2014

Accelerated domestication of Australian native grass species using molecular tools

Sylvia Malory

Southern Cross University

Publication details

Copyright S Malory 2014
Accelerated domestication of Australian native grass species using molecular tools

Sylvia MALORY
Bachelor of Horticulture (Honours)
Graduate Certificate in Resource Management

Thesis submitted to fulfill the requirements of Doctor of Philosophy
Southern Cross Plant Science,
Southern Cross University

19th March, 2014
Thesis Declaration

I certify that the work presented in this thesis is, to the best of my knowledge and belief, original, except as acknowledged in the text, and that the material has not been submitted, either in whole or in part, for a degree at this or any other university.

I acknowledge that I have read and understood the University's rules, requirements, procedures and policy relating to my higher degree research award and to my thesis. I certify that I have complied with the rules, requirements, procedures and policy of the University (as they may be from time to time).

Sylvia Malory

Signed ………………………………………………… Date ………………………………...
Abstract

*Microlaena stipoides*, commonly known as weeping grass, is a distant relative of rice (*Oryza sativa*). It is a drought, frost and shade tolerant perennial evergreen plant and produces seeds similar to rice. *M. stipoides* can be used as a crop plant and during the non-cropping season for pasture. This species responds well to nitrogen application and also regular irrigation, making commercial production possible and making it a target for domestication. Extensive sequencing and comparative mapping has established a high degree of conservation between rice and other grasses, allowing the isolation of the corresponding homologues of important rice genes in other grasses. In this study the rice genome sequence is being used for comparison with corresponding genes in *M. stipoides*. A high range of variability occurs within the natural populations for domestication traits and this can be harnessed for selective breeding programs. An induced mutation population has also been established to capture desired traits for crop improvement. Whole genome sequence data for *M. stipoides* was generated by Illumina Genome Analyzer II (GAII) and assembled by reference to the genome sequence of domesticated rice to generate species specific polymerase chain reaction (PCR) primers for putative BADH2, GW2 and Hd6 homologues in *M. stipoides*. Pooled PCR amplicons for these homologues generated from both chemically induced mutant and naturally occurring populations were then screened using the Illumina GAII to discover single nucleotide polymorphisms (SNPs). The SNPs generated by the Illumina GAII were then validated by Sequenom iPlex MassARRAY (SPM). Potentially useful SNPs can be used in establishing new breeding lines of *M. stipoides* suitable for domestication. Once domesticated, *M. stipoides* may become a new crop for commercial food production. This technique can also be utilised for other wild grasses to screen for desirable domestication traits and possibly to create new crops for food consumption.
Publications Arising from this Research
(at time of submission)

PEER REVIEW JOURNAL PUBLICATIONS:

CONFERENCE PRESENTATIONS (ORAL AND POSTER):
Accelerated domestication of Australian native grass species using molecular tools


I would like to express my gratitude to all my supervisors; Prof Graham King, Prof Robert Henry, Dr Frances Shapter, Dr Ian Chivers, and Dr Martin Elphinstone for their advice, support and patience throughout my study. Thank you to Dr Agnelo Furtado, who has been my mentor in my academic, professional and personal life.

Thank you to all the students and staff that I have met through my journey at Southern Cross Plant Science, especially at the Centre for Plant Conservation Genetics. I appreciate the friendship, training and support that have been provided to me through the good and bad times. Special mentions to Dr Timothy Sexton, Dr Shabana Kasem, Dr Julie Pattemore, Gary Abblett, Michael Cross, Dr Myralyn Abasolo, Allison Wiseman, Dr Dan Waters, Dr Cathy Nock, and Tiffeny Byrnes for the discussions, advice, motivation, and support that have been given to me.

I would like to thank my husband, extended family and close friends for their relentless encouragement, unwavering support and trust, and keeping my life in balance during the ups and downs. It was a long and sometimes lonely journey and meeting the people from the Northern Rivers Indonesian community and the International Multicultural Club kept me motivated; for that I am very grateful.

My big thank you to the Australian Research Council and Native Seeds Pty Ltd for their financial support and learning opportunities provided throughout my candidature.

Finally, thank you in advance for the anonymous examiners for spending their time reading, marking and providing critiques to this thesis.
Preamble

Since the domestication of plants thousands of years ago, farmers and plant breeders have been exploiting the natural variation present in crop plants by genetically manipulating developmental traits and physiological features related to adaptation to agriculture (Frary and Doganlar 2003; Alonso-Blanco, Aarts et al. 2009). Phenotypic selection has focused on target traits such as ease of cultivation, high yield, pest and disease resistance and high nutrient content, to develop crops with increasing productivity. Ultimately, plant domestication results in crop plants that look and behave significantly differently from their wild ancestors (Hancock 2005).

The grass family Poaceae contains over 10,000 species within 700 genera (Kellogg 1998; Grass Phylogeny Working Group 2001; Kellogg 2001). Poaceae is an important monocotyledon plant family for humans as it contains all the cereal crops including rice (Oryza spp.), wheat (Triticum spp.), maize (Zea spp.), barley (Hordeum spp.), oats (Avena spp.) and sorghum (Sorghum spp.), representing most of the world’s major food sources. Domestication of cereals began about 7,000 to 10,000 years ago (Harlan 1975; Gepts 2004; Doebley 2006). The most important domestication changes in cereals are the loss of grain dispersal such as panicle shattering at maturity, increase in grain size, loss of sensitivity to environmental cues for germination and flowering, synchronous tillering and flowering, compact growth habit and enhanced culinary chemistry (Brown, Jones et al. 2009; Panaud 2009).

Some of the phenotypic differences existing in wild or cultivated plants are caused by single-gene (monogenic) allelic variants, but most of the natural variation is quantitative and controlled by molecular polymorphisms at multiple loci and genes (multigenic). These are referred to as quantitative trait loci (QTL) and quantitative trait genes (QTGs) respectively (Alonso-Blanco, Aarts et al. 2009). In rice, two major QTL controlling seed shattering have been identified: shattering 4, (sh4) and seed shattering in chromosome 1, (qSH-1) (Konishi, Izawa et al. 2006; Li, Zhou et al. 2006). Seed shattering is caused by seed abscission (Ji, Chu et al. 2006) and in wild plants seeds are dispersed by shattering immediately after maturity (Cai and Morishima 2000). The non-shattering trait, where the seeds remain attached to the
pedicel after maturity, drastically improves grain yield and therefore is considered to be the hallmark of cereal domestication (Zhang, Zhu et al. 2009).

The Green Revolution, which started in the 1960s, marked what is arguably the most significant increase in global cereal crop production. It was mainly possible because of the worldwide adoption of improved varieties of crops with increased yields, mechanisation, and the use of fertilisers and pesticides (Peng, Richards et al. 1999; Khush 2001). One of the most important events during this revolution, which peaked in 2000 (Khush 2001), was the introduction of dwarfing genes into cereal crops, such as wheat, maize, and rice (Peng, Richards et al. 1999; Hedden 2003). The dwarfing genes resulted in less lodging and less energy spent on producing stalks, which in turn increased the productivity of the crops (Khush 2001; Hedden 2003). Major dwarfing genes have been identified in various cereal crops, such as: Reduced height, Rht, in wheat (Peng, Richards et al. 1999); maize dwarf-8, d8, in maize (Peng, Richards et al. 1999); and semidwarf1, sd1, in rice (Monna, Kitazawa et al. 2002; Sasaki, Ashikari et al. 2002).

It has been shown that substantial evolutionary changes can initially occur with selection processes operating on only a few genes (Paterson, Lin et al. 1995; Hancock 2005). Even though a large number of genes may be affected by artificial selection during domestication, a few major genes often influence a large part of the phenotypic variability (Paterson, Lin et al. 1995; Frary and Doganlar 2003; Hancock 2005). Molecular markers have provided a means to dissect the genetic basis of complex traits that are regulated by many individual QTL (Frary and Doganlar 2003; Hancock 2005). For example, the single tiller phenotype in tiller inhibition (tin3) mutants of *Triticum monococcum* has been mapped to a single QTL on the long arm of Chromosome 3. This simple recessive mutation from multiple to single tillers has associated effects of increasing seed size, seed number and tiller strength (Kuraparthy, Sood et al. 2007). QTL underlying important traits such as shattering and dwarfing have been found to be collinear between grass species (Paterson, Lin et al. 1995; Feuillet and Keller 2002). The conservation of genomes has been exploited to define sets of well-conserved anchor-probes, which are particularly useful when establishing genetic maps in grass species that had not previously been well studied (Feuillet and Keller 2002; Bolot, Abrouk et al. 2009; Wang, Roe et al. 2010).
There are three major components that make up the grain yield potential of rice: the number of panicles per plant (closely related to the tiller number per plant), the number of spikelets per panicle, and the 1000-grain weight (Zha, Luo et al. 2009). These are complex traits, with a large number of genes controlling grain yield interacting with each other (Hittalmani, Huang et al. 2003; Xie, Song et al. 2006). Only a few genes controlling grain yield have been cloned and characterised (Zha, Luo et al. 2009), a subset of which are shown on Table P1.

<table>
<thead>
<tr>
<th>Gene/protein name</th>
<th>Abbreviation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leucine-rich repeat Receptor-like Kinase</td>
<td>LRK1</td>
<td>(Zha, Luo et al. 2009)</td>
</tr>
<tr>
<td>Monoculm 1</td>
<td>MOCI</td>
<td>(Li, Qian et al. 2003)</td>
</tr>
<tr>
<td>QTL for Grain number 1A</td>
<td>Gna</td>
<td>(Ashikari, Sakakibara et al. 2005)</td>
</tr>
<tr>
<td>QTL for grain number, plant height and heading date 7</td>
<td>ghd7</td>
<td>(Xue, Xing et al. 2008)</td>
</tr>
<tr>
<td>QTL for Grain Width and weight 2</td>
<td>GW2</td>
<td>(Song, Huang et al. 2007)</td>
</tr>
<tr>
<td>QTL for Grain Size 3</td>
<td>GS3</td>
<td>(Fan, Xing et al. 2006)</td>
</tr>
<tr>
<td>QTL for Seed Width on chromosome 5</td>
<td>qSW5</td>
<td>(Shomura, Izawa et al. 2008)</td>
</tr>
</tbody>
</table>

Studies into the molecular diversity of crop plants and their wild relatives unravel fascinating facets of the domestication process and suggest ways of devising targeted approaches to access the diversity conserved in ex-situ germplasm collections (Glaszmann, Kilian et al. 2010). With improvements in the quantity of data and cost of genome sequencing technologies, soon it will be possible to determine and compare the whole sequence of hundreds of accessions (Glaszmann, Kilian et al. 2010; Cao, Schneeberger et al. 2011). With the help of molecular and genetic engineering technologies, crop wild relatives (CWR) can provide new genetic resources to improve existing crop types, and accelerate the development of new strategies for efficient and directed germplasm utilisation (Hoisington, Khairallah et al. 1999; Feuillet, Langridge et al. 2008; Negrão, Oliveira et al. 2008). Thus, a high priority which has been identified for cereal improvement worldwide is to enrich the cultivated gene pools by incorporating favourable alleles, genes or gene complexes from wild relatives (Feuillet, Langridge et al. 2008; Negrão, Oliveira et al. 2008).

One of the most significant advantages which CWR have provided to plant breeders has been the incorporation of resistance to pests and diseases from CWR to crop plants. The search for extended gene pools for genes that convey resistance to major crop pests and diseases is...
ongoing (Rao, Reddy et al. 2003; Hajjar and Hodgkin 2007). In addition, other beneficial traits such as abiotic stress, yield, quality and male sterility or fertility restoration have also been successfully adapted from CWR to commercial crops (Rao, Reddy et al. 2003; Hajjar and Hodgkin 2007; Feuillet, Langridge et al. 2008; Singh, Sharma et al. 2008).

Australian native floras contain more than 1300 species of Poaceae (Groves, Hill et al. 2002). Four major characteristics of the Australian native grasses, as summarised by Lodge (1994), are: ability to survive periods of moisture stress; adaptation to low soil fertility; adaptation to low grazing pressure by soft-footed marsupial herbivores and low trampling pressure; and adaptation to a high fire frequency. Apart from the adaption to low grazing pressure by soft-footed marsupial herbivores and low trampling pressure, the characteristics are extremely valuable to the pastoral industry in terms of adaptation to relatively unfavourable environments, but on the downside they tend to have low leaf-to-stem ratios and high lignin content (Lodge 1994). There is a growing interest by farmers to have pastures incorporating native grasses and also a growing interest for other non-agricultural uses such as amenity use and landscaping, roadside revegetation, soil stabilisation, land reclamation, and the restoration of mined areas (Lodge 1994; Smith and Whalley 2002; Firn 2007; Reed, Nie et al. 2008). Given Australia’s extensive natural genetic resources, the potential for exploiting beneficial traits, either by incorporation into existing cereals or creating new cereals, is significant.

*Microlaena stipoides*, also known as weeping grass, is a perennial grass that is found in all states of Australia, especially in the southern regions (Lodge 1996; Allan and Whalley 2004). *M. stipoides* is drought tolerant and in suitable environments will usually increase in abundance in pastures after drought, and it is also shade tolerant (Lodge 1996; Mitchell 2003). *M. stipoides* produces seeds that are similar to rice in appearance and in size (Davies, Waugh et al. 2005; O’Neill 2007). These characteristics make *M. stipoides* an ideal candidate for domestication, and so was selected as the target for this study.

The strategy adopted involved using seeds from the *M. stipoides* breeding line AR1 (provided by Native Seeds Pty Ltd.) and treating them with ethyl methanesulfonate (EMS) to induce point mutations. Induced mutations have been used to improve many major crops. A large number of induced mutant plants have been released as cultivars, while some others have been used as parents in the pedigree of a number of leading cultivars (Ahloowalia and
Maluszynski 2001). EMS is especially popular because it produces high point mutational densities but low levels of chromosomal breaks that would otherwise result in aneuploidy, reduced fertility, and dominant lethality (Greene, Codomo et al. 2003). EMS mutagenesis delivers more than 99% G/C-to-A/T transition mutation, which can be categorised as missense, truncation, or silent depending on how the codons are affected (Greene, Codomo et al. 2003).

A number of well characterised rice domestication-related genes were targeted in Chapter Two of this study. Attempts were made to characterise these genes in *Microlaena stipoides*, utilising sequences conserved between rice and other grasses to design universal primers. These universal primers were used to generate homologue sequence of the rice genes in *M. stipoides*.

Massively parallel sequencing (MPS) platforms such as Roche 454, Illumina Genome Analyzer and ABI SOLiD are capable of producing high-quality, genome-wide sequences at low cost. Nock, Waters et al. (2011) generated a partial genome sequence of *M. stipoides*. In Chapter Three this data was used to mine primers for rice domestication-related genes in *M. stipoides*, which was previously unavailable. The primers were used to generate the desired gene sequences that would form the base reference for polymorphism detection in the next stage of the study.

A number of studies at Southern Cross Plant Science have successfully used the combination of a MPS platform and Sequenom iPLEX MassARRAY (SPM) system to detect single nucleotide polymorphisms (SNPs) and the individual plants that carry the SNP (Bundock, Eliott et al. 2009; Sexton, Henry et al. 2010; Kharabian-Masouleh, Waters et al. 2011). In Chapter Four, the development of SNP screening of *M. stipoides* through the Illumina Genome Analyzer II platform combined with the SPM system to identify individuals is presented.
References:


Accelerated domestication of Australian native grass species using molecular tools


# Table of Contents

Chapter 1 - Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement ....... 1

1.1 The importance of cereals ................................................................. 1

1.2 Genomic evolution of grasses and the consequences of synteny ............. 2

1.3 Impact of domestication genes .......................................................... 4

1.4 Beyond QTLs; cloning and characterization of the domestication genes ...... 5

1.4.1 Rice ................................................................................................ 5

1.4.2 Wheat ........................................................................................... 9

1.4.3 Maize ............................................................................................ 11

1.4.4 Barley ............................................................................................ 12

1.4.5 Sorghum ......................................................................................... 13

1.5 Australian grasses: lack of domestication .............................................. 15

1.6 *Microlaena stipoides* ......................................................................... 16

1.7 Thesis Rationale .................................................................................. 18

1.8 References ........................................................................................... 19

Chapter 2 - Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides* ........................................................................................................... 29

2.1 Introduction .......................................................................................... 29

2.2 Materials and Methods ......................................................................... 32

2.2.1 Plant material and DNA extraction .................................................. 32

2.2.2 Gene information retrieval from Genbank ......................................... 32

2.2.3 Sequence alignment and universal PCR primer design ....................... 32

2.2.4 Preliminary PCR amplification using universal primers .................... 33

2.2.5 DNA sequencing ............................................................................. 34

2.2.6 Alignment of sequencing results to reference gene sequences ............ 34

2.2.7 Genome Walking ............................................................................. 34

2.2.8 Species specific whole gene PCR primer design ............................... 35
2.2.9 Cloning and DNA sequencing Alignment of sequencing results to reference gene sequences ........................................ 35
2.2.10 Whole gene PCR ............................................................................................................... 35
2.3 Results .................................................................................................................................. 36
2.3.1 Retrieval of candidate gene information from Genbank .................................................. 36
2.3.2 Universal PCR primer design .......................................................................................... 47
2.3.3 Sequence characterisation in M. stipoides ....................................................................... 62
2.3.4 Genome Walking .............................................................................................................. 65
2.3.5 Cloning to confirm sequence accuracy ............................................................................. 72
2.3.6 Alignment of sequencing results to reference gene sequences .......................................... 77
2.4 Discussion ........................................................................................................................... 77
2.5 References .......................................................................................................................... 78

Chapter 3 - Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes ................................................................. 80

3.1 Introduction ........................................................................................................................... 80
3.2 Materials and Methods ......................................................................................................... 86
3.2.1 Plant material and DNA extraction .................................................................................. 86
3.2.2 Identifying conserved regions between rice and other Poaceae for maximum possible primers placement for UPM ................................................................................................. 86
3.2.3 Identifying regions in M. stipoides aligned to rice reference genes for maximum possible primers for SSPM ........................................................................................................... 86
3.2.4 PCR primer design ........................................................................................................... 87
3.2.5 PCR amplification and sequencing preparation .................................................................. 87
3.3 Results .................................................................................................................................. 90
3.3.1 Maximum possible primers placement for UPM (prior to primer design) .................................................. 90
3.3.2 Maximum possible primers placement for SSPM ............................................................. 96
3.3.3 UPM PCR Primers ......................................................................................................... 97
3.3.4 SSPM PCR Primers ....................................................................................................... 98
3.3.5 UPM PCR Amplification and Sanger Sequencing Results ............................................. 99
3.3.6 SSPM PCR Amplification and Sanger Sequencing Results ........................................ 107
# Chapter 4 - SNP discovery in candidate domestication genes of *M. stipoides* .... 120

## 4.1 Introduction ................................................................. 120

## 4.2 Materials and Methods .................................................. 123

### 4.2.1 Plant materials .......................................................... 123

### 4.2.2 Seed treatment and sample preparation and DNA extraction ........ 123

### 4.2.3 PCR amplicon ............................................................ 124

### 4.2.4 MPS of pooled PCR amplicons ...................................... 124

### 4.2.5 Assembly to rice as a reference versus *de novo* assembly ........ 124

### 4.2.6 Illumina SNP detection and data analysis ....................... 125

### 4.2.7 SPM analysis ............................................................... 125

## 4.3 Results ........................................................................... 126

### 4.3.1 Illumina sequencing .................................................... 126

### 4.3.2 Assembly to rice as a reference versus de novo assembly .......... 126

### 4.3.3 SNP detection ............................................................. 129

### 4.3.4 SPM analysis ............................................................... 131

## 4.4 Discussion ........................................................................ 137

## 4.5 References ...................................................................... 139

# Chapter 5 - Conclusion ............................................................. 144

## 5.1 References ...................................................................... 146
List of Figures

Figure 2.3.1.1 Spidey alignment between *GW2* mRNA and *Oryza sativa* Japonica Group genomic DNA, chromosome 2, PAC clone: P0503B05 resulted in the identification of eight *GW2* exons ................................................................. 37

Figure 2.3.1.2 Spidey alignment between *Gn1a* mRNA and *Oryza sativa* Japonica Group genomic DNA, chromosome 1, BAC clone: B1046G12 resulted in the identification of four *Gn1a* exons .......................................................... 38

Figure 2.3.1.3 Alignment between EU333863 *Hordeum vulgare* subsp. vulgare Yrg1 mRNA and rice *GW2* mRNA (EF477275) showing conserved regions where primers can be designed .................................................. 40

Figure 2.3.1.4 Alignment between EU333863 *Hordeum vulgare* subsp. vulgare Yrg1 mRNA and rice *GW2* mRNA (EF477275) showing intron-exon structure and positions of primers .......................................................... 41

Figure 2.3.1.5 Alignment between *Sorghum bicolor* hypothetical protein (XM_002454958) and rice *Gn1a* mRNA (AB205193) showing conserved regions where primers can be designed .................................................. 44

Figure 2.3.1.6 Alignments between *Sorghum bicolor* hypothetical protein (XM_002454958) or sorghum trace and rice *Gn1a* mRNA (AB205193) showing intron-exon structure and positions of primers .................................................. 46

Figure 2.3.2.1 Position of forward primers gw2x1f1 and gw2x1f2 on exon 1 of *GW2* ... 48

Figure 2.3.2.2 Position of forward primer gw2x5f1 on exon 5 of *GW2* .... 49

Figure 2.3.2.3 Position of reverse primer gw2x5r1 on exon 5 of *GW2* .... 49

Figure 2.3.2.4 Position of reverse primer gw2x6r1 on exon 6 of *GW2* .... 49

Figure 2.3.2.5 Position of reverse primer gw2x6r2 on exon 6 of *GW2* .... 49

Figure 2.3.2.6 Position of forward primer gw2x7f1 on exon 7 of *GW2* .... 50

Figure 2.3.2.7 Position of reverse primer gw2x7r1 on exon 7 of *GW2* .... 50

Figure 2.3.2.8 Position of forward primer gw2x8f1 on exon 8 of *GW2* .... 51

Figure 2.3.2.9 Position of reverse primer gw2x8r1 on exon 8 of *GW2* .... 52

Figure 2.3.2.10 Position of reverse primer gw2x8r2 on exon 8 of *GW2* .... 53

Figure 2.3.2.11 Position of forward primer gn1ax1f4s on exon 1 of *Gn1a* .... 54

Figure 2.3.2.12 Position of reverse primer gn1ax1r1s on exon 1 of *Gn1a* .... 55

Figure 2.3.2.13 Position of forward primer gn1ax2f1z on exon 2 of *Gn1a* .... 56
Accelerated domestication of Australian native grass species using molecular tools

Figure 2.3.2.14 Position of reverse primer gn1ax2r1s on exon 2 of Gn1a .......................57
Figure 2.3.2.15 Position of forward primer gn1ax3f1s on exon 3 of Gn1a ......................57
Figure 2.3.2.16 Position of reverse primer gn1ax4r1s on exon 4 of Gn1a ......................58
Figure 2.3.2.17 Position of degenerate forward primer dgn1ax2r1 on exon 2 of Gn1a ....58
Figure 2.3.2.18 Position of degenerate forward primer dgn1ax3f1 on exon 3 of Gn1a ....59
Figure 2.3.2.19 Position of degenerate reverse primer dgn1ax4r1 on exon 4 of Gn1a .....59
Figure 2.3.2.20 PCR results for GW2 using primer pairs from Table 2.3.2.2 .....................61
Figure 2.3.2.21 PCR results for Gn1a using primer pairs from Table 2.3.2.2 ....................61
Figure 2.3.3.1 Sequence comparison of GW2 exon 1 in rice (x1GW2_Osat) and M. stipoides (x1GW2_Msti) resulting from sequencing PCR products 4 (Figure 2.3.2.20) ........................................................................................ 62
Figure 2.3.3.2 Sequence comparison of GW2 exon 5 in rice (x5GW2_Osat) and M. stipoides (x5GW2_Msti) resulting from sequencing PCR products 1, 2 and 4 (Figure 2.3.2.20) ........................................................................................ 63
Figure 2.3.3.3 Sequence comparison of GW2 exon 8 in rice (x8GW2_Osat) and M. stipoides (x8GW2_Msti) resulting from sequencing PCR products 3 and 9 (Figure 2.3.2.20) ........................................................................................ 63
Figure 2.3.3.4 Sequence comparison of Gn1a exon 1 in rice (x1Gn1a_Osat) and M. stipoides (x1Gn1a_Msti) resulting from sequencing PCR product 22 (Figure 2.3.2.21) ........................................................................................ 64
Figure 2.3.3.5 Sequence comparison of Gn1a exon 2 in rice (x2AB205193gn1a) and M. stipoides (gn1ax2dr1_F03) resulting from sequencing PCR product 23 (Figure 2.3.2.21) ........................................................................................ 65
Figure 2.3.4.1 Comparison between partial M. stipoides GW2 exon 1 sequence (Msgw2x1-5Tf_C01) and genome walking outer primer gw2x1op1 showing the position of the primer in the exon ..............................................66
Figure 2.3.4.2 Comparison between partial M. stipoides GW2 exon 1 sequence (Msgw2x1-5Tf_C01) and genome walking inner primer gw2x1ip5 showing the position of the primer in the exon ..............................................67
Figure 2.3.4.3 Comparison between partial M. stipoides GW2 exon 8 sequence (gw2x8r1_H01) and genome walking outer primer gw2x8op1 showing the position of the primer in the exon ..............................................67
Accelerated domestication of Australian native grass species using molecular tools

Figure 2.3.4.4  Comparison between partial *M. stipoides* GW2 exon 8 sequence (gw2x8r1_H01) and genome walking inner primer gw2x8ip3 showing the position of the primer in the exon ............................................................. 68

Figure 2.3.4.5  Comparison between partial *M. stipoides* Gn1a exon 1 sequence and genome walking outer primer gn1ax1op1 showing the position of the primer in the exon ............................................................. 68

Figure 2.3.4.6  Comparison between partial *M. stipoides* Gn1a exon 1 sequence and genome walking inner primer gn1ax1ip1 showing the position of the primer in the exon ..................................................................................... 69

Figure 2.3.4.7  Sequence comparison of GW2 exon 1 in rice (*x1EF447275Osat*) and sequence generated by cloning in *M. stipoides* (gw2x1-5c1m13f_A0) showing partial 5’ UTR region and partial intron 1 .................................. 70

