant Sciences* **171**: 1059-1071.

Hořejší, V., Drbal, K., Cebecauer, M., Černý, J., Brdička, T., Angelisová, P., and Stockinger, H. (1999) GPI-microdomains: a role in signalling via immunoreceptors. *Immunology Today* **20**: 356-361.

Huang, H.W., Program, N.C.S., Mullikin, J.C., and Hansen, N.F. (2015) Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics* **16**: 235.

Huang, X., Wang, J., Aluru, S., Yang, S.-P., and Hillier, L. (2003) PCAP: a whole-genome assembly program. *Genome Research* **13**: 2164-2170.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P.H., Graves, T.A., Alkan, C., and Dennis, M.Y. (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research* **24**: 688-696.

Hume, D., Popay, A., Cooper, B., Eerens, J., Lyons, T., Pennell, C., Tapper, B., Latch, G., and Baird, D. (2004) Effect of a novel endophyte on the productivity of perennial ryegrass (*Lolium perenne*) in New Zealand. *Proceedings of the 5th International Symposium on Neotyphodium / Grass Interactions*

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

Hume, D.E., (2005) Growth and management of endophytic grasses in pastoral agriculture. In: *Neotyphodium* in Cool-Season Grasses. pp. 201-226.

Hume, D.E., Card, S.D., and Rolston, M.P. (2013) Effects of storage conditions on endophyte and seed viability in pasture grasses. *Proceedings of the 22nd International Grassland Congress*: 405-408.

Hume, D.E., Ryan, D.L., Cooper, B.M., and Popay, A. (2007) Agronomic performance of AR37-infected ryegrass in northern New Zealand. *Proc. N. Z. Grass. Assoc.* **69**: 201-205.

Humphreys, M.W., Yadav, R., Cairns, A.J., Turner, L., Humphreys, J., and Skøt, L. (2006) A changing climate for grassland research. *New Phytologist* **169**: 9-26.

Hwang, S., Kim, E., Lee, I., and Marcotte, E.M. (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports* **5**: 17875.

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., and Akeson, M. (2015) Improved data analysis for the MinION nanopore sequencer. *Nature Methods* **12**: 351.

Jensen, C.S., Salchert, K., and Nielsen, K.K. (2001) A TERMINAL FLOWER1-like gene from perennial ryegrass involved in floral transition and axillary meristem identity. *Plant Physiology* **125**: 1517-1528.

Jensen, J., and Popay, A., (2004) *Perennial ryegrass infected with AR37 endophyte reduces survival of porina larvae*. New Zealand Plant Protection, New Zealand.

Johansson, B. (1972) Agarose gel electrophoresis. *Scandinavian Journal of Clinical and Laboratory Investigation* **29**: 7-19.

Johnson, L., Steringa, M., Koulman, A., Christensen, M., Johnson, R., Voisey, C., Bryan, G., Lamont, I., and Rasmussen, S. (2007) Biosynthesis of an extracellular siderophore is essential for maintenance of mutualistic endophyte-grass symbioses. *Proceedings of The Sixth International Symposium on Fungal Endophytes of Grasses* **13**: 177-179.

Johnson, L.J., de Bonth, A.C., Briggs, L.R., Caradus, J.R., Finch, S.C., Fleetwood, D.J., Fletcher, L.R., Hume, D.E., Johnson, R.D., and Popay, A.J. (2013) The exploitation of epichloae endophytes for agricultural benefit. *Fungal Diversity* **60**: 171-188.

Johnson, L.J., Johnson, R.D., Schardl, C.L., and Panaccione, D.G. (2003) Identification of differentially expressed genes in the mutualistic association of tall fescue with *Neotyphodium coenophialum*. *Physiological and Molecular Plant Pathology* **63**: 305-317.

Jones, D.R., and Varela-Nieto, I. (1998) The role of glycosyl–phosphatidylinositol in signal transduction. *The International journal of Biochemistry & Cell Biology* **30**: 313-326.

Jun, G., Wing, M.K., Abecasis, G.R., and Kang, H.M. (2015) An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Research*: gr. 176552.176114.

Kececioglu, J., and Ju, J. (2001) Separating repeats in DNA sequence assembly. *Proceedings of The Fifth Annual International Conference on Computational Biology*: 176-183.

Kelley, D.R., Schatz, M.C., and Salzberg, S.L. (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biology* **11**: R116.

Kilpatrick, R. (1961) *Juncus effusus*, a new host for *Epichloe typhina*. *Plant Disease Reporter* **45**: 899-&.

Kinoshita, T. (2016) Glycosylphosphatidylinositol (GPI) anchors: biochemistry and cell biology: Introduction to a thematic review series. *Journal of Lipid Research* **57**: 4-5.

Knaus, H.-G., McManus, O.B., Lee, S.H., Schmalhofer, W.A., Garcia-Calvo, M., Helms, L.M.H., Sanchez, M., Giangiacomo, K., and Reuben, J.P. (1994) Tremorgenic indole alkaloids potently inhibit smooth muscle high-conductance calcium-activated potassium channels. *Biochemistry* **33**: 5819-5828.

Koboldt, D.C., Larson, D.E., Chen, K., Ding, L., and Wilson, R.K., (2012) Massively parallel sequencing approaches for characterization of structural variation. In: Genomic Structural Variants. Springer, pp. 369-384.

Koga, H., Christensen, M., and Bennett, R. (1993) Incompatibility of some grass-*Acremonium* endophyte associations. *Mycological Research* **97**: 1237-1244.

Kojima, K., Nariai, N., Mimori, T., Takahashi, M., Yamaguchi-Kabata, Y., Sato, Y., and Nagasaki, M. (2013) A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads. *Bioinformatics* **29**: 2835-2843.

Kopecký, D., Havránková, M., Loureiro, J., Castro, S., Lukaszewski, A., Bartoš, J., Kopecká, J., and Doležel, J. (2010) Physical distribution of homoeologous recombination in individual chromosomes of *Festuca pratensis* in *Lolium multiflorum*. *Cytogenetic and Genome Research* **129**: 162-172.

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes. *Nature Methods* **6**: 291.

Kuldau, G., and Bacon, C. (2008) Clavicipitaceous endophytes: their ability to enhance resistance of grasses to multiple stresses. *Biological Control* **46**: 57-71.

Kundzewicz, Z.W., Mata, L., Arnell, N.W., Döll, P., Jimenez, B., Miller, K., Oki, T., Şen, Z., and Shiklomanov, I. (2008) The implications of projected climate change for freshwater resources and their management. *Hydrological Sciences Journal* **53**: 3-10.

Langmead, B., and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25.

Latch, G., and Christensen, M. (1985) Artificial infection of grasses with endophytes. *Annals of Applied Biology* **107**: 17-24.

Laurie, S., Fernandez-Callejo, M., Marco-Sola, S., Trotta, J.-R., Camps, J., Chacón, A., Espinosa, A., Gut, M., Gut, I., Heath, S., and Beltran, S. (2016) From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. *Human mutation* **37**: 1263-1271.

Lean, I.J. (2001) Association between feeding perennial ryegrass (*Lolium perenne* cultivar Grasslands Impact) containing high concentrations of ergovaline, and health and productivity in a herd of lactating dairy cows. *Australian Veterinary Journal* **79**: 262-264 %@ 0005-0423.

Legendre, M., Pochet, N., Pak, T., and Verstrepen, K.J. (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research* **17**: 1787-1796.

Leuchtmann, A. (1992) Systematics, distribution, and host specificity of grass endophytes. *Natural Toxins* **1**: 150-162.

Leuchtmann, A., Bacon, C.W., Schardl, C.L., White Jr, J.F., and Tadych, M. (2014) Nomenclatural realignment of Neotyphodium species with genus *Epichloë*. *Mycologia* **106**: 202-215.

Leuchtmann, A., and Clay, K. (1993) Nonreciprocal compatibility between *Epichloë typhina a*nd four host grasses. *Mycologia* **85**: 157-163.

Leuchtmann, A., and Schardl, C.L. (1998) Mating compatibility and phylogenetic relationships among two new species of *Epichloë* and other congeneric European species. *Mycological Research* **102**: 1169-1182.

Leuchtmann, A., Schardl, C.L., and Siegel, M.R. (1994) Sexual compatibility and taxonomy of a new species of *Epichloë s*ymbiotic with fine fescue grasses. *Mycologia* **86**: 802-812.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987-2993.

Li, H. (2012) Exploring single-sample SNP and INDEL calling with whole-genome *de novo* assembly. *Bioinformatics* **28**: 1838-1844.

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li, H., and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**: 589-595.

Li, H., and Homer, N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics* **11**: 473-483.

Li, H., Ruan, J., and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*: gr. 078212.078108.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**: 1966-1967.

Li, S., Li, R., Li, H., Lu, J., Li, Y., Bolund, L., Schierup, M.H., and Wang, J. (2013) SOAPindel: efficient identification of indels from short paired reads. *Genome Research* **23**: 195-200.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012) Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology* **2012**: 251364-251364.

Liu, X., Han, S., Wang, Z., Gelernter, J., and Yang, B.-Z. (2013) Variant callers for next-generation sequencing data: a comparison study. *PloS ONE* **8**: e75619.

Lopez-Llorca, L. (1990) Purification and properties of extracellular proteases produced by the nematophagous fungus *Verticillium suchlasporium*. *Canadian Journal of Microbiology* **36**: 530-537.

Lopez-Llorca, L.V., and Robertson, W.M. (1992) Immunocytochemical localization of a 32-kDa protease from the nematophagous fungus *Verticillium suchlasporium* in infected nematode eggs. *Experimental Mycology* **16**: 261-267.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., and Liu, Y. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**: 18.

Lyons, P.C., Evans, J.J., and Bacon, C.W. (1990) Effects of the fungal endophyte *Acremonium coenophialum* on nitrogen accumulation and metabolism in tall fescue. *Plant Physiology* **92**: 726.

Majewska-Sawka, A., and Nakashima, H. (2004) Endophyte transmission via seeds of *Lolium perenne* L.: immunodetection of fungal antigens. *Fungal Genetics and Biology* **41**: 534-541.

Malinowski, D.P., and Belesky, D.P. (2000) Adaptations of Endophyte-Infected Cool-Season Grasses to Environmental Stresses: Mechanisms of Drought and Mineral Stress Tolerance. *Crop Science* **40**: 923-940.

Marcotte, E.M., Pellegrini, M., Yeates, T.O., and Eisenberg, D. (1999) A census of protein repeats1. *Journal of Molecular Biology* **293**: 151-160.

Margraf, R.L., Durtschi, J.D., Dames, S., Pattison, D.C., Stephens, J.E., and Voelkerding, K.V. (2011) Variant identification in multi-sample pools by Illumina genome analyzer sequencing. *Journal of Biomolecular Techniques* **22**: 74.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., and Chen, Z. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376.

Martinez, D.A., and Nelson, M.A. (2010) The next generation becomes the now generation. *PLoS Genetics* **6**: e1000906.

Massicotte, R., and Angers, B. (2012) General-purpose genotype or how epigenetics extend the flexibility of a genotype. *Genetics Research International* **2012**.

McCourt, C.M., McArt, D.G., Mills, K., Catherwood, M.A., Maxwell, P., Waugh, D.J., Hamilton, P., O'Sullivan, J.M., and Salto-Tellez, M. (2013) Validation of next generation sequencing technologies in comparison to current diagnostic gold standards for BRAF, EGFR and KRAS mutational analysis. *PLoS One* **8**: e69604.

McCoy, R.C., Taylor, R.W., Blauwkamp, T.A., Kelley, J.L., Kertesz, M., Pushkarev, D., Petrov, D.A., and Fiston-Lavier, A.-S. (2014) Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PloS ONE* **9**: e106689.

McLellan, A. (1997) Structure and analysis of phenotypic and genetic variation in clonal plants. *The Ecology and Evolution of Clonal Plants*: 185-210.

Meader, S., Hillier, L.W., Locke, D., Ponting, C.P., and Lunter, G. (2010) Genome assembly quality: assessment and improvement using the neutral indel model. *Genome Research* **20**: 675-684.

Medvedev, P., Stanciu, M., and Brudno, M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**: S13.

Meijer, G., and Leuchtmann, A. (2001) Fungal genotype controls mutualism and sex in *Brachyposium sylvaticum* infected by *Epichloë sylvatica*. *Acta Biologica Hungarica* **52**: 249-263.

Miller, J.R., Koren, S., and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315-327.

Milne, G. (2007) Technology transfer of novel ryegrass endophytes in New Zealand. *Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses*: 237-239.

Minnee, E.M.K., (2011) An evaluation of tall fescue (*Lolium arundinaceum*) as an alternative to perennial ryegrass (*Lolium perenne*) for use on dairy farms in the Waikato. In.: University of Waikato, pp.

Moate, P., Williams, S., Grainger, C., Hannah, M., Mapleson, D., Auldist, M., Greenwood, J., Popay, A., Hume, D., and Mace, W. (2012) Effects of wild-type, AR1 and AR37 endophyte-infected perennial ryegrass on dairy production in Victoria, Australia. *Animal Production Science* **52**: 1117-1130.

Moon, C.D., Craven, K.D., Leuchtmann, A., Clement, S.L., and Schardl, C.L. (2004) Prevalence of interspecific hybrids amongst asexual fungal endophytes of grasses. *Molecular Ecology* **13**: 1455-1467.

Moon, C.D., Miles, C.O., Järlfors, U., and Schardl, C.L. (2002) The evolutionary origins of three new *Neotyphodium* endophyte species from grasses indigenous to the Southern Hemisphere. *Mycologia* **94**: 694-711.

Moon, C.D., Scott, B., Schardl, C.L., and Christensen, M.J. (2000) The evolutionary origins of Epichloë endophytes from annual ryegrasses. *Mycologia* **92**: 1103-1118.

Mortimer, P., and Di Menna, M.E. (1983) Ryegrass staggers: further substantiation of a *Lolium* endophyte aetiology and the discovery of weevil resistance of ryegrass pastures infected with *Lolium* endophyte. *Proceedings of the New Zealand Grassland Association* **44**: 240-243.

Mostrom, M.S., and Jacobsen, B.J. (2011) Ruminant mycotoxicosis. *Veterinary Clinics: Food Animal Practice* **27**: 315-344.

Moxon, R., Bayliss, C., and Hood, D. (2006) Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annual Review of Genetics* **40**: 307-333.

Munday-Finch, S.C., and Garthwaite, I. (1999) Toxicology of ryegrass endophyte in livestock. *Grassland Research and Practice Series* **7**: 63-67.

Murray, M., and Thompson, W.F. (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research* **8**: 4321-4326.

Nagarajan, N., and Pop, M. (2013) Sequence assembly demystified. *Nature Reviews Genetics* **14**: 157.

Narzisi, G., O'rawe, J.A., Iossifov, I., Fang, H., Lee, Y.-h., Wang, Z., Wu, Y., Lyon, G.J., Wigler, M., and Schatz, M.C. (2014) Accurate *de novo* and transmitted indel detection in exome-capture data using microassembly. *Nature Methods* **11**: 1033.

Nekrutenko, A., and Taylor, J. (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics* **13**: 667.

Nicholson, M.J., Eaton, C.J., Stärkel, C., Tapper, B.A., Cox, M.P., and Scott, B. (2015) Molecular cloning and functional analysis of gene clusters for the biosynthesis of

indole-diterpenes in *Penicillium crustosum* and *P. janthinellum*. *Toxins* **7**: 2701-2722.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**: 443.

Ning, Z., Cox, A.J., and Mullikin, J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Research* **11**: 1725-1729.

Niones, J.T., and Takemoto, D. (2014) An isolate of *Epichloë festucae*, an endophytic fungus of temperate grasses, has growth inhibitory activity against selected grass pathogens. *Journal of General Plant Pathology* **80**: 337-347.

Nishikawa, M., and Ogawa, K.i. (2004) Antimicrobial activity of a chelatable poly(arginyl-histidine) produced by the ergot fungus *Verticillium kibiense*. *Antimicrobial Agents and Chemotherapy* **48**: 229-235.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., and Johnson, W.E. (2013) Low concordance of multiple variant-calling pipelines:practical implications for exome and genome sequencing. *Genome Medicine* **5**: 28.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., and Johnson, W., (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. In.: BioMed Central Ltd, pp.

Ohno, S. (1970) Evolution by gene duplication.

Olson, N.D., Lund, S.P., Colman, R.E., Foster, J.T., Sahl, J.W., Schupp, J.M., Keim, P., Morrow, J.B., Salit, M.L., and Zook, J.M. (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in Genetics* **6**: 235.

Otto, C., Stadler, P.F., and Hoffmann, S. (2014) Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics* **30**: 1837-1843.

Out, A.A., van Minderhout, I.J., Goeman, J.J., Ariyurek, Y., Ossowski, S., Schneeberger, K., Weigel, D., van Galen, M., Taschner, P.E., and Tops, C.M. (2009) Deep sequencing to reveal new variants in pooled DNA samples. *Human Mutation* **30**: 1703-1712.

Oyola, S.O., Otto, T.D., Gu, Y., Maslen, G., Manske, M., Campino, S., Turner, D.J., MacInnis, B., Kwiatkowski, D.P., and Swerdlow, H.P. (2012) Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC genomics* **13**: 1.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., and Trajanoski, Z. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* **15**: 256-278.

Pan, J., Bhardwaj, M., Nagabhyru, P., Grossman, R.B., and Schardl, C.L. (2014) Enzymes from fungal and plant origin required for chemical diversification of insecticidal loline alkaloids in grass-*Epichloë* symbiota. *PLoS One* **9**: e115590.

Pan, J.J., and Clay, K. (2002) Infection by the systemic fungus *Epichloë glyceriae* and clonal growth of its host grass *Glyceria striata*. *Oikos* **98**: 37-46.

Panaccione, D.G. (2005) Origins and significance of ergot alkaloid diversity in fungi. *FEMS Microbiology Letters* **251**: 9-17.

Pańka, D., Piesik, D., Jeske, M., and Baturo-Cieśniewska, A. (2013) Production of phenolics and the emission of volatile organic compounds by perennial ryegrass (*Lolium perenne* L.)/*Neotyphodium lolii* association as a response to infection by *Fusarium poae*. *Journal of Plant Physiology* **170**: 1010-1019.

Pearson, W.R. (2013) An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics* **Chapter 3**: Unit3.1-Unit3.1.

Pennell, C., Popay, A., Ball, O., E. Hume, D., and Baird, D., (2005) *Occurrence and impact of pasture mealybug (Balanococcus poae) and root aphid (Aploneura lentisci) on ryegrass (Lolium spp.) with and without infection by Neotyphodium fungal endophytes,* p. 329-337.

Philipson, M.N., and Christey, M.C. (1986) The relationship of host and endophyte during flowering, seed formation, and germination of *Lolium perenne*. *New Zealand Journal of Botany* **24**: 125-134.

Pightling, A.W., Petronella, N., and Pagotto, F. (2014) Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses. *PLOS ONE* **9**: e104579.

Pirooznia, M., Kramer, M., Parla, J., Goes, F.S., Potash, J.B., McCombie, W.R., and Zandi, P.P. (2014) Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics* **8**: 14.

Popay, A., and Gerard, P. (2007) Cultivar and endophyte effects on a root aphid, Aploneura lentisci, in perennial ryegrass. *New Zealand Plant Protection* **60**: 223-227.

Popay, A., and Hume, D. (2011) Endophytes improve ryegrass persistence by controlling insects. *Pasture Persistence Symposium:Grassland Research and Practice Series No. 15*: 149-156.

Popay, A.J., Cotching, B., Moorhead, A., and Ferguson, C.M., (2012) AR37 endophyte effects on porina and root aphid populations and ryegrass damage in the field. In: Proceedings of the New Zealand Grassland Association. pp. 165-170.

Popay, A.J., and Cox, N.R. (2016) *Aploneura lentisci (Homoptera:* Aphididae*)* and its interactions with fungal endophytes in perennial ryegrass *(Lolium perenne)*. *Frontiers in Plant Science* **7**: 1395.

Popay, A.J., Silvester, W.B., and Gerard, P.J. (2004) New endophyte isolate suppresses root aphid, *Aploneura lentisci*, in perennial ryegrass. *Proceedings of the 5th International Symposium on Neotyphodium / Grass Interactions*: 317.

Popay, A.J., and Thom, E.R., (2009) Endophyte effects on major insect pests in Waikato dairy pasture. In: Proceedings of the New Zealand Grassland Association. pp. 121-126.

Popay, A.J., and Wyatt, R.T., (1995) Resistance to Argentine stem weevil in perennial ryegrass infected with endophytes producing different alkaloids. In: New Zealand Plant Protection Conference. Hastings: New Zealand Plant Protection Society, pp. 229-236.

Posch, A.E., Herwig, C., and Spadiut, O. (2013) Science-based bioprocess design for filamentous fungi. *Trends in Biotechnology* **31**: 37-44.

Prestidge, R. (1982) An association of *Lolium* endophyte with ryegrass resistance to Argentine stem weevil. *Proceedings of the New Zealand Weed and Pest Control Conference* **35**: 199-222.

Prestidge, R.A., and Gallagher, R.T. (1988) Endophyte fungus confers resistance to ryegrass: Argentine stem weevil larval studies. *Ecological Entomology* **13**: 429-435.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**: 341.

Rasmussen, S., Parsons, A.J., Bassett, S., Christensen, M.J., Hume, D.E., Johnson, L.J., Johnson, R.D., Simpson, W.R., Stacke, C., Voisey, C.R., Xue, H., and Newman, J.A. (2007) High nitrogen supply and carbohydrate content reduce fungal endophyte and alkaloid concentration in *Lolium perenne*. *New Phytologist* **173**: 787-797.

Rasmussen, S., Parsons, A.J., Fraser, K., Xue, H., and Newman, J.A. (2008) Metabolic profiles of *Lolium perenne* are differentially affected by nitrogen supply, carbohydrate content, and fungal endophyte infection. *Plant Physiology* **146**: 1440.

Rasmussen, S., Parsons, A.J., and Newman, J.A. (2009) Metabolomics analysis of the *Lolium perenne–Neotyphodium lolii* symbiosis: more than just alkaloids? *Phytochemistry Reviews* **8**: 535-550.

Rice, W.R. (2002) Evolution of sex: experimental tests of the adaptive significance of sexual recombination. *Nature Reviews Genetics* **3**: 241.

Rieber, N., Zapatka, M., Lasitschka, B., Jones, D., Northcott, P., Hutter, B., Jäger, N., Kool, M., Taylor, M., and Lichter, P. (2013) Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PloS ONE* **8**: e66621.

Rowan, D.D., Latch, G.C., Bacon, C., and White, J. (1994) Utilization of endophyte-infected perennial ryegrasses for increased insect resistance. *Biotechnology of Endophytic Fungi of Grasses*: 169-183.

Ruffalo, M., Koyutürk, M., Ray, S., and LaFramboise, T. (2012) Accurate estimation of short read mapping quality for next-generation genome sequencing. *Bioinformatics* **28**: i349-i355.

Ruffalo, M., LaFramboise, T., and Koyutürk, M. (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27**: 2790-2796.

Saikia, S., Nicholson, M.J., Young, C., Parker, E.J., and Scott, B. (2008) The genetic basis for indole-diterpene chemical diversity in filamentous fungi. *Mycological Research* **112**: 184-199.

Saikia, S., Parker, E.J., Koulman, A., and Scott, B. (2006) Four gene products are required for the fungal synthesis of the indole-diterpene, paspaline. *FEBS Letters* **580**: 1625-1630.

Saikia, S., Takemoto D Fau - Tapper, B.A., Tapper Ba Fau - Lane, G.A., Lane Ga Fau - Fraser, K., Fraser K Fau - Scott, B., and Scott, B. (2012) Functional analysis of an indole-diterpene gene cluster for lolitrem B biosynthesis in the grass endosymbiont *Epichloe festucae*. *FEBS Letter* **586**: 2563-2569.

Saikkonen, K., Ahlholm, J., Helander, M., Lehtimäki, S., and Niemeläinen, O. (2000) Endophytic fungi in wild and cultivated grasses in Finland. *Ecography* **23**: 360-366.

Saikkonen, K., Faeth, S.H., Helander, M., and J, S.T. (1998) Fungal endophytes: a continuum of interactions with host plants. *Annual Review of Ecology and Systematics* **29**: 319-343.

Saikkonen, K., Ruokolainen, K., Huitu, O., Gundel, P.E., Piltti, T., Hamilton, C.E., and Helander, M. (2013) Fungal endophytes help prevent weed invasions. *Agriculture, Ecosystems & Environment* **165**: 1-5.

Saikkonen, K., Saari, S., and Helander, M. (2010a) Defensive mutualism between plants and endophytic fungi? *Fungal Diversity* **41**: 101-113.

Saikkonen, K., Wäli, P., Helander, M., and Faeth, S.H. (2004) Evolution of endophyte–plant symbioses. *Trends in Plant Science* **9**: 275-280.

Saikkonen, K., Wäli, P.R., and Helander, M. (2010b) Genetic compatibility determines endophyte-grass combinations. *PloS ONE* **5**: e11395.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., and Roberts, M. (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* **22**: 557-567.

Sandmann, S., de Graaf, A.O., Karimi, M., van der Reijden, B.A., Hellström-Lindberg, E., Jansen, J.H., and Dugas, M. (2017) Evaluating variant calling tools for non-matched next-generation sequencing data. *Scientific Reports* **7**: 43169.

Schadt, E.E., Turner, S., and Kasarskis, A. (2010) A window into third-generation sequencing. *Human Molecular Genetics* **19**: R227-R240.

Schardl, C.L. (2001) Epichloë festucae and Related Mutualistic Symbionts of Grasses. *Fungal Genetics and Biology* **33**: 69-82.

Schardl, C.L., Florea, S., Pan, J., Nagabhyru, P., Bec, S., and Calie, P.J. (2013a) The epichloae: alkaloid diversity and roles in symbiosis with grasses. *Current Opinion in Plant Biology* **16**: 480-488.

Schardl, C.L., Grossman, R.B., Nagabhyru, P., Faulkner, J.R., and Mallik, U.P. (2007) Loline alkaloids: currencies of mutualism. *Phytochemistry* **68**: 980-996.

Schardl, C.L., and Leuchtmann, A. (1999) Three new species of *Epichloë* symbiotic with North American grasses. *Mycologia*: 95-107.

Schardl, C.L., Leuchtmann, A., Chung, K.-R., Penny, D., and Siegel, M.R. (1997) Coevolution by common descent of fungal symbionts (*Epichloë* spp.) and grass hosts. *Molecular Biology and Evolution* **14**: 133-143.

Schardl, C.L., Leuchtmann, A., and Spiering, M.J. (2004) Symbioses of grasses with seedborne fungal endophytes. *Annual Review of Plant Biology* **55**: 315-340.

Schardl, C.L., Scott, B., Florea, S., and Zhang, D. (2009) *Epichloë* endophytes: clavicipitaceous symbionts of grasses. *The Mycota*: 276-306.

Schardl, C.L., Young, C.A., Hesse, U., Amyotte, S.G., Andreeva, K., Calie, P.J., Fleetwood, D.J., Haws, D.C., Moore, N., Oeser, B., Panaccione, D.G., Schweri, K.K., Voisey, C.R., Farman, M.L., Jaromczyk, J.W., Roe, B.A., O'Sullivan, D.M., Scott, B., Tudzynski, P., An, Z., Arnaoudova, E.G., Bullock, C.T., Charlton, N.D., Chen, L., Cox, M., Dinkins, R.D., Florea, S., Glenn, A.E., Gordon, A., Güldener, U., Harris, D.R., Hollin, W., Jaromczyk, J., Johnson, R.D., Khan, A.K., Leistner, E., Leuchtmann, A., Li, C., Liu, J., Liu, J., Liu, M., Mace, W., Machado, C., Nagabhyru, P., Pan, J., Schmid, J., Sugawara, K., Steiner, U., Takach, J.E., Tanaka, E., Webb, J.S., Wilson, E.V., Wiseman, J.L., Yoshida, R., and Zeng, Z. (2013b) Plant-symbiotic fungi as chemical engineers: multi-genome analysis of the clavicipitaceae reveals dynamics of alkaloid loci. *PLOS Genetics* **9**: e1003323.

Schardl, C.L., Young, C.A., Pan, J., Florea, S., Takach, J.E., Panaccione, D.G., Farman, M.L., Webb, J.S., Jaromczyk, J., and Charlton, N.D. (2013c) Currencies of mutualisms: sources of alkaloid genes in vertically transmitted epichloae. *Toxins* **5**: 1064-1088.

Schatz, M.C., Delcher, A.L., and Salzberg, S.L. (2010) Assembly of large genomes using second-generation sequencing. *Genome Research* **20**: 1165-1173.

Schmid, J., and Christensen, M.J. (1999) Ryegrass endophyte: host/fungus interaction. *Grassland Research and Practice Series* **7**: 101-106.

Schmid, J., Day, R., Zhang, N., Dupont, P.-Y., Cox, M.P., Schardl, C.L., Minards, N., Truglio, M., Moore, N., and Harris, D.R. (2017) Host tissue environment directs activities of an *Epichloë* endophyte, while it induces systemic hormone and defense responses in its native perennial ryegrass host. *Molecular Plant-Microbe Interactions* **30**: 138-149 %@ 0894-0282.

Schmidt, A.L., and Mitter, V. (2004) Microsatellite mutation directed by an external stimulus. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **568**: 233-243.

Schreiber, J., Wescoe, Z.L., Abu-Shumays, R., Vivian, J.T., Baatar, B., Karplus, K., and Akeson, M. (2013) Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences* **110**: 18910-18915.

Schuster, S.C., Chi, K., Rusk, N., Kiermer, V., Wold, B., and Myers, R. (2008) Method of the year, next-generation DNA sequencing. *Functional Genomics and Medical Applications. Nature Methods* **5**: 11-21.

Scott, B., Becker, Y., Becker, M., and Cartwright, G., (2012) Morphogenesis, growth, and development of the grass symbiont *Epichlöe festucae*. In: Morphogenesis and Pathogenicity in Fungi. J. Pérez-Martín & A. Di Pietro (eds). Springer Berlin Heidelberg, pp. 243-264.

Scott, B., Takemoto, D., and Tanaka, A. (2007) Fungal endophyte production of reactive oxygen species is critical for maintaining the mutualistic symbiotic interaction between *Epichloë festucae* and perennial ryegrass. *Plant Signaling and Behavior* **2**: 171-173.

Selosse, M.-A., and Schardl, C.L. (2007) Fungal endophytes of grasses: hybrids rescued by vertical transmission? An evolutionary perspective. *New Phytologist* **173**: 452-458.

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., and Shen, B. (2014) Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *BioMed Research International* **2014**: 16.

Shiba, T., and Sugawara, K. (2005) Resistance to the rice leaf bug, *Trigonotylus caelestialium*, is conferred by *Neotyphodium* endophyte infection of perennial ryegrass, *Lolium perenne*. *Entomologia Experimentalis et Applicata* **115**: 387-392.

Siegel, M., Latch, G., Bush, L., Fannin, F., Rowan, D., Tapper, B., Bacon, C., and Johnson, M. (1990) Fungal endophyte-infected grasses: alkaloid accumulation and aphid response. *Journal of Chemical Ecology* **16**: 3301-3315.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210-3212.

Simons, K., and Toomre, D. (2000) Lipid rafts and signal transduction. *Nature Reviews Molecular Cell Biology* **1**: 31.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*: gr. 089532.089108.

Simpson, W., Christensen, M., Johnson, R., and Schmid, J. (2007) Spontaneous in planta changes in fungal endophytes impact symbiosis. *Proceedings of the 6th International Symposium on Fungal Endophytes of Grasses*: 191-194.

Simpson, W.R., Schmid, J., Singh, J., Faville, M.J., and Johnson, R.D. (2012) A morphological change in the fungal symbiont *Neotyphodium lolii* induces dwarfing in its host plant *Lolium perenne*. *Fungal Biology* **116**: 234-240.

Smith, J.D., Gamain, B., Baruch, D.I., and Kyes, S. (2001) Decoding the language of var genes and *Plasmodium falciparum* sequestration. *Trends in Parasitology* **17**: 538-545.

Snyder, L.A., Butcher, S.A., and Saunders, N.J. (2001) Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic *Neisseria* spp. *Microbiology* **147**: 2321-2332.

Spiering, M.J., Faulkner, J.R., Zhang, D.-X., Machado, C., Grossman, R.B., and Schardl, C.L. (2008) Role of the LolP cytochrome P450 monooxygenase in loline alkaloid biosynthesis. *Fungal Genetics and Biology* **45**: 1307-1314.

Spiering, M.J., Lane, G.A., Christensen, M.J., and Schmid, J. (2005a) Distribution of the fungal endophyte *Neotyphodium lolii* is not a major determinant of the distribution of fungal alkaloids in *Lolium perenne* plants. *Phytochemistry* **66**: 195-202.

Spiering, M.J., Moon, C.D., Wilkinson, H.H., and Schardl, C.L. (2005b) Gene clusters for insecticidal loline alkaloids in the grass-endophytic fungus *Neotyphodium uncinatum*. *Genetics* **169**: 1403-1414.

Stern, A., Brown, M., Nickel, P., and Meyer, T.F. (1986) Opacity genes in Neisseria gonorrhoeae: control of phase and antigenic variation. *Cell* **47**: 61-71.

Stewart, A.V., and Charlton, J.F.L. (2006) Pasture and forage plants for New Zealand. *New Zealand Grassland Association*.

Strzepek, K., Yohe, G., Neumann, J., and Boehlert, B. (2010) Characterizing changes in drought risk for the United States from climate change. *Environmental Research Letters* **5**: 044012.

Stuedemann, J.A., and Hoveland, C.S. (1988) Fescue Endophyte: History and Impact on Animal Agriculture. *Journal of Production Agriculture* **1**: 39-44.

Sweeney, P.M., and Danneberger, T.K. (1994) Random amplified polymorphic DNA in perennial ryegrass: a comparison of bulk samples vs. individuals. *HortScience* **29**: 624-626.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L.M., Kamoun, S., and Terauchi, R. (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by

whole genome resequencing of DNA from two bulked populations. *The Plant Journal* **74**: 174-183.

Takemoto, D., Kamakura, S., Saikia, S., Becker, Y., Wrenn, R., Tanaka, A., Sumimoto, H., and Scott, B. (2011) Polarity proteins Bem1 and Cdc24 are components of the filamentous fungal NADPH oxidase complex. *Proceedings of the National Academy of Sciences* **108**: 2861-2866.

Takemoto, D., Tanaka, A., and Scott, B. (2006) A p67Phox-like regulator is recruited to control hyphal branching in a fungal-grass mutualistic symbiosis. *The Plant Cell* **18**: 2807-2821.

Tan, Y.Y., Spiering, M.J., Scott, V., Lane, G.A., Christensen, M.J., and Schmid, J. (2001) *In planta* regulation of extension of an endophytic fungus and maintenance of high metabolic rates in its mycelium in the absence of apical extension. *Applied and Environmental Microbiology* **67**: 5377.

Tanaka, A., Cartwright, G.M., Saikia, S., Kayano, Y., Takemoto, D., Kato, M., Tsuge, T., and Scott, B. (2013) ProA, a transcriptional regulator of fungal fruiting body development, regulates leaf hyphal network development in the *Epichloë festucae–Lolium perenne* symbiosis. *Molecular Microbiology* **90**: 551-568.

Tanaka, A., Christensen, M., Takemoto, D., and Scott, B. (2007) Endophyte production of reactive oxygen species is critical for maintaining the mutualistic symbiotic interaction between Epichloë festucae and Pooid grasses. *Plant Signaling & Behavior* **2**: 171-173.

Tanaka, A., Christensen, M.J., Takemoto, D., Park, P., and Scott, B. (2006) Reactive oxygen species play a role in regulating a fungus-perennial ryegrass mutualistic interaction. *The Plant Cell* **18**: 1052-1066.

Tanaka, A., Takemoto, D., Chujo, T., and Scott, B., (2012) *Fungal endophytes of grasses,* p. 462-468.