Figure 2.3.4.8  Sequence comparison of GW2 exon 8 in rice (*x8EF447275Osat*) and sequence generated by cloning in *M. stipoides* (gw2x4-8c1m13f_G0) showing partial 3’ UTR region and partial intron 7 .................................. 71

Figure 2.3.5.1  Gel electrophoresis photo showing PCR results for GW2 using primer pairs from Table 2.3.5.2 (1 covers exons 1-5 and 2 covers exons 4-8)..... 73

Figure 2.3.5.2  Sequence comparison of GW2 exon 1 in *M. stipoides* (*x1Gw2Mstipoides*) and primer EMgw2x1f4 showing the position of the primer at the start of exon 1 ........................................................................................................ 73

Figure 2.3.5.3  Sequence comparison of GW2 exon 5 in *M. stipoides* (*x5Gw2Mstipoides*) and primer EMgw2x5r2 showing the position of the primer at the end of exon 5 ........................................................................................................ 74

Figure 2.3.5.4  Sequence comparison of GW2 exon 4 in *M. stipoides* (*x4Gw2Mstipoides*) and primer EMgw2x4f2 showing the position of the primer at the start of exon 4 ........................................................................................................ 74

Figure 2.3.5.5  Sequence comparison of GW2 exon 5 in *M. stipoides* (*x8Gw2Mstipoides*) and primer EMgw2x8r3 showing the position of the primer at the end of exon 8 ........................................................................................................ 75

Figure 3.1  A cladogram showing the relationship between the target species, *M. stipoides* and five major grass species, adapted from (Kellogg, 2009; Kellogg, 2001; Shapter, Eggler et al. 2009)................................. 84
| Figure 3.2.5 | Schematic representation of the workflow comparing homologues of known genes in poorly described genomes using UPM and SSPM...........89 |
| Figure 3.3.1 | Maximum possible primer placements for UPM and SSPM compared to actual primers designed for UPM and SSPM .................................................................................95 |
| Figure 3.3.5.1 | Photos of agarose gels showing the results of optimised UPM PCR amplifications (a) qsh-1, (b) sh4, (c) Gw2 exons 1-5, (d) Gw2 exons 5-6 and (e) Gw2 exons 7-8 .............................................................................102 |
| Figure 3.3.6.1 | Photos of agarose gels showing the results of optimized SSPM PCR amplifications (a) Wx exons 1-10, (b) Wx exons 9-14, (c) Hdl6 exons 1-5, (d) Hdl6 exons 4-10, (e) Sh4 exons 1-2, (f) BADH2 exons 1-5, (g) BADH2 exons 5-9, (h) BADH2 exons 8-15, (i) GW2 exons 1-5, (j) GW2 exons 4-8, and (k) GIF1 exons 3-7 ............................................................................111 |
| Figure 4.3.2.1 | Low and intermittent coverage of assembly of trimmed M. stipoides reads to rice reference genes for all three target genes. A1) BADH2 exons 1-5 amplicon, A2) exons 5-9 amplicon and A3) exons 8-15 amplicon. B1) GW2 exons 1-5 amplicon and B2) exons 1-5 amplicon. C1) Hdl6 exons 1-5 amplicon and C2) exons 4-10 amplicon. Top plot = gene structure in rice; Lower plot = coverage pattern. Axes: Y = coverage; X = base pairs......128 |
| Figure 4.3.4 | Cluster plot of the 18 SNPs genotyped using SPM. Axes: X= Low mass weight, Y= High mass weight. They show the distribution of the SPM for any particular expected SNP position in the population.......................136 |
List of Tables

Table P1  Genes, QTL and proteins controlling grain weight in rice......................... ix
Table 2.1  Details of genes affecting grain yield in rice........................................ 31
Table 2.3.2.1  Details of universal primers for \textit{GW2} and \textit{Gn1a}...................... 48
Table 2.3.2.2  Primer pairs and PCR details .......................................................... 60
Table 2.3.4  Details of genome walking primers for \textit{GW2} and \textit{Gn1a}................. 66
Table 2.3.5.1  PCR primers for \textit{GW2} ................................................................... 72
Table 2.3.5.2  Primer pairs and PCR details .......................................................... 72
Table 2.3.5.3  Exon sequence of \textit{GW2} determined by cloning overlapping PCR
  amplicons ........................................................................................................ 76
Table 2.3.6  Whole gene primers for \textit{GW2} .......................................................... 77
Table 3.1  Details of 18 annotated rice genes related to domestication traits analysed
  in this study ....................................................................................................... 85
Table 3.3.1  Comparison of mining homologues of rice genes in \textit{M. stipoides} using
  UPM and SSPM outlining the type of primers available for PCR
  amplification ..................................................................................................... 91
Table 3.3.2  Comparison of possible maximum percentage of gene covered using UPM
  and SSPM ........................................................................................................ 97
Table 3.3.3  Comparison of percentage of gene covered by primers designed using
  UPM and SSPM .............................................................................................. 99
Table 3.3.5.1  Details of primers used in this study and PCR results ...................... 103
Table 3.3.5.2  Percentage identity of putative homologues of rice genes successfully
  amplified in \textit{M. stipoides} using UPM and SSPM by Sanger sequencing of
  the PCR fragments ...................................................................................... 106
Table 4.3.1  Illumina sequencing results for a) control 109 lane; b) mutant 754 lane; c)
  mutant 109 lane ........................................................................................... 126
Table 4.3.3.1  Summary of SNP detection results using CLC Genomics Workbench
  version 4.0.3 ............................................................................................... 130
Table 4.3.3.2  Summary of SNP frequency ................................................................ 130
Table 4.3.4  Comparison between Illumina GAII SNP analysis in conjunction with the
  CLC Genomics Workbench and SPM results .............................................. 132
Chapter 1 - Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement

Understanding the molecular basis of domestication provides important tools to facilitate the efficient and rapid introgression of wild germplasm into crops varieties and even to allow domestication of completely new species (Henry 2012). Domestication events in cereals, an important food source for human and animals, have been well documented. In order to feed the growing population improvements in cereal production are urgently required.

1.1 The importance of cereals

Cereals belong to the grass family, Poaceae, which occur in all continents, inhabiting a vast range of environmental conditions (Kellogg 1998; Clayton, Harman et al. 2006 onwards). Currently, the five most important cereals in the world are rice, wheat, maize, barley and sorghum (Sang 2009). Cereals account for 50-80% of calories in human diets (Delseny, Salses et al. 2001; Khush 2001; Varshney, Hoisington et al. 2006) and the majority of stock feeds (Speedy 2003; Ashikari, Sakakibara et al. 2005). There is also a recent trend to use cereal crops as a source of biofuels following the rapid increase of petroleum-based fuel price and the uncertainty of supply in the future (Naylor, Liska et al. 2007; Jakob, Zhou et al. 2009; Henry 2010).

In the early days farmers recognised beneficial characteristics that could be chosen from the wild landraces, which were subsequently planted, harvested and reselected in order to build improved populations with a range of desirable characteristics (Vaughan, Balazs et al. 2007). The domesticated forms looked significantly different to their wild ancestors as a result of the selection of domestication genes (Vaughan, Balazs et al. 2007). The need to feed the world’s fast growing population led to the accelerated growth in food crop production,
known as Green Revolution. In comparison to the previous varieties, these new varieties were shorter and had a higher harvest index (proportion of the plant biomass harvested as grain) and as a result they were less prone to lodging allowing them to respond more efficiently to fertiliser applications and making them more resistant to damage by wind and rain (Peng, Richards et al. 1999; Evenson and Gollin 2003; Baenziger, Russell et al. 2006). The varieties that resulted from the Green Revolution have performed well and spread worldwide. Cereal prices have soared in the past few years due to low supply following poor harvests, contributing to food shortages in poverty-stricken parts of the world (FAO 2008). There is an urgent need to breed crops that can grow with little fertiliser input and nutrient poor soils, as resources are getting scarce worldwide (Lynch 2007).

1.2 Genomic evolution of grasses and the consequences of synteny

The Poaceae family diverged from other angiosperms 50 million years ago (Paterson, Bowers et al. 2004) and since that time have evolved into over 10,000 currently identified species (Clayton, Harman et al. 2006 onwards). During these complex evolutionary processes, the grass genome has undergone a number of genome duplications, recombinations and polyploidisation concurrently with gene redundancy and deletions (Bennetzen 2007). Besides polyploidisation and duplication, genome size differences are largely the result of differences in non-coding and repetitive DNA sequences (Caetano-Anollés 2005). The five primary mechanisms of genomic instability in all flowering plants are polyploidy, transposon amplification, chromosome breakage, unequal homologous recombination and illegitimate recombination, though they do not necessarily occur independently (Bennetzen 2007). Differences often result from mutational mechanisms of nucleic acid addition and loss, such as transposition, spontaneous insertions and deletions, and chromosomal rearrangements (Delseny, Salses et al. 2001; Caetano-Anollés 2005).

There is an extensive conservation of gene order and gene content among closely related grass species. However even largely collinear genomic regions have been shown to have a high degree of sequence diversity at the micro-structural level (Doebley, Stec et al. 1997). In the five main crop species of maize, rice, sorghum,
barley and wheat comparisons have been extended to the DNA sequence level (micro-colinearity), allowing study of the conservation of coding and non-coding regions as well as characterisation of molecular mechanisms of genome evolution in grasses (Feuillet and Keller 2002). More recently micro-colinearity has been used as a platform to determine genetic diversity existing in wild crop relatives of cultivated cereals (Shapter, Eggler et al. 2009; Zeid, Yu et al. 2010). The use of molecular markers derived from orthologous regions in different grass species has helped to increase the map density at specific genetic loci and facilitate map-based cloning (Feuillet and Keller 2002). Comparative studies have demonstrated a high degree of conservation of the exons between homologous grass genes. To some extent intron positions are also conserved but intron size can sometimes vary greatly (Feuillet and Keller 2002; Inada, Bashir et al. 2003).

After a separation of more than 50 million years, coding exon sequences and intron/exon boundaries are highly conserved, but intron sequences, 5’-leader sequences and 3’-trailer sequences show little convincing homology across grass species (Bennetzen and Ma 2003). Micro-colinearity studies provide key information on the genome structure and the mechanisms responsible for differences in genome size and evolution in grasses (He, Luo et al. 2006) and with increasing genomic sequence data, variation can be more easily investigated. Colinearity of a large region from barley (Hordeum vulgare) chromosome 5H and rice chromosome 3 has been demonstrated by mapping several common RFLP clones on both regions (Dubcovsky, Ramakrishna et al. 2001). Furthermore, rice chromosome 3 is evolutionarily conserved across the cultivated cereals and shares large blocks of synteny with maize and sorghum, which diverged from rice more than 50 million years ago (The Rice Chromosome 3 Sequencing Consortium 2005). With the exception of the centromere region, rice chromosome 3 is highly collinear with the short arm of maize chromosome 1 (1S) and the inverted long arm of maize chromosome 9 (9L), while rice 3L is highly conserved with the maize 1L and the inverted maize 5S (The Rice Chromosome 3 Sequencing Consortium 2005).

Nearly a decade ago, the japonica and indica rice genomes had been sequenced and annotated using the whole genome shotgun method (Goff, Ricke et al. 2002; Yu,
The complete sorghum genome was also sequenced using the whole genome shotgun method and other more traditional methods have been published (Paterson, Bowers et al. 2008; Paterson, Bowers et al. 2009). More recently, barley and maize genomes have been sequenced (Schnable, Ware et al. 2009; The International Barley Genome Sequencing Consortium 2012). A number of studies worldwide are well underway to sequence the highly complex wheat genome (Paux, Sourdille et al. 2008; Berkman, Skarszewski et al. 2011; Trick, Adamski et al. 2012) and sugarcane (Wang, Roe et al. 2010; Bundock, Casu et al. 2012). These advances have resulted in exponential increases in availability of genomic sequence data across the Poaceae and have enabled variation to be more easily investigated at the sequence level (Khush 2005). The genomic sequence information can be used for comparative genomic analysis that will help assign putative gene function to a gene according to what that gene does in another species (Khush 2005; Shapter, Eggler et al. 2009; Wicker, Taudien et al. 2009). By knowing the sequence of specific genes breeders can tap into natural genetic variation of a crop species along with its wild relatives and use it to develop new varieties (Khush 2005; Dillon, Shapter et al. 2007; Varshney and Tuberosa 2007; Feuillet, Langridge et al. 2008; Henry 2012).

1.3 Impact of domestication genes

Domesticated of crop species differ from their wild ancestors and relatives in a subset of key traits that contribute to their success in cultivation (Frary and Doganlar 2003). Independent domestication of four major cereal complexes, in Africa (sorghum and millet), Asia (rice), the near east (wheat, barley, oat and rye) and America (maize), produced similar results whereby small-seeded wild grasses with natural seed dispersal were converted into large-grained ones that depended on farmers to harvest and sow their seeds (Paterson, Lin et al. 1995; Paterson, Freeling et al. 2005).

Five primary cereal domestication characteristics are reduced plant height, yield increase, non-shattering, controlled dormancy and photoperiod insensitivity (Khush 2001). These characteristics would be detrimental for the survival of the crops in the natural environment, but aid in the ease of cultivation for farmers and have been
selected for when developing breeding lines (Doebley, Gaut et al. 2006; Tang and Shi 2007). Clustering of domestication-related genes in close proximity on the same chromosomes has been well documented in literature (Frary and Doganlar 2003; Hancock 2005; Lee, Oh et al. 2005). Numerous QTL studies of crop species have uncovered genes with pleiotropic effects. As a result, their selection would make a large-scale phenotypic change much more rapid than if the traits were evolving separately (Hancock 2005). A comparative QTL study suggested that there had been convergent selection for large seeds, reduced shattering and day length-insensitive flowering across rice, sorghum and maize and these studies led to the hypothesis that selections at a small number of loci can explain much of the phenotypic outcome of the domestication process (Thomson, Tai et al. 2003).

Tang & Shi (2007) reported three mechanisms by which artificial selection may have an impact on the genomes of domesticated species. First, many genes that are directly under artificial selection for their effects on the desired traits under domestication may have adverse effects on fitness in nature. Such genes may be under direct artificial selection for traits such as shattering, or selected as a consequence of pleiotropic effects. Second, the relaxation of selective constraint during domestication could allow the accumulation of mutations that may otherwise be deleterious in nature. Third, mutations that are deleterious even under domestication may hitchhike with genes that are strongly selected for in domestication. It is especially common in inbreeding plants, where recombination does not contribute to increased variation.

1.4 Beyond QTLs; cloning and characterization of the domestication genes

1.4.1 Rice

Rice (*Oryza sativa*) is regarded as one of the most important cereal crops. It is generally agreed that rice was domesticated around 10,000 years ago in China, although there is continuing debate as to its point or points of origin and whether its direct ancestral lineage stems from *O. rufipogon* (Kawakami, Ebana et al. 2007) and/or *O. nivara* (Tang and Shi 2007). There are also hypotheses for single, dual or
multiple domestication events (Londo, Chiang et al. 2006). Allelic variations occur between cultivated varieties and wild species.

The origins of different Oryza species are believed to be *O. sativa* in Asia and *O. glaberrima* in Africa (Delseny, Salses et al. 2001; Sweeney and McCouch 2007; Vaughan, Lu et al. 2008) and other species from other continents. Indica and japonica varieties cross-hybridize. However, this usually results in many of the progeny being sterile or partially sterile (Delseny, Salses et al. 2001; Kubo and Yoshimura 2002; Sweeney and McCouch 2007). All species in the *O. sativa* complex have the diploid AA genome as defined by cytogenesis whereas wild species, belonging to the *O. officinalis* complex are either diploid or allotetraploids with BB, CC, EE or BBCC or CCDD genomes (Delseny, Salses et al. 2001; Vaughan, Morishima et al. 2003).

Rice is considered as a model plant because it has a small genome of 389 mega base pairs (Mbp), sequence conservation with other cereal crops and is readily transformed (Devos and Gale 1997; Delseny, Salses et al. 2001; Ashikari, Sakakibara et al. 2005; International Rice Genome Sequencing Project 2005; Paterson, Freeling et al. 2005; Ashikari and Matsuoka 2006; He, Luo et al. 2006). Extensive comparative mapping has established that the gene order is essentially conserved between rice chromosomes and other cultivated cereals, allowing one to isolate important genes in rice first and then use the gene sequence as a probe to isolate the corresponding homologue in other cereals (Delseny, Salses et al. 2001; McIntosh, Pacey-Miller et al. 2005; Shapte, Eggler et al. 2009). The rice genome has twelve pairs of chromosomes which can be individually recognised (Delseny, Salses et al. 2001; Goff, Ricke et al. 2002). The completion of the rice genome sequencing confirms that rice, similar to other cereals has a rather GC-rich genome and highly repetitive DNA (Delseny, Salses et al. 2001; Yu, Hu et al. 2002).

In rice breeding, high grain yield is paramount and is determined by three yield components; panicle per plant, grain weight and grain number per panicle (Tian, Zhu et al. 2006). Plant height is also an important trait that directly contributes to grain productivity (Ashikari, Sakakibara et al. 2005). These characteristics seem to
be regulated by QTL, as they are approximately normally distributed in mapping populations (Ashikari, Sakakibara et al. 2005).

Grain yield is a complex trait dependent on various growth and component traits and is affected by the environment (Hittalmani, Huang et al. 2003). Xie, Song et al. (2006) found 26 QTL for grain weight from crosses between *O. rufipogon* and diverse cultivars. To date, one of the most effective QTL for increasing grain number is *gn1a* which corresponds to the cytokinin oxidase/dehydrogenase (OsCKX2) on chromosome 1, accounting for 44% of phenotypic variation in a breeding population (Ashikari, Sakakibara et al. 2005). Transgenic rice with antisense OsCKX2 cDNA has shown an increase in grain development (Ashikari, Sakakibara et al. 2005). This region of rice chromosome 1 shows sequence similarities with chromosome 3, 6 and 8 in maize, where QTL for grain yield traits have previously been mapped (Ashikari, Sakakibara et al. 2005). If *gn1a* in rice corresponds to these maize QTLs and orthologous CKX genes in other species it may also regulate yield in other cereal crops (Ashikari, Sakakibara et al. 2005).

Variation in grain size amongst japonica and indica varieties serve different functional needs and are selected for in breeding programs accordingly. A QTL with major effect on grain size was consistently detected around the centromeric region of chromosome 3 in numerous studies across different backgrounds and environments (Li, Thomson et al. 2004; Fan, Xing et al. 2006). Earlier QTL studies identified *Gw3.1*, a grain weight-related QTL (Li, Thomson et al. 2004) and *GS3*, a major QTL for grain length and weight and a minor QTL for grain width and thickness (Fan, Xing et al. 2006). A putative *GS3* gene has been identified, with five exons encoding 232 amino acids (Fan, Xing et al. 2006). A nonsense mutation in the second exon of the putative *GS3* gene is shared among all the large grain varieties tested by Fan, Xing et al. (2006) and the mutation causes a 178-aa truncation in the C-terminus of the predicted protein, inferring that *GS3* may function as a negative regulator for grain size.

Song, Huang et al. (2007) mapped a major QTL for grain width, *GW2*, on chromosome 2. *GW2* encodes a previously unknown RING(Real Interesting New
Gene-type protein with E3 ubiquitin ligase activity, which is known to function in the degradation by the ubiquitin-proteasome pathway and reduced expression of $GW2$ increases grain size (mainly grain width), resulting in enhanced grain weight, whereas overexpression decreases grain size and weight (Song, Huang et al. 2007). $GW2$ was also shown to have pleiotropic effects on panicle number per plant, days to heading and main panicle length, in addition to grain number per main panicle (Song, Huang et al. 2007). Song, Huang et al. (2007) successfully cloned $GW2$.

Seed dehiscence or ‘shattering’ is caused by seed abscission (Ji, Chu et al. 2006). In rice, several loci have been associated with this trait: sh1 on chromosome 11, sh2 on chromosome 1, sh3 on chromosome 4 and sh4 on chromosome 3 (Ji, Chu et al. 2006). Sh4 explained 69% of phenotypic variance whereby sh4 allele of the wild species caused shattering and was dominant (Li, Zhou et al. 2006). More recent studies have also identified the $qSH-1$ as controlling this phenotype (Konishi, Izawa et al. 2006). It appears these QTLs have differing modes of action and a null allele at either major loci, sh4 or $qSH-1$, can cause the non-shattering phenotype (Konishi, Izawa et al. 2006). Natural reversion to shattering habit is increasingly becoming an issue in Asian rice crops and has increased characterisation studies of these genes (Vaughan, Lu et al. 2008). This should provide a clearer understanding of the mechanisms for controlling the shattering habit in rice in the near future.

Several other QTL associated with domestication and increasing productivity in rice have been characterised. The rice heading date or photoperiod sensitivity (PS) trait has been attributed to four major QTLs; Hd1, Hd3a, Hd6 and Ehd1 (Ashikari and Matsuoka 2006). Hd1 is a major rice PS QTL and a homologue of CONSTANS (CO) from Arabidopsis, composing two exons and encodes a protein with a zinc finger domain (Yano, Katayose et al. 2000). It is speculated that Hd1 affects the transcription of genes for which expression is controlled by photoperiod changes (Yano, Katayose et al. 2000). The Hd3a locus contains a candidate gene that shows high similarity to the FLOWERING LOCUS T (FT) gene, which promotes flowering in Arabidopsis and the introduction of the gene caused an early-heading phenotype in rice (Kojima, Takahashi et al. 2002). The amount of Hd3a mRNA is up-regulated by Hd1 under short day conditions, suggesting that
Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement

Hd3a promotes heading under the control of Hd1 and through their research, Kojima et al. (2002) deduced that Hd3a encodes a protein closely related to Arabidopsis FT and that the function and regulatory relationship with Hd1 and CO, respectively, of Hd3a and FT are conserved between rice (a short day (SD) plant) and Arabidopsis (a long-day (LD) plant). The Hd6 locus encodes an α subunit of protein kinase CK2 and has been narrowed to a 26.4 kilo base pairs (kbp) genomic region and the CK2a allele of Nipponbare has a premature stop codon that is not found in the same position in the Kasalath allele (Takahashi, Shomura et al. 2001). The Nipponbare mRNA encodes only a truncated protein and is non-functional whereas genetic complementation analysis revealed that the Kasalath allele of CK2a increases days-to-heading (Takahashi, Shomura et al. 2001). The Early heading date 1 (Ehd1) gene confers SD promotion of flowering in the absence of a functional allele of Hd1 (Doi, Izawa et al. 2004). Doi, Izawa et al. (2004) hypothesised that Ehd1 promotes flowering by inducing FT-like gene expression only under SD conditions.

Fragrance and gelatinisation temperature (GT) are two other important rice characteristics targeted by human selection. In rice, fragrance is a recessive trait shown to be caused by an eight base pairs (bp) deletion and three SNPs in a gene on chromosome 8 which encodes a putative betaine aldehyde dehydrogenase 2 (BAD2 or BADH2, Bradbury, Henry et al. 2005). Gelatinisation temperature is influenced by starch synthase IIa (SSIIa) and two important SNPs at exon 8 of SSIIa divide either high or low GT in 70 diverse genotypes in rice (Waters, Henry et al. 2006).

1.4.2 Wheat

Ploidy levels of domesticated wheat species range from diploid (2n=14) to hexaploid (6n=42), and wheats of all ploidy levels have been domesticated at some point in time (Salamini, Ozkan et al. 2002). Bread wheat (Triticum aestivum L.) is an allohexaploid with three different genomes (A, B and D) and in which 2n=6x=42 (Faris and Gill 2002; Franck and Appels 2002). Wild emmer (T. turgidum ssp dicoccoides syn. T. dicoccoides) is an allopolyploid developed from hybridization of T. urartu (AA) with a B genome diploid of unknown origin and the freethreshing form of this gave rise to durum wheat (T. turgidum ssp durum syn T.
Wheat species with tetraploid genomes were involved in accidental crosses with the wild diploid species *T. tauschii*, the donor of the D genome, and it gave rise to hexaploid wheat *T. aestivum* also known as bread wheat (Devos and Gale 1997). There are different theories behind the origin of the AABB genomes in tetraploid wheat, one of them is that tetraploid wheat was derived from a single tetraploid progenitor *T. diccoides*, the donor of A & B genomes (Salamini, Ozkan et al. 2002; Reif, Zhang et al. 2005) and the other is tetraploid *T. turgidum* subsp. durum AABB (Devos and Gale 1997; Peleg, Fahima et al. 2011).

Mutations generated new alleles while recombination created novel allele combinations that formed the original gene pool from which modern wheat cultivars have been developed (Reif, Zhang et al. 2005). This expanded genetic variation has subsequently been reduced by genetic drift and selection, resulting in landrace cultivars adapted to specific conditions and in the past century, these landrace cultivars have been continually replaced by modern wheat cultivars (Reif, Zhang et al. 2005). In recent years this trend has been reversing, with a growing interest in returning to these landraces as a source of novel germplasm for breeding programs (Reif, Zhang et al. 2005; Reynolds, Drecce et al. 2007; Peleg, Fahima et al. 2011).

The wheat varieties developed and used in the Green Revolution were shorter because they responded abnormally to the plant growth hormone gibberellin (Peng, Richards et al. 1999). The reduced response to gibberellin is conferred by mutant dwarfing alleles at one of two Reduced height-1 (*Rht-B1* and *Rht-D1*) loci (Peng, Richards et al. 1999). Breeders used 'Norin 10' that carried the genes *Rht1* and *Rht2* and incorporated it into many breeding programs throughout the world (Börner, Plaschke et al. 1996).

Brittle rachis1 (*Br1*) is a wheat shattering locus that is located on the short arm of wheat group-3 chromosomes and they are syntenic with the short arm of rice chromosome 1 (Li and Gill 2006). In wild ancestral wheats, seed shattering is caused by a brittle rachis, which is conferred by a dominant Br allele and a
Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement

recessive mutant allele br at this locus producing a non-brittle spike (Li and Gill 2006). Li and Gill (2006) hypothesised that Br1 is not orthologous to the shattering loci detected in other grasses and is possibly unique to wheat.