Tanaka, A., Takemoto, D., Hyon, G.S., Park, P., and Scott, B. (2008) NoxA activation by the small GTPase RacA is required to maintain a mutualistic symbiotic association between *Epichloë festucae* and *perennial ryegrass*. *Molecular Microbiology* **68**: 1165-1178.

Tanaka, A., Tapper, B.A., Popay, A., Parker, E.J., and Scott, B. (2005) A symbiosis expressed non-ribosomal peptide synthetase from a mutualistic fungal endophyte of perennial ryegrass confers protection to the symbiotum from insect herbivory. *Molecular Microbiology* **57**: 1036-1050.

Tapper, B., and Lane, G. (2004) Janthitrems found in a *Neotyphodium* endophyte of perennial ryegrass. *5th International Symposium on Neotyphodium / Grass Interactions* **301**.

Tapper, B.A., Cooper, B.M., Easton, H.S., Fletcher, L.R., Hume, D.E., Lane, G.A., Latch, G.C.M., Pennell, C.G.L., Popay, A.J., and Christensen, M.J., (2011) Grass endophytes. In.: Google Patents, pp.

Telleria, E.L., Araújo, A.P.O.d., Secundino, N.F., d'Avila-Levy, C.M., and Traub-Csekö, Y.M. (2010) Trypsin-Like Serine Proteases in Lutzomyia longipalpis – Expression, Activity and Possible Modulation by *Leishmania infantum chagasi*. *PLoS ONE* **5**: e10697.

TePaske, M.R., Powell, R.G., and Clement, S.L. (1993) Analyses of selected endophyte-infected grasses for the presence of loline-type and ergot-type alkaloids. *Journal of Agricultural and Food Chemistry* **41**: 2299-2303.

Thankaswamy-Kosalai, S., Sen, P., and Nookaew, I. (2017) Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics* **109**: 186-191.

Thom, E.R., Popay, A.J., Waugh, C.D., and Minneé, E.M.K. (2014) Impact of novel endophytes in perennial ryegrass on herbage production and insect pests from pastures under dairy cow grazing in northern New Zealand. *Grass and Forage Science* **69**: 191-204.

Thom, E.R., Waugh, C.D., and McCabe, R.J. (1998) Growth and persistence of perennial and hybrid ryegrasses when grazed by dairy cows in the central Waikato region of New Zealand. *New Zealand Journal of Agricultural Research* **41**: 477-486.

Tian, P., Le, T., Ludlow, E.J., Smith, K., Forster, J., Guthridge, K., and Spangenberg, G., (2013a) *Characterisation of novel perennial ryegrass host–Neotyphodium endophyte associations,* p. 716-725.

Tian, P., Le, T.-N., Smith, K., Forster, J., Guthridge, K., and Spangenberg, G. (2013b) Stability and viability of novel perennial ryegrass host–*Neotyphodium* endophyte associations. *Crop and Pasture Science* **64**: 39-50.

Torres, M.S., White, J.F., Zhang, X., Hinton, D.M., and Bacon, C.W. (2012) Endophyte-mediated adjustments in host morphology and physiology and effects on host fitness traits in grasses. *Fungal Ecology* **5**: 322-330.

Tsai, H.-F., Liu, J.-S., Staben, C., Christensen, M.J., Latch, G., Siegel, M.R., and Schardl, C.L. (1994) Evolutionary diversification of fungal endophytes of tall fescue grass by hybridization with *Epichloë* species. *Proceedings of the National Academy of Sciences* **91**: 2542-2546.

van Belkum, A. (1999) Short sequence repeats in microbial pathogenesis and evolution. *Cellular and Molecular Life Sciences* **56**: 729-734.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., and Thibault, J. (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43**: 11.10. 11-11.10. 33.

Van Ho, A., Ward, D.M., and Kaplan, J. (2002) Transition metal transport in yeast. *Annual Reviews in Microbiology* **56**: 237-261.

Verstrepen, K.J., and Fink, G.R. (2009) Genetic and epigenetic mechanisms underlying cell-surface variability in protozoa and fungi. *Annual Review of Genetics* **43**: 1-24.

Verstrepen, K.J., Jansen, A., Lewitter, F., and Fink, G.R. (2005) Intragenic tandem repeats generate functional variability. *Nature Genetics* **37**: 986-990.

Vijay, N., Poelstra, J.W., Künstner, A., and Wolf, J.B.W. (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments. *Molecular Ecology* **22**: 620-634.

Voisey, C.R., Christensen, M.T., Johnson, L.J., Forester, N.T., Gagic, M., Bryan, G.T., Simpson, W.R., Fleetwood, D.J., Card, S.D., and Koolaard, J.P. (2016) cAMP signaling regulates synchronised growth of symbiotic *Epichloe* fungi with the host grass *Lolium perenne*. *Frontiers in Plant Science* **7**: 1546.

Wajid, B., and Serpedin, E. (2012) Review of general algorithmic features for genome assemblers for next generation sequencers. *Genomics, Proteomics & Bioinformatics* **10**: 58-73.

Warden, C.D., Adamson, A.W., Neuhausen, S.L., and Wu, X. (2014) Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ-Life and Environment* **2**: e600.

Webb, D.M., and Knapp, S.J. (1990) DNA extraction from a previously recalcitrant plant genus. *Plant Molecular Biology Reporter* **8**: 180.

Wei, Z., Wang, W., Hu, P., Lyon, G.J., and Hakonarson, H. (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research* **39**: e132-e132.

Wescoe, Z.L., Schreiber, J., and Akeson, M. (2014) Nanopores discriminate among five C5-cytosine variants in DNA. *Journal of the American Chemical Society* **136**: 16582-16587.

White, J.F., and Reddy, P.V. (1998) Examination of structure and molecular phylogenetic relationships of some graminicolous symbionts in genera *Epichloë* and *Parepichloë*. *Mycologia* **90**: 226-234.

Wilkins, P. (1991) Breeding perennial ryegrass for agriculture. *Euphytica* **52**: 201-214.

Wilkinson, H.H., Siegel, M.R., Blankenship, J.D., Mallory, A.C., Bush, L.P., and Schardl, C.L. (2000) Contribution of fungal loline alkaloids to protection from aphids in a grass-endophyte mutualism. *Molecular Plant-Microbe Interactions* **13**: 1027-1033.

Winter, D.J., Ganley, A.R.D., Young, C.A., Liachko, I., Schardl, C.L., Dupont, P.-Y., Berry, D., Ram, A., Scott, B., and Cox, M.P. (2018) Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus *Epichloë festucae*. *PLoS Genetics* **14**: e1007467.

Wong, P.B., Wiley, E.O., Johnson, W.E., Ryder, O.A., O'Brien, S.J., Haussler, D., Koepfli, K.-P., Houck, M.L., Perelman, P., and Mastromonaco, G. (2012) Tissue sampling methods and standards for vertebrate genomics. *GigaScience* **1**: 8.

Woodfield, D.R., and Easton, H.S. (2004) Advances in pasture plant breeding for animal productivity and health. *New Zealand Veterinary Journal* **52**: 300-310.

Wyllie, M. (2013) Comprehensive analysis of clinical trials data shows unequivocally that Phosphodiesterase Inhibitors (PDEi) improve orgasm. The power of meta-analysis? *BJU international* **111**: 190-191.

Xu, C. (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal* **16**: 15-24.

Xu, F., Wang, W., Wang, P., Li, M.J., Sham, P.C., and Wang, J. (2012) A fast and accurate SNP detection algorithm for next-generation sequencing data. *Nature Communications* **3**: 1258.

Ye, L., Hillier, L.W., Minx, P., Thane, N., Locke, D.P., Martin, J.C., Chen, L., Mitreva, M., Miller, J.R., and Haub, K.V. (2011) A vertebrate case study of the quality of assemblies derived from next-generation sequences. *Genome Biology* **12**: R31.

Young, C., Hume, D., and McCulley, R. (2013) Forages and pastures symposium: fungal endophytes of tall fescue and perennial ryegrass: pasture friend or foe? *Journal of Animal Science* **91**: 2379-2394.

Young, C.A., Tapper, B.A., May, K., Moon, C.D., Schardl, C.L., and Scott, B. (2009) Indole-diterpene biosynthetic capability of epichloe endophytes as predicted by *ltm* gene analysis. *Applied and Environmental Microbiology* **75**: 2200-2211.

Yu, X., Guda, K., Willis, J., Veigl, M., Wang, Z., Markowitz, S., Adams, M.D., and Sun, S. (2012) How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Mining* **5**: 6.

Yu, X., and Sun, S. (2013) Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics* **14**: 274.

Zerbino, D., and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*: gr. 074492.074107.

Zhang, N., Scott, V., Al-Samarrai, T.H., Tan, Y.Y., Spiering, M.J., McMillan, L.K., Lane, G.A., Scott, D.B., Christensen, M.J., and Schmid, J. (2006) Transformation of the ryegrass endophyte *Neotyphodium lolii* can alter its in planta mycelial morphology. *Mycological Research* **110**: 601-611.

Zhang, N., Wheeler, D., Truglio, M., Lazzarini, C., Upritchard, J., McKinney, W., Rogers, K., Prigitano, A., Tortorano, A.M., and Cannon, R.D. (2018) Multi-locus next-generation sequence typing of DNA extracted from pooled colonies detects multiple unrelated *Candida albicans* strains in a significant proportion of patient samples. *Frontiers in Microbiology* **9**: 1179.

Zhang, N., Zhang, S., Borchert, S., Richardson, K., and Schmid, J. (2011a) 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. *Molecular Plant-Microbe Interactions* **24**: 984-992.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., and Shen, B. (2011b) A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PloS ONE* **6**: e17915.

Zhang, W., Ng, H.W., Shu, M., Luo, H., Su, Z., Ge, W., Perkins, R., Tong, W., and Hong, H. (2015) Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium. *Journal of Genetics* **94**: 731-740.

Zhu, Y., Bergland, A.O., González, J., and Petrov, D.A. (2012) Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*. *PloS ONE* **7**: e41901.

Zogli, P., and Libault, M. (2017) Plant response to biotic stress: Is there a common epigenetic response during plant-pathogenic and symbiotic interactions? *Plant Science* **263**: 89-93.

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology* **32**: 246.

5. APPENDIX

Supplementary Table 1

count selected										count selected									
406		14	63	359	406	0	18	401	406	0	406		212		406				

		population				caller			*E. festucae* Genome					AR37 assembly SNP impact predition			
snp No	SNP query	AR37	KLP	SAM	Sum	CRISP	FB	Sum	ORF hit	unique ORF hit (one per ORF)	hit ID	ORF/UPDOWN		POS	REF	ALT	Impact short
1	NODE_1_length_513618_cov_58.2815:121152-124152			1	1		1	1	EfM3.067170.mRNA-1	EfM3.067170.mRNA-1	EfM3.067170.mRNA-1 6823	ORF	UP,DOWN	122652	C	T	intron
1	NODE_1_length_513618_cov_58.2815:121152-124152			1	1		1	1	EfM3.067170.mRNA-1		EfM3.067170.mRNA-1 8604	UP,DOWN	UP,DOWN	122652	C	T	intron
1	NODE_1_length_513618_cov_58.2815:121152-124152			1	1		1	1	EfM3.067170.mRNA-1		EfM3.067170.mRNA-1 8655.8	UP,DOWN	UP,DOWN	122652	C	T	intron
1	NODE_1_length_513618_cov_58.2815:121152-124152			1	1		1	1	EfM3.067180.mRNA-1	EfM3.067180.mRNA-1	EfM3.067180.mRNA-1 4323	UP,DOWN	UP,DOWN	122652	C	T	intron
1	NODE_1_length_513618_cov_58.2815:121152-124152			1	1		1	1	EfM3.067180.mRNA-1		EfM3.067180.mRNA-1 6647	UP,DOWN	UP,DOWN	122652	C	T	intron
2	NODE_10_length_269079_cov_54.4166:172772-175772			1	1		1	1	EfM3.004830.mRNA-1	EfM3.004830.mRNA-1	EfM3.004830.mRNA-1 8414	ORF	UP,DOWN	180961	G	A	stop
2	NODE_10_length_269079_cov_54.4166:172772-175772			1	1		1	1	EfM3.069530.mRNA-1	EfM3.069530.mRNA-1	EfM3.069530.mRNA-1 2463.7	UP,DOWN	UP,DOWN	180961	G	A	stop
2	NODE_10_length_269079_cov_54.4166:172772-175772			1	1		1	1	EfM3.004820.mRNA-1	EfM3.004820.mRNA-1	EfM3.004820.mRNA-1 3989	UP,DOWN	UP,DOWN	180961	G	A	stop
4	NODE_10_length_269079_cov_54.4166:207044-210044			1	1		1	1	EfM3.004910.mRNA-1	EfM3.004910.mRNA-1	EfM3.004910.mRNA-1 3158.8	UP,DOWN	UP,DOWN	180961	G	A	stop
4	NODE_10_length_269079_cov_54.4166:207044-210044			1	1		1	1	EfM3.004900.mRNA-1		EfM3.004900.mRNA-1 1372	UP,DOWN	UP,DOWN	180961	G	A	stop
52	NODE_26_length_209350_cov_63.8136:90112-93112	1		1	1		1	1	EfM3.052650.mRNA-1	EfM3.052650.mRNA-1	EfM3.052650.mRNA-1 7832	1000bp +- hit ends 50 bp before SNP in ORF	UP,DOWN	91612	G	A	down
52	NODE_26_length_209350_cov_63.8136:90112-93112	1		1	1		1	1	EfM3.052640.mRNA-1		EfM3.052640.mRNA-1 13877		UP,DOWN	91612	G	A	down
52	NODE_26_length_209350_cov_63.8136:90112-93112	1		1	1		1	1	EfM3.052640.mRNA-1	EfM3.052640.mRNA-1	EfM3.052640.mRNA-1 10410		UP,DOWN	91612	G	A	down
52	NODE_26_length_209350_cov_63.8136:90112-93112	1		1	1		1	1	EfM3.052650.mRNA-1		EfM3.052650.mRNA-1 9664		ORF	91612	G	A	down

4	NODE_10_length_269079_cov_54.4166:207044-210044	1	1		1	1		EfM3.004900.mRNA-1	EfM3.004900.mRNA-1	EfM3.004900.mRNA-1 1151	UP,DOWN	UP,DOWN	180961	G	A	stop
4	NODE_10_length_269079_cov_54.4166:207044-210044	1	1		1	1		EfM3.004910.mRNA-1		EfM3.004910.mRNA-1 3690	UP,DOWN	UP,DOWN	180961	G	A	stop
4	NODE_10_length_269079_cov_54.4166:207044-210044	1	1		1	1		EfM3.004910.mRNA-1		EfM3.004910.mRNA-1 3730.2	UP,DOWN	UP,DOWN	180961	G	A	stop
4	NODE_10_length_269079_cov_54.4166:207044-210044	1	1		1	1		EfM3.004910.mRNA-1		EfM3.004910.mRNA-1 3834	ORF	UP,DOWN	180961	G	A	stop
4	NODE_10_length_269079_cov_54.4166:207044-210044	1	1		1	1		EfM3.004910.mRNA-1		EfM3.004910.mRNA-1 4061	UP,DOWN	UP,DOWN	180961	G	A	stop
5	NODE_102_length_101183_cov_61.0069:66238-69238	1	1		1	1		EfM3.029160.mRNA-1	EfM3.029160.mRNA-1	EfM3.029160.mRNA-1 4631	UP,DOWN	UP,DOWN	67738	T	C	NS
5	NODE_102_length_101183_cov_61.0069:66238-69238	1	1		1	1		EfM3.029160.mRNA-1		EfM3.029160.mRNA-1 5815	UP,DOWN	UP,DOWN	67738	T	C	NS
5	NODE_102_length_101183_cov_61.0069:66238-69238	1	1		1	1		EfM3.029160.mRNA-1		EfM3.029160.mRNA-1 7967	UP,DOWN	UP,DOWN	67738	T	C	NS
5	NODE_102_length_101183_cov_61.0069:66238-69238	1	1		1	1		EfM3.029170.mRNA-1	EfM3.029170.mRNA-1	EfM3.029170.mRNA-1 4667	UP,DOWN	UP,DOWN	67738	T	C	NS
5	NODE_102_length_101183_cov_61.0069:66238-69238	1	1		1	1		EfM3.029170.mRNA-1		EfM3.029170.mRNA-1 5130	ORF	UP,DOWN	67738	T	C	NS
5	NODE_102_length_101183_cov_61.0069:66238-69238	1	1		1	1		EfM3.029170.mRNA-1		EfM3.029170.mRNA-1 6241	UP,DOWN	UP,DOWN	67738	T	C	NS
6	NODE_113_length_94814_cov_53.6398:25823-28823	1	1		1	1		EfM3.032590.mRNA-1	EfM3.032590.mRNA-1	EfM3.032590.mRNA-1 1239	UP,DOWN	UP,DOWN	27323	T	C	syn
6	NODE_113_length_94814_cov_53.6398:25823-28823	1	1		1	1		EfM3.032590.mRNA-1		EfM3.032590.mRNA-1 4897	UP,DOWN	UP,DOWN	27323	T	C	syn
6	NODE_113_length_94814_cov_53.6398:25823-28823	1	1		1	1		EfM3.032590.mRNA-1		EfM3.032590.mRNA-1 5108	UP,DOWN	UP,DOWN	27323	T	C	syn
12	NODE_125_length_88790_cov_49.7683:43810-46810	1	1	2	1	1	2	EfM3.044460.mRNA-1	EfM3.044460.mRNA-1	EfM3.044460.mRNA-1 6447	UP,DOWN	UP,DOWN	45310	C	T	down
12	NODE_125_length_88790_cov_49.7683:43810-46810	1	1	2	1	1	2	EfM3.044460.mRNA-1		EfM3.044460.mRNA-1 6997	UP,DOWN	UP,DOWN	45310	C	T	down
12	NODE_125_length_88790_cov_49.7683:43810-46810	1	1	2	1	1	2	EfM3.044470.mRNA-1	EfM3.044470.mRNA-1	EfM3.044470.mRNA-1 9212	UP,DOWN	UP,DOWN	45310	C	T	down
11	NODE_123_length_90057_cov_57.6174:67340-70340	1		1		1	1	EfM3.030490.mRNA-1	EfM3.030490.mRNA-1	EfM3.030490.mRNA-1 6339	1207 bp +-	UP,DOWN	29191	G	A	stop
6	NODE_113_length_94814_cov_53.6398:25823-28823	1	1		1	1		EfM3.032590.mRNA-1		EfM3.032590.mRNA-1 8916	ORF	UP,DOWN	27323	T	C	syn
6	NODE_113_length_94814_cov_53.6398:25823-28823	1	1		1	1		EfM3.032600.mRNA-1	EfM3.032600.mRNA-1	EfM3.032600.mRNA-1 7029	UP,DOWN	UP,DOWN	27323	T	C	syn
6	NODE_113_length_94814_cov_53.6398:25823-28823	1	1		1	1		EfM3.032600.mRNA-1		EfM3.032600.mRNA-1 7777	UP,DOWN	UP,DOWN	27323	T	C	syn
7	NODE_114_length_94547_cov_53.7569:32629-35629	1	1		1	1		EfM3.009800.mRNA-1	EfM3.009800.mRNA-1	EfM3.009800.mRNA-1 3695	UP,DOWN	UP,DOWN	34129	C	A	up
7	NODE_114_length_94547_cov_53.7569:32629-35629	1	1		1	1		EfM3.009820.mRNA-1	EfM3.009820.mRNA-1	EfM3.009820.mRNA-1 8700	UP,DOWN	UP,DOWN	34129	C	A	up
7	NODE_114_length_94547_cov_53.7569:32629-35629	1	1		1	1		EfM3.009810.mRNA-1	EfM3.009810.mRNA-1	EfM3.009810.mRNA-1 6051	UP,DOWN	UP,DOWN	34129	C	A	up
7	NODE_114_length_94547_cov_53.7569:32629-35629	1	1		1	1		EfM3.009820.mRNA-1		EfM3.009820.mRNA-1 8897	UP,DOWN	UP,DOWN	34129	C	A	up

9	NODE_12_length_253888_cov_53.1707:83316-86316	1		1	2		1	1	EfM3.064630.mRNA-1	EfM3.064630.mRNA-1	EfM3.064630.mRNA-1 13101	1357 bp +-	UP,DOWN	84816	C	T	NS
11	NODE_123_length_90057_cov_57.6174:67340-70340		1		1		1	1	EfM3.030480.mRNA-1		EfM3.030480.mRNA-1 6651	1333 bp +-	UP,DOWN	29191	G	A	stop
9	NODE_12_length_253888_cov_53.1707:83316-86316	1		1	2		1	1	EfM3.064620.mRNA-1		EfM3.064620.mRNA-1 4320	1276 bp+-	UP,DOWN	84816	C	T	NS
9	NODE_12_length_253888_cov_53.1707:83316-86316	1		1	2		1	1	EfM3.064620.mRNA-1		EfM3.064620.mRNA-1 6849	551 bp +-	UP,DOWN	84816	C	T	NS
11	NODE_123_length_90057_cov_57.6174:67340-70340		1		1		1	1	EfM3.030480.mRNA-1	EfM3.030480.mRNA-1	EfM3.030480.mRNA-1 11246	ORF	ORF	29191	G	A	stop
9	NODE_12_length_253888_cov_53.1707:83316-86316	1		1	2		1	1	EfM3.064620.mRNA-1	EfM3.064620.mRNA-1	EfM3.064620.mRNA-1 11617	in ORF	ORF	84816	C	T	NS
13	NODE_127_length_88254_cov_59.0102:27300-30300			1	1		1	1	EfM3.022710.mRNA-1	EfM3.022710.mRNA-1	EfM3.022710.mRNA-1 3639	UP,DOWN	UP,DOWN	28800	A	G	IG
14	NODE_13_length_250774_cov_58.0492:244038-247038			1	1		1	1	EfM3.005370.mRNA-1	EfM3.005370.mRNA-1	EfM3.005370.mRNA-1 12069	ORF	UP,DOWN	89654	A	T	IG
14	NODE_13_length_250774_cov_58.0492:244038-247038			1	1		1	1	EfM3.005370.mRNA-1		EfM3.005370.mRNA-1 13372	UP,DOWN	UP,DOWN	89654	A	T	IG
14	NODE_13_length_250774_cov_58.0492:244038-247038			1	1		1	1	EfM3.005370.mRNA-1		EfM3.005370.mRNA-1 14354	UP,DOWN	UP,DOWN	89654	A	T	IG
14	NODE_13_length_250774_cov_58.0492:244038-247038			1	1		1	1	EfM3.005370.mRNA-1		EfM3.005370.mRNA-1 4692	UP,DOWN	UP,DOWN	89654	A	T	IG
14	NODE_13_length_250774_cov_58.0492:244038-247038			1	1		1	1	EfM3.005370.mRNA-1		EfM3.005370.mRNA-1 5653	UP,DOWN	UP,DOWN	89654	A	T	IG
14	NODE_13_length_250774_cov_58.0492:244038-247038			1	1		1	1	EfM3.005370.mRNA-1		EfM3.005370.mRNA-1 7248	UP,DOWN	UP,DOWN	89654	A	T	IG
14	NODE_13_length_250774_cov_58.0492:244038-247038			1	1		1	1	EfM3.005370.mRNA-1		EfM3.005370.mRNA-1 7688	UP,DOWN	UP,DOWN	89654	A	T	IG
14	NODE_13_length_250774_cov_58.0492:244038-247038			1	1		1	1	EfM3.005370.mRNA-1		EfM3.005370.mRNA-1 8056	UP,DOWN	UP,DOWN	89654	A	T	IG
15	NODE_136_length_83357_cov_56.0992:38653-41653			1	1		1	1	EfM3.007980.mRNA-1	EfM3.007980.mRNA-1	EfM3.007980.mRNA-1 3304	UP,DOWN	UP,DOWN	40153	C	T	up
15	NODE_136_length_83357_cov_56.0992:38653-41653			1	1		1	1	EfM3.007980.mRNA-1		EfM3.007980.mRNA-1 3525	UP,DOWN	UP,DOWN	40153	C	T	up
20	NODE_152_length_71826_cov_62.9213:66401-69401	1		1			1	1	EfM3.066230.mRNA-1		EfM3.066230.mRNA-1 3633	ORF	ORF	67901	A	T	IG
15	NODE_136_length_83357_cov_56.0992:38653-41653			1	1		1	1	EfM3.007980.mRNA-1		EfM3.007980.mRNA-1 3829	UP,DOWN	UP,DOWN	40153	C	T	up
63	NODE_323_length_24384_cov_80.5497:11269-14269	1	1	2		1		1	EfM3.027950.mRNA-1		EfM3.027950.mRNA-1 2335	UP,DOWN	UP,DOWN	12769	G	A	IG
63	NODE_323_length_24384_cov_80.5497:11269-14269	1	1	2		1		1	EfM3.034810.mRNA-1	EfM3.034810.mRNA-1	EfM3.034810.mRNA-1 5081.8	UP,DOWN	UP,DOWN	12769	G	A	IG
63	NODE_323_length_24384_cov_80.5497:11269-14269	1	1	2		1		1	EfM3.053010.mRNA-1	EfM3.053010.mRNA-1	EfM3.053010.mRNA-1 5752.9	UP,DOWN	UP,DOWN	12769	G	A	IG
63	NODE_323_length_24384_cov_80.5497:11269-14269	1	1	2		1		1	EfM3.053010.mRNA-1		EfM3.053010.mRNA-1 6319	UP,DOWN	UP,DOWN	12769	G	A	IG
63	NODE_323_length_24384_cov_80.5497:11269-14269	1	1	2		1		1	EfM3.080310.mRNA-1	EfM3.080310.mRNA-1	EfM3.080310.mRNA-1 5191	UP,DOWN	UP,DOWN	12769	G	A	IG
8	NODE_116_length_93301_cov_57.0796:62430-65430	1	1	2		1	1	2	EfM3.016880.mRNA-1	EfM3.016880.mRNA-1	EfM3.016880.mRNA-1 4137.3	UP,DOWN	UP,DOWN	63930	T	C	IG

8	NODE_116_length_93301_cov_57.0796:62430-65430	1	1	2	1	1	2	EfM3.053290.mRNA-1	EfM3.053290.mRNA-1	EfM3.053290.mRNA-1 470	UP,DOWN	UP,DOWN	63930	T	C	IG
16	NODE_14_length_249874_cov_54.4362:134049-137049		1	1		1	1	EfM3.029670.mRNA-1	EfM3.029670.mRNA-1	EfM3.029670.mRNA-1 5769	UP,DOWN	UP,DOWN	135549	C	T	NS
16	NODE_14_length_249874_cov_54.4362:134049-137049		1	1		1	1	EfM3.029670.mRNA-1		EfM3.029670.mRNA-1 6133	UP,DOWN	UP,DOWN	135549	C	T	NS
16	NODE_14_length_249874_cov_54.4362:134049-137049		1	1		1	1	EfM3.029670.mRNA-1		EfM3.029670.mRNA-1 6307	UP,DOWN	UP,DOWN	135549	C	T	NS
16	NODE_14_length_249874_cov_54.4362:134049-137049		1	1		1	1	EfM3.029670.mRNA-1		EfM3.029670.mRNA-1 7850	ORF	UP,DOWN	135549	C	T	NS
17	NODE_148_length_73013_cov_50.88:25550-28550		1	1		1	1	EfM3.080790.mRNA-1	EfM3.080790.mRNA-1	EfM3.080790.mRNA-1 6569	UP,DOWN	UP,DOWN	27050	A	T	NS
17	NODE_148_length_73013_cov_50.88:25550-28550		1	1		1	1	EfM3.080800.mRNA-1	EfM3.080800.mRNA-1	EfM3.080800.mRNA-1 2541	UP,DOWN	UP,DOWN	27050	A	T	NS
18	NODE_15_length_248045_cov_53.9114:34212-37212	1	1	2	1	1	2	EfM3.066230.mRNA-1		EfM3.066230.mRNA-1 6374	UP,DOWN	UP,DOWN	35712	A	G	IG
17	NODE_148_length_73013_cov_50.88:25550-28550		1	1		1	1	EfM3.080800.mRNA-1		EfM3.080800.mRNA-1 3498	UP,DOWN	UP,DOWN	27050	A	T	NS
17	NODE_148_length_73013_cov_50.88:25550-28550		1	1		1	1	EfM3.080800.mRNA-1		EfM3.080800.mRNA-1 8207	ORF	UP,DOWN	27050	A	T	NS
20	NODE_152_length_71826_cov_62.9213:66401-69401	1		1		1	1	EfM3.081040.mRNA-1	EfM3.081040.mRNA-1	EfM3.081040.mRNA-1 3214	1240 bp +-	UP,DOWN	67901	A	T	IG
20	NODE_152_length_71826_cov_62.9213:66401-69401	1		1		1	1	EfM3.071280.mRNA-1	EfM3.071280.mRNA-1	EfM3.071280.mRNA-1 3055	1302 bp +-	UP,DOWN	67901	A	T	IG
28	NODE_2_length_482800_cov_55.4142:33376-36376		1		1	1		EfM3.067790.mRNA-1	EfM3.067790.mRNA-1	EfM3.067790.mRNA-1 6081	1047 bp +-	UP,DOWN	480703	T	A	NS
28	NODE_2_length_482800_cov_55.4142:33376-36376		1		1	1		EfM3.067770.mRNA-1		EfM3.067770.mRNA-1 2568	869 bp +-	UP,DOWN	480703	T	A	NS
33	NODE_208_length_51299_cov_60.2239:41098-44098	1			1	1	1	EfM3.075030.mRNA-1	EfM3.075030.mRNA-1	EfM3.075030.mRNA-1 2331	518 bp +-	UP,DOWN	25369	C	T	IG
19	NODE_150_length_72625_cov_57.4354:26032-29032		1	1		1	1	EfM3.027420.mRNA-1	EfM3.027420.mRNA-1	EfM3.027420.mRNA-1 3791	UP,DOWN	UP,DOWN	27532	C	T	NS
19	NODE_150_length_72625_cov_57.4354:26032-29032		1	1		1	1	EfM3.027410.mRNA-1		EfM3.027410.mRNA-1 8455	UP,DOWN	UP,DOWN	27532	C	T	NS
19	NODE_150_length_72625_cov_57.4354:26032-29032		1	1		1	1	EfM3.027430.mRNA-1	EfM3.027430.mRNA-1	EfM3.027430.mRNA-1 4601	UP,DOWN	UP,DOWN	27532	C	T	NS
19	NODE_150_length_72625_cov_57.4354:26032-29032		1	1		1	1	EfM3.027420.mRNA-1		EfM3.027420.mRNA-1 4127	ORF	UP,DOWN	27532	C	T	NS
19	NODE_150_length_72625_cov_57.4354:26032-29032		1	1		1	1	EfM3.027410.mRNA-1	EfM3.027410.mRNA-1	EfM3.027410.mRNA-1 6777	UP,DOWN	UP,DOWN	27532	C	T	NS
21	NODE_153_length_71363_cov_54.1628:27565-30565		1	1		1	1	EfM3.019810.mRNA-1	EfM3.019810.mRNA-1	EfM3.019810.mRNA-1 4230	UP,DOWN	UP,DOWN	29065	C	T	NS
21	NODE_153_length_71363_cov_54.1628:27565-30565		1	1		1	1	EfM3.019800.mRNA-1		EfM3.019800.mRNA-1 1229	UP,DOWN	UP,DOWN	29065	C	T	NS
21	NODE_153_length_71363_cov_54.1628:27565-30565		1	1		1	1	EfM3.019800.mRNA-1	EfM3.019800.mRNA-1	EfM3.019800.mRNA-1 1108	UP,DOWN	UP,DOWN	29065	C	T	NS
21	NODE_153_length_71363_cov_54.1628:27565-30565		1	1		1	1	EfM3.019810.mRNA-1		EfM3.019810.mRNA-1 5434	ORF	UP,DOWN	29065	C	T	NS
21	NODE_153_length_71363_cov_54.1628:27565-30565		1	1		1	1	EfM3.019820.mRNA-1	EfM3.019820.mRNA-1	EfM3.019820.mRNA-1 10376	UP,DOWN	UP,DOWN	29065	C	T	NS

22	NODE_16_length_242150_cov_52.3555:107295-110295		1	1		1	1	EfM3.006620.mRNA-1	EfM3.006620.mRNA-1	EfM3.006620.mRNA-1 15310	ORF	UP,DOWN	108795	C	T	NS
22	NODE_16_length_242150_cov_52.3555:107295-110295		1	1		1	1	EfM3.006620.mRNA-1		EfM3.006620.mRNA-1 8498	UP,DOWN	UP,DOWN	108795	C	T	NS
22	NODE_16_length_242150_cov_52.3555:107295-110295		1	1		1	1	EfM3.006620.mRNA-1		EfM3.006620.mRNA-1 8709	UP,DOWN	UP,DOWN	108795	C	T	NS
23	NODE_16_length_242150_cov_52.3555:134442-137442		1	1		1	1	EfM3.006530.mRNA-1	EfM3.006530.mRNA-1	EfM3.006530.mRNA-1 6656	UP,DOWN	UP,DOWN	108795	C	T	NS
23	NODE_16_length_242150_cov_52.3555:134442-137442		1	1		1	1	EfM3.006530.mRNA-1		EfM3.006530.mRNA-1 6731.6	UP,DOWN	UP,DOWN	108795	C	T	NS
23	NODE_16_length_242150_cov_52.3555:134442-137442		1	1		1	1	EfM3.006540.mRNA-1	EfM3.006540.mRNA-1	EfM3.006540.mRNA-1 5213	UP,DOWN	UP,DOWN	108795	C	T	NS
24	NODE_166_length_64395_cov_60.6658:53820-56820		1	1		1	1	EfM3.058100.mRNA-1	EfM3.058100.mRNA-1	EfM3.058100.mRNA-1 2212	UP,DOWN	UP,DOWN	55320	T	A	up
24	NODE_166_length_64395_cov_60.6658:53820-56820		1	1		1	1	EfM3.058100.mRNA-1		EfM3.058100.mRNA-1 2824	UP,DOWN	UP,DOWN	55320	T	A	up
24	NODE_166_length_64395_cov_60.6658:53820-56820		1	1		1	1	EfM3.058100.mRNA-1		EfM3.058100.mRNA-1 2923	UP,DOWN	UP,DOWN	55320	T	A	up
28	NODE_2_length_482800_cov_55.4142:33376-36376	1		1		1	1	EfM3.067790.mRNA-2	EfM3.067790.mRNA-2	EfM3.067790.mRNA-2 6641	1047 bp +-	UP,DOWN	480703	T	A	NS
28	NODE_2_length_482800_cov_55.4142:33376-36376	1		1		1	1	EfM3.067770.mRNA-1	EfM3.067770.mRNA-1	EfM3.067770.mRNA-1 1457	1388 bp +-	UP,DOWN	480703	T	A	NS
28	NODE_2_length_482800_cov_55.4142:33376-36376	1		1		1	1	EfM3.067790.mRNA-1		EfM3.067790.mRNA-1 6724	1339 bp +-	UP,DOWN	480703	T	A	NS
28	NODE_2_length_482800_cov_55.4142:33376-36376		1	1		1	1	EfM3.067780.mRNA-1	EfM3.067780.mRNA-1	EfM3.067780.mRNA-1 7721	ORF	ORF	480703	T	A	NS
28	NODE_2_length_482800_cov_55.4142:33376-36376		1	1		1	1	EfM3.067790.mRNA-2		EfM3.067790.mRNA-2 6724	1339 bp +-	UP,DOWN	480703	T	A	NS
41	NODE_225_length_46794_cov_49.8022:1183-4183		1	1		1	1	EfM3.066230.mRNA-1		EfM3.066230.mRNA-1 994	1220 bp +-	UP,DOWN	2683	A	G	IG
50	NODE_253_length_38838_cov_49.4969:5257-8257		1	1		1	1	EfM3.081920.mRNA-1	EfM3.081920.mRNA-1	EfM3.081920.mRNA-1 15079	ORF	ORF	6757	C	T	stop
55	NODE_284_length_32063_cov_68.7758:11775-14775		1	1		1	1	EfM3.050025.mRNA-1	EfM3.050025.mRNA-1	EfM3.050025.mRNA-1 4500	522 bp +-	UP,DOWN	19809	G	A	IG
24	NODE_166_length_64395_cov_60.6658:53820-56820			1	1	1	1	EfM3.058100.mRNA-1		EfM3.058100.mRNA-1 3132	UP,DOWN	UP,DOWN	55320	T	A	up
24	NODE_166_length_64395_cov_60.6658:53820-56820			1	1	1	1	EfM3.058100.mRNA-1		EfM3.058100.mRNA-1 3393	UP,DOWN	UP,DOWN	55320	T	A	up
25	NODE_168_length_63307_cov_53.9501:44337-47337			1	1	1	1	EfM3.026750.mRNA-1	EfM3.026750.mRNA-1	EfM3.026750.mRNA-1 10597	UP,DOWN	UP,DOWN	45837	A	T	syn
25	NODE_168_length_63307_cov_53.9501:44337-47337			1	1	1	1	EfM3.026750.mRNA-1		EfM3.026750.mRNA-1 11561	UP,DOWN	UP,DOWN	45837	A	T	syn
25	NODE_168_length_63307_cov_53.9501:44337-47337			1	1	1	1	EfM3.026750.mRNA-1		EfM3.026750.mRNA-1 14809	ORF	UP,DOWN	45837	A	T	syn
26	NODE_170_length_63252_cov_54.6198:53160-56160			1	1	1	1	EfM3.041300.mRNA-1	EfM3.041300.mRNA-1	EfM3.041300.mRNA-1 8893	ORF	UP,DOWN	54660	C	G	NS
27	NODE_173_length_62854_cov_46.9374:37261-40261	1		1	2	1	1	EfM3.078720.mRNA-1	EfM3.078720.mRNA-1	EfM3.078720.mRNA-1 10353	in ORF	ORF	38761	T	A	up
29	NODE_2_length_482800_cov_55.4142:479203-482203			1	1	1	1	EfM3.010640.mRNA-1	EfM3.010640.mRNA-1	EfM3.010640.mRNA-1 8495	ORF	UP,DOWN	480703	T	A	NS