The mutation that gave rise to \( Q \) allele, which provides the square-headed phenotype and free-threshing character, occurred just once however most cultivated wheat varieties possess the \( Q \) allele (Faris and Gill 2002; Simons, Fellers et al. 2006). Wild wheats tend to have the \( q \) allele and have speltoid spikes that are not free-threshing (Faris and Gill 2002; Simons, Fellers et al. 2006). The early farmers found more efficient grain harvest with the emergence of the free-threshing character concomitant with reduced rachis fragility and glume tenacity (Simons, Fellers et al. 2006). \( Q \) allele is incompletely dominant to \( q \) and hence the Qq plants have spikes that are intermediate between the speltoid and square-headed phenotypes (Faris and Gill 2002). The \( Q \) allele regulates a diverse set of developmental traits in plants, but especially traits related to inflorescence structure and flowering (Doebley 2006).

One gene affecting seed size in wheat is grain protein content (GPC) is \( GPC-B1 \), an early regulator of senescence with pleiotropic effects on grain nutrient content (Uauy, Distelfeld et al. 2006). To date, except for \( Q \) and \( GPC-B1 \), no other genes relevant to the wheat domestication syndrome have been isolated (Dubcovsky and Dvorak 2007).

1.4.3 Maize

There are contradictory views of the ploidy of maize. One view contends that maize is a tetraploid, probably of relatively recent origin since most DNA probes hybridise to duplicated loci (Devos and Gale 1997). It has been proposed that genome size in the progenitors of Zea and Tripsacum doubled as a result of a segmental allotetraploid event, which occurred sometime between the divergence of sorghum and the rediploidization of the ancestors of maize (Caetano-Anollés 2005). The alternative view is that maize (n=10) is diploid by deleting most of the duplicated centromere regions and also deleting or tolerating degeneration of one member of most of its paired gene sets, sometimes fragmenting ancestral gene
orders across multiple chromosomes and obscuring similarities in gene order that existed among its ancestors (Paterson, Freeling et al. 2005).

The progenitor of maize (Zea mays spp. mays) is believed to be teosinte (Z. mays spp. parviglumis), with the main phenotypic difference being that teosinte has multiple branches while modern maize typically has a single main stem (Devos and Gale 1997; Doebley, Stec et al. 1997). Maize was domesticated from teosinte in southern Mexico between 6,600-9,000 years ago (Buckler, Gaut et al. 2006). The teosinte branched1 (tb1) gene is largely responsible for a major-effect QTL on chromosome 1 that controls differences in plant architecture between maize and teosinte (Doebley, Stec et al. 1997; Clark, Linton et al. 2004). Tb1 decreases the number of primary lateral branches and tillers on the plants by allowing the outgrowth of buds rather than by increasing the number of auxillary buds formed on the main stalk (Doebley, Stec et al. 1997). The wild-type gene in maize acts to repress bud outgrowth due to being overexpressed compared with the orthologue in teosinte. Loss of tb1 function in maize results in loss of apical dominance and outgrowth of axillary buds.

In domesticated maize, the diversity is roughly 30% below that of the closest wild relative, although the drop in diversity can be substantially greater in some of the genes directly involved in domestication (Buckler and Thornsberry 2002). The 2.3-gigabase (gbp) maize genome was sequenced recently and it was found that uneven gene losses between duplicated regions were involved in returning an ancient allotetraploid to a genetically diploid state (Schnable, Ware et al. 2009).

1.4.4 Barley

Barley (Hordeum vulgare) was domesticated 8,000-10,000 years ago from H. spontaneum in the Fertile Crescent (Badr, Muller et al. 2000). There are contradictory views regarding whether barley was domesticated once (Badr, Muller et al. 2000; Salamini, Ozkan et al. 2002) or twice (Morrell and Clegg 2007). Major domestication traits in barley are non-brittle or non-shattering grain, two or six rows spike and presence or absence of hulls (Saisho and Purugganan 2007).
Wild relatives of barley disperse their seeds at maturity by means of their brittle rachis, but in domesticated barley, brittleness of the rachis was lost (Komatsuda, Maxim et al. 2004). \textit{Btr1} and \textit{Btr2} are two complimentary and closely linked shattering genes in barley located on the short arm of chromosome 3H (Li and Gill 2006). Molecular mapping showed that wheat \textit{Br1} and \textit{Br2} may not be orthologous to either \textit{Btr1} or \textit{Btr2} of barley, although more common markers between the two loci are needed for detailed comparative mapping (Komatsuda, Maxim et al. 2004; Komatsuda, Pourkheirandish et al. 2007).

Both two- and six-rowed barleys emerged in Greece between 8,000 and 6,000 years before present (yBP), although six-rowed barley dominated in the Balkans and Central Europe (5,000–2,000 yBP) and in Southern Europe and North Africa (7,000–4,000 yBP, Zohary and Hopf 2000). A recent study using positional cloning isolated \textit{vrs1}, the gene responsible for the six-rowed spike in barley. Loss of function of \textit{Vrs1} (wild type) resulted in complete conversion of the rudimentary lateral spikelets in two-rowed barley into fully developed fertile spikelets with the six-rowed phenotype (Komatsuda, Pourkheirandish et al. 2007). Phylogenetic analysis demonstrated that the six-rowed phenotype originated repeatedly, at different times and in different regions, through independent mutations of \textit{Vrs1} (Komatsuda, Pourkheirandish et al. 2007).

Very recently, a physical, genetic and functional sequence assembly of the barley genome was released (The International Barley Genome Sequencing Consortium 2012). Barley is a diploid cereal with a large haploid genome of 5.1 gbp (The International Barley Genome Sequencing Consortium 2012). The genome contains a very large number of SNPs loci and studying the variations in the genome provides an essential reference for genetic research and breeding (The International Barley Genome Sequencing Consortium 2012).

1.4.5 \textbf{Sorghum}

Sorghum (\textit{Sorghum bicolor}) is genetically suited to hot and dry agroecologies where it is difficult to grow most food grains and is the major food crop in the semi-arid zones of Western and Central Africa (Casa, Mitchell et al. 2005). It is a
Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement

diploid C4 grass and has a relatively small genome (735 Mbp, Dillon, Shapter et al. 2007). Sorghum was domesticated in north-eastern Africa from its progenitor *S. bicolor* spp. verticilliflorum, although the hypothesised number of events varies from a single domestication to three domestication events (Londo, Chiang et al. 2006). Initial domestication of sorghum would have focused primarily on converting wild types with small, shattering seed to types with larger, non-shattering seeds (Dillon, Shapter et al. 2007). Stable, high yielding sorghum have been developed by selecting traits such as photoperiod insensitivity, reduced height, drought tolerance, and pest and disease resistance (Ramesh, Reddy et al. 2006). There are four major dwarfing genes found in sorghum (*Dw1-Dw4*), however, to date there is only one that has been cloned, *Dw3* (Brown, Rooney et al. 2008). *Dw3* encodes a phosphoglycoprotein auxin efflux carrier orthologous to *PGP1* in Arabidopsis (Brown, Rooney et al. 2008). A comprehensive analysis of genetic diversity in sorghum landraces based on race, altitude of origin, photoperiod, seed quality, agronomic traits and DNA markers has demonstrated sorghum has considerable polymorphism that has been poorly exploited in terms of crop improvement (Dillon, Shapter et al. 2007).

There is a striking conserved sorghum-rice microsynteny that corresponded closely with regions of ancient whole-genome duplication (Bowers, Arias et al. 2005). Gene orders also appear to be largely conserved between sorghum and maize and only a limited number of rearrangements have been identified (Devos and Gale 1997; Paterson, Bowers et al. 2009).

Sugarcane (*Saccharum officinarum*) is the closest domesticated grass to sorghum, both belonging to the Andropogoneae tribe of the Poaceae (Dillon, Shapter et al. 2007). The sugarcane genome is highly complex and large, and its similarity to sorghum has been utilised by researchers to decipher the sugarcane genome (Wang, Roe et al. 2010; Bundock, Casu et al. 2012). The sorghum genome was recently sequenced (Paterson, Bowers et al. 2009) and shown to have about 61% repeat content. This event paves a way for researchers to develop greater insight into the sorghum genome and use this knowledge to enhance breeding strategies and/or as a
Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement

model grass genome, to infer information for other lesser known but more complex related grass species.

1.5 **Australian grasses: lack of domestication**

Although the domestication of native Australian grasses has been proposed since the mid 1890’s (Turner 1895), registrations of cultivars has been sparse (Lodge 1996). Waters, Whalley et al. (2002) listed 21 varieties of Australian native grasses registered for Plant Breeders’ Rights and/or Plant Varieties Rights. Most of these grasses were selected for forage and pasture, land reclamation, revegetation, amenity, or turf grass applications (Lodge 1996; Waters, Whalley et al. 2002). The number of grasses registered in Australia steadily grows, although they contribute very little to the total 7105 plants registered on the Plant Breeders Rights website (IP Australia).

One of the most important characteristics of Australian native grasses is drought tolerance, as these grasses have evolved under either continuous or periodic drought. Similarly, they evolved on soils of generally low fertility and are better adapted to Australia’s conditions (Lodge 1996; Mitchell 2003). Unlike introduced grasses, even under low fertility conditions, these grasses are capable of providing high quality fodders if grazed intermittently and allowed substantial recovery periods (Robinson and Archer 1988). Unfortunately, many native grasses cannot outcompete introduced species under high fertility conditions nor when subject to the continuous grazing pressure that has been imposed since European settlement (Lodge 1996).

In general, native grass seed production per plant is intrinsically lower than cultivated cereals. When combined with a lack of synchrony of maturity and a shattering habit, typical of undomesticated Poaceae, this results in harvested grain yields being comparatively low and an increased cost of production (Lodge 1996; Oram and Lodge 2003). There is also limited knowledge of their breeding systems and the main focus of native grass improvement programs to date has been seed retention and seed production (Lodge 1996; Cole and Johnston 2006).
Australian native grass species include a diverse array of wild cereal relatives which are adapted to a wider range of environmental conditions than current commercial cereals and may contain novel alleles which have utility in commercial production systems (Shapter, Henry et al. 2008; Henry, Rice et al. 2010). The wild cereal relatives exhibit similar types of starch granules to that of their respective cultivated species, though in general the wild species retain a greater proportion of the endosperm cell wall at maturity (Shapter, Henry et al. 2008; Kasem, Waters et al. 2010). A recent study found useful variations in the endosperm morphology of wild grass species and these could provide valuable genetic resources for enhancing the nutritional and functional value of cereals into the future (Shapter, Henry et al. 2008).

Ploidy levels among Australian native grasses vary greatly, ranging from diploid grasses such as *Australopyrum* spp. and *Deschampsia* spp. to 38x in *Poa* spp. (Murray, De Lange et al. 2005). Genome doubling and dosage effects are the major consequences of autopolyploidy, although genome merger has a greater effect than genome doubling on genomic and gene expression changes in allopolyploids (Jackson and Chen 2010). Polyploid-induced changes can generate new individuals that have the capability to exploit new niches or to outcompete progenitor species (Leitch and Leitch 2008). There are three well documented advantages of polyploidy: heterosis, gene redundancy, and asexual reproduction (Comai 2005). Heterosis causes polyploids to be more vigorous than their diploid progenitors, gene redundancy can shield polyploids from the deleterious effect of mutations and asexual reproduction allows polyploids to reproduce in the absence of sexual mates (Comai 2005).

### 1.6 *Microlaena stipoides*

*Microlaena stipoides* (Weeping grass) is a perennial grass native to Australia occurring in all Australian States except the Northern Territory. It is widespread in south-eastern regions including Tasmania and also the south-west of Western Australia (Lodge 1996; Groves, Hill et al. 2002; Allan and Whalley 2004). Different ecotypes have been shown to be relatively frost tolerant; maintain
greenness in winter, are drought and shade tolerant and highly palatable to grazing animals (Lodge 1996; Chivers and Aldous 2005). *M. stipoides* responds well to nitrogen application and also regular irrigation, making commercial production possible, although different cultivars have differing needs (Chivers and Aldous 2005).

A high degree of variability has been found in seed yield and component traits, and this suggests there is sufficient variation within the species to make selections for higher yielding lines (Davies, Waugh et al. 2005; Davies, Waugh et al. 2005). Commercial seed production is already underway for *M. stipoides* (Martin 2004; Chivers and Aldous 2005). One of the problems encountered in seed production is the high cost associated with seed harvesting, since it is done in multiple sessions and is laborious (Martin 2004). Some varieties have seed size similar to commercial rice varieties and there may be potential for establishing larger seed size *M. stipoides* through molecular breeding (O’Neill 2007).

*M. stipoides* has been identified as a potential grain crop for its seed size and also due to its wide distribution and environmental adaptations (Davies, Waugh et al. 2005a; Davies, Waugh et al. 2005b). *M. stipoides* seed protein content has been reported as high as 22%, compared with only 9% for rice, 12–14% for the best hard wheats, and up to 18% for high-protein hybrid maize varieties (Native Seeds Pty Ltd 2002; O’Neill 2007). *M. stipoides* was identified as retaining its endosperm cell wall structure at maturity, as opposed to the majority of cereals which only have remnants at maturity (Shapter, Henry et al. 2008). The starch granule morphology of *M. stipoides* was found to be very similar to rice (Shapter, Henry et al. 2008). *M. stipoides* is an allotetraploid grass with 2n=48 (Murray, De Lange et al. 2005). The duplication of loci of an allopolyploid can contribute to increased allelic diversity available for selection. This grass has potential, within the current Australian agricultural system to provide a dual purpose crop for grazing and grain. Currently the major limitations to its use as a crop include seed shattering and comparatively low seed yield. *M. stipoides* has the potential to produce higher yields through appropriate selection programs and the use of our increasing understanding of the genetics surrounding domestication.
1.7 Thesis Rationale

Although cereals were domesticated independently in different parts of the world, the characteristics selected by farmers and breeders were strikingly similar. These characteristics ease cultivation and harvest. The Green Revolution gave rise to elite varieties that are highly productive for food production. The advances in molecular breeding and whole genome sequencing aid the understanding of the underlying genetic explanations that generate the cereal characteristics. The synteny between grasses can be utilised in inferring known genetic information from one species to another, usually more complex and lesser understood. The transfer of genes between one species to another aids in breeding for better crops and even new crops to better feed the world’s population.

The aim of the study was to accelerate the domestication an Australian native grass, *M. stipoides*. As shown in Figure 3.1 (Chapter 3), *M. stipoides* is closely related to the five major grass (Poaceae) crop species (rice, wheat, maize, barley and sorghum). Utilising sequence homology between the grasses as a genomic platform, domestication genes that have been identified in rice and other domesticated grasses would be characterised in *M. stipoides* and markers and/or screening techniques developed for selection and breeding programs. A concurrent study targeting shattering genes was already in place; therefore shattering genes were not the focus of this thesis. It was proposed to use chemical induced mutations to create genetic variation in *M. stipoides* followed by molecular screening to identify individuals with genes of interest. They would be used as parental lines to create domesticated varieties for the commercial market.

A number of rice varieties have been created by induced mutations. Many of the induced mutants have been released as cultivars and some others have been used as parents for further breeding of leading cultivars (Ahloowalia and Maluszynski 2001). The advantage of using mutation for breeding is that it is not considered as transgenic and therefore is much more widely accepted by consumers (Baenziger, Russell et al. 2006). Techniques that combine mutagenesis with novel and sensitive methods of detecting induced DNA sequence variation are establishing a new niche for mutagenesis. In the ever growing area of plant functional genomics, this
provides a bridge from gene or SNP discovery in model species to application in other crops (Waugh, Leader et al. 2006).

1.8 References


Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement


Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement


Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement


Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement


Martin, P. M. (2004). The potential of native grasses for use as managed turf. 4th International Crop Science Congress, Brisbane.


Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement


Chapter 1: Literature review: Molecular events in the domestication of the major cereal crops and their significance in crop improvement


Chapter 2 - Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

2.1 Introduction

This chapter focuses on a method for characterising the genes of the grass species *Microlaena stipoides* homologous to those affecting grain yield in rice (*Oryza sativa*). This approach is based on utilising sequences conserved between rice and other grasses to design universal primers. Rice has been extensively used as a model for many grasses because it has a relatively small genome (~400 Mb) and a high gene density (Bennetzen and Chen 2008). Universal primers are commonly used when species-specific sequence of the target species is limited or unavailable (Verma and Singh 2003; McIntosh, Pacey-Miller et al. 2005; Heinze 2007; Shapter, Eggler et al. 2009). This approach is possible due to relative sequence conservation, synteny and micro-colinearity amongst grasses (Devos and Gale 1997; Gale and Devos 1998; Feuillet and Keller 2002). Universal primers were designed within conserved regions and used in polymerase chain reaction (PCR) to amplify the desired gene sequences in the target species, *M. stipoides*. Ideally, species specific primers derived from the genomic sequence of the target species are used to amplify desired regions. However, in cases such as *M. stipoides* where little genome sequence data are available, it may be possible to characterise homologues of rice grain yield genes by exploiting the sequence conservation and synteny between rice and other Poaceae.

Many important rice agronomic traits, such as grain yield are expressed in continuous phenotypic variation (Ashikari, Sakakibara et al. 2005) and are predominantly controlled by quantitative trait loci (QTL, Hittalmani, Shashidhar et al. 2002). Inheritance patterns of QTL are complex and dependant on
environmental factors (Hittalmani, Shashidhar et al. 2002; Song and Ashikari 2008). The three major genetic factors affecting rice grain yield include 1000 kernel weight, grain number per panicle and grain weight per panicle (Li, Pinson et al. 1997), with additional interactions between these genes. Due to the complexity of these interactions, understanding how each component gene contributes to a trait will contribute to a better understanding of how to manage such genetic information to produce improved crops. The focus of this chapter is on genes affecting grain yield.

Genes associated with grain yield in rice are well documented. Five of these genes (Table 2.1) related to grain number and weight, were used as reference genes. Summary of *Gn1a*, *GW2* and *GS3* were presented in the literature review (Chapter 1). In addition to the three genes, *qSW5* and *GW5* have also been characterised (Shomura, Izawa et al. 2008; Weng, Gu et al. 2008). Shomura, Izawa et al. (2008) discovered a QTL for rice seed width on chromosome 5 (*qSW5*), which explained 38.5% of natural variation in the F2 population. A deletion in *qSW5* resulted in a significant increase in sink size owing to an increase in cell number in the outer glume of the rice flower (Shomura, Izawa et al. 2008). *GW5* is a major QTL underlying rice width and weight, and is likely to act in the ubiquitin-proteasome pathway to regulate cell division during seed development (Weng, Gu et al. 2008). It has been mapped to a recombination hotspot on rice chromosome 5 (Weng, Gu et al. 2008).
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

**Table 2.1 Details of genes affecting grain yield in rice**

<table>
<thead>
<tr>
<th>Gene/protein name</th>
<th>Abbreviation</th>
<th>Grain yield trait</th>
<th>Publication</th>
<th>Genbank entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTL for grain number 1A</td>
<td><em>Gn1a</em></td>
<td>grain number</td>
<td>(Ashikari, Sakakibara et al. 2005)</td>
<td>AB205193</td>
</tr>
<tr>
<td>QTL for seed width on chromosome 5</td>
<td><em>qSW5</em></td>
<td>grain width</td>
<td>(Shomura, Izawa et al. 2008)</td>
<td>AB433345</td>
</tr>
<tr>
<td>QTL for grain size 3</td>
<td><em>GS3</em></td>
<td>grain length and weight</td>
<td>(Fan, Xing et al. 2006)</td>
<td>DQ355996</td>
</tr>
<tr>
<td>QTL for grain width and weight 2</td>
<td><em>GW2</em></td>
<td>grain width and weight</td>
<td>(Song, Huang et al. 2007)</td>
<td>EF447275</td>
</tr>
<tr>
<td>QTL for grain width and weight 5</td>
<td><em>GW5</em></td>
<td>grain width and weight</td>
<td>(Weng, Gu et al. 2008)</td>
<td>Not available</td>
</tr>
</tbody>
</table>

*O. sativa* and *M. stipoides* are both within the Erhartoideae, have a base chromosome number of 12 and have similar grain properties. It was hypothesised that due to gene conservation across the Poaceae generally and more specifically between *O. sativa* and *M. stipoides*, that grain yield genes identified in *O. sativa* would exist in *M. stipoides*. This had previously been shown to be the case for the *waxy* gene (Shapter, Eggler et al. 2009). The aim of this study was to characterise homologues of the grain yield related genes (Table 2.1) in *M. stipoides*.

Where possible target gene sequences were acquired through public gene banks for at least one Poaceae, outside the *Oryza* spp., and were used to identify conserved regions when aligned with the *O. sativa* gene sequence. These conserved regions were used as the basis to design universal primers. Following this assumption PCRs were performed using universal primers to characterise the gene sequence in *M. stipoides*. Once the gene sequence in *M. stipoides* was established, species specific primers were designed to confirm the authenticity of the sequence in *M. stipoides*. 
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

2.2 **Materials and Methods**

2.2.1 **Plant material and DNA extraction**

Native Seeds Pty Ltd ([http://nativeseeds.com.au/](http://nativeseeds.com.au/), last accessed 30\(^{th}\) October 2012) provided *M. stipoides* seed from their AR1 breeding line. Fresh leaf tissue was harvested from a single plant for DNA extraction. The method of DNA extraction was based on Carroll, Klimyuk et al. (1995) with the following exception: 3 grams of leaf tissue was ground to a fine powder in liquid nitrogen, using a mortar and pestle, and the quantity of reagents adjusted proportionally.

2.2.2 **Gene information retrieval from Genbank**

Sequences for rice grain yield related genes (Table 2.1) were obtained from NCBI’s Genbank website ([http://www.ncbi.nlm.nih.gov/genbank/](http://www.ncbi.nlm.nih.gov/genbank/), last accessed 30\(^{th}\) October 2012). Exon-intron boundaries were defined in the Genbank records. In cases where full genic structure or sequence information was not available, as with mRNA entries, searches using nucleotide blast (blastn) optimised for highly similar sequences (megablast), ([http://blast.ncbi.nlm.nih.gov/Blast.cgi](http://blast.ncbi.nlm.nih.gov/Blast.cgi), last accessed 9th November 2012) were performed on each gene to find the rice genomic sequence. Spidey was ([http://www.ncbi.nlm.nih.gov/spidey/](http://www.ncbi.nlm.nih.gov/spidey/), last accessed 8\(^{th}\) November 2012) used to align genomic sequences to the mRNA sequences, in order to identify exon-intron boundaries.

2.2.3 **Sequence alignment and universal PCR primer design**

Searches using nucleotide blast (blastn) optimised for highly similar sequences (megablast) were performed on each gene to find homologues in other Poaceae. When blastn for highly similar sequences did not yield suitable alignments, stringency was reduced to identify sequences with reduced similarity. The sorghum trace archive ([http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PROGRAM=blastn&BLAST_SPEC=TraceArchive&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch](http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PROGRAM=blastn&BLAST_SPEC=TraceArchive&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch), last accessed 10\(^{th}\) November 2012) was also used where the blastn for somewhat
similar sequences was insufficient. Alignments between the exons from the rice reference gene and its homologues were conducted using Sequencher® V4.6 (Gene Codes Corporation, Ann Arbor, MI, USA). Where more than one homologue was found, the alignment with the most sequence matches was accepted. Regions with a conserved sequence of at least 17 contiguous bases and a maximum of two bases with degeneracy at the start and the end of the coding region of each gene were identified as possible sites for primer placements. This allowed for the amplification of putative gene homologues of the rice reference genes in M. stipoides. Primer design was conducted using Clone Manager Professional 9 (Scientific & Educational Software, Cary, NC, USA). PCR amplicons were limited to 3 kbp, as recommended by Roche Taq DNA Polymerase protocol. Clone Manager Professional 9 primer criteria for primers were set as follows: a minimum of 17 bases at 45-65% GC content, Tm between 55-65°C and the Tm difference between forward and reverse primer ±4 °C, stability of 5’ versus 3’ at 1.2 kcals, less than three matches for 3’ dimers, less than seven for any dimer, less than five base runs, less than three repeats of di-nucleotide pairs, less than eight protein degeneracy within 6 bp of 3’ end, less than or equal to one G’s or C’s at 3’ end. A maximum of two bases of degeneracy was allowed. Primers that met these criteria were chosen whenever available and a maximum of two unmet criteria was accepted.

2.2.4 Preliminary PCR amplification using universal primers

PCRs were conducted using Roche Taq DNA Polymerase (Roche Products Pty Limited, Dee Why, NSW, Australia) at 1 unit/reaction with 1X PCR Buffer, primers at 0.2µM each (when a degenerate primer was used the concentration was increased to 0.4µM), with 2mM magnesium chloride and 10 ng of genomic DNA. The PCR program was set for one cycle of denaturation at 96°C for two minutes, followed by 10 touchdown cycles of denaturation at 96°C for 30 seconds, annealing at the touchdown temperature for respective primer pairs for 30 seconds, extension at 68°C for one minute per kbp of PCR product. A further 30 cycles of denaturation at 96°C for 30 seconds, annealing at 10°C lower than the temperature required by respective primer pairs for 30 seconds, extension at 68°C for one minute per kbp of PCR product. A final cycle at 72°C for five minutes allowed for further extension.
Optimisation was done using the addition of 3% DMSO and 3% glycerol and gradient PCR. The PCR products were run on 1% agarose gel electrophoresis stained with ethidium bromide to confirm amplification and then sequenced to confirm homology. Where a single PCR band was amplified in a reaction, Exosap-IT (GE Healthcare Bio-Sciences Pty. Ltd., Rydalmere, NSW, Australia) was used to purify the PCR product prior to direct Sanger sequencing. Multiple PCR bands in a reaction were separated via gel electrophoresis and individually excised then purified using QIAquick Gel Extraction Kit (Qiagen Pty Ltd, Doncaster, VIC, Australia) prior to direct sequencing.

2.2.5 DNA sequencing

Sequencing reactions were prepared using ABI PRISM® Big DyeTM Terminator (BDT)V3.1 (Applied Biosystems, Mulgrave, VIC, Australia) at one eighth reactions and then analysed by capillary electrophoresis on an ABI 3730 Genetic Analyzer, performed by Southern Cross Plant Genetics (SCPG, Lismore, NSW, Australia).

2.2.6 Alignment of sequencing results to reference gene sequences

The resulting sequences provided by SCPG were edited and aligned back to the rice reference genes using Sequencher® V4.6, ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/, last accessed 30th October 2012) and Clone Manager Professional 9.