29	NODE_2_length_482800_cov_55.4142:479203-482203		1	1		1	1	EfM3.034570.mRNA-1	EfM3.034570.mRNA-1	EfM3.034570.mRNA-1 4380.8	UP,DOWN	UP,DOWN	480703	T	A	NS
61	NODE_321_length_24719_cov_56.5756:11233-14233	1		1		1	1	EfM3.081760.mRNA-1	EfM3.081760.mRNA-1	EfM3.081760.mRNA-1 11673	530 bp +-	UP,DOWN	12733	C	T	down
65	NODE_334_length_22958_cov_62.8208:46-3046	1		1		1	1	EfM3.052580.mRNA-1	EfM3.052580.mRNA-1	EfM3.052580.mRNA-1 5615	ORF	ORF	1546	T	A	IG
76	NODE_45_length_167596_cov_51.1068:132954-135954	1		1		1	1	EfM3.011480.mRNA-1	EfM3.011480.mRNA-1	EfM3.011480.mRNA-1 1693	1029 bp +-	UP,DOWN	85370	C	T	syn
76	NODE_45_length_167596_cov_51.1068:132954-135954	1		1		1	1	EfM3.011470.mRNA-1		EfM3.011470.mRNA-1 8174.4	1427 bp +-	UP,DOWN	85370	C	T	syn
76	NODE_45_length_167596_cov_51.1068:132954-135954	1		1		1	1	EfM3.011470.mRNA-1		EfM3.011470.mRNA-1 8486	595 bp +-	UP,DOWN	85370	C	T	syn
76	NODE_45_length_167596_cov_51.1068:132954-135954	1		1		1	1	EfM3.011470.mRNA-1	EfM3.011470.mRNA-1	EfM3.011470.mRNA-1 4879	256 bp +-	UP,DOWN	85370	C	T	syn
76	NODE_45_length_167596_cov_51.1068:132954-135954	1		1		1	1	EfM3.011480.mRNA-1		EfM3.011480.mRNA-1 4719	299 bp +-	UP,DOWN	85370	C	T	syn
29	NODE_2_length_482800_cov_55.4142:479203-482203		1	1		1	1	EfM3.010630.mRNA-1	EfM3.010630.mRNA-1	EfM3.010630.mRNA-1 8598	UP,DOWN	UP,DOWN	480703	T	A	NS
29	NODE_2_length_482800_cov_55.4142:479203-482203		1	1		1	1	EfM3.010650.mRNA-1	EfM3.010650.mRNA-1	EfM3.010650.mRNA-1 1321	UP,DOWN	UP,DOWN	480703	T	A	NS
29	NODE_2_length_482800_cov_55.4142:479203-482203		1	1		1	1	EfM3.010650.mRNA-2	EfM3.010650.mRNA-2	EfM3.010650.mRNA-2 1321	UP,DOWN	UP,DOWN	480703	T	A	NS
29	NODE_2_length_482800_cov_55.4142:479203-482203		1	1		1	1	EfM3.034570.mRNA-1		EfM3.034570.mRNA-1 4392.8	UP,DOWN	UP,DOWN	480703	T	A	NS
29	NODE_2_length_482800_cov_55.4142:479203-482203		1	1		1	1	EfM3.034570.mRNA-1		EfM3.034570.mRNA-1 4404.8	UP,DOWN	UP,DOWN	480703	T	A	NS
30	NODE_204_length_52389_cov_54.0348:30337-33337	1		1		1	1	EfM3.053600.mRNA-1	EfM3.053600.mRNA-1	EfM3.053600.mRNA-1 6672	UP,DOWN	UP,DOWN	31837	A	T	up
30	NODE_204_length_52389_cov_54.0348:30337-33337	1		1		1	1	EfM3.053580.mRNA-1	EfM3.053580.mRNA-1	EfM3.053580.mRNA-1 6710	UP,DOWN	UP,DOWN	31837	A	T	up
30	NODE_204_length_52389_cov_54.0348:30337-33337	1		1		1	1	EfM3.053590.mRNA-1	EfM3.053590.mRNA-1	EfM3.053590.mRNA-1 12902	UP,DOWN	UP,DOWN	31837	A	T	up
31	NODE_205_length_52307_cov_59.9204:24815-27815	1		1		1	1	EfM3.028860.partial-mRNA-1	EfM3.028860.partial-mRNA-1	EfM3.028860.partial-mRNA-1 2409.7	UP,DOWN	UP,DOWN	26315	G	A	NS
31	NODE_205_length_52307_cov_59.9204:24815-27815	1		1		1	1	EfM3.050930.mRNA-1	EfM3.050930.mRNA-1	EfM3.050930.mRNA-1 10525	UP,DOWN	UP,DOWN	26315	G	A	NS
31	NODE_205_length_52307_cov_59.9204:24815-27815	1		1		1	1	EfM3.050930.mRNA-1		EfM3.050930.mRNA-1 3843	UP,DOWN	UP,DOWN	26315	G	A	NS
31	NODE_205_length_52307_cov_59.9204:24815-27815	1		1		1	1	EfM3.050930.mRNA-1		EfM3.050930.mRNA-1 6136	ORF	UP,DOWN	26315	G	A	NS
31	NODE_205_length_52307_cov_59.9204:24815-27815	1		1		1	1	EfM3.050930.mRNA-1		EfM3.050930.mRNA-1 6722	UP,DOWN	UP,DOWN	26315	G	A	NS
32	NODE_207_length_51378_cov_52.2233:9276-12276	1		1		1	1	EfM3.008170.mRNA-1	EfM3.008170.mRNA-1	EfM3.008170.mRNA-1 1378	UP,DOWN	UP,DOWN	10776	G	A	up
32	NODE_207_length_51378_cov_52.2233:9276-12276	1		1		1	1	EfM3.008170.mRNA-1		EfM3.008170.mRNA-1 2175	UP,DOWN	UP,DOWN	10776	G	A	up
32	NODE_207_length_51378_cov_52.2233:9276-12276	1		1		1	1	EfM3.008180.mRNA-1	EfM3.008180.mRNA-1	EfM3.008180.mRNA-1 4352	UP,DOWN	UP,DOWN	10776	G	A	up

32	NODE_207_length_51378_cov_52.2233:9276-12276	1	1	1	1	EfM3.008180.mRNA-1		EfM3.008180.mRNA-1 6916	UP,DOWN	UP,DOWN	10776	G	A	up
32	NODE_207_length_51378_cov_52.2233:9276-12276	1	1	1	1	EfM3.008180.mRNA-1		EfM3.008180.mRNA-1 7671	UP,DOWN	UP,DOWN	10776	G	A	up
32	NODE_207_length_51378_cov_52.2233:9276-12276	1	1	1	1	EfM3.008180.mRNA-1		EfM3.008180.mRNA-1 8455	UP,DOWN	UP,DOWN	10776	G	A	up
32	NODE_207_length_51378_cov_52.2233:9276-12276	1	1	1	1	EfM3.008180.mRNA-2	EfM3.008180.mRNA-2	EfM3.008180.mRNA-2 4352	UP,DOWN	UP,DOWN	10776	G	A	up
32	NODE_207_length_51378_cov_52.2233:9276-12276	1	1	1	1	EfM3.008180.mRNA-2		EfM3.008180.mRNA-2 6916	UP,DOWN	UP,DOWN	10776	G	A	up
32	NODE_207_length_51378_cov_52.2233:9276-12276	1	1	1	1	EfM3.008180.mRNA-2		EfM3.008180.mRNA-2 7671	UP,DOWN	UP,DOWN	10776	G	A	up
32	NODE_207_length_51378_cov_52.2233:9276-12276	1	1	1	1	EfM3.008180.mRNA-2		EfM3.008180.mRNA-2 8455	UP,DOWN	UP,DOWN	10776	G	A	up
34	NODE_210_length_50417_cov_51.3888:5523-8523	1	1	1	1	EfM3.027980.mRNA-1	EfM3.027980.mRNA-1	EfM3.027980.mRNA-1 5582	UP,DOWN	UP,DOWN	2465	A	C	IG
34	NODE_210_length_50417_cov_51.3888:5523-8523	1	1	1	1	EfM3.027980.mRNA-1		EfM3.027980.mRNA-1 5882	UP,DOWN	UP,DOWN	2465	A	C	IG
34	NODE_210_length_50417_cov_51.3888:5523-8523	1	1	1	1	EfM3.027980.mRNA-1		EfM3.027980.mRNA-1 6318	UP,DOWN	UP,DOWN	2465	A	C	IG
34	NODE_210_length_50417_cov_51.3888:5523-8523	1	1	1	1	EfM3.027980.mRNA-1		EfM3.027980.mRNA-1 6710	UP,DOWN	UP,DOWN	2465	A	C	IG
34	NODE_210_length_50417_cov_51.3888:5523-8523	1	1	1	1	EfM3.027980.mRNA-1		EfM3.027980.mRNA-1 6916.5	UP,DOWN	UP,DOWN	2465	A	C	IG
35	NODE_212_length_49764_cov_53.3715:39262-42262	1	1	1	1	EfM3.055700.mRNA-1	EfM3.055700.mRNA-1	EfM3.055700.mRNA-1 3777	UP,DOWN ORF	UP,DOWN	40762	C	A	IG
35	NODE_212_length_49764_cov_53.3715:39262-42262	1	1	1	1	EfM3.055700.mRNA-1		EfM3.055700.mRNA-1 4666	UP,DOWN	UP,DOWN	40762	C	A	IG
35	NODE_212_length_49764_cov_53.3715:39262-42262	1	1	1	1	EfM3.055710.mRNA-1	EfM3.055710.mRNA-1	EfM3.055710.mRNA-1 3152.8	UP,DOWN	UP,DOWN	40762	C	A	IG
36	NODE_215_length_49240_cov_49.4966:3322-6322	1	1	1	1	EfM3.018360.mRNA-1	EfM3.018360.mRNA-1	EfM3.018360.mRNA-1 14106	ORF	UP,DOWN	4822	A	G	syn
36	NODE_215_length_49240_cov_49.4966:3322-6322	1	1	1	1	EfM3.018360.mRNA-1		EfM3.018360.mRNA-1 9854	UP,DOWN	UP,DOWN	4822	A	G	syn
36	NODE_215_length_49240_cov_49.4966:3322-6322	1	1	1	1	EfM3.018360.mRNA-2	EfM3.018360.mRNA-2	EfM3.018360.mRNA-2 5367	UP,DOWN	UP,DOWN	4822	A	G	syn
36	NODE_215_length_49240_cov_49.4966:3322-6322	1	1	1	1	EfM3.018360.mRNA-2		EfM3.018360.mRNA-2 8547	ORF	UP,DOWN	4822	A	G	syn
37	NODE_218_length_48865_cov_50.9297:33785-36785	1	1	1	1	EfM3.020730.mRNA-1	EfM3.020730.mRNA-1	EfM3.020730.mRNA-1 938	UP,DOWN	UP,DOWN	35285	T	A	up
37	NODE_218_length_48865_cov_50.9297:33785-36785	1	1	1	1	EfM3.020730.mRNA-2	EfM3.020730.mRNA-2	EfM3.020730.mRNA-2 938	UP,DOWN	UP,DOWN	35285	T	A	up
37	NODE_218_length_48865_cov_50.9297:33785-36785	1	1	1	1	EfM3.020720.mRNA-1	EfM3.020720.mRNA-1	EfM3.020720.mRNA-1 9453	UP,DOWN	UP,DOWN	35285	T	A	up
38	NODE_22_length_221081_cov_53.8872:162429-165429	1	1	1	1	EfM3.004300.mRNA-1	EfM3.004300.mRNA-1	EfM3.004300.mRNA-1 10130	ORF	UP,DOWN	163929	T	A	NS
38	NODE_22_length_221081_cov_53.8872:162429-165429	1	1	1	1	EfM3.004290.mRNA-1		EfM3.004290.mRNA-1 6389.4	UP,DOWN	UP,DOWN	163929	T	A	NS
38	NODE_22_length_221081_cov_53.8872:162429-165429	1	1	1	1	EfM3.004290.mRNA-1	EfM3.004290.mRNA-1	EfM3.004290.mRNA-1 6198	UP,DOWN	UP,DOWN	163929	T	A	NS

38	NODE_22_length_221081_cov_53.8872:162429-165429	1	1	1	1	EfM3.004300.mRNA-1		EfM3.004300.mRNA-1 5352	UP,DOWN	UP,DOWN	163929	T	A	NS
39	NODE_224_length_47513_cov_51.0753:22839-25839	1	1	1	1	EfM3.002080.mRNA-1	EfM3.002080.mRNA-1	EfM3.002080.mRNA-1 4586	UP,DOWN	UP,DOWN	24339	T	C	NS
39	NODE_224_length_47513_cov_51.0753:22839-25839	1	1	1	1	EfM3.002070.mRNA-1		EfM3.002070.mRNA-1 2527.1	UP,DOWN	UP,DOWN	24339	T	C	NS
39	NODE_224_length_47513_cov_51.0753:22839-25839	1	1	1	1	EfM3.002070.mRNA-1	EfM3.002070.mRNA-1	EfM3.002070.mRNA-1 2468	UP,DOWN	UP,DOWN	24339	T	C	NS
39	NODE_224_length_47513_cov_51.0753:22839-25839	1	1	1	1	EfM3.002080.mRNA-1		EfM3.002080.mRNA-1 4914	ORF	UP,DOWN	24339	T	C	NS
39	NODE_224_length_47513_cov_51.0753:22839-25839	1	1	1	1	EfM3.002080.mRNA-1		EfM3.002080.mRNA-1 5334	UP,DOWN	UP,DOWN	24339	T	C	NS
40	NODE_224_length_47513_cov_51.0753:26495-29495	1	1	1	1	EfM3.002090.mRNA-1	EfM3.002090.mRNA-1	EfM3.002090.mRNA-1 4145	ORF	UP,DOWN	24339	T	C	NS
40	NODE_224_length_47513_cov_51.0753:26495-29495	1	1	1	1	EfM3.002090.mRNA-1		EfM3.002090.mRNA-1 8584	UP,DOWN	UP,DOWN	24339	T	C	NS
40	NODE_224_length_47513_cov_51.0753:26495-29495	1	1	1	1	EfM3.002100.mRNA-1	EfM3.002100.mRNA-1	EfM3.002100.mRNA-1 9512	UP,DOWN	UP,DOWN	24339	T	C	NS
42	NODE_23_length_219583_cov_56.1617:103833-106833	1	1	1	1	EfM3.022250.mRNA-1	EfM3.022250.mRNA-1	EfM3.022250.mRNA-1 11907	ORF	UP,DOWN	105333	A	G	syn
42	NODE_23_length_219583_cov_56.1617:103833-106833	1	1	1	1	EfM3.022240.mRNA-1		EfM3.022240.mRNA-1 4772	UP,DOWN	UP,DOWN	105333	A	G	syn
42	NODE_23_length_219583_cov_56.1617:103833-106833	1	1	1	1	EfM3.022240.mRNA-1	EfM3.022240.mRNA-1	EfM3.022240.mRNA-1 1845	UP,DOWN	UP,DOWN	105333	A	G	syn
42	NODE_23_length_219583_cov_56.1617:103833-106833	1	1	1	1	EfM3.022240.mRNA-2		EfM3.022240.mRNA-2 5180	UP,DOWN	UP,DOWN	105333	A	G	syn
42	NODE_23_length_219583_cov_56.1617:103833-106833	1	1	1	1	EfM3.022240.mRNA-2	EfM3.022240.mRNA-2	EfM3.022240.mRNA-2 2253	UP,DOWN	UP,DOWN	105333	A	G	syn
42	NODE_23_length_219583_cov_56.1617:103833-106833	1	1	1	1	EfM3.022250.mRNA-2	EfM3.022250.mRNA-2	EfM3.022250.mRNA-2 8663	UP,DOWN	UP,DOWN	105333	A	G	syn
43	NODE_232_length_44650_cov_61.2149:10564-13564	1	1	1	1	EfM3.015360.mRNA-1	EfM3.015360.mRNA-1	EfM3.015360.mRNA-1 11269	ORF	UP,DOWN	12064	A	G	NS
43	NODE_232_length_44650_cov_61.2149:10564-13564	1	1	1	1	EfM3.015360.mRNA-1		EfM3.015360.mRNA-1 1160	UP,DOWN	UP,DOWN	12064	A	G	NS
43	NODE_232_length_44650_cov_61.2149:10564-13564	1	1	1	1	EfM3.015360.mRNA-1		EfM3.015360.mRNA-1 11726	UP,DOWN	UP,DOWN	12064	A	G	NS
44	NODE_237_length_43406_cov_51.6467:18358-21358	1	1	1	1	EfM3.047780.mRNA-1	EfM3.047780.mRNA-1	EfM3.047780.mRNA-1 8429	ORF	UP,DOWN	19858	C	T	up
44	NODE_237_length_43406_cov_51.6467:18358-21358	1	1	1	1	EfM3.047770.mRNA-1	EfM3.047770.mRNA-1	EfM3.047770.mRNA-1 8938	UP,DOWN	UP,DOWN	19858	C	T	up
44	NODE_237_length_43406_cov_51.6467:18358-21358	1	1	1	1	EfM3.047780.mRNA-2	EfM3.047780.mRNA-2	EfM3.047780.mRNA-2 5773	UP,DOWN	UP,DOWN	19858	C	T	up
45	NODE_24_length_218463_cov_54.4866:202651-205651	1	1	1	1	EfM3.057950.mRNA-1	EfM3.057950.mRNA-1	EfM3.057950.mRNA-1 3214	UP,DOWN	UP,DOWN	204151	T	A	IG
45	NODE_24_length_218463_cov_54.4866:202651-205651	1	1	1	1	EfM3.051030.mRNA-1	EfM3.051030.mRNA-1	EfM3.051030.mRNA-1 2581.8	UP,DOWN	UP,DOWN	204151	T	A	IG
45	NODE_24_length_218463_cov_54.4866:202651-205651	1	1	1	1	EfM3.060940.mRNA-1	EfM3.060940.mRNA-1	EfM3.060940.mRNA-1 9675	UP,DOWN	UP,DOWN	204151	T	A	IG
46	NODE_242_length_42019_cov_52.6894:10222-13222	1	1	1	1	EfM3.047620.mRNA-1	EfM3.047620.mRNA-1	EfM3.047620.mRNA-1 7824	UP,DOWN	UP,DOWN	11722	G	A	NS

46	NODE_242_length_42019_cov_52.6894:10222-13222	1	1	1	1	EfM3.047620.mRNA-1		EfM3.047620.mRNA-1 8059	UP,DOWN	UP,DOWN	11722	G	A	NS
46	NODE_242_length_42019_cov_52.6894:10222-13222	1	1	1	1	EfM3.047620.mRNA-1		EfM3.047620.mRNA-1 9495	UP,DOWN	UP,DOWN	11722	G	A	NS
46	NODE_242_length_42019_cov_52.6894:10222-13222	1	1	1	1	EfM3.047620.mRNA-1		EfM3.047620.mRNA-1 9552	ORF	UP,DOWN	11722	G	A	NS
47	NODE_246_length_40352_cov_52.7719:8380-11380	1	1	1	1	EfM3.041210.mRNA-1	EfM3.041210.mRNA-1	EfM3.041210.mRNA-1 9467	ORF	UP,DOWN	9880	G	A	NS
47	NODE_246_length_40352_cov_52.7719:8380-11380	1	1	1	1	EfM3.041200.mRNA-1	EfM3.041200.mRNA-1	EfM3.041200.mRNA-1 6889	UP,DOWN	UP,DOWN	9880	G	A	NS
48	NODE_248_length_39928_cov_78.1307:35766-38766	1	1	1	1	EfM3.027950.mRNA-1		EfM3.027950.mRNA-1 6368	UP,DOWN	UP,DOWN	37266	G	A	IG
48	NODE_248_length_39928_cov_78.1307:35766-38766	1	1	1	1	EfM3.064570.mRNA-1	EfM3.064570.mRNA-1	EfM3.064570.mRNA-1 1297	UP,DOWN	UP,DOWN	37266	G	A	IG
48	NODE_248_length_39928_cov_78.1307:35766-38766	1	1	1	1	EfM3.067010.mRNA-1	EfM3.067010.mRNA-1	EfM3.067010.mRNA-1 1380	UP,DOWN	UP,DOWN	37266	G	A	IG
48	NODE_248_length_39928_cov_78.1307:35766-38766	1	1	1	1	EfM3.028530.mRNA-1	EfM3.028530.mRNA-1	EfM3.028530.mRNA-1 1320	UP,DOWN	UP,DOWN	37266	G	A	IG
49	NODE_25_length_213189_cov_60.4124:54464-57464	1	1	1	1	EfM3.036280.mRNA-1	EfM3.036280.mRNA-1	EfM3.036280.mRNA-1 7658	ORF	UP,DOWN	55964	C	T	down
49	NODE_25_length_213189_cov_60.4124:54464-57464	1	1	1	1	EfM3.036280.mRNA-1		EfM3.036280.mRNA-1 7719	UP,DOWN	UP,DOWN	55964	C	T	down
49	NODE_25_length_213189_cov_60.4124:54464-57464	1	1	1	1	EfM3.036280.mRNA-1		EfM3.036280.mRNA-1 7786	UP,DOWN	UP,DOWN	55964	C	T	down
49	NODE_25_length_213189_cov_60.4124:54464-57464	1	1	1	1	EfM3.036280.mRNA-1		EfM3.036280.mRNA-1 7844	UP,DOWN	UP,DOWN	55964	C	T	down
49	NODE_25_length_213189_cov_60.4124:54464-57464	1	1	1	1	EfM3.036280.mRNA-1		EfM3.036280.mRNA-1 7864	UP,DOWN	UP,DOWN	55964	C	T	down
49	NODE_25_length_213189_cov_60.4124:54464-57464	1	1	1	1	EfM3.036280.mRNA-1		EfM3.036280.mRNA-1 7875.1	UP,DOWN	UP,DOWN	55964	C	T	down
49	NODE_25_length_213189_cov_60.4124:54464-57464	1	1	1	1	EfM3.036280.mRNA-1		EfM3.036280.mRNA-1 8059	UP,DOWN	UP,DOWN	55964	C	T	down
51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052800.mRNA-2	EfM3.052800.mRNA-2	EfM3.052800.mRNA-2 1048	UP,DOWN	UP,DOWN	91612	G	A	down
51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052800.mRNA-1		EfM3.052800.mRNA-1 1803	UP,DOWN	UP,DOWN	91612	G	A	down
51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052800.mRNA-1	EfM3.052800.mRNA-1	EfM3.052800.mRNA-1 1284	UP,DOWN	UP,DOWN	91612	G	A	down
51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052800.mRNA-2		EfM3.052800.mRNA-2 1803	UP,DOWN	UP,DOWN	91612	G	A	down
51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052820.mRNA-1	EfM3.052820.mRNA-1	EfM3.052820.mRNA-1 6780	UP,DOWN	UP,DOWN	91612	G	A	down
51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052810.mRNA-1		EfM3.052810.mRNA-1 4524	UP,DOWN	UP,DOWN	91612	G	A	down
51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052810.mRNA-1		EfM3.052810.mRNA-1 5630	UP,DOWN	UP,DOWN	91612	G	A	down
51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052810.mRNA-1	EfM3.052810.mRNA-1	EfM3.052810.mRNA-1 3652	UP,DOWN	UP,DOWN	91612	G	A	down
51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052820.mRNA-1		EfM3.052820.mRNA-1 7091	UP,DOWN	UP,DOWN	91612	G	A	down

51	NODE_26_length_209350_cov_63.8136:142130-145130	1	1	1	1	EfM3.052820.mRNA-1		EfM3.052820.mRNA-1 7113	UP,DOWN	UP,DOWN	91612	G	A	down
53	NODE_27_length_207524_cov_51.7517:102331-105331	1	1	1	1	EfM3.016840.mRNA-1	EfM3.016840.mRNA-1	EfM3.016840.mRNA-1 3768	ORF	UP,DOWN	103831	C	T	NS
53	NODE_27_length_207524_cov_51.7517:102331-105331	1	1	1	1	EfM3.016830.mRNA-2	EfM3.016830.mRNA-2	EfM3.016830.mRNA-2 7295	UP,DOWN	UP,DOWN	103831	C	T	NS
53	NODE_27_length_207524_cov_51.7517:102331-105331	1	1	1	1	EfM3.016830.mRNA-1	EfM3.016830.mRNA-1	EfM3.016830.mRNA-1 8732	UP,DOWN	UP,DOWN	103831	C	T	NS
53	NODE_27_length_207524_cov_51.7517:102331-105331	1	1	1	1	EfM3.016840.mRNA-1		EfM3.016840.mRNA-1 6974	UP,DOWN	UP,DOWN	103831	C	T	NS
54	NODE_272_length_33880_cov_81.5648:11307-14307	1	1	1	1	EfM3.062340.mRNA-1	EfM3.062340.mRNA-1	EfM3.062340.mRNA-1 3814	ORF	UP,DOWN	12807	C	T	up
54	NODE_272_length_33880_cov_81.5648:11307-14307	1	1	1	1	EfM3.062340.mRNA-1		EfM3.062340.mRNA-1 6221	UP,DOWN	UP,DOWN	12807	C	T	up
54	NODE_272_length_33880_cov_81.5648:11307-14307	1	1	1	1	EfM3.077170.mRNA-1	EfM3.077170.mRNA-1	EfM3.077170.mRNA-1 5073	UP,DOWN	UP,DOWN	12807	C	T	up
54	NODE_272_length_33880_cov_81.5648:11307-14307	1	1	1	1	EfM3.062340.mRNA-2	EfM3.062340.mRNA-2	EfM3.062340.mRNA-2 1629	UP,DOWN	UP,DOWN	12807	C	T	up
56	NODE_289_length_31007_cov_69.0943:6595-9595	1	1	1	1	EfM3.027950.mRNA-1	EfM3.027950.mRNA-1	EfM3.027950.mRNA-1 2152	UP,DOWN	UP,DOWN	3716	T	A	IG
56	NODE_289_length_31007_cov_69.0943:6595-9595	1	1	1	1	EfM3.027950.mRNA-1		EfM3.027950.mRNA-1 6293	UP,DOWN	UP,DOWN	3716	T	A	IG
57	NODE_29_length_200305_cov_61.0043:101068-104068	1	1	1	1	EfM3.064900.mRNA-1	EfM3.064900.mRNA-1	EfM3.064900.mRNA-1 3923	UP,DOWN	UP,DOWN	102568	G	A	up
57	NODE_29_length_200305_cov_61.0043:101068-104068	1	1	1	1	EfM3.064900.mRNA-1		EfM3.064900.mRNA-1 7443.6	UP,DOWN	UP,DOWN	102568	G	A	up
57	NODE_29_length_200305_cov_61.0043:101068-104068	1	1	1	1	EfM3.064900.mRNA-1		EfM3.064900.mRNA-1 7794	UP,DOWN	UP,DOWN	102568	G	A	up
57	NODE_29_length_200305_cov_61.0043:101068-104068	1	1	1	1	EfM3.064910.mRNA-1	EfM3.064910.mRNA-1	EfM3.064910.mRNA-1 2988	UP,DOWN	UP,DOWN	102568	G	A	up
58	NODE_3_length_374172_cov_53.5127:340794-343794	1	1	1	1	EfM3.044520.mRNA-1	EfM3.044520.mRNA-1	EfM3.044520.mRNA-1 1782.3	UP,DOWN	UP,DOWN	342294	C	T	down
58	NODE_3_length_374172_cov_53.5127:340794-343794	1	1	1	1	EfM3.033720.mRNA-1		EfM3.033720.mRNA-1 1980	UP,DOWN	UP,DOWN	342294	C	T	down
58	NODE_3_length_374172_cov_53.5127:340794-343794	1	1	1	1	EfM3.033720.mRNA-1	EfM3.033720.mRNA-1	EfM3.033720.mRNA-1 1896	UP,DOWN	UP,DOWN	342294	C	T	down
58	NODE_3_length_374172_cov_53.5127:340794-343794	1	1	1	1	EfM3.044520.mRNA-1		EfM3.044520.mRNA-1 2161.4	UP,DOWN	UP,DOWN	342294	C	T	down
58	NODE_3_length_374172_cov_53.5127:340794-343794	1	1	1	1	EfM3.059200.mRNA-1	EfM3.059200.mRNA-1	EfM3.059200.mRNA-1 1825	UP,DOWN	UP,DOWN	342294	C	T	down
58	NODE_3_length_374172_cov_53.5127:340794-343794	1	1	1	1	EfM3.055400.mRNA-1	EfM3.055400.mRNA-1	EfM3.055400.mRNA-1 2065	UP,DOWN	UP,DOWN	342294	C	T	down
58	NODE_3_length_374172_cov_53.5127:340794-343794	1	1	1	1	EfM3.051540.mRNA-1	EfM3.051540.mRNA-1	EfM3.051540.mRNA-1 7940	UP,DOWN	UP,DOWN	342294	C	T	down
60	NODE_31_length_198054_cov_59.78:78484-81484	1	1	1	1	EfM3.075900.mRNA-1	EfM3.075900.mRNA-1	EfM3.075900.mRNA-1 1607	UP,DOWN	UP,DOWN	79984	C	T	syn
62	NODE_321_length_24719_cov_56.5756:12470-15470	1	1	1	1	EfM3.081760.mRNA-1		EfM3.081760.mRNA-1 12700	ORF	UP,DOWN	12733	C	T	down
64	NODE_33_length_189858_cov_52.4144:114867-117867	1	1	1	1	EfM3.038310.mRNA-1	EfM3.038310.mRNA-1	EfM3.038310.mRNA-1 11928	ORF	UP,DOWN	116367	C	T	NS

10	NODE_123_length_90057_cov_57.6174:27691-30691	1	1	2	1	1	2	EfM3.000150.mRNA-1	EfM3.000150.mRNA-1	EfM3.000150.mRNA-1 5346	ORF	UP,DOWN	29191	G	A	stop
10	NODE_123_length_90057_cov_57.6174:27691-30691	1	1	2	1	1	2	EfM3.000150.mRNA-1		EfM3.000150.mRNA-1 6540	UP,DOWN	UP,DOWN	29191	G	A	stop
10	NODE_123_length_90057_cov_57.6174:27691-30691	1	1	2	1	1	2	EfM3.000150.mRNA-1		EfM3.000150.mRNA-1 9507	UP,DOWN	UP,DOWN	29191	G	A	stop
10	NODE_123_length_90057_cov_57.6174:27691-30691	1	1	2	1	1	2	EfM3.000160.mRNA-1	EfM3.000160.mRNA-1	EfM3.000160.mRNA-1 6151.1	UP,DOWN	UP,DOWN	29191	G	A	stop
77	NODE_45_length_167596_cov_51.1068:83870-86870	1		1		1	1	EfM3.011630.mRNA-1	EfM3.011630.mRNA-1	EfM3.011630.mRNA-1 10939	ORF	ORF	85370	C	T	syn
88	NODE_60_length_150168_cov_64.7687:117591-120591	1		1		1	1	EfM3.027950.mRNA-1		EfM3.027950.mRNA-1 4435	134 bp +-	UP,DOWN	17815	T	A	IG
92	NODE_62_length_148379_cov_50.3182:73200-76200	1		1		1	1	EfM3.076860.mRNA-1	EfM3.076860.mRNA-1	EfM3.076860.mRNA-1 5388	1132 bp +-	UP,DOWN	5261	A	T	up
64	NODE_33_length_189858_cov_52.4144:114867-117867		1	1		1	1	EfM3.038310.mRNA-1		EfM3.038310.mRNA-1 12003	UP,DOWN	UP,DOWN	116367	C	T	NS
64	NODE_33_length_189858_cov_52.4144:114867-117867		1	1		1	1	EfM3.038310.mRNA-1		EfM3.038310.mRNA-1 12323	UP,DOWN	UP,DOWN	116367	C	T	NS
64	NODE_33_length_189858_cov_52.4144:114867-117867		1	1		1	1	EfM3.038310.mRNA-3	EfM3.038310.mRNA-3	EfM3.038310.mRNA-3 10516	UP,DOWN	UP,DOWN	116367	C	T	NS
3	NODE_10_length_269079_cov_54.4166:179461-182461	1	1	2	1	1	2	EfM3.004850.mRNA-1	EfM3.004850.mRNA-1	EfM3.004850.mRNA-1 12104	ORF	UP,DOWN	180961	G	A	stop
3	NODE_10_length_269079_cov_54.4166:179461-182461	1	1	2	1	1	2	EfM3.004850.mRNA-1		EfM3.004850.mRNA-1 5837	UP,DOWN	UP,DOWN	180961	G	A	stop
3	NODE_10_length_269079_cov_54.4166:179461-182461	1	1	2	1	1	2	EfM3.004850.mRNA-1		EfM3.004850.mRNA-1 6276	UP,DOWN	UP,DOWN	180961	G	A	stop
64	NODE_33_length_189858_cov_52.4144:114867-117867		1	1		1	1	EfM3.038310.mRNA-2		EfM3.038310.mRNA-2 14724	ORF	UP,DOWN	116367	C	T	NS
64	NODE_33_length_189858_cov_52.4144:114867-117867		1	1		1	1	EfM3.038310.mRNA-2	EfM3.038310.mRNA-2	EfM3.038310.mRNA-2 12443	UP,DOWN	UP,DOWN	116367	C	T	NS
64	NODE_33_length_189858_cov_52.4144:114867-117867		1	1		1	1	EfM3.038310.mRNA-3		EfM3.038310.mRNA-3 10545	ORF	UP,DOWN	116367	C	T	NS
64	NODE_33_length_189858_cov_52.4144:114867-117867		1	1		1	1	EfM3.038310.mRNA-3		EfM3.038310.mRNA-3 12003	UP,DOWN	UP,DOWN	116367	C	T	NS
64	NODE_33_length_189858_cov_52.4144:114867-117867		1	1		1	1	EfM3.038310.mRNA-3		EfM3.038310.mRNA-3 12221	UP,DOWN	UP,DOWN	116367	C	T	NS
66	NODE_347_length_21560_cov_59.4685:2948-5948		1	1		1	1	EfM3.012040.mRNA-1	EfM3.012040.mRNA-1	EfM3.012040.mRNA-1 5721	UP,DOWN	UP,DOWN	4448	C	T	up
66	NODE_347_length_21560_cov_59.4685:2948-5948		1	1		1	1	EfM3.012040.mRNA-1		EfM3.012040.mRNA-1 7493	UP,DOWN	UP,DOWN	4448	C	T	up
92	NODE_62_length_148379_cov_50.3182:73200-76200	1		1		1	1	EfM3.076870.mRNA-1	EfM3.076870.mRNA-1	EfM3.076870.mRNA-1 3426.9	62 bp +-	UP,DOWN	5261	A	T	up
67	NODE_37_length_186844_cov_52.9967:11833-14833		1	1		1	1	EfM3.047010.mRNA-1	EfM3.047010.mRNA-1	EfM3.047010.mRNA-1 6162	ORF	UP,DOWN	13333	G	A	up
67	NODE_37_length_186844_cov_52.9967:11833-14833		1	1		1	1	EfM3.047030.mRNA-1	EfM3.047030.mRNA-1	EfM3.047030.mRNA-1 7247	UP,DOWN	UP,DOWN	13333	G	A	up
67	NODE_37_length_186844_cov_52.9967:11833-14833		1	1		1	1	EfM3.047020.mRNA-1	EfM3.047020.mRNA-1	EfM3.047020.mRNA-1 7247	UP,DOWN	UP,DOWN	13333	G	A	up
68	NODE_380_length_14556_cov_62.4796:11530-14530		1	1		1	1	EfM3.052320.mRNA-1	EfM3.052320.mRNA-1	EfM3.052320.mRNA-1 3382.6	UP,DOWN	UP,DOWN	13030	C	T	IG