2.2.7 Genome Walking

To extend the gene sequence from the universal primer sites out to the 5’ and 3’ UTR, the BD GenomeWalkerTM Universal Kit (BD Biosciences, San Jose, CA, USA) was used, according to the manufacturer’s instructions. Genome walking primers were designed using Clone Manager Professional 9.
2.2.8 Species specific whole gene PCR primer design

Species specific whole gene PCR primers were designed in the 5’ and 3’ UTR. Primer design was conducted using Clone Manager Professional 9 on *M. stipoides* sequence generated from Genome Walking results. Primers were designed using the same primer criteria in universal primer design with the exception that no degeneracy was allowed. PCR optimisation was done using temperature gradient to produce robust PCR products.

2.2.9 Cloning and DNA sequencing Alignment of sequencing results to reference gene sequences

Direct sequencing of PCR products often results in ambiguous base calls. Cloning was performed to identify specific base calls associated with specific alleles. PCR products were run on 1% low-melting agarose gel. Products were excised individually to make sure that individual bands were extracted prior to cloning. The individual gel slice was melted in a thermocycler at 72°C for 1 minute and then held at 37°C immediately prior to ligation into the vector. Three µl of molten gel was used in the standard ligation reaction mix of pGEM®-T Easy Vector II System (Promega Corporation, Madison, WI, USA) as per manufacturer’s protocol. Clone screening for size was done by PCR using Platinum Taq in the manufacturer’s recommended reaction mix. Desired clones were isolated and amplified using the Templiphi™ System (GE Healthcare, Rydalmere, NSW, Australia) following the manufacturer’s instructions. Sequencing reactions were prepared using the standard BDT V3.1 one eighth reactions and 5 µl of diluted Templiphi™ product as template. Introduction in Body text

2.2.10 Whole gene PCR

Long range PCRs to amplify the whole gene of GW2 in *M. stipoides* were attempted using three different DNA polymerases, Herculase II Fusion DNA Polymerases (Agilent Technologies Inc., Santa Clara, CA, USA), Accuprime GC-Rich DNA Polymerase (Invitrogen Australia Pty. Ltd., Mulgrave, VIC, Australia), and Iproof (Bio-rad Life Science, Hercules, CA, USA) under a range of different
cycling conditions as recommended by respective protocols of the DNA polymerases.

2.3 Results

2.3.1 Retrieval of candidate gene information from Genbank

Available GW2 and Gn1a Genbank entries included mRNA sequences where exon-intron boundary information was not available. For GW2, Genbank accession AP005004 (Oryza sativa Japonica Group genomic DNA, chromosome 2, PAC clone: P0503B05) was the best genomic match. Alignment between GW2 and AP005004 using Spidey identified eight exons in the span of 5881 bp (Figure 2.3.1.1). For Gn1a, Genbank accession AP003200 (Oryza sativa Japonica Group genomic DNA, chromosome 1, BAC clone: B1046G12) was the best genomic match. Alignment between Gn1a and AP003200 using Spidey identified four exons in the span of 4688 bp (Figure 2.3.1.2). GS3 Genbank entry showed five exons in the span of 5968 bp. No Genbank entry was found for GW5. For qSW5, the Genbank entry showed one exon 396 bp long.
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

**Figure 2.3.1.1** Spidey alignment between *GW2* mRNA and *Oryza sativa* Japonica Group genomic DNA, chromosome 2, PAC clone: P0503B05 resulted in the identification of eight *GW2* exons.
Figure 2.3.1.2 Spidey alignment between Gn1a mRNA and Oryza sativa Japonica Group genomic DNA, chromosome 1, BAC clone: B1046G12 resulted in the identification of four Gn1a exons

The highest blastn score for GW2 was EU333863 Hordeum vulgare subsp. vulgare Yrg1 mRNA, complete cds. Primers were designed to regions conserved between GW2 (EF477275) and EU333863 (Figure 2.3.1.3 and 2.3.1.4). The highest blastn score for Gn1a was Sorghum bicolor hypothetical protein, mRNA (XM_002454958.1). Primers were designed to regions conserved between Gn1a (AB205193) and XM_002454958.1 (Figure 2.3.1.5). For GS3, blastn results were inadequate in 2008, when the search was conducted. Sorghum trace archive results were used to align with GS3 and primers were designed to regions conserved between GS3 and the sorghum trace archive results (Figure 2.3.1.6). For qSW5, the best blastn result was XM_002439359 (Sorghum bicolor hypothetical protein, mRNA).
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Figure 2.3.1.3  Alignment between EU333863 *Hordeum vulgare* subsp. vulgare Yrg1 mRNA and rice *GW2* mRNA (EF477275) showing conserved regions where primers can be designed
Figure 2.3.1.4  Alignment between EU333863 *Hordeum vulgare* subsp. vulgare Yrg1 mRNA and rice GW2 mRNA (EF477275) showing intron-exon structure and positions of primers
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised M. stipoides
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Figure 2.3.1.5  Alignment between *Sorghum bicolor* hypothetical protein (XM_002454958) and rice *Gn1a* mRNA (AB205193) showing conserved regions where primers can be designed
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Figure 2.3.1.6 Alignments between *Sorghum bicolor* hypothetical protein (XM_002454958) or sorghum trace and rice *Gn1a* mRNA (AB205193) showing intron-exon structure and positions of primers
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

### 2.3.2 Universal PCR primer design

Primers that met the previously set criteria were designed for *GW2* and *Gn1a* (Table 2.3.2.1). These primers were used in the preliminary PCRs. The positions of the primers are shown on Figures 2.3.2.1-2.3.2.19. Details of the primer pairs, PCR conditions and PCR results are summarised in Table 2.3.2.2. Photos of agarose gel electrophoresis of the PCR products from Table 2.3.2.2 are shown on Figures 2.3.2.20 and 2.3.2.21.

For *GS3*, exons 1 to 4 had sufficient conserved region with sorghum. Exon 5 did not align with the sorghum sequence and as a result no conserved region was found. The first two exons of the gene were GC-rich, causing primers to have high Tm, whereas exons three and four were GC-poor, causing primers to have low Tm. There were insufficient primer pairs that met the criteria and PCR was not conducted. There were no adequate conserved regions between *qSW5* and sorghum accession and so PCR primers were not able to be designed. In summary, primer pairs were available for *GW2* and *Gn1a*, whereas there was insufficient information from which to design primers for *GS3* and *qSW5*. 
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised M. stipoides

Table 2.3.2.1 Details of universal primers for GW2 and Gn1a

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Direction</th>
<th>Sequence 5' to 3' Nucleotides</th>
<th>% GC</th>
<th>Tm °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>gw2x1f1</td>
<td>forward</td>
<td>CAGGGGCTGTACGAGGCA</td>
<td>18</td>
<td>66</td>
</tr>
<tr>
<td>gw2x1f2</td>
<td>forward</td>
<td>AGGGGCTGTACGAGGCA</td>
<td>17</td>
<td>64</td>
</tr>
<tr>
<td>gw2x5f1</td>
<td>forward</td>
<td>ATGCGCRCAGCAAGCACCTTC</td>
<td>19</td>
<td>57</td>
</tr>
<tr>
<td>gw2x5r1</td>
<td>reverse</td>
<td>GGTATGTGCTGCAATTATCTTC</td>
<td>20</td>
<td>45</td>
</tr>
<tr>
<td>gw2x6r1</td>
<td>reverse</td>
<td>CATTACGTTGGTCGCCAG</td>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>gw2x6r2</td>
<td>reverse</td>
<td>TATGCCTAGAAAGGAAAG</td>
<td>18</td>
<td>38</td>
</tr>
<tr>
<td>gw2x7f1</td>
<td>forward</td>
<td>TGGTTATGGAAGCGATGTGGG</td>
<td>20</td>
<td>45</td>
</tr>
<tr>
<td>gw2x7r1</td>
<td>reverse</td>
<td>GAACGCAAAATCCTGTCACC</td>
<td>18</td>
<td>55</td>
</tr>
<tr>
<td>gw2x8f1</td>
<td>forward</td>
<td>TGCAACATTGCTGGTGGAAG</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>gw2x8r1</td>
<td>reverse</td>
<td>TTGGTGCTATCCCGCTCCAG</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td>gw2x8r2</td>
<td>reverse</td>
<td>TGTTTCGATGCTCCTC</td>
<td>17</td>
<td>47</td>
</tr>
<tr>
<td>gn1ax1f4s</td>
<td>forward</td>
<td>ACATCAGCGCCGCTGCTG</td>
<td>17</td>
<td>70</td>
</tr>
<tr>
<td>gn1ax2r1s</td>
<td>reverse</td>
<td>GTGGGTGCTCGTGCTGTAC</td>
<td>19</td>
<td>63</td>
</tr>
<tr>
<td>gn1ax3f1s</td>
<td>forward</td>
<td>TCTGGGAGCTGCCCCAC</td>
<td>18</td>
<td>83</td>
</tr>
<tr>
<td>gn1ax4r1s</td>
<td>reverse</td>
<td>CGGCTGGCCGCTCAGGG</td>
<td>17</td>
<td>76</td>
</tr>
<tr>
<td>dgn1ax2r1</td>
<td>reverse</td>
<td>CTACGTCGAGGGTCCGGT</td>
<td>18</td>
<td>66</td>
</tr>
<tr>
<td>dgn1ax3f1</td>
<td>forward</td>
<td>GTTCGYGTTCGYYGCAAGCAGC</td>
<td>21</td>
<td>61</td>
</tr>
<tr>
<td>dgn1ax4r1</td>
<td>reverse</td>
<td>GCCCAGGAGGATTTTCWGG</td>
<td>19</td>
<td>55</td>
</tr>
</tbody>
</table>

Figure 2.3.2.1 Position of forward primers gw2x1f1 and gw2x1f2 on exon 1 of GW2
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

---

**Figure 2.3.2.2** Position of forward primer gw2x5f1 on exon 5 of *GW2*

**Figure 2.3.2.3** Position of reverse primer gw2x5r1 on exon 5 of *GW2*

**Figure 2.3.2.4** Position of reverse primer gw2x6r2 on exon 6 of *GW2*

**Figure 2.3.2.5** Position of reverse primer gw2x6r2 on exon 6 of *GW2*
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

**Figure 2.3.2.6**  Position of forward primer gw2x7f1 on exon 7 of *GW2*

**Figure 2.3.2.7**  Position of reverse primer gw2x7r1 on exon 7 of *GW2*
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

<table>
<thead>
<tr>
<th>Accession</th>
<th>Name</th>
<th>Sequence</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>x8EF4472750s</td>
<td>gw2x8f1</td>
<td>GAGCAGGGAAATAGGGGATCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCGGCTGAAATAGGGGATCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTTCTGTGCTGATGCGGTAATAGGGGATCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATGCTGTGCAACTAGGCGATTCAGATGCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGGAGGCAATAGGGGATCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TGCAATGCGGATGCGGATCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>360</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCAGGAGGGCGAATAATAGGGGATCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>420</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CAAGCTGATGCTGATGCGGATGCGGATCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>480</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GAGAGCGAGATGATGCTGATGCGGATGCGGATCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>540</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TGCAAGGAGGGATGCGGATGCGGATCGCTGCTGATGCAACCTTTATGCTGTAACCTGAGCCATCT</td>
<td>564</td>
</tr>
</tbody>
</table>

**Figure 2.3.2.8** Position of forward primer gw2x8f1 on exon 8 of *GW2*
Figure 2.3.2.9  Position of reverse primer gw2x8r1 on exon 8 of GW2
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Figure 2.3.2.10 Position of reverse primer gw2x8r2 on exon 8 of *GW2*
Figure 2.3.2.11  Position of forward primer gn1ax1f4s on exon 1 of Gn1a
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

### Figure 2.3.2.12 Position of reverse primer gn1ax1r1s on exon 1 of Gn1a

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1AE205193gn1a</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td>60</td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>120</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>130</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>240</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>300</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>360</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>420</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>490</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>540</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>600</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
<tr>
<td>x1AE205193gn1a</td>
<td>646</td>
</tr>
<tr>
<td>gn1ax1r1s</td>
<td><img src="image" alt="Alignment" /></td>
</tr>
</tbody>
</table>

---

55
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

**Figure 2.3.2.13** Position of forward primer gn1ax2f1z on exon 2 of *Gn1a*
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

---

**Figure 2.3.2.14** Position of reverse primer gn1ax2r1s on exon 2 of *Gn1a*

---

**Figure 2.3.2.15** Position of forward primer gn1ax3f1s on exon 3 of *Gn1a*
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Figure 2.3.2.16  Position of reverse primer gn1ax4r1s on exon 4 of *Gn1a*

![CLUSTAL 2.1 multiple sequence alignment](image)

Figure 2.3.2.17  Position of degenerate forward primer dgn1ax2r1 on exon 2 of *Gn1a*
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

---

**Figure 2.3.2.18** Position of degenerate forward primer dgn1ax3f1 on exon 3 of *Gn1a*

**Figure 2.3.2.19** Position of degenerate reverse primer dgn1ax4r1 on exon 4 of *Gn1a*
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Table 2.3.2.2  Primer pairs and PCR details

<table>
<thead>
<tr>
<th>Pair</th>
<th>Forward</th>
<th>Reverse</th>
<th>Tm °C</th>
<th>Tm °C</th>
<th>Distance (bp)</th>
<th>Denaturation</th>
<th>Annealing</th>
<th>Extension</th>
<th>Denaturation</th>
<th>Annealing</th>
<th>Extension</th>
<th>Extension</th>
<th>Stop</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gw2x5f1</td>
<td>gw2x6r1</td>
<td>55</td>
<td>259</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>50-40°C 30 secs</td>
<td>68°C 30 secs</td>
<td>96°C 30 secs</td>
<td>40°C 30 secs</td>
<td>68°C 30 secs</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>single band at expected size</td>
</tr>
<tr>
<td>2</td>
<td>gw2x5f1</td>
<td>gw2x6r2</td>
<td>52</td>
<td>321</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>50-40°C 30 secs</td>
<td>68°C 30 secs</td>
<td>96°C 30 secs</td>
<td>40°C 30 secs</td>
<td>68°C 30 secs</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>single band at expected size</td>
</tr>
<tr>
<td>3</td>
<td>gw2x8f1</td>
<td>gw2x8r2</td>
<td>55</td>
<td>263</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>50-40°C 30 secs</td>
<td>68°C 30 secs</td>
<td>96°C 30 secs</td>
<td>40°C 30 secs</td>
<td>68°C 30 secs</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>single band at expected size</td>
</tr>
<tr>
<td>4</td>
<td>gw2x1f1</td>
<td>gw2x5r1</td>
<td>58</td>
<td>2028</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>5</td>
<td>gw2x1f1</td>
<td>gw2x6r1</td>
<td>55</td>
<td>2164</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>6</td>
<td>gw2x1f1</td>
<td>gw2x6r2</td>
<td>52</td>
<td>2226</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>7</td>
<td>gw2x5f1</td>
<td>gw2x7r1</td>
<td>61</td>
<td>2081</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>8</td>
<td>gw2x7f1</td>
<td>gw2x8r1</td>
<td>66</td>
<td>1602</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>9</td>
<td>gw2x7f1</td>
<td>gw2x8r2</td>
<td>55</td>
<td>1795</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 2 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>10</td>
<td>gw2x1f1</td>
<td>gw2x7r1</td>
<td>61</td>
<td>3986</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 4 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 4 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>11</td>
<td>gw2x5f1</td>
<td>gw2x8r1</td>
<td>66</td>
<td>3658</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 4 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 4 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>12</td>
<td>gw2x5f1</td>
<td>gw2x8r2</td>
<td>55</td>
<td>3851</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 4 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 4 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>13</td>
<td>gw2x1f1</td>
<td>gw2x8r1</td>
<td>66</td>
<td>5563</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 5 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 5 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>14</td>
<td>gw2x1f1</td>
<td>gw2x8r2</td>
<td>55</td>
<td>5756</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>52-42°C 30 secs</td>
<td>68°C 5 mins</td>
<td>96°C 30 secs</td>
<td>42°C 30 secs</td>
<td>68°C 5 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>15</td>
<td>gn1ax1f4s</td>
<td>gn1ax2r1s</td>
<td>65</td>
<td>624</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>64-54°C 30 secs</td>
<td>68°C 1 min</td>
<td>96°C 30 secs</td>
<td>54°C 30 secs</td>
<td>68°C 1 min</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>16</td>
<td>gn1ax3f1s</td>
<td>gn1ax4r1s</td>
<td>67</td>
<td>674</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>64-54°C 30 secs</td>
<td>68°C 1 min</td>
<td>96°C 30 secs</td>
<td>54°C 30 secs</td>
<td>68°C 1 min</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>no visible band</td>
</tr>
<tr>
<td>17</td>
<td>gn1ax1f4s</td>
<td>gn1ax2r1l</td>
<td>-</td>
<td>731</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>64-54°C 30 secs</td>
<td>68°C 1 min</td>
<td>96°C 30 secs</td>
<td>54°C 30 secs</td>
<td>68°C 1 min</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>18</td>
<td>dgn1ax3f1 - gn1ax4r1s</td>
<td>67</td>
<td>757</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>64-54°C 30 secs</td>
<td>68°C 1 min</td>
<td>96°C 30 secs</td>
<td>54°C 30 secs</td>
<td>68°C 1 min</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>dgn1ax3f1 - dgn1ax4r1l</td>
<td>-</td>
<td>981</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>64-54°C 30 secs</td>
<td>68°C 1 min</td>
<td>96°C 30 secs</td>
<td>54°C 30 secs</td>
<td>68°C 1 min</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>gn1ax1f4s</td>
<td>gn1ax4r1s</td>
<td>67</td>
<td>4185</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>64-54°C 30 secs</td>
<td>68°C 5 mins</td>
<td>96°C 30 secs</td>
<td>54°C 30 secs</td>
<td>68°C 5 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>non-specific bands</td>
</tr>
<tr>
<td>21</td>
<td>gn1ax1f4s</td>
<td>gn1ax4r1l</td>
<td>-</td>
<td>4409</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>64-54°C 30 secs</td>
<td>68°C 5 mins</td>
<td>96°C 30 secs</td>
<td>54°C 30 secs</td>
<td>68°C 5 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>no visible band</td>
</tr>
<tr>
<td>22</td>
<td>gn1ax1f4s</td>
<td>gn1ax1r1s</td>
<td>69</td>
<td>322</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>64-54°C 30 secs</td>
<td>68°C 30 secs</td>
<td>96°C 30 secs</td>
<td>54°C 30 secs</td>
<td>68°C 30 secs</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>single band at expected size</td>
</tr>
<tr>
<td>23</td>
<td>gn1ax2f1z</td>
<td>dgn1ax2r1l</td>
<td>-</td>
<td>252</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>64-54°C 30 secs</td>
<td>68°C 30 secs</td>
<td>96°C 30 secs</td>
<td>54°C 30 secs</td>
<td>68°C 30 secs</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td>two bands at expected size</td>
</tr>
</tbody>
</table>
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

**Figure 2.3.2.20**  PCR results for *GW2* using primer pairs from Table 2.3.2.2

**Figure 2.3.2.21**  PCR results for *Gn1a* using primer pairs from Table 2.3.2.2
Visible bands from Figures 2.3.2.20 and 2.3.2.21 were sequenced using direct PCR sequencing when there was a single visible band. Otherwise bands were extracted from the gel and then sequenced. PCR products 1, 2, 3, 4, 9, 22 and 23 were successful in generating *M. stipoides* sequences.

### 2.3.3 Sequence characterisation in *M. stipoides*

*GW2* universal primers were designed from sequence regions conserved between rice and barley (details in Tables 2.3.2.1 and 2.3.2.2). PCR products resulting from these primers were able to capture partial sequence of exons 1, 5 and 8 in *M. stipoides* (Figures 2.3.3.1-2.3.3.3). The corresponding sequences of *M. stipoides* were highly similar to those in rice.

**Figure 2.3.3.1**  Sequence comparison of GW2 exon 1 in rice (*x1GW2_Osat*) and *M. stipoides* (*x1GW2_Msti*) resulting from sequencing PCR products 4 (Figure 2.3.2.20)
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

**Figure 2.3.3.2** Sequence comparison of GW2 exon 5 in rice (x5GW2_Osat) and *M. stipoides* (x5GW2_Msti) resulting from sequencing PCR products 1, 2 and 4 (Figure 2.3.2.20)

**Figure 2.3.3.3** Sequence comparison of GW2 exon 8 in rice (x8GW2_Osat) and *M. stipoides* (x8GW2_Msti) resulting from sequencing PCR products 3 and 9 (Figure 2.3.2.20)
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

*Gn1a* sequence characterisation in *M. stipoides* using universal primers was able to capture partial sequence of exons 1 and 2 in *M. stipoides* (Figures 2.3.3.4 and 2.3.3.5). The sequence in *M. stipoides* was strongly homologous to rice. Primers placed in exons 3 and 4 failed to capture any *Gn1a* sequence in *M. stipoides*.

---

**Figure 2.3.3.4** Sequence comparison of *Gn1a* exon 1 in rice (*x1Gn1a_Osat*) and *M. stipoides* (*x1Gn1a_Msti*) resulting from sequencing PCR product 22 (Figure 2.3.2.21)

---
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

**Figure 2.3.3.5** Sequence comparison of **Gn1a** exon 2 in rice (x2AB205193gn1a) and *M. stipoides* (gn1ax2dr1_F03) resulting from sequencing PCR product 23 (Figure 2.3.2.21)

### 2.3.4 Genome Walking

The species specific genome walking primers were able to be designed from the universal primer amplification derived sequence, of both **Gn1a** and **GW2** (Table 2.3.4). The positions of the genome walking primers (Table 2.3.4) in the exons are shown on Figures 2.3.4.1 to 2.3.4.8.

Genome walking was able to capture the sequence of the first half of exon 1 and 5’ UTR (Figure 2.3.4.1) and second half of exon 8 and 3’ UTR (Figure 2.3.4.2) in **GW2**. This enabled primers to be designed so that they were able to amplify the entire **GW2** gene. However, genome walking was unable to capture additional sequence for **Gn1a** and this gene was subsequently excluded from further analysis.
### Table 2.3.4 Details of genome walking primers for GW2 and Gn1a

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Direction</th>
<th>Sequence 5’ to 3’</th>
<th>Nucleotides</th>
<th>% GC</th>
<th>Tm °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>gw2x1op1</td>
<td>Reverse</td>
<td>CGAGGAGTGCCCCCATCTGCTTCT</td>
<td>24</td>
<td>62</td>
<td>72</td>
</tr>
<tr>
<td>gw2x1ip5</td>
<td>Reverse</td>
<td>CTGCTACCCGGGGGCGGCAGATGC</td>
<td>24</td>
<td>78</td>
<td>76</td>
</tr>
<tr>
<td>gw2x8op1</td>
<td>Forward</td>
<td>AGTTCAACCAGATAGCTGGGATGGGATAG</td>
<td>28</td>
<td>50</td>
<td>68</td>
</tr>
<tr>
<td>gw2x8ip3</td>
<td>Forward</td>
<td>GAGAGGAAGGAGAGTGCTCAACTGAC</td>
<td>26</td>
<td>53</td>
<td>68</td>
</tr>
<tr>
<td>gn1ax1op1</td>
<td>Reverse</td>
<td>CTCCGTCTCGGTGATGGGCCGCTACATCGA</td>
<td>30</td>
<td>63</td>
<td>77</td>
</tr>
<tr>
<td>gn1ax1ip1</td>
<td>Reverse</td>
<td>CAGCGGCTCTCCGTCGTCGATG</td>
<td>26</td>
<td>76</td>
<td>80</td>
</tr>
</tbody>
</table>

**Figure 2.3.4.1** Comparison between partial *M. stipoides* GW2 exon 1 sequence (Msgw2x1-5Tf_C01) and genome walking outer primer gw2x1op1 showing the position of the primer in the exon
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Figure 2.3.4.2  Comparison between partial *M. stipoides* GW2 exon 1 sequence (Msgw2x1-5Tf_C01) and genome walking inner primer gw2x1ip5 showing the position of the primer in the exon

---

Figure 2.3.4.3  Comparison between partial *M. stipoides* GW2 exon 8 sequence (gw2x8r1_H01) and genome walking outer primer gw2x8op1 showing the position of the primer in the exon
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

---

**Figure 2.3.4.4** Comparison between partial *M. stipoides* GW2 exon 8 sequence (gw2x8r1_H01) and genome walking inner primer gw2x8ip3 showing the position of the primer in the exon.

---

**Figure 2.3.4.5** Comparison between partial *M. stipoides* Gn1a exon 1 sequence and genome walking outer primer gn1ax1op1 showing the position of the primer in the exon.
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Figure 2.3.4.6 Comparison between partial *M. stipoides* Gn1a exon 1 sequence and genome walking inner primer gn1axlip1 showing the position of the primer in the exon
Figure 2.3.4.7 Sequence comparison of GW2 exon 1 in rice (x1EF447275Osat) and sequence generated by cloning in M. stipoides (gw2x1-5c1m13f_A0) showing partial 5' UTR region and partial intron 1
Figure 2.3.4.8  Sequence comparison of GW2 exon 8 in rice (x8EF447275Osat) and sequence generated by cloning in M. stipoides (gw2x4-8c1m13f_G0) showing partial 3’ UTR region and partial intron 7
2.3.5 Cloning to confirm sequence accuracy

Two PCR amplifications were needed to generate products that spanned all the exons of GW2. The 5’ amplicon was generated using primer pair EMgw2x1f4 and EMgw2x5r2 and the 3’ amplicon used EMgw2x4f2 and EMgw2x8r3 (Table 2.3.5.1). Table 2.3.5.2 shows the details for the primer pairs used, PCR conditions and results. Figure 2.3.5.1 shows the bands produced by the primer pairs 1 and 2 from Table 2.3.5.2. Positions of the primers EMgw2x1f4, EMgw2x5r2, EMgw2x4f2 and EMgw2x8r3 are shown on Figures 2.3.5.2-2.3.5.5.