69	NODE_39_length_180293_cov_50.0251:161389-164389	1	1	1	1	EfM3.037650.mRNA-1	EfM3.037650.mRNA-1	EfM3.037650.mRNA-1 3974	UP,DOWN	UP,DOWN	162889	C	T	syn
69	NODE_39_length_180293_cov_50.0251:161389-164389	1	1	1	1	EfM3.037650.mRNA-1		EfM3.037650.mRNA-1 5729	UP,DOWN	UP,DOWN	162889	C	T	syn
69	NODE_39_length_180293_cov_50.0251:161389-164389	1	1	1	1	EfM3.037660.mRNA-1	EfM3.037660.mRNA-1	EfM3.037660.mRNA-1 5387	UP,DOWN	UP,DOWN	162889	C	T	syn
69	NODE_39_length_180293_cov_50.0251:161389-164389	1	1	1	1	EfM3.037660.mRNA-1		EfM3.037660.mRNA-1 5550	UP,DOWN	UP,DOWN	162889	C	T	syn
69	NODE_39_length_180293_cov_50.0251:161389-164389	1	1	1	1	EfM3.037660.mRNA-1		EfM3.037660.mRNA-1 6003	UP,DOWN	UP,DOWN	162889	C	T	syn
69	NODE_39_length_180293_cov_50.0251:161389-164389	1	1	1	1	EfM3.037660.mRNA-1		EfM3.037660.mRNA-1 6113	UP,DOWN	UP,DOWN	162889	C	T	syn
69	NODE_39_length_180293_cov_50.0251:161389-164389	1	1	1	1	EfM3.037660.mRNA-1		EfM3.037660.mRNA-1 6538	UP,DOWN	UP,DOWN	162889	C	T	syn
69	NODE_39_length_180293_cov_50.0251:161389-164389	1	1	1	1	EfM3.037660.mRNA-1		EfM3.037660.mRNA-1 6984	ORF	UP,DOWN	162889	C	T	syn
70	NODE_4_length_349919_cov_62.769:156537-159537	1	1	1	1	EfM3.026150.mRNA-1	EfM3.026150.mRNA-1	EfM3.026150.mRNA-1 6659	UP,DOWN	UP,DOWN	79819	C	T	up
70	NODE_4_length_349919_cov_62.769:156537-159537	1	1	1	1	EfM3.026150.mRNA-1		EfM3.026150.mRNA-1 7846	ORF	UP,DOWN	79819	C	T	up
70	NODE_4_length_349919_cov_62.769:156537-159537	1	1	1	1	EfM3.026150.mRNA-1		EfM3.026150.mRNA-1 8577	UP,DOWN	UP,DOWN	79819	C	T	up
70	NODE_4_length_349919_cov_62.769:156537-159537	1	1	1	1	EfM3.026150.mRNA-1		EfM3.026150.mRNA-1 8929	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319	1	1	1	1	EfM3.013240.mRNA-2	EfM3.013240.mRNA-2	EfM3.013240.mRNA-2 2565.1	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319	1	1	1	1	EfM3.013230.mRNA-1		EfM3.013230.mRNA-1 6091	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319	1	1	1	1	EfM3.013230.mRNA-1		EfM3.013230.mRNA-1 8466	UP,DOWN	UP,DOWN	79819	C	T	up
59	NODE_31_length_198054_cov_59.78:188245-191245	1		1	1	1	EfM3.028530.mRNA-1	EfM3.028530.mRNA-1 1896	1040 bp +-	UP,DOWN	79984	C	T	syn
59	NODE_31_length_198054_cov_59.78:188245-191245	1		1	1	1	EfM3.064570.mRNA-1	EfM3.064570.mRNA-1 1879	1044 bp +-	UP,DOWN	79984	C	T	syn
59	NODE_31_length_198054_cov_59.78:188245-191245	1		1	1	1	EfM3.067010.mRNA-1	EfM3.067010.mRNA-1 1918	1040 bp +-	UP,DOWN	79984	C	T	syn
59	NODE_31_length_198054_cov_59.78:188245-191245	1		1	1	1	EfM3.067010.mRNA-1	EfM3.067010.mRNA-1 2318	679 bp +-	UP,DOWN	79984	C	T	syn
71	NODE_4_length_349919_cov_62.769:78319-81319	1	1	1	1	EfM3.013240.mRNA-3	EfM3.013240.mRNA-3	EfM3.013240.mRNA-3 2892	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319	1	1	1	1	EfM3.013230.mRNA-2		EfM3.013230.mRNA-2 6091	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319	1	1	1	1	EfM3.013230.mRNA-2		EfM3.013230.mRNA-2 8466	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319	1	1	1	1	EfM3.013230.mRNA-1	EfM3.013230.mRNA-1	EfM3.013230.mRNA-1 4422	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319	1	1	1	1	EfM3.013240.mRNA-1		EfM3.013240.mRNA-1 5170	ORF	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319	1	1	1	1	EfM3.013240.mRNA-1		EfM3.013240.mRNA-1 5272	UP,DOWN	UP,DOWN	79819	C	T	up

71	NODE_4_length_349919_cov_62.769:78319-81319		1	1		1	1	EfM3.013230.mRNA-2	EfM3.013230.mRNA-2	EfM3.013230.mRNA-2 4422	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319		1	1		1	1	EfM3.013240.mRNA-2		EfM3.013240.mRNA-2 2604.5	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319		1	1		1	1	EfM3.013240.mRNA-2		EfM3.013240.mRNA-2 3229	UP,DOWN	UP,DOWN	79819	C	T	up
92	NODE_62_length_148379_cov_50.3182:73200-76200	1		1		1	1	EfM3.076870.mRNA-1		EfM3.076870.mRNA-1 3538	41 bp +-	UP,DOWN	5261	A	T	up
92	NODE_62_length_148379_cov_50.3182:73200-76200	1		1		1	1	EfM3.076870.mRNA-1		EfM3.076870.mRNA-1 3548.8	187 bp +-	UP,DOWN	5261	A	T	up
92	NODE_62_length_148379_cov_50.3182:73200-76200	1		1		1	1	EfM3.076870.mRNA-1		EfM3.076870.mRNA-1 3862	382 bp +-	UP,DOWN	5261	A	T	up
10 1	NODE_70_length_137115_cov_55.8162:27530-30530	1	1	2		1	1	EfM3.058290.mRNA-1	EfM3.058290.mRNA-1	EfM3.058290.mRNA-1 9078	539 bp +-	UP,DOWN	29030	C	T	IG
10 1	NODE_70_length_137115_cov_55.8162:27530-30530	1	1	2		1	1	EfM3.010310.mRNA-1	EfM3.010310.mRNA-1	EfM3.010310.mRNA-1 4986	539 bp +-	UP,DOWN	29030	C	T	IG
10 1	NODE_70_length_137115_cov_55.8162:27530-30530	1	1	2		1	1	EfM3.063400.mRNA-1	EfM3.063400.mRNA-1	EfM3.063400.mRNA-1 7975	621 bp +-	UP,DOWN	29030	C	T	IG
71	NODE_4_length_349919_cov_62.769:78319-81319		1	1		1	1	EfM3.013240.mRNA-2		EfM3.013240.mRNA-2 4210	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319		1	1		1	1	EfM3.013240.mRNA-1	EfM3.013240.mRNA-1	EfM3.013240.mRNA-1 4286	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319		1	1		1	1	EfM3.013240.mRNA-3		EfM3.013240.mRNA-3 3331	UP,DOWN	UP,DOWN	79819	C	T	up
71	NODE_4_length_349919_cov_62.769:78319-81319		1	1		1	1	EfM3.013240.mRNA-3		EfM3.013240.mRNA-3 4312	UP,DOWN	UP,DOWN	79819	C	T	up
72	NODE_40_length_173182_cov_58.9455:31946-34946	1	1		1	1		EfM3.010560.mRNA-1	EfM3.010560.mRNA-1	EfM3.010560.mRNA-1 11039	ORF	UP,DOWN	33446	C	T	NS
72	NODE_40_length_173182_cov_58.9455:31946-34946	1	1		1	1		EfM3.010560.mRNA-1		EfM3.010560.mRNA-1 3389	UP,DOWN	UP,DOWN	33446	C	T	NS
73	NODE_407_length_11509_cov_68.1964:1665-4665	1	1		1	1		EfM3.027950.mRNA-1		EfM3.027950.mRNA-1 5718	ORF	UP,DOWN	3165	G	T	IG
74	NODE_41_length_172435_cov_51.8826:153050-156050	1	1		1	1		EfM3.065740.mRNA-1	EfM3.065740.mRNA-1	EfM3.065740.mRNA-1 3663.2	UP,DOWN	UP,DOWN	99914	T	A	IG
74	NODE_41_length_172435_cov_51.8826:153050-156050	1	1		1	1		EfM3.019330.mRNA-1		EfM3.019330.mRNA-1 1796.6	UP,DOWN	UP,DOWN	99914	T	A	IG
74	NODE_41_length_172435_cov_51.8826:153050-156050	1	1		1	1		EfM3.019330.mRNA-1	EfM3.019330.mRNA-1	EfM3.019330.mRNA-1 1587.3	UP,DOWN	UP,DOWN	99914	T	A	IG
74	NODE_41_length_172435_cov_51.8826:153050-156050	1	1		1	1		EfM3.034690.mRNA-1		EfM3.034690.mRNA-1 6601	ORF	UP,DOWN	99914	T	A	IG
74	NODE_41_length_172435_cov_51.8826:153050-156050	1	1		1	1		EfM3.034690.mRNA-1		EfM3.034690.mRNA-1 6926	UP,DOWN	UP,DOWN	99914	T	A	IG
74	NODE_41_length_172435_cov_51.8826:153050-156050	1	1		1	1		EfM3.075020.mRNA-1	EfM3.075020.mRNA-1	EfM3.075020.mRNA-1 1843.3	UP,DOWN	UP,DOWN	99914	T	A	IG
74	NODE_41_length_172435_cov_51.8826:153050-156050	1	1		1	1		EfM3.045940.mRNA-1	EfM3.045940.mRNA-1	EfM3.045940.mRNA-1 2054	UP,DOWN	UP,DOWN	99914	T	A	IG
74	NODE_41_length_172435_cov_51.8826:153050-156050	1	1		1	1		EfM3.034700.mRNA-1	EfM3.034700.mRNA-1	EfM3.034700.mRNA-1 12916	UP,DOWN	UP,DOWN	99914	T	A	IG
74	NODE_41_length_172435_cov_51.8826:153050-156050	1	1		1	1		EfM3.034690.mRNA-1	EfM3.034690.mRNA-1	EfM3.034690.mRNA-1 2513	UP,DOWN	UP,DOWN	99914	T	A	IG

75	NODE_44_length_169116_cov_55.4188:89800-92800	1	1	1	1	EfM3.008570.mRNA-1	EfM3.008570.mRNA-1	EfM3.008570.mRNA-1 7935	ORF	UP,DOWN	91300	G	A	NS
75	NODE_44_length_169116_cov_55.4188:89800-92800	1	1	1	1	EfM3.008560.mRNA-1	EfM3.008560.mRNA-1	EfM3.008560.mRNA-1 4526	UP,DOWN	UP,DOWN	91300	G	A	NS
75	NODE_44_length_169116_cov_55.4188:89800-92800	1	1	1	1	EfM3.008580.mRNA-1	EfM3.008580.mRNA-1	EfM3.008580.mRNA-1 7742	UP,DOWN	UP,DOWN	91300	G	A	NS
78	NODE_45_length_167596_cov_51.1068:87374-90374	1	1	1	1	EfM3.011620.mRNA-1	EfM3.011620.mRNA-1	EfM3.011620.mRNA-1 12320	ORF	UP,DOWN	85370	C	T	syn
78	NODE_45_length_167596_cov_51.1068:87374-90374	1	1	1	1	EfM3.011620.mRNA-1		EfM3.011620.mRNA-1 7469	UP,DOWN	UP,DOWN	85370	C	T	syn
78	NODE_45_length_167596_cov_51.1068:87374-90374	1	1	1	1	EfM3.048310.mRNA-1	EfM3.048310.mRNA-1	EfM3.048310.mRNA-1 2040	UP,DOWN	UP,DOWN	85370	C	T	syn
79	NODE_486_length_4771_cov_73.2313:2146-4771	1	1	1	1	EfM3.027950.mRNA-1		EfM3.027950.mRNA-1 7120	UP,DOWN	UP,DOWN	3646	A	G	IG
80	NODE_49_length_165114_cov_56.6826:35043-38043	1	1	1	1	EfM3.082750.mRNA-1	EfM3.082750.mRNA-1	EfM3.082750.mRNA-1 12364	UP,DOWN	UP,DOWN	36543	G	A	splice
80	NODE_49_length_165114_cov_56.6826:35043-38043	1	1	1	1	EfM3.082740.mRNA-1	EfM3.082740.mRNA-1	EfM3.082740.mRNA-1 7945	UP,DOWN	UP,DOWN	36543	G	A	splice
81	NODE_50_length_163666_cov_60.5291:37717-40717	1	1	1	1	EfM3.016550.mRNA-1	EfM3.016550.mRNA-1	EfM3.016550.mRNA-1 8547	ORF	UP,DOWN	39217	C	A	NS
81	NODE_50_length_163666_cov_60.5291:37717-40717	1	1	1	1	EfM3.016560.mRNA-1	EfM3.016560.mRNA-1	EfM3.016560.mRNA-1 11076	UP,DOWN	UP,DOWN	39217	C	A	NS
81	NODE_50_length_163666_cov_60.5291:37717-40717	1	1	1	1	EfM3.016560.mRNA-2	EfM3.016560.mRNA-2	EfM3.016560.mRNA-2 10781	UP,DOWN	UP,DOWN	39217	C	A	NS
81	NODE_50_length_163666_cov_60.5291:37717-40717	1	1	1	1	EfM3.016540.mRNA-1	EfM3.016540.mRNA-1	EfM3.016540.mRNA-1 6372	UP,DOWN	UP,DOWN	39217	C	A	NS
82	NODE_52_length_160413_cov_62.0488:45353-48353	1	1	1	1	EfM3.034510.mRNA-1	EfM3.034510.mRNA-1	EfM3.034510.mRNA-1 11710	ORF	UP,DOWN	46853	T	A	syn
82	NODE_52_length_160413_cov_62.0488:45353-48353	1	1	1	1	EfM3.034520.mRNA-1	EfM3.034520.mRNA-1	EfM3.034520.mRNA-1 7287	UP,DOWN	UP,DOWN	46853	T	A	syn
82	NODE_52_length_160413_cov_62.0488:45353-48353	1	1	1	1	EfM3.034520.mRNA-1		EfM3.034520.mRNA-1 8008	UP,DOWN	UP,DOWN	46853	T	A	syn
83	NODE_54_length_157999_cov_52.6672:114344-117344	1	1	1	1	EfM3.022910.mRNA-1	EfM3.022910.mRNA-1	EfM3.022910.mRNA-1 3887	UP,DOWN	UP,DOWN	115844	C	T	NS
83	NODE_54_length_157999_cov_52.6672:114344-117344	1	1	1	1	EfM3.022900.mRNA-1		EfM3.022900.mRNA-1 8639	ORF	UP,DOWN	115844	C	T	NS
83	NODE_54_length_157999_cov_52.6672:114344-117344	1	1	1	1	EfM3.022900.mRNA-1	EfM3.022900.mRNA-1	EfM3.022900.mRNA-1 7969	UP,DOWN	UP,DOWN	115844	C	T	NS
84	NODE_54_length_157999_cov_52.6672:133830-136830	1	1	1	1	EfM3.022860.mRNA-1	EfM3.022860.mRNA-1	EfM3.022860.mRNA-1 19838	UP,DOWN	UP,DOWN	115844	C	T	NS
84	NODE_54_length_157999_cov_52.6672:133830-136830	1	1	1	1	EfM3.022860.mRNA-1		EfM3.022860.mRNA-1 21601	ORF	UP,DOWN	115844	C	T	NS
85	NODE_55_length_155895_cov_57.0108:15582-18582	1	1	1	1	EfM3.035050.mRNA-1	EfM3.035050.mRNA-1	EfM3.035050.mRNA-1 9485	UP,DOWN	UP,DOWN	17082	T	A	up
86	NODE_56_length_155850_cov_245.381:43885-46885	1	1	1	1	EfM3.070750.mRNA-1	EfM3.070750.mRNA-1	EfM3.070750.mRNA-1 2625	UP,DOWN	UP,DOWN	45385	C	T	up
86	NODE_56_length_155850_cov_245.381:43885-46885	1	1	1	1	EfM3.070750.mRNA-1		EfM3.070750.mRNA-1 2921.1	UP,DOWN	UP,DOWN	45385	C	T	up
86	NODE_56_length_155850_cov_245.381:43885-46885	1	1	1	1	EfM3.070750.mRNA-1		EfM3.070750.mRNA-1 5994	ORF	UP,DOWN	45385	C	T	up

87	NODE_57_length_154763_cov_54.4333:6160-9160	1	1	1	1	EfM3.066230.mRNA-1		EfM3.066230.mRNA-1 2768	UP,DOWN	UP,DOWN	7660	A	G	IG
89	NODE_60_length_150168_cov_64.7687:16315-19315	1	1	1	1	EfM3.066230.mRNA-1	EfM3.066230.mRNA-1	EfM3.066230.mRNA-1 1835	UP,DOWN	UP,DOWN	17815	T	A	IG
90	NODE_61_length_148900_cov_61.3899:63596-66596	1	1	1	1	EfM3.014570.mRNA-1	EfM3.014570.mRNA-1	EfM3.014570.mRNA-1 13186	UP,DOWN	UP,DOWN	65096	G	T	NS
90	NODE_61_length_148900_cov_61.3899:63596-66596	1	1	1	1	EfM3.014570.mRNA-1		EfM3.014570.mRNA-1 14171	UP,DOWN	UP,DOWN	65096	G	T	NS
90	NODE_61_length_148900_cov_61.3899:63596-66596	1	1	1	1	EfM3.014570.mRNA-1		EfM3.014570.mRNA-1 17337	ORF	UP,DOWN	65096	G	T	NS
91	NODE_62_length_148379_cov_50.3182:3761-6761	1	1	1	1	EfM3.076650.mRNA-1	EfM3.076650.mRNA-1	EfM3.076650.mRNA-1 1780	UP,DOWN	UP,DOWN	5261	A	T	up
91	NODE_62_length_148379_cov_50.3182:3761-6761	1	1	1	1	EfM3.076650.mRNA-1		EfM3.076650.mRNA-1 2902	UP,DOWN	UP,DOWN	5261	A	T	up
91	NODE_62_length_148379_cov_50.3182:3761-6761	1	1	1	1	EfM3.076660.mRNA-1	EfM3.076660.mRNA-1	EfM3.076660.mRNA-1 7908	UP,DOWN	UP,DOWN	5261	A	T	up
93	NODE_63_length_147872_cov_68.758:52760-55760	1	1	1	1	EfM3.022030.mRNA-1	EfM3.022030.mRNA-1	EfM3.022030.mRNA-1 596.3	UP,DOWN	UP,DOWN	54260	T	A	stop
93	NODE_63_length_147872_cov_68.758:52760-55760	1	1	1	1	EfM3.022030.mRNA-1		EfM3.022030.mRNA-1 8810	ORF	UP,DOWN	54260	T	A	stop
93	NODE_63_length_147872_cov_68.758:52760-55760	1	1	1	1	EfM3.022030.mRNA-1		EfM3.022030.mRNA-1 9416	UP,DOWN	UP,DOWN	54260	T	A	stop
93	NODE_63_length_147872_cov_68.758:52760-55760	1	1	1	1	EfM3.022040.mRNA-1	EfM3.022040.mRNA-1	EfM3.022040.mRNA-1 11477	UP,DOWN	UP,DOWN	54260	T	A	stop
94	NODE_64_length_146070_cov_52.0445:35222-38222	1	1	1	1	EfM3.057280.mRNA-1	EfM3.057280.mRNA-1	EfM3.057280.mRNA-1 5280	UP,DOWN	UP,DOWN	36722	C	T	NS
94	NODE_64_length_146070_cov_52.0445:35222-38222	1	1	1	1	EfM3.057280.mRNA-1		EfM3.057280.mRNA-1 6995	ORF	UP,DOWN	36722	C	T	NS
95	NODE_65_length_143904_cov_51.3299:12705-15705	1	1	1	1	EfM3.049500.mRNA-1	EfM3.049500.mRNA-1	EfM3.049500.mRNA-1 10791	ORF	UP,DOWN	14205	G	A	NS
10 1	NODE_70_length_137115_cov_55.8162:27530-30530	1	1	2	1	1 EfM3.026630.mRNA-1		EfM3.026630.mRNA-1 4229	120 bp +-	UP,DOWN	29030	C	T	IG
10 1	NODE_70_length_137115_cov_55.8162:27530-30530	1	1	2	1	1 EfM3.063400.mRNA-2	EfM3.063400.mRNA-2	EfM3.063400.mRNA-2 7867	621 bp +-	UP,DOWN	29030	C	T	IG
10 1	NODE_70_length_137115_cov_55.8162:27530-30530	1	1	2	1	1 EfM3.023690.mRNA-1	EfM3.023690.mRNA-1	EfM3.023690.mRNA-1 3523	ORF	ORF	29030	C	T	IG
10 1	NODE_70_length_137115_cov_55.8162:27530-30530	1	1	2	1	1 EfM3.026630.mRNA-1	EfM3.026630.mRNA-1	EfM3.026630.mRNA-1 3565	ORF	ORF	29030	C	T	IG
10 6	NODE_83_length_116007_cov_58.7017:63830-66830	1		1	1	1 EfM3.028630.mRNA-1	EfM3.028630.mRNA-1	EfM3.028630.mRNA-1 7310	1317 bp +-	UP,DOWN	22150	A	T	IG
95	NODE_65_length_143904_cov_51.3299:12705-15705	1	1	1	1	EfM3.021840.mRNA-1		EfM3.021840.mRNA-1 6792.8	UP,DOWN	UP,DOWN	14205	G	A	NS
95	NODE_65_length_143904_cov_51.3299:12705-15705	1	1	1	1	EfM3.021840.mRNA-1		EfM3.021840.mRNA-1 6809.7	UP,DOWN	UP,DOWN	14205	G	A	NS
95	NODE_65_length_143904_cov_51.3299:12705-15705	1	1	1	1	EfM3.021840.mRNA-1	EfM3.021840.mRNA-1	EfM3.021840.mRNA-1 6786.8	UP,DOWN	UP,DOWN	14205	G	A	NS
95	NODE_65_length_143904_cov_51.3299:12705-15705	1	1	1	1	EfM3.049510.mRNA-1	EfM3.049510.mRNA-1	EfM3.049510.mRNA-1 2004	UP,DOWN	UP,DOWN	14205	G	A	NS
95	NODE_65_length_143904_cov_51.3299:12705-15705	1	1	1	1	EfM3.049510.mRNA-1		EfM3.049510.mRNA-1 2072	UP,DOWN	UP,DOWN	14205	G	A	NS

95	NODE_65_length_143904_cov_51.3299:12705-15705	1	1	1	1	EfM3.049510.mRNA-1		EfM3.049510.mRNA-1 2288	UP,DOWN	UP,DOWN	14205	G	A	NS
96	NODE_66_length_142472_cov_57.5482:131721-134721	1	1	1	1	EfM3.065190.mRNA-1	EfM3.065190.mRNA-1	EfM3.065190.mRNA-1 10767	ORF	UP,DOWN	133221	T	A	up
96	NODE_66_length_142472_cov_57.5482:131721-134721	1	1	1	1	EfM3.065200.mRNA-1	EfM3.065200.mRNA-1	EfM3.065200.mRNA-1 5172	UP,DOWN	UP,DOWN	133221	T	A	up
96	NODE_66_length_142472_cov_57.5482:131721-134721	1	1	1	1	EfM3.065200.mRNA-1		EfM3.065200.mRNA-1 5469	UP,DOWN	UP,DOWN	133221	T	A	up
97	NODE_68_length_142066_cov_53.4584:6682-9682	1	1	1	1	EfM3.010940.mRNA-1	EfM3.010940.mRNA-1	EfM3.010940.mRNA-1 5446	UP,DOWN	UP,DOWN	8182	A	G	NS
97	NODE_68_length_142066_cov_53.4584:6682-9682	1	1	1	1	EfM3.010940.mRNA-1		EfM3.010940.mRNA-1 6026	UP,DOWN	UP,DOWN	8182	A	G	NS
97	NODE_68_length_142066_cov_53.4584:6682-9682	1	1	1	1	EfM3.010940.mRNA-1		EfM3.010940.mRNA-1 7169	UP,DOWN	UP,DOWN	8182	A	G	NS
97	NODE_68_length_142066_cov_53.4584:6682-9682	1	1	1	1	EfM3.010940.mRNA-1		EfM3.010940.mRNA-1 7480	ORF	UP,DOWN	8182	A	G	NS
98	NODE_7_length_312897_cov_54.6218:309087-312087	1	1	1	1	EfM3.071150.mRNA-1	EfM3.071150.mRNA-1	EfM3.071150.mRNA-1 4076.8	UP,DOWN	UP,DOWN	60329	G	T	up
98	NODE_7_length_312897_cov_54.6218:309087-312087	1	1	1	1	EfM3.071150.mRNA-1		EfM3.071150.mRNA-1 7883	UP,DOWN	UP,DOWN	60329	G	T	up
98	NODE_7_length_312897_cov_54.6218:309087-312087	1	1	1	1	EfM3.071160.mRNA-1	EfM3.071160.mRNA-1	EfM3.071160.mRNA-1 6758	UP,DOWN	UP,DOWN	60329	G	T	up
98	NODE_7_length_312897_cov_54.6218:309087-312087	1	1	1	1	EfM3.071160.mRNA-1		EfM3.071160.mRNA-1 6979	UP,DOWN	UP,DOWN	60329	G	T	up
99	NODE_7_length_312897_cov_54.6218:58829-61829	1	1	1	1	EfM3.064010.mRNA-1	EfM3.064010.mRNA-1	EfM3.064010.mRNA-1 4523	UP,DOWN	UP,DOWN	60329	G	T	up
99	NODE_7_length_312897_cov_54.6218:58829-61829	1	1	1	1	EfM3.064000.mRNA-1	EfM3.064000.mRNA-1	EfM3.064000.mRNA-1 6375	UP,DOWN	UP,DOWN	60329	G	T	up
99	NODE_7_length_312897_cov_54.6218:58829-61829	1	1	1	1	EfM3.063990.mRNA-1	EfM3.063990.mRNA-1	EfM3.063990.mRNA-1 10091	UP,DOWN	UP,DOWN	60329	G	T	up
100	NODE_7_length_312897_cov_54.6218:81722-84722	1	1	1	1	EfM3.063880.mRNA-1	EfM3.063880.mRNA-1	EfM3.063880.mRNA-1 6903	UP,DOWN	UP,DOWN	60329	G	T	up
100	NODE_7_length_312897_cov_54.6218:81722-84722	1	1	1	1	EfM3.063880.mRNA-1		EfM3.063880.mRNA-1 8931	UP,DOWN	UP,DOWN	60329	G	T	up
100	NODE_7_length_312897_cov_54.6218:81722-84722	1	1	1	1	EfM3.063900.mRNA-1	EfM3.063900.mRNA-1	EfM3.063900.mRNA-1 2888	UP,DOWN	UP,DOWN	60329	G	T	up
100	NODE_7_length_312897_cov_54.6218:81722-84722	1	1	1	1	EfM3.063890.mRNA-1		EfM3.063890.mRNA-1 8437	UP,DOWN	UP,DOWN	60329	G	T	up
100	NODE_7_length_312897_cov_54.6218:81722-84722	1	1	1	1	EfM3.063890.mRNA-1	EfM3.063890.mRNA-1	EfM3.063890.mRNA-1 5043	UP,DOWN	UP,DOWN	60329	G	T	up
100	NODE_7_length_312897_cov_54.6218:81722-84722	1	1	1	1	EfM3.063910.mRNA-1	EfM3.063910.mRNA-1	EfM3.063910.mRNA-1 3187	UP,DOWN	UP,DOWN	60329	G	T	up
106	NODE_83_length_116007_cov_58.7017:63830-66830	1	1	1	1	EfM3.028640.mRNA-1	EfM3.028640.mRNA-1	EfM3.028640.mRNA-1 9575	ORF	ORF	22150	A	T	IG
107	NODE_84_length_115316_cov_51.0035:95831-98831	1	1	1	1	EfM3.065810.mRNA-1	EfM3.065810.mRNA-1	EfM3.065810.mRNA-1 496	1398 bp +-	UP,DOWN	97331	C	T	IG
107	NODE_84_length_115316_cov_51.0035:95831-98831	1	1	1	1	EfM3.065810.mRNA-1		EfM3.065810.mRNA-1 568.5	1300 bp +-	UP,DOWN	97331	C	T	IG
107	NODE_84_length_115316_cov_51.0035:95831-98831	1	1	1	1	EfM3.065830.mRNA-1	EfM3.065830.mRNA-1	EfM3.065830.mRNA-1 6208	1429 bp +-	UP,DOWN	97331	C	T	IG

10 7	NODE_84_length_115316_cov_51.0035 :95831-98831	1	1	1	1	EfM3.065820. mRNA-1		EfM3.065820.mRNA- 1 5182	570 bp +-	UP,DOW N	97331	C	T	IG
10 7	NODE_84_length_115316_cov_51.0035 :95831-98831	1	1	1	1	EfM3.065820. mRNA-1		EfM3.065820.mRNA- 1 5692	738 bp +-	UP,DOW N	97331	C	T	IG
10 7	NODE_84_length_115316_cov_51.0035 :95831-98831	1	1	1	1	EfM3.065820. mRNA-1	EfM3.065820.mR NA-1	EfM3.065820.mRNA- 1 4928	362 bp +-	UP,DOW N	97331	C	T	IG
10 2	NODE_70_length_137115_cov_55.8162 :95629-98629	1	1	1	1	EfM3.071530. mRNA-1	EfM3.071530.mR NA-1	EfM3.071530.mRNA- 1 3916	UP,DOW N	UP,DOW N	29030	C	T	IG
10 2	NODE_70_length_137115_cov_55.8162 :95629-98629	1	1	1	1	EfM3.071530. mRNA-1		EfM3.071530.mRNA- 1 5739	UP,DOW N	UP,DOW N	29030	C	T	IG
10 3	NODE_73_length_128172_cov_60.9869 :116337-119337	1	1	1	1	EfM3.036960. mRNA-1	EfM3.036960.mR NA-1	EfM3.036960.mRNA- 1 5166	UP,DOW N	UP,DOW N	117837	G	A	up
10 3	NODE_73_length_128172_cov_60.9869 :116337-119337	1	1	1	1	EfM3.036950. mRNA-1	EfM3.036950.mR NA-1	EfM3.036950.mRNA- 1 5067	UP,DOW N	UP,DOW N	117837	G	A	up
10 3	NODE_73_length_128172_cov_60.9869 :116337-119337	1	1	1	1	EfM3.036960. mRNA-1		EfM3.036960.mRNA- 1 5741	UP,DOW N	UP,DOW N	117837	G	A	up
10 3	NODE_73_length_128172_cov_60.9869 :116337-119337	1	1	1	1	EfM3.036960. mRNA-1		EfM3.036960.mRNA- 1 8369	UP,DOW N	UP,DOW N	117837	G	A	up
10 4	NODE_76_length_122656_cov_51.3915 :57168-60168	1	1	1	1	EfM3.032280. mRNA-1	EfM3.032280.mR NA-1	EfM3.032280.mRNA- 1 5345	ORF	UP,DOW N	58668	G	A	up
10 4	NODE_76_length_122656_cov_51.3915 :57168-60168	1	1	1	1	EfM3.032280. mRNA-1		EfM3.032280.mRNA- 1 6554	UP,DOW N	UP,DOW N	58668	G	A	up
10 4	NODE_76_length_122656_cov_51.3915 :57168-60168	1	1	1	1	EfM3.032290. mRNA-1	EfM3.032290.mR NA-1	EfM3.032290.mRNA- 1 6176	UP,DOW N	UP,DOW N	58668	G	A	up
10 5	NODE_8_length_282811_cov_53.518:1 68834-171834	1	1	1	1	EfM3.073210. mRNA-1	EfM3.073210.mR NA-1	EfM3.073210.mRNA- 1 9270	UP,DOW N	UP,DOW N	170334	T	C	NS
10 5	NODE_8_length_282811_cov_53.518:1 68834-171834	1	1	1	1	EfM3.073220. mRNA-1	EfM3.073220.mR NA-1	EfM3.073220.mRNA- 1 6178	UP,DOW N	UP,DOW N	170334	T	C	NS
10 5	NODE_8_length_282811_cov_53.518:1 68834-171834	1	1	1	1	EfM3.073220. mRNA-1		EfM3.073220.mRNA- 1 9096	ORF	UP,DOW N	170334	T	C	NS
10 8	NODE_87_length_113559_cov_49.1042 :43962-46962	1	1	1	1	EfM3.045900. mRNA-1	EfM3.045900.mR NA-1	EfM3.045900.mRNA- 1 5893	UP,DOW N	UP,DOW N	45462	G	A	up
10 8	NODE_87_length_113559_cov_49.1042 :43962-46962	1	1	1	1	EfM3.045900. mRNA-1		EfM3.045900.mRNA- 1 7356	UP,DOW N	UP,DOW N	45462	G	A	up
10 9	NODE_94_length_109506_cov_54.4607 :2670-5670	1	1	1	1	EfM3.082360. mRNA-1	EfM3.082360.mR NA-1	EfM3.082360.mRNA- 1 11469	ORF	UP,DOW N	4170	C	T	syn

IG: intergenic
NS: non-synonymous
Syn: synonymous
Up: upstream
Down: downstream