Table 2.3.5.1 PCR primers for GW2

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Direction</th>
<th>Sequence 5' to 3'</th>
<th>Nucleotides</th>
<th>% GC</th>
<th>Tm °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMgw2x1f4</td>
<td>forward</td>
<td>ATGGGGAACAGGATAGGG</td>
<td>18</td>
<td>55</td>
<td>59</td>
</tr>
<tr>
<td>EMgw2x1r2</td>
<td>reverse</td>
<td>GAAGCAGATGGGGCCTCC</td>
<td>19</td>
<td>63</td>
<td>64</td>
</tr>
<tr>
<td>EMgw2x2f1</td>
<td>forward</td>
<td>TACTACCCAAGCTTAAACCG</td>
<td>20</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>EMgw2x4f2</td>
<td>forward</td>
<td>TGCCCCATTGCGAAAACTC</td>
<td>19</td>
<td>47</td>
<td>60</td>
</tr>
<tr>
<td>EMgw2x6r1</td>
<td>reverse</td>
<td>TTATGCCTAATGGGAAAGG</td>
<td>20</td>
<td>40</td>
<td>56</td>
</tr>
<tr>
<td>EMgw2x8f3</td>
<td>forward</td>
<td>GCAGGAATATGCGGAAATAG</td>
<td>23</td>
<td>43</td>
<td>60</td>
</tr>
<tr>
<td>EMgw2x8r3</td>
<td>reverse</td>
<td>GCAAGGATTGGCATGTGTGAG</td>
<td>22</td>
<td>50</td>
<td>64</td>
</tr>
</tbody>
</table>

Table 2.3.5.2 Primer pairs and PCR details

<table>
<thead>
<tr>
<th>Pair</th>
<th>Forward</th>
<th>Tm °C</th>
<th>Reverse</th>
<th>Tm °C</th>
<th>Distance (bp)</th>
<th>Denaturation Forward</th>
<th>Denaturation Reverse</th>
<th>Annealing</th>
<th>Extension Forward</th>
<th>Extension Reverse</th>
<th>Stop</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 EMgw2x1f4</td>
<td>59</td>
<td>63</td>
<td>2161</td>
<td>60</td>
<td>59°C 2 mins</td>
<td>96°C 30 secs</td>
<td>59°C 30 secs</td>
<td>68°C 2 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td></td>
<td>two bands at expected size</td>
</tr>
<tr>
<td>2 EMgw2x4f2</td>
<td>60</td>
<td>64</td>
<td>4215</td>
<td>60</td>
<td>96°C 2 mins</td>
<td>96°C 30 secs</td>
<td>57°C 30 secs</td>
<td>68°C 4 mins</td>
<td>72°C 5 mins</td>
<td>15°C 1 hour</td>
<td></td>
<td>single band at expected size</td>
</tr>
</tbody>
</table>
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Figure 2.3.5.1 Gel electrophoresis photo showing PCR results for *GW2* using primer pairs from Table 2.3.5.2 (1 covers exons 1-5 and 2 covers exons 4-8)

Visible bands from Figure 2.3.5.1 were sequenced using direct PCR sequencing when there was a single visible band or otherwise bands were extracted from the gel and then sequenced. The sequence results confirmed that the bands were capturing the correct regions of the genome.

**Figure 2.3.5.2** Sequence comparison of *GW2* exon 1 in *M. stipoides* (x1Gw2Mstipoides) and primer EMgw2x1f4 showing the position of the primer at the start of exon 1
Figure 2.3.5.3 Sequence comparison of *GW2* exon 5 in *M. stipoides* (x5Gw2Mstipoides) and primer EMgw2x5r2 showing the position of the primer at the end of exon 5

Figure 2.3.5.4 Sequence comparison of *GW2* exon 4 in *M. stipoides* (x4Gw2Mstipoides) and primer EMgw2x4f2 showing the position of the primer at the start of exon 4
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Figure 2.3.5.5  Sequence comparison of GW2 exon 5 in *M. stipoides* (x8Gw2Mstipoides) and primer EMgw2x8r3 showing the position of the primer at the end of exon 8

Cloning produced complete putative exon sequence for seven of the eight exons, the exception being exon 7 (Table 2.3.5.3). Whole gene sequence was unable to be determined at this stage because of the high variability in the intronic regions. All the putative exons in *M. stipoides* showed strong identity to the corresponding exons in rice, differing by only a few bases (between 87-96% sequence identities).
## Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Table 2.3.5.3  Exon sequence of GW2 determined by cloning overlapping PCR amplicons

<table>
<thead>
<tr>
<th>Exon</th>
<th>Sequence</th>
<th>% Identity with rice exon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ATGGGGGACAGGATAGGGGGAGGAGAAAGCCGGGGGTGGAGGAGCGGTACACGCGCCAGGGGACTGT</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>CCGGGGGCCCGACGATCCGCCGCTGGCCGCGCCCTGGAAGTGCGGCTTACGCTTCTTGAGGGAGATCAT</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>TACTACCCAAGGCCTTAACCGATCAAAGTGTTTGCTCGAAAGGGGTATGCACTG</td>
<td>92</td>
</tr>
<tr>
<td>3</td>
<td>AGTGCTTCATGCAAGAACCAAATGCATACTGCTGCTAGCCTACACCA</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>ATGCCCATCTGCAAAAATCCAGTTATGCTGTTGAGATCGTGGTGTGAAAGAAAGGAGGAGGAGCAT</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>AGAGCAATTT</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GAAAGAAGAACAAATAGGTTGTCTTTCTTAGCAGAAACACTCAAAGCAGACTGAAGTGAGATCGAGATAT</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>TTGCAACAGACATTCTTTTCAG</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>TGCCATCATACCGATGTACTGAAGCAAGAACAAATGCTGTCATGCAACCTCTGCTGCCCCAGACTAG</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>CATGCGTCTCATTACCATGCAATCG</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>GACGAGGGAGTACAAGGAGAATTCCCTGTTTGGCGAGTTTTATGCTGTAATTGGACCTCCCTGAGTTGAGTC</td>
<td>92</td>
</tr>
<tr>
<td>8</td>
<td>AATCGTTCCTCCAGCTGTGGCTCTCTAGAAATACCTCAATCCCTGGAGATTCTTTCTTATGCTGCTTGGAT</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>AGCGACAGCGACAAACGCTAGTTCTTCTTTCTTACATTGCGGAGCTGACAGTTCTCCCTGACGAGTCC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCCGACAGCTCAACATTCTGCTGTTGGAAGCATGTGAATCCTGAGGACATGGTCACAGAAGGAGGATGAGT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAAATTTACATTGCTGCTGCTGCAACACTGAGCTGACAGAGCCGAGACTGTCATCGTTACCTTTATCTTG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAAATTTACATTGCTGCTGCTGCAACACTGAGCTGACAGAGCCGAGACTGTCATCGTTACCTTTATCTTG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAAATTTACATTGCTGCTGCTGCAACACTGAGCTGACAGAGCCGAGACTGTCATCGTTACCTTTATCTTG</td>
<td></td>
</tr>
</tbody>
</table>
2.3.6 Alignment of sequencing results to reference gene sequences

Six primers were designed in the UTR and exonic regions for whole gene PCR in a single amplicon (Table 2.3.6). A number of combinations of primer pairings were trialled using different PCR reactions. PCR amplifications using whole genome primers located either within the exonic regions or the 5’ and 3’ UTR, were not successful in generating a single specific product that was able to be sequenced.

Table 2.3.6 Whole gene primers for GW2

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Direction</th>
<th>Sequence 5’ to 3’</th>
<th>Nucleotides</th>
<th>% GC</th>
<th>Tm °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>wg-gw2x1f1</td>
<td>forward</td>
<td>TACAGTGAGGTGGTGCGGAAG</td>
<td>21</td>
<td>57</td>
<td>66</td>
</tr>
<tr>
<td>wg-gw2x1f3</td>
<td>forward</td>
<td>ATCTACACTGAGGCGGTCTG</td>
<td>20</td>
<td>55</td>
<td>63</td>
</tr>
<tr>
<td>wg-gw2x8r1</td>
<td>reverse</td>
<td>CTATAGGGCACGCGGTGGTC</td>
<td>19</td>
<td>63</td>
<td>64</td>
</tr>
<tr>
<td>wg-gw2x8r3</td>
<td>reverse</td>
<td>CAACAACACGTAAAGGAAAGG</td>
<td>21</td>
<td>47</td>
<td>61</td>
</tr>
<tr>
<td>EMgw2x1f4</td>
<td>forward</td>
<td>ATGGGGAACAGGATAGGG</td>
<td>18</td>
<td>55</td>
<td>59</td>
</tr>
<tr>
<td>EMgw2x8r3</td>
<td>reverse</td>
<td>CTACAACCATGCCAATCCTGC</td>
<td>22</td>
<td>50</td>
<td>64</td>
</tr>
</tbody>
</table>

Note: A primer name starting with wg indicates that it is located in the UTR region whereas primer names starting with EM indicate that they are located in the exons.

2.4 Discussion

Using the synteny between rice and other grasses, this study was able to obtain partial gene sequence of GW2 in M. stipoides. The method is time consuming and was relatively inefficient, using the sequence data available. Based on this approach, only one of the five genes targeted was successfully characterised. This appears primarily to be due to the lack of available data at the time, and relative degree of sequence divergence in the target genes between different grass taxa. Although the corpus of publically available sequence data has since increased rapidly, driven by reduced costs of sequencing and better sequencing technology, it was determined that an improved method for characterising these gene homologues in M. stipoides was required.

Apart from the rationale discussed above, the increased number of target loci associated with tetraploid, M. stipoides could have contributed to the failure of
Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*

Some of the PCRs and/or sequencing of PCR products. Primers that amplify rice, barley and *M. stipoides* may amplify more than one homologue. This could have caused the direct sequencing to result in ambiguous base calls (page 66). Other means of capturing homologues in a polyploid target such as barcoded amplification tags (BATs) in PCR products, short tandem repeat markers (STRs), or next generation sequencing are necessary in order to characterise the gene homologues of interest.

2.5 References


Chapter 2: Characterisation of gene homologues of rice quality traits in undomesticated and previously genetically uncharacterised *M. stipoides*


Chapter 3 - Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

Published in abridged form as Malory, Shapter et al. (2011). "Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes." Plant Biotechnology Journal 9(9): 1131-1140.

3.1 Introduction

Plant domestication resulted in changes in the appearance, biochemical properties and genetics of crops to assist in the ease of cultivation. In their review, Tang, Sezen et al. (2010) summarised the genomic changes that are associated with the domestication of crops, such as amino acid substitution, deletion and truncation, transposon insertion, regulatory change, splice site mutation and gene duplication. Farmers and plant breeders have been selecting beneficial characteristics to create better crops to feed the ever increasing world population. Selections involved in crop domestication, whether natural or artificial, result in genetic bottlenecks (Doebley, Gaut et al. 2006; Tang, Sezen et al. 2010). Wild crop relatives can provide alternative gene pools to counteract the genetic bottlenecks (Dillon, Shapter et al. 2007; Singh, Sharma et al. 2008), however genomic characterisation of wild species is generally poor (Zeid, Yu et al. 2010). Variants of known genes with potentially novel functionality have been found by exploring related species by utilising synteny and homology (Feltus, Singh et al. 2006; Zeid, Yu et al. 2010). The common method of screening biodiversity for useful gene homologues has often been based upon amplification of targeted genes using available sequence from cultivated or model species to design universal primers that amplify the target
Amplification of gene homologues using universal primers is made possible by designing and positioning primers in the regions with conserved sequence between two or more species, usually within the same family. This universal primer method (UPM) is useful for amplifying a known gene in another related species, where the sequence is unknown (McIntosh, Pacey-Miller et al. 2005; Ramirez, Fleming et al. 2004; Shapter, Eggler et al. 2009; Zeid, Yu et al. 2010). The crucial requirement of this approach is the presence of sequences with sufficient conservation across species to allow for design of universal primers. This UPM approach is often not successful with diverse organisms or highly variable genes. For example, homologues of the desired genes may have been characterised in only one species, making the identification of conserved regions to allow for the design of universal primers impossible; secondly, sequence conservation between gene homologues in related species may not be sufficient for the design of effective universal primers, leading to failure to amplify the gene homologues in the target species.

Traditionally, gene homologues amplified by polymerase chain reaction (PCR) have their identities confirmed by Sanger sequencing (Sanger, Nicklen et al. 1977). Automation fluorescence technology, capillary electrophoresis and computer-driven laser detection have enabled this technology to achieve high throughput and high quality sequencing, making this method the gold standard sequencing platform to date (Men, Wilson et al. 2008). Each high quality Sanger read can extend to 1 kilobase pairs (kbp) in length. There are several drawbacks associated with this method, such as the difficulty of obtaining accurate sequence data for polyploids without cloning, the lack of ability to handle and analyse allele frequencies and the cost of sequencing, which, at $1 per kbp, makes it impossible for researchers with limited budget to sequence whole genomes (Men, Wilson et al. 2008).

Improvements in massively parallel sequencing (MPS) make it an increasingly viable method for whole genome sequencing (Amaral, Megens et al. 2009; Hyten, Cannon et al. 2010; Rubin, Zody et al. 2010). There are a number of platforms available for this sequencing method. The most commonly used platforms are
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

Roche 454 (454 Life Sciences, Branford, CT, USA), Illumina Genome Analyzer (Illumina, Inc., San Diego, CA, USA) and ABI SOLiD (Applied Biosystems, Foster City, CA, USA). They generate a large number of short reads, typically 450 base pairs (bp) for Roche 454 and 50 to 100 bp for both Illumina Genome Analyzer and ABI SOLiD (DiGuistini, Liao et al. 2009). Reviews comparing the different platforms indicate that high quality genome wide sequence with a broad range of applications can be rapidly achieved using MPS (Harismendy, Ng et al. 2009; Mardis, 2008; Metzker, 2010). Although genome assembly remains difficult with this technology due a number of factors such as poor ability to deal with repetitive sequences (DiGuistini, Liao et al. 2009; Metzker, 2010), it provides a viable option for characterising homologues of known genes in poorly described genomes. Sequence data generated by MPS can be used to design species specific primers that allow for the amplification on homologues of known genes in related species.

There are over 10,000 species in the grass family Poaceae (Hsiao, Jacobs et al. 1999). Major crop plants belonging in Poaceae such as rice, maize and sorghum have been well characterised at the genomic level (Goff, Ricke et al. 2002; Paterson, Bowers et al. 2009; Schnable, Ware et al. 2009; Yu, Hu et al. 2002) and research is underway to characterise barley, a complex diploid (Schulte, Close et al. 2009) and complex polyploid grasses such as sugar cane and wheat (Bundock, Eliott et al. 2009; Paux, Sourdille et al. 2008). Members of the Poaceae are highly conserved at the genomic level. A number of researchers have used the DNA sequence conservation between species within the Poaceae to identify homologues of known genes from one species in others (Campbell, Zhu et al. 2007; Shapter, Eggler et al. 2009; Zeid, Yu et al. 2010). Recently, _Brachypodium distachyon_ became the first wild grass to have its whole genome sequenced, allowing it to be used as another model system for grasses (The International Brachypodium Initiative, 2010). Completed genome sequences can be used as templates for the design of molecular markers in poorly described species (Feltus, Singh et al. 2006).

There is a need to genetically characterise poorly known species for a number of purposes such as biodiversity, biodiscovery, and breeding for crop improvement. The availability of molecular markers for these species is very limited because marker development is laborious, time-consuming, and expensive (Zeid, Yu et al.
In general, resources and research are directed towards major crops. Very few Poaceae taxa in the world have been well characterised. In those species that have been well characterised, the sample sizes have been small, not reflecting the total diversity of each individual species. Understanding the genetic diversity of known taxa will also aid the discovery of unknown species (David, Lemeunier et al. 2007). Characterising poorly known species will contribute to the knowledge of the gene pool available for use for crop improvement through advancing plant breeding techniques and genetically modified organism technology (Dillon, Shapter et al. 2007; Singh, Sharma et al. 2008; Tang, Sezen et al. 2010).

*Microlaena stipoides*, a wild relative of rice, *Oryza* spp. (Figure 3.1) was used as the target species in this paper. It responds well to irrigation and nitrogen application (Chivers and Aldous, 2005) and has been targeted for commercialisation and domestication as a perennial grain crop (Davies, Waugh et al. 2005a). *M. stipoides* seeds have similar grain ultrastructure to rice (Shapter, Henry et al. 2008) and in some ecotypes the seed size is approaching rice (Davies, Waugh et al. 2005b). *M. stipoides* is a tetraploid (Murray, De Lange et al. 2005) with a reported genome size of approximately 869 mega base pairs (Nock, Waters et al. 2010).
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

In this paper we propose a novel method for characterising homologues of known genes in poorly described genomes. We attempted to amplify homologues of well characterised genes from a model species in a related but poorly described species by utilising species specific primers derived from MPS genomic data of the poorly described species. MPS of genomic DNA was used to obtain *M. stipoides* sequence information for 18 genes related to domestication characteristics in *Oryza* (Table 3.1). The characteristics controlled by these genes are photoperiod sensitivity, heading date, grain yield (grain weight, size, number and filling rate), tillering, seed shattering, prostrate growth, pericarp colour, waxy grains, dwarfism and grain fragrance (Table 3.1). Species specific primers were designed based on the alignments of *M. stipoides* sequence and the 18 *Oryza* reference genes sequence. A comparison between characterising homologues of known genes in poorly described genomes using UPM and species specific primer method (SSPM) is presented here.

---

**Figure 3.1** A cladogram showing the relationship between the target species, *M. stipoides* and five major grass species, adapted from (Kellogg, 2009; Kellogg, 2001; Shapter, Eggler et al. 2009)
### Table 3.1  Details of 18 annotated rice genes related to domestication traits analysed in this study

<table>
<thead>
<tr>
<th>Gene/protein name</th>
<th>Abbreviation</th>
<th>Domestication trait</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTL for heading date 1</td>
<td><em>Hd1</em></td>
<td>photoperiod sensitivity and heading date</td>
<td>Yano, Katayose et al. (2000)</td>
</tr>
<tr>
<td>QTL of seed shattering in chromosome 1</td>
<td><em>qSH-1</em></td>
<td>seed shattering</td>
<td>Konishi, Izawa et al. (2006)</td>
</tr>
<tr>
<td>QTL for grain number 1A</td>
<td><em>Gn1a</em></td>
<td>grain number</td>
<td>Ashikari, Sakakibara et al. (2005)</td>
</tr>
<tr>
<td>Red pericarp</td>
<td><em>Rc</em></td>
<td>red pericarp</td>
<td>Sweeney, Thomson et al. (2006)</td>
</tr>
<tr>
<td>QTL for seed width on chromosome 5</td>
<td><em>qSW5</em></td>
<td>grain width</td>
<td>Shomura, Izawa et al. (2008)</td>
</tr>
<tr>
<td>Waxy*, also known as granule bound starch synthase I (GBSSI)</td>
<td><em>Wx</em></td>
<td>waxy grains</td>
<td>Hirano, Euguchi et al. (1998)</td>
</tr>
<tr>
<td>Semi dwarf 1</td>
<td><em>Sd-1</em></td>
<td>dwarfism</td>
<td>Ashikari, Sasaki et al. (2002)</td>
</tr>
<tr>
<td>Monoculm 1</td>
<td><em>MOC1</em></td>
<td>tillering</td>
<td>Li, Qian et al. (2003)</td>
</tr>
<tr>
<td>QTL for heading date 3a</td>
<td><em>Hd3a</em></td>
<td>heading date</td>
<td>Kojima, Takahashi et al. (2002)</td>
</tr>
<tr>
<td>Early heading date 1*</td>
<td><em>Ehd1</em></td>
<td>heading date</td>
<td>Doi, Izawa et al. (2004)</td>
</tr>
<tr>
<td>QTL for heading date 6</td>
<td><em>Hd6</em></td>
<td>heading date</td>
<td>Yamamoto, Lin et al. (2000)</td>
</tr>
<tr>
<td>QTL for grain size 3</td>
<td><em>GS3</em></td>
<td>grain length and weight</td>
<td>Fan, Xing et al. (2006)</td>
</tr>
<tr>
<td>QTL for shattering 4, also known as shattering 1 (SHA1)</td>
<td><em>sh4</em></td>
<td>seed shattering</td>
<td>Li, Zhou et al. (2006)</td>
</tr>
<tr>
<td>Betaine aldehyde dehydrogenase 2</td>
<td><em>BADH2</em></td>
<td>fragrance</td>
<td>Bradbury, Fitzgerald et al. (2005)</td>
</tr>
<tr>
<td>QTL for grain width and weight 2</td>
<td><em>GW2</em></td>
<td>grain weight and width</td>
<td>Song, Huang et al. (2007)</td>
</tr>
<tr>
<td>Grain incomplete filling 1</td>
<td><em>GIF1</em></td>
<td>grain filling rate</td>
<td>Wang, Wang et al. (2008)</td>
</tr>
<tr>
<td>QTL for grain number, plant height and heading date 7</td>
<td><em>Ghd7</em></td>
<td>grain number, plant height, and heading date</td>
<td>Xue, Xing et al. (2008)</td>
</tr>
<tr>
<td>Prostrate growth 1</td>
<td><em>PROG1</em></td>
<td>prostrate growth</td>
<td>Jin, Huang et al. (2008)</td>
</tr>
</tbody>
</table>

* indicates that first exon and intron are located on the UTR region of the gene.
3.2 **Materials and Methods**

3.2.1 **Plant material and DNA extraction**

Native Seeds Pty Ltd (http://nativeseeds.com.au/, last accessed 9th November 2012) provided *M. stipoides* seed from their AR1 breeding line. Fresh leaf tissue was harvested from a plant for DNA extraction. The method of DNA extraction was based on Carroll, Klimyuk et al. (1995) with the following exceptions: 3 grams of leaf tissue in liquid nitrogen was ground to a fine powder using mortar and pestle and the quantity of reagents were adjusted proportionally.

3.2.2 **Identifying conserved regions between rice and other Poaceae for maximum possible primers placement for UPM**

Sequences for 18 rice domestication genes (Table 3.1) were obtained from NCBI’s Genbank website (http://www.ncbi.nlm.nih.gov/genbank/, last accessed 10th November 2012). Searches using nucleotide blast (blastn) optimised for highly similar sequences (megablast) (http://blast.ncbi.nlm.nih.gov/Blast.cgi, last accessed 10th November 2012) were performed on each gene to find homologue(s) in other species within Poaceae. Alignments between the exons from the rice reference gene and the homologue(s) found through blastn were conducted using Sequencher® V4.6 (Gene Codes Corporation, Ann Arbor, MI, USA). Where there was more than one homologue found, the alignment with the most sequence matches was accepted. Regions with conserved sequence for at least 17 bases and a maximum of two bases with degeneracy at the start and the end of the coding region of each gene were identified as possible sites for primer placements to allow for the amplification of putative gene homologues of the rice reference genes in *M. stipoides*.

3.2.3 **Identifying regions in *M. stipoides* aligned to rice reference genes for maximum possible primers for SSPM**

Sequences for 18 rice domestication genes (Table 3.1) were obtained from NCBI Genbank website. *M. stipoides* sequence data were obtained from a study by Nock, Waters et al. (2010). Single and paired-end short reads produced by the MPS
method using CLC Genomics Workbench 3.6.5 (CLC bio, Aarhus N, Denmark) were assembled against reference Oryza gene sequence data. Sequence data were trimmed using CLC Genomics Workbench version 4.0.3 (www.clcbio.com, last accessed 02/11/12). Reference assembly was undertaken with a mismatch cost of 2, indel costs of 3, length fraction of 0.8 and similarity of 0.8, minimum distance for paired end reads of 90bp with a maximum of 180bp, and non-specific matches randomised. Consensus sequences identified were exported to Sequencher® V4.6 and aligned back to the reference genes from rice. Regions of M. stipoides sequence with at least 17 consecutive bases at the start and at the end of the coding region of each gene were identified as possible sites for primer placements to allow for the amplification of putative gene homologues of the rice reference genes in M. stipoides.

3.2.4 PCR primer design

Clone Manager Professional 9 (Scientific & Educational Software, Cary, NC, USA) was used for both UPM and SSPM primer design. PCR amplicons were limited to 4 kbp. Clone Manager Professional 9 primer criteria for both UPM and SSPM primers were set as follows: a minimum of 17 bases at 45-65% GC content, Tm between 55-65°C and the Tm difference between forward and reverse primer ±4 ºC, stability of 5’ versus 3’ at 1.2 kcals, less than three matches for 3’ dimers, less than seven for any dimer, less than five base runs, less than three repeats of di-nucleotide pairs, less than eight protein degeneracy within 6 bp of 3’ end, less than or equal to one G’s or C’s at 3’ end. For UPM primers, a maximum of two bases of degeneracy was allowed. Primers that met these criteria were chosen whenever available and a maximum of two unmet criteria was accepted.

3.2.5 PCR amplification and sequencing preparation

Universal primers and species specific primers used for PCR amplification are listed in Table 3.3.5.1. PCRs were conducted using Platinum® Taq DNA Polymerase (Invitrogen Australia Pty. Ltd., Mulgrave, VIC, Australia) at 1 unit/reaction with 1X PCR Buffer, primers at 0.2uM each , with 3% DMSO, 3% glycerol, 2mM magnesium chloride and 10 ng of genomic DNA. The PCR program was set for one cycle of denaturation at 96°C for two minutes, followed by 35
cycles of denaturation at 96°C for 30 seconds, annealing at the temperature required by respective primer pairs for 30 seconds, extension at 68°C for one minute per kbp of PCR product. A single cycle at 72°C for five minutes allowed for further extension. Optimisation of PCR was done using temperature gradient of annealing step down to 5°C below and 5°C above the average annealing temperature of the primer pair. The PCR products were run on 1% agarose gel electrophoresis stained with ethidium bromide to confirm amplification and then sequenced to confirm identity. Where a single PCR band was amplified in a reaction, Exosap-IT (GE Healthcare Bio-Sciences Pty. Ltd., Rydalmere, NSW, Australia) was used to clean up the PCR product prior to direct Sanger sequencing. Multiple PCR bands in a reaction were separated via gel electrophoresis and individually excised then cleaned using QIAquick Gel Extraction Kit (Qiagen Pty Ltd, Doncaster, VIC, Australia) prior to direct Sanger sequencing. Figure 3.2.5 summarises the workflow involved in comparing the two methods for characterising homologues of known genes in poorly described genomes using UPM and SSPM.
Homologues mining in M. stipoides for domestication related genes from Oryza

UPM
- 18 Genbank entries from genus Oryza
  - Blastn search for 18 Oryza genes
    - 1 gene without homologous sequence in other Poaceae
    - 17 genes with homologous sequence(s) in Poaceae
      - 11 genes sufficient for primer design
      - 6 genes insufficient for primer design
        - 6 genes non-degenerate universal primers
        - 5 genes non-degenerate & degenerate universal primers

SSPM
- 18 Genbank entries from genus Oryza
  - Align M. stipoides MPS sequence data to Oryza reference genes
    - 18 genes with between 30-73% homology to the reference genes
      - 2 genes insufficient for primer design
      - 16 genes sufficient for primer design
        - 16 genes non-degenerate species specific primers

Figure 3.2.5 Schematic representation of the workflow comparing homologues of known genes in poorly described genomes using UPM and SSPM
3.3 Results

3.3.1 Maximum possible primers placement for UPM (prior to primer design)

Seventeen had homologous sequences identified within Poaceae, the exception being Ghd7 (Table 3.3.1). Ehd1 and qSW5 did not have sufficient areas of conservation to allow for the alignment of the available species with Oryza reference gene for any universal primer placement (Table 3.3.1). Rc, GS3, and PROG1 had only limited areas of conservation which restricted the proportion of the gene potentially able to be captured to such an extent that universal primer amplification was deemed to be uninformative (Table 3.3.2), therefore the total number of genes where universal primers could not be placed was six. There was sufficient conservation between species for universal primer design within the remaining 12 genes (Hd1, qSH-1, Gln1a, Wx, Sd-1, MOC1, Hd3a, Hd6, sh4, BADH2, GW2, and GIF1). Figure 3.3.1 shows the possible universal primer pair placements indicating the maximum gene capture. Supplementary data of alignments of the candidate rice genes to corresponding Poaceae orthologous sequences (obtained from performing blastn search), presented as a Sequencher® file are provided on the CD accompanying this thesis in the folder Appendix 1 - Supplementary 3.3.1.

A complete M. stipoides homologue (EF600044) was already published for the Waxy gene (Shapter, Eggler et al. 2009). For the purpose of this paper, the M. stipoides waxy homologue from Genbank was disregarded and the search for homologue(s) of Wx gene in other Poaceae species was performed as per the other genes.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Genbank entry</th>
<th>UPM</th>
<th>SSPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Relevant putative homologues in other species within Poaceae</td>
<td>Best minimum match percentage at 20 bp overlap (Sequencher®) Primer site(s)³ Number of PCR amplicons ⁴ Percentage of reference sequence covered by MPS ² Minimum match percentage at 20 bp overlap (Sequencher®) Primer site(s)³ Number of PCR amplicons ⁴</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hd1</td>
<td>AB041837</td>
<td><em>Phyllostachys edulis</em>, <em>Brachypodium distachyon</em>, Zea mays, <em>Sorghum bicolor</em>, <em>S. arundinaceum</em>, <em>Triticum aestivum</em>, <em>Hordeum vulgare</em>, <em>Lolium perenne</em>, <em>Schedonorus pratensis</em> &amp; <em>Festuca arundinacea</em></td>
<td>67 ND 1 34.7 82 ND 1</td>
</tr>
<tr>
<td>qSH-1</td>
<td>AB071333</td>
<td><em>Triticum aestivum</em>, <em>Hordeum vulgare</em> &amp; <em>Zea mays</em></td>
<td></td>
</tr>
<tr>
<td>Gn1a</td>
<td>AB205193</td>
<td><em>S. bicolor</em>, <em>P. edulis</em>, <em>Z. mays</em>, <em>T. aestivum</em> &amp; <em>Bambusa oldhamii</em></td>
<td></td>
</tr>
<tr>
<td>Rc</td>
<td>AB250059</td>
<td><em>S. bicolor</em> &amp; <em>T. aestivum</em></td>
<td></td>
</tr>
<tr>
<td>qSW5</td>
<td>AB433345</td>
<td><em>S. bicolor</em> &amp; <em>P. edulis</em></td>
<td></td>
</tr>
<tr>
<td>Wx</td>
<td>AF031162</td>
<td>Found in other major Poaceae including <em>Sorghum</em> spp., <em>Z. mays</em>, <em>Triticum</em> spp. and <em>Hordeum</em> spp*. Also found in</td>
<td></td>
</tr>
</tbody>
</table>

91
### Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

Various other Poaceae (too many to list here), including target species *M. stipoides*.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sdl-1</strong></td>
<td>AF465255</td>
<td><em>Luziola leiocarpa</em>, <em>Chikusichloa aquatica</em>, <em>Rhynchoryza subulata</em>, <em>Ehrharta erecta</em>, <em>Leersia tisserantii</em>, <em>Zea mays</em> &amp; <em>Sorghum bicolor</em></td>
<td>78</td>
<td>ND</td>
<td>1</td>
</tr>
<tr>
<td><strong>MOC1</strong></td>
<td>AY242058</td>
<td><em>S. bicolor</em></td>
<td>78</td>
<td>ND</td>
<td>1</td>
</tr>
<tr>
<td><strong>Hd3a</strong></td>
<td>BD169069</td>
<td><em>H. vulgare</em>, <em>T. turgidum</em>, <em>T. aestivum</em>, <em>Z. mays</em>, <em>Lolium perenne</em>, <em>T. monococcum</em>, <em>S. bicolor</em>, <em>P. meyeri</em>, <em>Streptogyna crinita</em>, <em>Streptogyna americana</em>, <em>Aulonemia subpectinata</em>, <em>Sasa senanensis</em>, <em>Dendrocalamus asper</em>, <em>Sasa kurilensis</em>, <em>Guaduella foliosa</em>, <em>G. marantifolia</em>, <em>Sasa tsuboiana</em>, <em>Sasa jotanii</em>, <em>Dianthryya bicolor</em> &amp; <em>Lithachne pauciflora</em></td>
<td>80</td>
<td>ND</td>
<td>1</td>
</tr>
<tr>
<td><strong>Ehd1</strong></td>
<td>BD407931</td>
<td><em>S. bicolor</em></td>
<td>less than 60</td>
<td>NS</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Hd6</strong></td>
<td>DQ157464</td>
<td><em>Z. mays</em>, <em>S. bicolor</em>, <em>H. vulgare</em> &amp; <em>Lolium perenne</em></td>
<td>81</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td><strong>GS3</strong></td>
<td>DQ355996</td>
<td><em>Z. mays</em> &amp; <em>S. bicolor</em></td>
<td>less than 60</td>
<td>NS</td>
<td>N/A</td>
</tr>
<tr>
<td>Gene</td>
<td>Accession</td>
<td>Description</td>
<td>Score</td>
<td>TV</td>
<td>E-value</td>
</tr>
<tr>
<td>--------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------</td>
<td>---</td>
<td>---------</td>
</tr>
<tr>
<td>sh4</td>
<td>DQ383373</td>
<td><em>Echinochloa</em> spp., <em>Zea mays</em> &amp; <em>S. bicolor</em></td>
<td>75</td>
<td>ND</td>
<td>1</td>
</tr>
<tr>
<td>BADH2</td>
<td>DQ910546</td>
<td><em>S. bicolor</em>, <em>Z. mays</em>, <em>H. vulgare</em>, <em>Leymus chinensis</em>, <em>Zoysia tenuifolia</em>, <em>T. aestivum</em>, <em>Pascopyrum smithii</em>, <em>Schedonorus arundinaceus</em>, <em>Elymus trachycaulus</em>, <em>Phalaris arundinacea</em>, <em>Psathyrostachys juncea</em> &amp; <em>Agropyron cristatum</em></td>
<td>79</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>GW2</td>
<td>EF447275</td>
<td><em>H. vulgare</em>, <em>Z. mays</em> &amp; <em>S. bicolor</em></td>
<td>81</td>
<td>ND</td>
<td>3</td>
</tr>
<tr>
<td>GIF1</td>
<td>EU095553</td>
<td><em>Dendrocalamopsis oldhamii</em>, <em>Z. mays</em>, <em>S. bicolor</em>, <em>T. aestivum</em>, <em>P. edulis</em>, <em>Saccharum hybrid</em>, <em>L. perenne</em> &amp; <em>B. distachyon</em></td>
<td>77</td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>Ghd7</td>
<td>EU286801</td>
<td>not found</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>PROG1</td>
<td>FJ155665</td>
<td><em>H. vulgare</em>, <em>P. edulis</em>, <em>Z. mays</em>, <em>T. aestivum</em> &amp; <em>S. bicolor</em></td>
<td>62</td>
<td>NS</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\(^1\)BLASTN search was done using the Genbank accession number. The parameters were set as: choose search set: database others, program selection: optimise for highly similar sequences (megablast). The algorithm parameters were set as default with the exception of max target sequences changed to 20000. * indicates the BLASTN result that gave the best alignment and used for universal primer placement for UPM.
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

2Sequencher® allows for 0-100 bp minimum match percentage to perform alignment. The default is set at 20 bp. Alignment was not possible for match below 60%.

3Types of primer possible were NS – not sufficient, ND – sufficient for non-degenerate primers, C – sufficient for combination of non-degenerate and degenerate primers, and N/A – not applicable.

4Number of PCR amplicons for amplifying the gene, maximum amplicon size is limited to 4 kbp. N/A – not applicable.

5Calculation is based on the length of consensus sequence divided by the length of reference sequence.
Figure 3.3.1  Maximum possible primer placements for UPM and SSPM compared to actual primers designed for UPM and SSPM
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

3.3.2 Maximum possible primers placement for SSPM

MPS performed by Nock, Waters et al. (2010) using Illumina GAII produced over 23 million 36 bp paired end reads with a total of 1.6 giga base pairs (gbp). Based on the size of *M. stipoides* genome (Murray, De Lange et al. 2005), the maximum possible coverage of the whole genome was 2X. Whole genome sequencing resulted in patchy coverage of the *M. stipoides* genome when aligned back to rice as a reference. Supplementary data of assemblies of the MPS data to the rice target genes is provided on the CD accompanying this thesis in the folder Appendix 2 - Supplementary 3.3.2. Coverage of putative homologues of targeted genes varied as shown in Table 3.2.6. On average 44% of the *M. stipoides* sequence (exons only) could be aligned to the targeted reference gene sequences. Despite the low coverage, it was sufficient for species specific non-degenerate primer placement within the putative exon boundaries for all the targeted genes except one, *Ehd1*, where the capture of exons was too low (three exons out of the total of five exons). The remaining 17 genes had possible placements of non-degenerate species specific primer pairs for maximum gene capture (Figure 3.3.1).

On average, 94.6% of the rice reference sequences could potentially be captured with the SSPM. Maximum possible gene capture by SSPM ranged from a minimum of 32.3% for *Ehd1* and a maximum of 100% for *GW2* (Table 3.3.2). For comparison, across the 15 genes for which areas of sequence conservation were identified between an *Oryza* spp. and other Poaceae sufficient for universal primer design an average 85.3% of the rice reference gene’s sequence could be captured (Table 3.3.2). This ranged from 100% capture of the coding sequence for *Hd1* and *GW2* to a minimum capture of 19.6% and 37.6% in *PROG1* and *Rc* respectively.
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Genbank entry</th>
<th>UPM</th>
<th>Maximum exons covered</th>
<th>SSPM</th>
<th>Maximum exons covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hd1</td>
<td>AB041837</td>
<td>100.0</td>
<td>all (1-2)</td>
<td>98.1</td>
<td>all (1-2)</td>
</tr>
<tr>
<td>qSH-1</td>
<td>AB071333</td>
<td>88.7</td>
<td>all (1-4)</td>
<td>98.3</td>
<td>all (1-4)</td>
</tr>
<tr>
<td>Gn1a</td>
<td>AB205193</td>
<td>96.2</td>
<td>all (1-4)</td>
<td>97.2</td>
<td>all (1-4)</td>
</tr>
<tr>
<td>Rc</td>
<td>AB250059</td>
<td>37.6</td>
<td>exons 5 &amp; 8 out of 8</td>
<td>99.9</td>
<td>all (1-8)</td>
</tr>
<tr>
<td>qSW5</td>
<td>AB433345</td>
<td>none</td>
<td></td>
<td>93.7</td>
<td>all (1)</td>
</tr>
<tr>
<td>Wx</td>
<td>AF031162</td>
<td>99.8</td>
<td>all (2-14)</td>
<td>99.8</td>
<td>all (2-14)</td>
</tr>
<tr>
<td>Sd-1</td>
<td>AF465255</td>
<td>94.8</td>
<td>all (1-3)</td>
<td>98.4</td>
<td>all (1-3)</td>
</tr>
<tr>
<td>MOC1</td>
<td>AY242058</td>
<td>85.4</td>
<td>all (1)</td>
<td>96.6</td>
<td>all (1)</td>
</tr>
<tr>
<td>Hd3a</td>
<td>BD169069</td>
<td>99.3</td>
<td>exons 1-2 &amp; 4, exon 3 did not align</td>
<td>99.1</td>
<td>all (1-4)</td>
</tr>
<tr>
<td>Ehd1</td>
<td>BD407931</td>
<td>did not align</td>
<td></td>
<td>32.3</td>
<td>exons 3 &amp; 5, exons 1, 2 &amp; 4 did not align</td>
</tr>
<tr>
<td>Hd6</td>
<td>DQ157464</td>
<td>99.4</td>
<td>all (1-10) only exons 1-4, exon 5 did not align</td>
<td>99.9</td>
<td>all (1-10)</td>
</tr>
<tr>
<td>GS3</td>
<td>DQ355996</td>
<td>63.1</td>
<td></td>
<td>99.5</td>
<td>all (1-5)</td>
</tr>
<tr>
<td>sh4</td>
<td>DQ383373</td>
<td>96.8</td>
<td>all (1-2)</td>
<td>98.1</td>
<td>all (1-2)</td>
</tr>
<tr>
<td>BADH2</td>
<td>DQ910546</td>
<td>99.8</td>
<td>all (1-15)</td>
<td>99.8</td>
<td>all (1-15)</td>
</tr>
<tr>
<td>GW2</td>
<td>EF447275</td>
<td>100.0</td>
<td>all (1-8)</td>
<td>100.0</td>
<td>all (1-8)</td>
</tr>
<tr>
<td>GIF1</td>
<td>EU095553</td>
<td>99.1</td>
<td>all (1-7)</td>
<td>98.2</td>
<td>all (1-7)</td>
</tr>
<tr>
<td>Ghd7</td>
<td>EU286801</td>
<td>none</td>
<td></td>
<td>94.9</td>
<td>all (1-2)</td>
</tr>
<tr>
<td>PROG1</td>
<td>FJ155665</td>
<td>19.6</td>
<td>poor alignment of exon 1</td>
<td>98.0</td>
<td>all (1)</td>
</tr>
</tbody>
</table>

Note: The table above shows the maximum possible genes or exons covered prior to primers being designed.

### 3.3.3 UPM PCR Primers

Primer design was not attempted for Rc, qSW5, Ehd1, GS3, Ghd7 and PROG1 for the reasons that were mentioned earlier. PCR primer design was attempted within the identified maximum placement of UPM primers for the remaining 12 genes (Hd1, qSH-1, Gn1a, Wx, Sd-1, MOC1, Hd3a, Hd6, sh4, BADH2, GW2, and GIF1). Using the criteria as described in the materials and methods section, non-degenerate primers were successfully designed for six genes (Hd1, Sd-1, MOC1, Hd3a, sh4 and GW2), whereas a combination of non-degenerate primers and degenerate
primers were needed for five genes (qSH-1, Wx, Hd6, BADH2 and GIF1). No forward primer meeting the set criteria was found for Gn1a, however a primer pair encompassing exons two to four were identified. In summary, the UPM method allowed for PCR primer design encompassing the gene coding region for 11 genes out of the 18 genes (Figure 3.3.1).

### 3.3.4 SSPM PCR Primers

Primer design was not attempted for Ehd1 for the reason mentioned earlier. PCR primer designed was done within the identified maximum placement of SSPM primers for the remaining 17 genes (Hd1, qSH-1, Gn1a, Rc, qSW5, Wx, Sd-1, MOC1, Hd3a, Hd6, GS3, sh4, BADH2, GW2, GIF1, Ghd7 and PROG1). Using the criteria as described in the materials and methods section, non-degenerate species specific primers were successfully designed for all of the genes with the exception of GS3. Compatible primers were identified only on exon one and exon five for GS3, the size of the PCR amplicon exceeded 4 kbp. There was insufficient coverage of exons two, three and four to allow for primer design. GS3 was omitted from the subsequent steps. In summary, the SSPM method allowed for PCR primer design encompassing the gene coding region for 16 genes out of the 18 genes (Figure 3.3.1).

Figure 3.3.1 shows the position of primers using both UPM and SSPM within the identified maximum possible primer placements. SSPM primers covered more portion of the genes compared to UPM primers. SSPM primers covered all exons for Gn1a, Rc, qSW5, Ehd1, GS3, Ghd7 and PROG1 where UPM primers for these genes were either not possible, did not cover all exons or provided coverage that was too low to give meaningful sequence information. Table 3.3.3 shows the percentage of genes covered by UPM and SSPM primers designed. Using the UPM, the highest proportion of gene covered by the primers was 99.2% (Wx), the lowest was 17.1% (PROG1) and on average 79.7% of the genes were covered (Table 3.3.3). Using the SSPM, the whole of GW2 was able to be covered by the primers, followed by Rc (99.9%), Wx (99.8%) and Hd6 (99.5%), the lowest was 32.3% (Ehd1) and on average 92% of the genes were covered (Table 3.3.3).
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

Table 3.3.3  Comparison of percentage of gene covered by primers designed using UPM and SSPM

<table>
<thead>
<tr>
<th>Gene</th>
<th>Genbank entry</th>
<th>UPM Percentage gene captured by primers</th>
<th>Exons covered</th>
<th>SSPM Percentage gene captured by primers</th>
<th>Exons covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hd1</td>
<td>AB041837</td>
<td>92.9</td>
<td>all (1-2)</td>
<td>97.7</td>
<td>all (1-2)</td>
</tr>
<tr>
<td>qSH-1</td>
<td>AB071333</td>
<td>82.9</td>
<td>all (1-4)</td>
<td>89.2</td>
<td>all (1-4)</td>
</tr>
<tr>
<td>Gn1a</td>
<td>AB205193</td>
<td>83.1</td>
<td>exons 2-4 of 4</td>
<td>96.3</td>
<td>all (1-4)</td>
</tr>
<tr>
<td>Rc</td>
<td>AB250059</td>
<td>32.7</td>
<td>exons 5-8 of 8</td>
<td>99.9</td>
<td>all (1-8)</td>
</tr>
<tr>
<td>qSW5</td>
<td>AB433345</td>
<td>none</td>
<td></td>
<td>77.3</td>
<td>all (1)</td>
</tr>
<tr>
<td>Wx</td>
<td>AF031162</td>
<td>99.2</td>
<td>all (2-14)</td>
<td>99.8</td>
<td>all (2-14)</td>
</tr>
<tr>
<td>Sd-1</td>
<td>AF465255</td>
<td>92.1</td>
<td>all (1-3)</td>
<td>98.4</td>
<td>all (1-3)</td>
</tr>
<tr>
<td>MOC1</td>
<td>AY242058</td>
<td>61.9</td>
<td>all (1)</td>
<td>95.9</td>
<td>all (1)</td>
</tr>
<tr>
<td>Hd3a</td>
<td>BD169069</td>
<td>92.3</td>
<td>all (1-4)</td>
<td>97.2</td>
<td>all (1-4)</td>
</tr>
<tr>
<td>Ehd1</td>
<td>BD407931</td>
<td>none</td>
<td></td>
<td>32.3</td>
<td>exons 4-6 of 6</td>
</tr>
<tr>
<td>Hd6</td>
<td>DQ157464</td>
<td>98.6</td>
<td>all (1-10)</td>
<td>99.5</td>
<td>all (1-10)</td>
</tr>
<tr>
<td>GS3</td>
<td>DQ355996</td>
<td>60.0</td>
<td>exons 1-3 of 5</td>
<td>94.2</td>
<td>all (1-5)*</td>
</tr>
<tr>
<td>sh4</td>
<td>DQ383373</td>
<td>90.2</td>
<td>all (1-2)</td>
<td>89.5</td>
<td>all (1-2)</td>
</tr>
<tr>
<td>BADH2</td>
<td>DQ910546</td>
<td>98.8</td>
<td>all (1-15)</td>
<td>99.6</td>
<td>all (1-15)</td>
</tr>
<tr>
<td>GW2</td>
<td>EF447275</td>
<td>97.7</td>
<td>all (1-8)</td>
<td>100.0</td>
<td>all (1-8)</td>
</tr>
<tr>
<td>GIF1</td>
<td>EU095553</td>
<td>96.1</td>
<td>all (1-7)</td>
<td>98.2</td>
<td>all (1-7)</td>
</tr>
<tr>
<td>Gh7</td>
<td>EU286801</td>
<td>none</td>
<td></td>
<td>94.7</td>
<td>all (1-2)</td>
</tr>
<tr>
<td>PROG1</td>
<td>FJ155665</td>
<td>17.1</td>
<td>poor alignment of exon 1 of 1</td>
<td>96.0</td>
<td>all (1)</td>
</tr>
</tbody>
</table>

*For GS3, good alignments were only available for exons one and five and primers were available for these exons. Using primers from exons one and five would result in a PCR amplicon that exceeds 4 kbp in size and may not be achievable using our method. Exons two, three and four did not have enough M. stipoides sequence information to allow for SSPM primer design. GS3, along with Ehd1, which did not have enough coverage to allow for all exons to be amplified were excluded from further analysis.

Note: the table above shows the maximum possible genes or exons covered after primers were designed.

### 3.3.5 UPM PCR Amplification and Sanger Sequencing Results

Of the 11 genes with universal primers designed, PCRs were conducted for three genes; qSH-1, sh4 and GW2 (Table 3.3.5.1). Post-PCR agarose gel separation and visualisation of the qSH-1 amplicon (exons one to four) produced a distinct band at the expected size (Figure 3.3.5.1). Partial sequencing and alignment to exons one and three of the reference gene showed 85% and 86% identity respectively to rice.
Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

$qSH-1$ (Table 3.3.5.2). The primer pair for sh4 produced a distinct major band at the expected size with three smaller faint bands which were separated by gel extraction (Figure 3.3.5.1). The major band was excised and sequenced and exons one and two showed 81% identity to the sh4 shattering gene in rice.

Three pairs of primers were designed for GW2, the first pair encompassed exons one to five, the second pair exons five to six and the third pair exons seven to eight. PCR amplification of exons one to five produced two distinct fragments at the expected size (Figure 3.3.5.1). These two fragments were sequenced and showed two distinct but homologous sequence reads in M. stipoides. Sequencing of both fragments showed identity of 92% with rice GW2 exons one to five (Table 3.3.5.2). PCR amplifications of exons five to six and seven to eight both produced single fragments (Figure 3.3.5.1). The exon 5/6 amplicon had 94% and the exon 7/8 amplicon had 86% identity to their respective rice GW2 exons (Table 3.3.5.2). In summary, of the five universal primer pairs tested, five were successfully Sanger sequenced and confirmed as putative homologues in M. stipoides.
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

(a)

(b)

(c)

Optimised Gw2 exons 1-5
Figure 3.3.5.1  Photos of agarose gels showing the results of optimised UPM PCR amplifications (a) *qsh-1*, (b) *sh4*, (c) *Gw2* exons 1-5, (d) *Gw2* exons 5-6 and (e) *Gw2* exons 7-8
<table>
<thead>
<tr>
<th>Gene</th>
<th>Primer Name</th>
<th>Direction</th>
<th>Sequence 5' to 3'</th>
<th>PCR result</th>
<th>Pair name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>qSH-1</td>
<td>qsh1x1f1</td>
<td>forward</td>
<td>CGCGAGAAGCTCCGGTTCC</td>
<td>single bright band</td>
<td>qsh1x1-4</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>qsh1x1r1</td>
<td>reverse</td>
<td>YGACAAGGGCCACGAGWA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sh4</td>
<td>sh4x1f2</td>
<td>forward</td>
<td>CGCGGGACTACCGCAAG</td>
<td>a bright band with a few weaker non-specific bands</td>
<td>sh4x1-2</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>sh4x2r4</td>
<td>reverse</td>
<td>GCGGTGTCGGAGACGAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GW2</td>
<td>gw2x1f1</td>
<td>forward</td>
<td>CAGGGGCTGTACGAGCAAC</td>
<td>double bright bands</td>
<td>gw2x1-5</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>gw2x5r1</td>
<td>reverse</td>
<td>GGATGTGCTGCAATAATGCTC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>gw2x5f1</td>
<td>forward</td>
<td>ATGGCCAGCAAGCAGACTTC</td>
<td>single bright band</td>
<td>gw2x5-6</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>gw2x6r1</td>
<td>reverse</td>
<td>CATTCAGTTTCTTGCTC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>gw2x7f1</td>
<td>forward</td>
<td>TGGTTATGGAAGCGATTTGG</td>
<td>single bright band</td>
<td>gw2x7-8</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>gw2x8r2</td>
<td>reverse</td>
<td>TGGTCTTTCGATGCTTC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hosl</td>
<td>Mshd1x1f1</td>
<td>forward</td>
<td>GTTGGAGTTGAGGGAGCCAAGG</td>
<td>multiple non-specific bands</td>
<td>hdx1-2</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Mshd1x2r2</td>
<td>reverse</td>
<td>GAACATTGGAACGATACCATAGC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qSH-1</td>
<td>Msqsh1x1f1</td>
<td>forward</td>
<td>TCCTGCTCCACCACCATC</td>
<td>multiple non-specific bands</td>
<td>qsh1x1-4</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Msqsh1x1r1</td>
<td>reverse</td>
<td>CCTCCAAATGAGCCCATCAC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gn1a</td>
<td>Mgsn1ax1f1</td>
<td>forward</td>
<td>TCGGCTCCTTCCTGCA</td>
<td>multiple non-specific bands</td>
<td>gn1ax1-2</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Mgsn1ax2r2</td>
<td>reverse</td>
<td>TCGGCTCCTTGAAAGG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mgsn1ax3f1</td>
<td>forward</td>
<td>CGCGAGAGCTACGGCTCCCG</td>
<td>multiple non-specific bands</td>
<td>gn1ax3-4</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Mgsn1ax4r4</td>
<td>reverse</td>
<td>CTTGGCCAGAGTACGAC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rc</td>
<td>Msrcx1f1</td>
<td>forward</td>
<td>ATGCGCGGGCGCGAAG</td>
<td>multiple non-specific bands</td>
<td>rcx1-2</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Msrcx2r1</td>
<td>reverse</td>
<td>AACCACCCCTCGCTCCCGTC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Msrcx3f2</td>
<td>forward</td>
<td>CGCGAAGAGGGATTTCCTGCAAGG</td>
<td>multiple non-specific bands</td>
<td>rcx3-8</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Msrcx8r1</td>
<td>reverse</td>
<td>CTGCGTACTTCACTTCTTGG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qSW5</td>
<td>Msqsw5x1f1</td>
<td>forward</td>
<td>GCAAGCGAGAGGAGGATGG</td>
<td>multiple non-specific bands</td>
<td>qsw5x1</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>Msqsw5x1r1</td>
<td>reverse</td>
<td>CACTGCGGCGGCGGTCCTTCC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wx</td>
<td>Mswxx1f1</td>
<td>forward</td>
<td>ATGTCGGCTCTGACCAC</td>
<td>single bright band</td>
<td>wxx1-10</td>
<td>S</td>
</tr>
<tr>
<td>Gene</td>
<td>Forward primer</td>
<td>Reverse primer</td>
<td>Band configuration</td>
<td>Accession</td>
<td>Comment</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>----------------</td>
<td>----------------</td>
<td>--------------------</td>
<td>-----------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Mswxx10r2</td>
<td>reverse</td>
<td>GAACGATCTGAACGTCCTC</td>
<td>single bright band</td>
<td>wxx9-14</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Mswxx9f1</td>
<td>forward</td>
<td>TGAAGGCTGGAATCCTGG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mswxx14r1</td>
<td>reverse</td>
<td>GCAGCAGCTGTTTCTTG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sd-1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mssd1x1f1</td>
<td>forward</td>
<td>CACCGCCCATGGACTCCAC</td>
<td>multiple non-specific bands</td>
<td>sdx1-3</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Mssd3x3r2</td>
<td>reverse</td>
<td>CTGGCCTGCTGACCTG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MOC1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Msmoc1x1f1</td>
<td>forward</td>
<td>GGTGGCTGCAAAGAAATGG</td>
<td>multiple non-specific bands</td>
<td>moc1x1</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Msmoc1x1r1</td>
<td>reverse</td>
<td>GACGACGACGCTGCAAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hd3a</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mshd3ax1f1</td>
<td>forward</td>
<td>ACCCTCTGGTGCTGGTAG</td>
<td>multiple non-specific bands</td>
<td>hd3ax1-4</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Mshd3ax4r1</td>
<td>reverse</td>
<td>TGGCGCTGCTGAGAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hd6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mshd6x1f1</td>
<td>forward</td>
<td>TACGCGACGTCACGTG</td>
<td>double bright bands</td>
<td>hd6x1-5</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Mshd6x5r1</td>
<td>reverse</td>
<td>CACCGAAATGCAACAAAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mshd6x4f1</td>
<td>forward</td>
<td>GCATTAGATTACTGCCATTCC</td>
<td>double bright bands</td>
<td>hd6x4-10</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Mshd6x10r1</td>
<td>reverse</td>
<td>CTGGCCTGCTGCTATTCCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>sh4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mssh4x1f1</td>
<td>forward</td>
<td>GGACTACCAGCAAGGGCAAC</td>
<td>a bright band with a few weaker non-specific bands</td>
<td>sh4x1-2</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Mssh4x2r1</td>
<td>reverse</td>
<td>GAGACGAGGCGTGGATG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BADH2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Msbadh2x1f3</td>
<td>forward</td>
<td>CAGCGGCAGCTCTCGTC</td>
<td>single bright band</td>
<td>badh2x1-5</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Msbadh2x5r3</td>
<td>reverse</td>
<td>GGAAGCAATCCAGTGGTTTTAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Msbadh2x5f1</td>
<td>forward</td>
<td>TGGCAACATGGAAGGTAGC</td>
<td>double bright bands</td>
<td>badh2x5-9</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Msbadh2x9r1</td>
<td>reverse</td>
<td>TAAAGCAAGCATGGTCACAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Msbadh2x8f1</td>
<td>forward</td>
<td>CCTGGTTCACTGCCAATTGG</td>
<td>single bright band</td>
<td>badh2x8-15</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Msbadh2x15r2</td>
<td>reverse</td>
<td>TTGGATGGGGATTTGTACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GW2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Msgw2x1f4</td>
<td>forward</td>
<td>ATGGGGAACAGGATAGGG</td>
<td>double bright bands</td>
<td>gw2x1-5</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Msgw2x5r2</td>
<td>reverse</td>
<td>AAGGATGTGGCTGCAATATCTCG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Msgw2x4f2</td>
<td>forward</td>
<td>TGCCATTTGCAAAACTC</td>
<td>single bright band</td>
<td>gw2x4-8</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Msgw2x8r2</td>
<td>reverse</td>
<td>TACAACATGGCAATTC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GIF1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Msgif1x1f1</td>
<td>forward</td>
<td>TGCTGCTCAGACGCTC</td>
<td>multiple non-specific bands</td>
<td>gif1x1-3</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Msgif1x3r2</td>
<td>reverse</td>
<td>GCAGTTGTCAGCGCTTC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

104
<table>
<thead>
<tr>
<th></th>
<th>Primer Set</th>
<th>Forward/Reverse</th>
<th>Sequence</th>
<th>Result</th>
<th>Code</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Msgif1x3f1</strong></td>
<td>forward</td>
<td>TGTGGGGGAACATCGTG</td>
<td>single bright band</td>
<td>gif1x3-7</td>
<td>S</td>
<td>UPM</td>
</tr>
<tr>
<td><strong>Msgif1x7r1</strong></td>
<td>reverse</td>
<td>ATCTCCCATGCCTGAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ghd7</strong></td>
<td><strong>Msghd7x1f1</strong></td>
<td>forward</td>
<td>AGCAGCCGGAGAAGGATG</td>
<td>multiple non-specific bands</td>
<td>ghd7x1-2</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>reverse</td>
<td>ATAGGCTTTTCTGGACGATC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PROG1</strong></td>
<td><strong>Msprog1x1f1</strong></td>
<td>forward</td>
<td>ATCCCTCATCGGCTTTCTTG</td>
<td>multiple non-specific bands</td>
<td>prog1x1</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>reverse</td>
<td>AGGACAAGCCTCCTGCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Primers were not designed for *Ehd1* and *GS3* as the coverage was not sufficient for effective primer designs.

¹ U indicates UPM-derived primers and S indicates SSPM-derived primers
Table 3.3.5.2  Percentage identity of putative homologues of rice genes successfully amplified in *M. stipoides* using UPM and SSPM by Sanger sequencing of the PCR fragments

<table>
<thead>
<tr>
<th>Method</th>
<th>Gene</th>
<th>Amplicon</th>
<th>Oryza exon(^1)</th>
<th>Percentage identity with <em>Oryza</em> exons(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPM</td>
<td><em>qSH-1</em></td>
<td>qsh1x1-4</td>
<td>1</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>85</td>
</tr>
<tr>
<td>UPM</td>
<td><em>sh4</em></td>
<td>sh4x1-2</td>
<td>1 to 2</td>
<td>81</td>
</tr>
<tr>
<td>UPM</td>
<td><em>GW2</em></td>
<td>gw2x1-5 larger fragment</td>
<td>1 to 5</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gw2x1-5 smaller fragment</td>
<td>1 to 5</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gw2x5-6</td>
<td>5 to 6</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gw2x7-8</td>
<td>8</td>
<td>86</td>
</tr>
<tr>
<td>SSPM</td>
<td><em>Wx</em></td>
<td>wxx2-10</td>
<td>2</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>wxx9-14</td>
<td>9 to 13</td>
<td>89</td>
</tr>
<tr>
<td>SSPM</td>
<td><em>Hd6</em></td>
<td>hd6x1-5 larger fragment</td>
<td>4</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hd6x1-5 medium fragment*</td>
<td>4</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hd6x1-5 smaller fragment</td>
<td>4</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hd6x4-10 larger fragment</td>
<td>4</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hd6x4-10 larger fragment 7 to 9</td>
<td>7 to 9</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hd6x4-10 smaller fragment</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hd6x4-10 smaller fragment 7 to 9</td>
<td>7 to 9</td>
<td>92</td>
</tr>
<tr>
<td>SSPM</td>
<td><em>sh4</em></td>
<td>sh4x1-2</td>
<td>1 to 2</td>
<td>82</td>
</tr>
<tr>
<td>SSPM</td>
<td><em>BADH2</em></td>
<td>badh2x1-5</td>
<td>2 to 4</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>badh2x5-9 larger fragment*</td>
<td>6 to 7</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>badh2x5-9 smaller fragment</td>
<td>6 to 7</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>badh2x8-15</td>
<td>12 to 14</td>
<td>91</td>
</tr>
<tr>
<td>SSPM</td>
<td><em>GW2</em></td>
<td>gw2x1-5 larger fragment*</td>
<td>1</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gw2x1-5 larger fragment*</td>
<td>3 to 4</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gw2x1-5 smaller fragment*</td>
<td>1 to 5</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gw2x4-8</td>
<td>8</td>
<td>85</td>
</tr>
<tr>
<td>SSPM</td>
<td><em>GIFI</em></td>
<td>gif1x3-7</td>
<td>3</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>87</td>
</tr>
</tbody>
</table>

\(^*\) indicates that during gel extraction the fragments did not completely separate, resulting in ambiguous base calls during Sanger sequencing.

\(^1\) Some of the exons were complete and some were partial.

\(^2\) Percentage identity with *Oryza* exons was done by aligning the *M. stipoides* exon(s) with reference exons from *Oryza* using ClustalW (http://www.ebi.ac.uk/clustalw/, last accessed 31st December 2010).
Species specific primers were designed for 16 out of the 18 genes using the MPS *M. stipiodes* data. A total of 24 PCR primer pairs were designed, some genes requiring more than one primer pair for full coverage (Figure 3.3.1 and Table 3.3.5.1). PCR amplification using 13 of the 24 primer pairs yielded multiple nonspecific fragments which could not be sequenced. The remaining 11 primer pairs produced amplicons that showed greater than 75% sequence identity in putative coding regions when aligned back to *Oryza*. Photos of agarose gels of the successful PCR ampicons are shown on Figures 3.3.6.1. Two of these 11 primer pairs produced additional fragments which showed low identity (49% and 55%) to *Oryza* (Table 3.3.5.2).
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

(a) Optimised Wx exons 1-10

(b) Optimised Wx exons 9-14

(c) Optimised Hdl6 exons 1-5
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

(d)

(e)

(f)

(g)
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

(h)

(i)

(j)
Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes

Figure 3.3.6.1  Photos of agarose gels showing the results of optimized SSPM PCR amplifications (a) Wx exons 1-10, (b) Wx exons 9-14, (c) Hd6 exons 1-5, (d) Hd6 exons 4-10, (e) Sh4 exons 1-2, (f) BADH2 exons 1-5, (g) BADH2 exons 5-9, (h) BADH2 exons 8-15, (i) GW2 exons 1-5, (j) GW2 exons 4-8, and (k) GIF1 exons 3-7

3.4 Discussion

The utility of UPM is well documented (McIntosh, Pacey-Miller et al. 2005; Ramirez, Fleming et al. 2004; Shapter, Eggler et al. 2009; Zeid, Yu et al. 2010), however there are three major limitations to this design method: firstly, identifying sufficient areas of conservation between species for potential primer design which capture enough of the target gene homologue; secondly, whether these areas of conservation meet the minimum criteria for PCR primer design; and lastly, whether the universal primers are specific enough to amplify the target homologue successfully. The use of degenerate primers in areas of mediocre conservation is a common strategy for working around the first limitation of the UPM, however this increases the risk of amplification of unrelated sequences and/or other members of a gene family, due to the reduced specificity of the primers (Linhart and Shamir, 2005).
In this study, primer sites were identified for 11 of the 18 putative gene homologues using the UPM. Due to the established utility of the UPM we performed amplifications on just three of the genes, \( qSH-1, \, sh4 \) and \( GW2 \). Of the five primer pairs utilised, all produced PCR fragments which were identified as putative homologues in \( M. \, stipoides \). Universal primer design was possible for six gene homologues using non-degenerate primers and five gene homologues using a combination of non-degenerate and degenerate primers. Although the UPM may not have resulted in the successful amplification of all of the gene homologues, where degenerate primers were designed, the successful amplification using all of the primer pairs suggests that the UPM of homologue capture is highly successful were primer location is not a major limitation.

Using the SSPM, 16 of the 18 genes targeted had non-degenerate primer pairs identified, allowing for the primer design for an additional five genes compared to the UPM and theoretically improved specificity for all the primers designed. However, of the 24 PCR primer pairs that were tested in \( M. \, stipoides \) only 11 were successful, capturing six putative homologues of the desired genes in \( M. \, stipoides \). Sanger sequencing confirmed more than 75% sequence identity with the reference gene exons for all 11 primer pairs (Table 3.3.5.1), although two additional fragments with low sequence identity were identified.

Massively parallel sequencing is a rapidly evolving technology, particularly in terms of the capacity of the short read platforms such as Illumina GAII. In the past year sequence output potential of the platform has more than doubled in terms of length of individual reads and total read number per lane. The coverage bias which leads to parts of the genome/genes being over-represented and other parts being consequently under represented (Harismendy, Ng et al. 2009) currently associated with the short read platforms is also a limiting factor for this methodology. While complete genome or gene coverage of target species is not necessary to mine gene homologues using the SSPM the efficacy of the SSPM will be greatly enhanced as the read length, read number and evenness of genomic coverage improves rapidly over the next generation of chemistry and hardware. We were able to capture six genes from a theoretical 2X genome coverage using 36 bp paired end reads on the Illumina Genome Analyzer platform. The latest improvements in the platform
allows for 150 bp paired end read lengths and up to 640 million paired-end reads per flow cell (http://www.illumina.com/systems/genome_analyzer_iix.ilmn, last accessed 10th November 2012), equating to 80 million paired-end reads per lane, which would result in a theoretical coverage of the *M. stipoides* genome of 27X. Obviously the greater proportion of a gene’s sequence available for primer design the more rigorous the design criteria can be and this should improve the amplification success from this methodology.

Rice as a model cereal plant has a relatively small genome size, approximately 420-466 mega base pairs (Goff, Ricke et al. 2002; Yu, Hu et al. 2002). Other crop and model plants have much bigger genomes such as maize at 2.3 gbp (Schnable, Ware et al. 2009), barley at larger than 5 gbp (Schulte, Close et al. 2009) and hexaploid wheat at 17 gbp (Paux, Sourdille et al. 2008). The improvements in the MPS technology may not be sufficient to produce gap-free whole genome sequence of these species at low cost, especially with the presence of coverage bias as previously mentioned. There are a number of methods to increase coverage of gene regions in whole genome sequencing. Some of the commonly used methods are methylation filtering, High C\(_t\) selection and microarrays. Whitelaw, Barbazuk et al. (2003) successfully combined methylation filtering and High C\(_t\) selection in the enrichment of gene-coding sequences in maize. Chou, Liu et al. (2003) enriched DNA samples with high-density oligonucleotide microarray to enrich a targeted gene. The combination of gene enrichment methods with MPS technology has the potential of improving targeted gene sequences that will allow for more species specific primer design, resulting in capturing desired gene homologues in the target species.

The UPM has been used for rapid amplification of gene homologues in another species when the target species has little or no known DNA sequence data. When the target species has DNA sequence data generated from MPS, the SSPM can be used a method for the rapid amplification of important gene homologues. Even with the limitations of the platform that were current at the time of data collection for this study, the SSPM allowed for more primer combinations with greater gene coverage of most of the target genes. The SSPM method also eliminated the need for obtaining target gene sequence in more than one species and the need to design
Another advantage of being able to design outside highly conserved areas of the gene is that they are less susceptible to mispriming with members of a gene family or closely related gene. SSPM primers were tested under ‘first pass’ conditions and with a single primer pair per amplicon therefore the success rate of the PCR amplifications reported here for SSPM represents a minimum baseline of success which should be easily improved upon with the design of alternative primer pairs available from the MPS data, and optimisation of PCR reagents and conditions.

Both UPM and SSPM methods involve direct sequencing of PCR products from an allotetraploid species. It would be expected that most primers pairs would amplify both homologues and that polymorphisms would therefore be observed in the sequence data – indeed, any indels would prevent the sequence from being read from the traces beyond the site of the indels.

The SSPM is expected to produce species specific DNA sequence; however our data show that this method still produced non-specific DNA sequence. This could be explained by a number of possible causes. Low coverage of our target genome may have led to poor initial alignment with the reference sequences leading to incorrect primer site identification. It is expected that improved MPS technology will result in higher coverage and partially resolve this problem. Primer design stringency may have been too low, however higher MPS coverage would allow for the design of longer and more specific primers.

MPS data can be used to mine desired genes from poorly characterised species using a single well described species’ sequence data as a reference. Universal primers designed from sequence from multiple species have a number of limitations and involves more steps than utilising MPS. Improvements in the MPS technology will enable more and better quality data to be generated rapidly. This method will be applicable for gene homologue mining in any poorly described species where an appropriate single reference gene is available. In this study, MPS facilitated the amplification of homologues of six previously undescribed genes in a species with limited genetic data available, illustrating that the SSPM will be a valuable complementary approach to the discovery of useful genes in wild crop relatives.
3.5 References


Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes


Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes


Chapter 3: Characterising homologues of crop domestication genes in poorly described wild relatives by high-throughput sequencing of whole genomes


Chapter 4 - SNP discovery in candidate domestication genes of *M. stipoides*

4.1 Introduction

The need to produce more food to feed the world has driven plant breeders to look for sources to improve the quality and yield of crops. Crop wild relatives (CWR) have been used as the source of novel alleles to improve characteristics such as pest and disease resistance, yield and nutrition content, drought tolerance, of established crops (Cox, Bender et al. 2002; Hajjar and Hodgkin 2007; Azhar and Heslop-Harrison 2008; Feuillet, Langridge et al. 2008; Singh, Sharma et al. 2008; Flint-Garcia, Bodnar et al. 2009). In this study, the term homoeologues (the correct term for homologues derived from the diploid parents of a tetraploid) is replaced by alleles, as previously used by other researchers in this field. The domestication of CWR’s also has the potential to generate new crops possessing less resource requirements.

Severe genetic bottlenecks have occurred during domestication and selection of crops (Eyre-Walker, Gaut et al. 1998; Hamblin, Casa et al. 2006). Many genetic variations that existed in crops prior to domestication were removed from the gene pool. Induced mutations can rapidly produce phenotypes that are rare in wild populations, bypassing the need to generate many generations through selection and crossings. Mutated crops have been more readily accepted than crops generated by transgenic means (Baenziger, Russell et al. 2006; Parry, Madgwick et al. 2009). These induced mutation individuals can therefore form the basis of the breeding lines of crop and CWR.

To ensure that any gene of interest carries desired mutations, the populations of induced mutations may need to be very large; the size required is dependent on the dosage of mutagen and the level of gene duplication in the genome (Parry, Madgwick et al. 2009). Polyploid species tend to have a very high tolerance of mutations due to the complementation of important genes by homologues, but this
genetic buffering causes recessive mutations in single homologues of genes to be less likely to show a phenotype, and it becomes necessary to identify mutations in each homologue of the target gene and bring these together by crossing (Parry, Madgwick et al. 2009). A popular chemical substance used for inducing mutation is ethyl methanesulfonate (EMS), and of the mutations it induces, >99% are G/C-to-A/T transitions (Greene, Codomo et al. 2003; Kim, Schumaker et al. 2006). Some advantages of using EMS are: ability to generate loss- or gain-of-function mutants enabling an understanding of the role of specific amino acid residues in protein function. Generating weak nonlethal alleles can provide useful information for understanding the functions of essential genes (Kim, Schumaker et al. 2006). EMS mutagenesis has been extensively used for generating breeding lines (Maluszynski, Ahloowalia et al. 1995; Ahloowalia, Maluszynski et al. 2004; Kim, Schumaker et al. 2006).

The advancement in sequencing technology through massively parallel sequencing (MPS) has accelerated the availability of genomic information for complex and non-model organisms (Morozova and Marra 2008; Trick, Long et al. 2009). MPS rapidly generates large amounts of data with relatively cheap cost compared to Sanger sequencing (You, Huo et al. 2011). MPS was successfully used to design primers to mine rice domestication gene homologues in *M. stipoides* (Malory, Shapter et al. 2011). As a follow up, MPS of the best genes homologues captured was performed on the samples of both natural and EMS-induced mutation populations of *M. stipoides*.

MPS technology is capable of producing millions of reads allowing high throughput sequencing of whole genomes with low cost and a short timeline. It still costs thousands of dollars per lane, so it is an expensive way to analyse variants in individuals, especially because in many cases (depending on the size of the genome), more than one run is required to completely sequence an individual. It is not currently feasible to analyse variations in a large population, or to study non-model and complex organisms for research groups with a small budget (Ekblom and Galindo 2011). As an alternative, MPS of pooled DNA samples have been used successfully by a number of researchers to identify specific SNPs in the species that they studied (Ingman and Gyllensten 2008; Druley, Vallania et al. 2009; Erlich,
Chapter 4: SNP discovery in candidate domestication genes of M. stipoides


Pooling schemes offer insights into allele frequencies, but do not offer the identity of an allele carrier (Prabhu and Pe'er 2009; Sexton, Henry et al. 2010). Some of the methods used to facilitate identifying allele-carrying individuals include barcoding (Erlich, Chang et al. 2009; Smith, Heisler et al. 2010; Bybee, Bracken-Grissom et al. 2011; Gholami, Bekele et al. 2012) or use of a framework of an overlapping pool design, where each individual sample is resequenced in several pools (Erlich, Chang et al. 2009; Prabhu and Pe'er 2009). Both methods are time consuming and very expensive. In this study, a different strategy was conducted: combining MPS of pooled samples and a genotyping platform, Sequenom iPLEX MassARRAY (SPM, Sequenom Inc., San Diego, CA), as previously done by other researchers (Bundock, Eliott et al. 2009; Buggs, Chamala et al. 2010; Liu, Chen et al. 2010; Sexton, Henry et al. 2010; Kharabian-Masouleh, Waters et al. 2011; Uh, Deelen et al. 2012). SPM was chosen to genotype the SNPs generated by the MPS because the platform has features that allow accurate custom SNP genotyping assay with modest multiplexing and minimal assay setup costs due to unmodified oligonucleotide primers (Gabriel, Ziaugra, et al. 2009). The platform involves homogeneous reaction format with a single extension primer to generate allele-specific products with distinct masses, multiplexed PCR reactions, a single termination mix and universal reaction conditions for all SNPs, small reagent volumes, and is capable of a throughput of >100,000 genotypes/day/system (Gabriel, Ziaugra, et al. 2009).

To achieve good coverage and cost effectiveness, specific polymerase chain reaction (PCR) amplicons were sequenced using MPS, instead of whole genomes.
Other researchers have used selected partial regions instead of whole genomes for the same reasons (Harismendy and Frazer 2009; Brown, Lo et al. 2010; Durstewitz, Polley et al. 2010; Sexton, Henry et al. 2010; Bybee, Bracken-Grissom et al. 2011; De Leeneer, Hellemans et al. 2011; Gholami, Bekele et al. 2012).

Extensive progress has been achieved in mapping short reads from different MPS platforms to a reference genome, and then identifying variants between individual sequences and the reference genome. However, only a few attempts have been reported utilising MPS to discover SNPs when there is no reference genome (Everett, Grau et al. 2011; You, Huo et al. 2011). The short reads produced by MPS remain a challenge for assembly; many studies have focused on resequencing and gene expression studies in organisms with a sequenced genome (Everett, Grau et al. 2011). Some of the difficulties in assembling the short reads generated by MPS of non-model organisms are gene duplications, genetic polymorphism, alternative splicing and transcription noise (Cahais, Gayral et al. 2012).

SPM has been commonly used by researchers to genotype diploid organisms (Daelemans, Ritchie et al. 2010; Pattemore, Rice et al. 2010; Sexton, Henry et al. 2010; Kharabian-Masouleh, Waters et al. 2011). In this study, MPS using Illumina Genome Analyzer IIx platform (Illumina GA II, Illumina, San Diego, CA, USA) of pooled PCR amplicons of putative homologues of rice domestication genes in *M. stipoides* was performed to screen for SNPs. It was followed by validation of a subset of the SNPs using the SPM platform.

4.2 Materials and Methods

4.2.1 Plant materials


4.2.2 Seed treatment and sample preparation and DNA extraction

The protocol was as described in Shapter, Cross et al. (2013). In summary, seeds were treated with 145 mM EMS to induce mutations, whereas the control seeds
were treated with deionised water. The seeds were planted immediately and grown in a glasshouse until maturity. Seeds were harvested, planted and grown in a glasshouse. Young leaves were collected from 863 juvenile M2 seedlings and 109 S1 individuals for DNA extraction. Prior to PCR amplification DNA was pooled from five individuals and 10ng of pooled template was used per PCR reaction. DNA was quantified, normalised and pooled in equimolar proportions at each step in an attempt to maintain relative allele frequencies in the subsequent Illumina GAII sequence data.

4.2.3 PCR amplicon

The three gene homologues chosen were a candidate gene for the QTL grain width and weight 2, GW2 (Song, Huang et al. 2007); betaine aldehyde dehydrogenase 2, BADH2 (Bradbury, Fitzgerald et al. 2005) and a candidate gene for the QTL heading date 6, Hd6 (Yamamoto, Lin et al. 2000). There were two amplicons for GW2, covering exons 1 to 5 and exons 4 to 8 respectively. For BADH2, there were three amplicons, covering exons 1 to 5, exons 5 to 9 and exons 8 to 15 respectively. For Hd6, there were two amplicons, covering exons 1 to 5 and exons 4 to 10 respectively. In total seven mega pools of PCR amplicons being sequenced using Illumina GAII. Primer details and PCR amplification protocol used were as described in Malory et al. (2011).

4.2.4 MPS of pooled PCR amplicons

There were a total of three lanes of MPS performed. One lane consisted of 109 control samples, one lane of 754 mutant samples, and another lane of a subset of 109 mutant samples. The protocol was as described in Shapter, Cross et al. (2013). Southern Cross Plant Genomics performed the MPS. The calculation of pooled PCR products is shown in Appendix 3.

4.2.5 Assembly to rice as a reference versus de novo assembly

Sequence data were trimmed using CLC Genomics Workbench version 4.0.3 (www.clcbio.com, last accessed 02/11/12). Reads with a quality score of less than 0.001 were discarded and paired-end reads were trimmed to a minimum of 30 base
pairs (bp). Reference assembly was undertaken with a mismatch cost of 2, indel costs of 3, length fraction of 0.8 and similarity of 0.8, minimum distance for paired end reads of 180bp with a maximum of 340bp, and non-specific matches ignored.

The *de novo* assembly parameters were: length fraction of 0.8 and similarity of 0.95, minimum distance for paired end reads of 140bp with a maximum of 320bp, and non-specific matches randomised. Minimum acceptable average coverage was set at 100x and all the resulting contigs were aligned in Sequencher® V4.6. Only large contigs (over 100 bp) were selected to be matched back to exon sequences of each gene homologue. Sequencher® file of the de novo assemblies of the reads combined according to PCR contigs are shown on Appendix 4 on the CD accompanying this thesis. Reference sequence created as the result of *de novo* assembly contained exons and intron sequences and were allocated as per amplicons.

4.2.6 **Illumina SNP detection and data analysis**

SNP detection parameters were set at; window length -21, maximum number of gaps or mismatches -2, SNP minimum quality score -30 and quality score for the surrounding bases- 30, minimum coverage required -100X , with a minimum variant frequency of 1%, was designed to capture all high quality polymorphisms. Analysis of the CLC SNP discovery output was conducted using Microsoft Excel 2007. Based on comparative alignment of these gene homologues and their splice junction sites to rice, putative exon/intron boundaries were assigned to the *M. stipoides* reference sequence from *de novo* assembly and this was used to assign putative functionality of the SNPs. Only SNPs in the exons were further analysed.

4.2.7 **SPM analysis**

A subset of SNPs identified by MPS sequencing was assessed using the Assay Design software, version 4.0 (Sequenom Inc., San Diego, CA). A MassARRAY was designed and optimised for the highest number of putative SNPs that could lead to non-synonymous protein changes. SPM output was processed using Typer 4.0 software. Individual’s allelic variations were identified from the report produced by Typer 4.0. The SNPs included in the SPM are shown in Appendix 7.
4.3 Results

4.3.1 Illumina sequencing

The Illumina sequencing resulted in between 51.3 million base pairs (mbp) to 69.9 mbp of raw sequence. Trimmed sequence was between 46.2 mbp to 62.6 mbp (Table 4.3.1).

Table 4.3.1 Illumina sequencing results for a) control 109 lane; b) mutant 754 lane; c) mutant 109 lane

<table>
<thead>
<tr>
<th>Read type</th>
<th>No. reads</th>
<th>Ave. length</th>
<th>No. reads after trim</th>
<th>% trimmed</th>
<th>Ave. length after trim</th>
</tr>
</thead>
<tbody>
<tr>
<td>a paired</td>
<td>60,597,256</td>
<td>73.9</td>
<td>55,295,850</td>
<td>91.25</td>
<td>66.9</td>
</tr>
<tr>
<td></td>
<td>single</td>
<td>162,276</td>
<td>66.1</td>
<td>115,701</td>
<td>71.30</td>
</tr>
<tr>
<td>b paired</td>
<td>51,397,886</td>
<td>71.0</td>
<td>46,224,954</td>
<td>89.94</td>
<td>63.4</td>
</tr>
<tr>
<td></td>
<td>single</td>
<td>198,424</td>
<td>47.5</td>
<td>65,644</td>
<td>33.08</td>
</tr>
<tr>
<td>c paired</td>
<td>69,877,758</td>
<td>63.0</td>
<td>62,595,954</td>
<td>89.58</td>
<td>55.1</td>
</tr>
<tr>
<td></td>
<td>single</td>
<td>105,427</td>
<td>56.9</td>
<td>49,845</td>
<td>47.28</td>
</tr>
</tbody>
</table>

4.3.2 Assembly to rice as a reference versus de novo assembly

4.3.2.1 Assembly to rice as a reference

In this study, of all three of the target genes, only putative whole or partial exon sequences have been characterised in *M. stipoides* (Malory, Shapter et al. 2011, Chapter 3). In our preliminary analysis, assembly to complete rice gene references was not effective for SNP detection because of the extensive polymorphisms and large number of insertions and deletions (indels), especially in the introns. Real SNPs and SNPs caused by error in alignment back to rice could not be differentiated.

Assembly of all the reads from the three pools to rice gene sequence resulted in a very high number of SNPs and indels. Coverage was very poor and intermittent (Figure 4.3.2.1). It was decided that de novo assembly using all the reads was necessary to achieve good SNP detection.
Figure 4.3.2.1  Low and intermittent coverage of assembly of trimmed *M. stipoides* reads to rice reference genes for all three target genes. A1) *BADH2* exons 1-5 amplicon, A2) exons 5-9 amplicon and A3) exons 8-15 amplicon. B1) *GW2* exons 1-5 amplicon and B2) exons 1-5 amplicon. C1) *Hd6* exons 1-5 amplicon and C2) exons 4-10 amplicon. Top plot = gene structure in rice; Lower plot = coverage pattern. Axes: Y = coverage; X = base pairs.
4.3.2.2 De novo assembly

*De novo* assembly of the reads was performed to generate *M. stipoides* putative reference gene homologue sequence. Only long contigs over 100 bp with high coverage that match back to rice gene reference sequence were used as reference. SNP analyses were conducted on the reads assembled to these *de novo* generated references. Details of *de novo* assembly are shown in Appendix 4.

The control 109 pool resulted in 869 reference contigs, mutant 109 resulted in 796 reference contigs and the mutant 754 resulted in 186 references contigs. All these references underwent a secondary reference assembly in Sequencher® V4.6. A total of two distinct *GW2* references, two distinct *BADH2* references and four distinct *Hd6* references were generated after the contigs were aligned back to rice exonic sequences. Most of the polymorphisms and indels were in the intron regions, with few in the exons. These distinct references were treated as putative alleles and were used for SNP detection. Genbank files (.gbk files) showing the sequence information are provided on Appendix 5. These files can be viewed by free software such as CLC Sequence Viewer V.6.7.1 (http://www.clcbio.com/products/clc-sequence-viewer/, last accessed 11th November 2012).

4.3.3 SNP detection

Table 4.3.3.1 shows the summary of SNP detection using all reads to the de novo reference gene amlicons. As mentioned before, only a small proportion of the SNPs occurred in the exons, with even smaller proportion that could lead to non-synonymous protein change. Details of all SNPs detected by CLC Genomics Workbench are shown in Appendix 6. Table 4.3.3.2 shows the summary of SNP frequency.
Table 4.3.3.1 Summary of SNP detection results using CLC Genomics Workbench version 4.0.3

<table>
<thead>
<tr>
<th>Homologue</th>
<th>Putative allele</th>
<th>Amplicon</th>
<th>Total SNPs</th>
<th>SNPs in exon</th>
<th>Exon size (bp)</th>
<th>Non synonymous SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BADH2</strong></td>
<td>1</td>
<td>Exons 1-5</td>
<td>63</td>
<td>20</td>
<td>522</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 1-5</td>
<td>102</td>
<td>23</td>
<td>503</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Exons 5-9</td>
<td>69</td>
<td>19</td>
<td>399</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 5-9</td>
<td>50</td>
<td>18</td>
<td>399</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Exons 8-15</td>
<td>120</td>
<td>32</td>
<td>740</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 8-15</td>
<td>93</td>
<td>31</td>
<td>727</td>
<td>9</td>
</tr>
<tr>
<td><strong>GW2</strong></td>
<td>1</td>
<td>Exons 1-5</td>
<td>56</td>
<td>8</td>
<td>536</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 1-5</td>
<td>91</td>
<td>7</td>
<td>425</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Exons 4-8</td>
<td>118</td>
<td>30</td>
<td>978</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 4-8</td>
<td>128</td>
<td>15</td>
<td>503</td>
<td>9</td>
</tr>
<tr>
<td><strong>Hd6</strong></td>
<td>1</td>
<td>Exons 1-5</td>
<td>128</td>
<td>48</td>
<td>620</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 1-5</td>
<td>94</td>
<td>33</td>
<td>620</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Exons 2-5</td>
<td>111</td>
<td>26</td>
<td>569</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Exons 2-5</td>
<td>73</td>
<td>23</td>
<td>569</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Exons 4-10</td>
<td>83</td>
<td>22</td>
<td>584</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Exons 4-7</td>
<td>55</td>
<td>14</td>
<td>398</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.3.3.2 Summary of SNP frequency

<table>
<thead>
<tr>
<th>Homologue</th>
<th>Putative allele</th>
<th>Amplicon</th>
<th>Amplicon size</th>
<th>Total SNPs</th>
<th>Individuals</th>
<th>Base pairs scanned</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BADH2</strong></td>
<td>1</td>
<td>Exons 1-5</td>
<td>2173</td>
<td>63</td>
<td>972</td>
<td>2112156</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 1-5</td>
<td>1873</td>
<td>102</td>
<td>972</td>
<td>1820556</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Exons 5-9</td>
<td>1738</td>
<td>69</td>
<td>972</td>
<td>1689336</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 5-9</td>
<td>1666</td>
<td>50</td>
<td>972</td>
<td>1619352</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Exons 8-15</td>
<td>3035</td>
<td>120</td>
<td>972</td>
<td>2950020</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 8-15</td>
<td>2967</td>
<td>93</td>
<td>972</td>
<td>2883924</td>
</tr>
<tr>
<td><strong>GW2</strong></td>
<td>1</td>
<td>Exons 1-5</td>
<td>2094</td>
<td>56</td>
<td>972</td>
<td>2035368</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 1-5</td>
<td>2383</td>
<td>91</td>
<td>972</td>
<td>2316276</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Exons 4-8</td>
<td>4229</td>
<td>118</td>
<td>972</td>
<td>4110588</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 4-8</td>
<td>4173</td>
<td>128</td>
<td>972</td>
<td>4056156</td>
</tr>
<tr>
<td><strong>Hd6</strong></td>
<td>1</td>
<td>Exons 1-5</td>
<td>2911</td>
<td>128</td>
<td>972</td>
<td>2829492</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Exons 1-5</td>
<td>2732</td>
<td>94</td>
<td>972</td>
<td>2655504</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Exons 2-5</td>
<td>2331</td>
<td>111</td>
<td>972</td>
<td>2265732</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Exons 2-5</td>
<td>2400</td>
<td>73</td>
<td>972</td>
<td>2332800</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Exons 4-10</td>
<td>2089</td>
<td>83</td>
<td>972</td>
<td>2030508</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Exons 4-7</td>
<td>1380</td>
<td>55</td>
<td>972</td>
<td>1341360</td>
</tr>
</tbody>
</table>

Total bp scanned 39049128
Total SNPs 1434
SNPs/kbp 0.036723
After consolidating all the putative alleles, there were a total of 298 SNPs in exons identified (Table 4.3.3.1 and Appendix 6). There was no premature stop codon detected in the samples. There was also no intron splice site variant detected. As shown in Table 4.3.3.2, in total 39,049,128 bp of sequence was scanned resulting in 1434 SNPs (inside and outside exons) resulting in SNP frequency of 0.037 SNPs/kb.

Out of the 298 SNPs, 210 were analysed using the Sequenom Assay Design software, version 4.0 (Appendix 6). Most SNPs identified could not be genotyped because of the high SNP density in the genomic region they were identified. Ignoring this information would lead to mispriming and subsequent assay failure (Jones, Chu et al. 2009; Sexton, Henry et al. 2010). The best assay design possible was for 18 SNPs (seven for BADH2, six for GW2 and five for Hd6). These SNPs were: B1.1, B6.3, B9.1, B10.1, B11.2, B12.2, B13.3 MAJOR, G4.1, G5.1, G8v1.1.2 MAJOR, G8v1.2.1, G8v1.2.3, G8v1.3.5, H2v12.7 MAJOR, H3v34.2 MAJOR, H7v3.1, H9v3.3, and H10v3.1 MAJOR. These 18 SNPs were non-synonymous SNP, predicted to modify the amino acid, and hence protein structure.

### 4.3.4 SPM analysis

Of the 18 SNPs analysed, the SPM detection ranged from 18.9% to 98%. SNPs B1.1, B13.3 MAJOR and G8v1.2.1 had to be excluded from further analysis because SPM detection was lower than 50%, and too many no calls were recorded by SPM (Table 4.3.4). Furthermore, G8v1.2.1 and H9v3.3 had possible contamination because two of the blank controls showed allele readings (Table 4.3.4).

Most of the SNPs were detected by SPM as having both alleles present, except for H3v34.2 MAJOR and H10v3.1 MAJOR, where only the major alleles were present in samples successfully genotyped. SPM results and Illumina SNP analysis results do not correspond to each other in terms of allele frequency (Table 4.3.4). Illumina sequencing amplified selected PCR amplicons whereas SPM scans whole genomes using its own primer sets. The tetraploidy of *M. stipoides* complicates differentiations of the alleles present.
Table 4.3.4 Comparison between Illumina GAII SNP analysis in conjunction with the CLC Genomics Workbench and SPM results

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Heterozygosity</th>
<th>CLC prediction</th>
<th>% het</th>
<th>% major</th>
<th>% minor</th>
<th>% no call</th>
<th>% SPM hit</th>
<th>Comment</th>
<th>Blank controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1.1</td>
<td>GT</td>
<td>95/4.8</td>
<td>18.9</td>
<td>0</td>
<td>0</td>
<td>80.8</td>
<td>18.9</td>
<td>low success rate with sequenom, all het</td>
<td>3: all null</td>
</tr>
<tr>
<td>B6.3</td>
<td>GC</td>
<td>62.2/37.7</td>
<td>91.6</td>
<td>0.1</td>
<td>4.4</td>
<td>4</td>
<td>96.1</td>
<td>high success rate with sequenom, all het</td>
<td>3: all null</td>
</tr>
<tr>
<td>B9.1</td>
<td>GA</td>
<td>58.5/41.4</td>
<td>95.7</td>
<td>0.2</td>
<td>0.3</td>
<td>3.8</td>
<td>96.2</td>
<td>high success rate with sequenom, mostly het</td>
<td>3: all null</td>
</tr>
<tr>
<td>B10.1</td>
<td>TA</td>
<td>60.0/38.9</td>
<td>96</td>
<td>0.3</td>
<td>0.3</td>
<td>3.3</td>
<td>96.7</td>
<td>high success rate with sequenom, mostly het</td>
<td>3: all null</td>
</tr>
<tr>
<td>B11.2</td>
<td>GA</td>
<td>63.9/36.1</td>
<td>95.6</td>
<td>0.7</td>
<td>0.1</td>
<td>3.6</td>
<td>96.4</td>
<td>high success rate with sequenom, mostly het</td>
<td>3: all null</td>
</tr>
<tr>
<td>B12.2</td>
<td>CT</td>
<td>65.1/34.8</td>
<td>96</td>
<td>0.6</td>
<td>0.3</td>
<td>3.2</td>
<td>96.8</td>
<td>high success rate with sequenom, mostly het</td>
<td>3: all null</td>
</tr>
<tr>
<td>B13.3</td>
<td>GA</td>
<td>98.8/1.2</td>
<td>0.7</td>
<td>30.1</td>
<td>0.1</td>
<td>69.2</td>
<td>30.9</td>
<td>low success rate with sequenom, mostly major</td>
<td>3: all null</td>
</tr>
<tr>
<td>G4.1</td>
<td>GA</td>
<td>69.1/30.8</td>
<td>57.7</td>
<td>0.3</td>
<td>0.1</td>
<td>41.8</td>
<td>58.2</td>
<td>medium success rate with sequenom, mostly het</td>
<td>3: all null</td>
</tr>
<tr>
<td>G5.1</td>
<td>CT</td>
<td>71.8/28.1</td>
<td>96.2</td>
<td>1.1</td>
<td>0</td>
<td>2.6</td>
<td>97.4</td>
<td>high success rate with sequenom, mostly het</td>
<td>3: all null</td>
</tr>
<tr>
<td>G8v1.1.2 MAJOR</td>
<td>CT</td>
<td>88.8/11.2</td>
<td>86.8</td>
<td>0</td>
<td>0</td>
<td>13.2</td>
<td>86.8</td>
<td>high success rate with sequenom, all het</td>
<td>3: all null</td>
</tr>
<tr>
<td>G8v1.2.1</td>
<td>CT</td>
<td>67.7/32.3</td>
<td>25.4</td>
<td>0.1</td>
<td>2.1</td>
<td>72.4</td>
<td>27.6</td>
<td>low success rate with sequenom, mostly het</td>
<td>1 null, 2 showing minor</td>
</tr>
<tr>
<td>G8v1.2.3</td>
<td>GA</td>
<td>68/32</td>
<td>95.5</td>
<td>0</td>
<td>0.1</td>
<td>4.4</td>
<td>95.5</td>
<td>high success rate with sequenom, mostly het</td>
<td>3: all null</td>
</tr>
<tr>
<td>G8v1.3.5</td>
<td>CG</td>
<td>75.1/24.9</td>
<td>91.5</td>
<td>0.1</td>
<td>4.5</td>
<td>3.9</td>
<td>96.1</td>
<td>high success rate with sequenom, mostly het</td>
<td>3: all null</td>
</tr>
<tr>
<td>H2v12.7 MAJOR</td>
<td>AG</td>
<td>59.6/40.3</td>
<td>95.1</td>
<td>0.1</td>
<td>0</td>
<td>4.7</td>
<td>95.3</td>
<td>high success rate with sequenom, mostly het</td>
<td>3: all null</td>
</tr>
<tr>
<td>H3v34.2 MAJOR</td>
<td>AG</td>
<td>89.4/10.6</td>
<td>96.5</td>
<td>0</td>
<td>3.5</td>
<td>96.5</td>
<td>high success rate with sequenom, all major</td>
<td>3: all null</td>
<td></td>
</tr>
<tr>
<td>H7v3.1</td>
<td>GA</td>
<td>67.7/32.3</td>
<td>0.2</td>
<td>87.7</td>
<td>0</td>
<td>12</td>
<td>88.0</td>
<td>high success rate with sequenom, mostly major</td>
<td>3: all null</td>
</tr>
<tr>
<td>H9v3.3</td>
<td>CA</td>
<td>69.5/30.3</td>
<td>95.1</td>
<td>0</td>
<td>2.8</td>
<td>2.1</td>
<td>98.0</td>
<td>high success rate with sequenom, mostly het</td>
<td>1 null, 2 showing minor</td>
</tr>
<tr>
<td>H10v3.1 MAJOR</td>
<td>CT</td>
<td>61.2/38.8</td>
<td>95.8</td>
<td>0</td>
<td>4.2</td>
<td>95.8</td>
<td>high success rate with sequenom, all major</td>
<td>3: all null</td>
<td></td>
</tr>
</tbody>
</table>

SPM call cluster plot are shown in Figure 4.3.4. The cluster plots generated by Typer 4.0 show the distribution of the alleles detected by SPM for each SNP. Individuals carrying the allele(s) are shown on Appendix 8.
Chapter 4: SNP discovery in candidate domestication genes of *M. stipoides*
Chapter 4: SNP discovery in candidate domestication genes of M. stipoides
Chapter 4: SNP discovery in candidate domestication genes of M. stipoides
Figure 4.3.4  Cluster plot of the 18 SNPs genotyped using SPM. Axes: X= Low mass weight, Y= High mass weight. They show the distribution of the SPM for any particular expected SNP position in the population.
4.4 Discussion

The SNPs generated by Illumina GAIIx were detected by aligning the reads from all three lanes of sequencing against the *de novo* reference homologue amplicons generated by CLC Genomics Workbench (version 4.0.3). SPM was able to detect these SNPs in the individual samples. This demonstrated that CLC Workbench was capable of producing de novo alignment for a poorly characterised gene sequence.

Prabhu and Pe’er (2009) identified three practical concerns for pooled MPS strategies: (1) false-positives created by errors generated during amplification and sequencing; (2) false-negatives due to under sampling particular alleles made worse by non-uniform coverage; and consequently, (3) ambiguous identification of individual carriers in the presence of errors. In this study, strategies were in place to address these concerns. The first two are common with current MPS sequencing technology. Careful quantification of genomic DNA samples prior to pooling can reduce imbalances (Bansal, Tewhey et al. 2011). In this study, SNPs validated were not rare SNPs, only SNPs above 1% in frequency were analysed to avoid mistaking error or false positive (Druley, Vallania et al. 2009; Out, Minderhout et al. 2009). To address the third concern, a subset of SNPs identified by MPS was validated using the genotyping platform SPM.

Our target species, *M. stipodes* is a tetraploid. The presence of two genomes makes identification and analysis of SNPs more challenging than in diploid species, as for a given locus usually four versions of a DNA sequence have to be analysed simultaneously (Durstewitz, Polley et al. 2010). Identifying the alleles that carry the SNPs may not be conclusive. In this study, there is a large heterogeneity in the starting material. It is not clear whether the breeding line was derived from single seed descent or by bulk selection of plants from a population. To complicate matters, outcrossing events may have occurred during favourable conditions. The combination of tetraploidy and the variations in the breeding lines made the differentiation of natural and EMS-induced SNPs not possible. Recently some studies have shown that SPM can also be informative for the complex polyploid sugarcane (Bundock, Eliott et al. 2009; Serang, Mollinari et al. 2012). Sugarcane has been bred only very recently from wild progenitors. Like sugarcane, our
breeding line was in its infancy of domestication, having only been under selection for seven generations. However, both the Bundock, Eliott et al. (2009) and Serang, Mollinari et al. (2012) studies had the advantage of extensive knowledge of the parental lines, which was not the case in this study. The techniques applied by the previous two studies could therefore not be applied to this study.

In their research, Liu et al. (2010) noted that although Sequenom genotyping platform has the potential to provide quantitative data on the relative frequency of the two alleles, many SNPs actually exhibit nonlinear relationships between the input ratio and the detected allele frequency. This study produced similar deviations from linearity for some SNPs of the codominant category. It is further complicated by the tetraploid nature of the species being studied.

The results from SNPs detected by Illumina GAII and subsequent genotyping of a subset of the SNPs using MPS could be used as parental information for *M. stipoides*. In hindsight, the separation of PCR amplicons based on alleles using barcodes could have made the identifications of alleles easier (Jones, Chu et al. 2009), but it would have come at a high cost of the barcodes. In the future, cloning and genetic transformation of *M. stipoides* may be conducted to prove the function of the target homologues were the same as the genes in rice.
4.5 References


Chapter 4: SNP discovery in candidate domestication genes of M. stipoides


The target species of this study, *M. stipoides*, is a tetraploid grass native to Australia. Australian native grass species include a diverse array of wild cereal relatives which are adapted to a wider range of environmental conditions than current commercial cereals and may contain novel alleles which have utility in commercial production systems.

A series of genetic bottlenecks, which arose through recurrent selection, has eliminated a number of useful genomic resources from the major cereal crop gene pools (Doebley, Gaut et al. 2006; Tang, Sezen et al. 2010). Crop wild relatives (CWR) have the potential to provide new genetic resources to improve existing crop types, and to accelerate the development of new strategies for efficient and directed germplasm utilisation (Hoisington, Khairallah et al. 1999; Feuillet, Langridge et al. 2008; Negrão, Oliveira et al. 2008). Extensive conservation of gene order and gene content exists among closely related grass species (Devos and Gale 1997; Gale and Devos 1998; Feuillet and Keller 2002). For nearly two decades the high degree of synteny between grasses has facilitated isolation of gene homologues known to be important in well characterised crops such as rice. However, this approach is time consuming, highly dependent on sequence availability in multiple species, and may not be suitable in complex and uncharacterised target species. Chapter Two of this thesis shows that the approach was not an effective method for *M. stipoides*, and that therefore a more efficient method is needed.

Whole genome sequencing of crop plants using Sanger sequencing used to cost millions of dollars and took decade(s) to complete involving collaborations between many centres throughout the world (Goff, Ricke et al. 2002; Yu, Hu et al. 2002; Tuskan, DiFazio et al. 2006; Velasco, Zharkikh et al. 2007). Because of these hurdles and the immense size and complexity of cereal crop genomes such as wheat, maize and barley, sequence information of cereal crops was limited and scarce. Sequence information for CWR have until recently been scarcer than that of crop species. Thus, although CWR have the potential of improving existing crops or becoming new crops in their own right, most of the relevant underlying genetic information has not been available.
Recent advances in massively parallel sequencing (MPS) have allowed for rapid and relatively cheaper whole genome sequencing of numerous crops and wild species. In Chapter Three, *M. stipoides* whole genome sequence provided by Nock, Waters et al. (2011) was used to obtain *M. stipoides*-specific sequence information for 18 rice genes related to domestication characteristics. This approach proved to be more successful and enabled more gene homologues to be successfully isolated. Low and intermittent coverage of whole genomic sequence was sufficient to characterise a number of homologues of important rice domestication related genes in *M. stipoides*.

In Chapter Four, PCR amplicons were sequenced using MPS (instead of whole genomic data) and this provided good coverage at relatively reasonable cost. De novo assembly was able to generate different putative alleles of the gene homologues sequenced using MPS. Thousands of SNPs were detected when aligning the MPS reads to the de novo reference sequences, although most of the SNPs occurred within intronic regions. Sequenom iPLEX MassARRAY (SPM) was able to detect the subset of SNPs chosen, however it was not able to differentiate at allelic level, due to the complications of the tetraploidy of *M. stipoides*.

This study demonstrates that there is additional background research required in order to accelerate the domestication of *M. stipoides* as a new crop species for the Australian environment. However, the study was able to identify in *M. stipoides* previously unknown valuable gene homologues of a number of rice domestication related genes. There is still much additional characterisation required for these gene homologues, such as their functional role in *M. stipoides*. The information generated by this thesis provides baseline information for *M. stipoides* that, using molecular tools, can be utilised in the future accelerated domestication of this species as a crop plant. The defining characteristics of *M. stipoides* include drought, frost and shade tolerant; all of which are likely to be sought after in future crop plants.


Appendix 1- Alignment of rice 18 rice domestication related genes to its respective blastn results showing primer positions .................................................. 148

Appendix 2- Assemblies of the MPS data to the rice target genes ...................... 149

Appendix 3- Calculation on pooled PCR amplicons....................................... 150

Appendix 4- De novo alignment of MPS reads ............................................. 151

Appendix 5- Genbank files (files with .gbk extensions)................................ 152

Appendix 6- Results of MPS SNP detection using CLC Genomics Workbench ..... 153

Appendix 7- SNPs for MassARRAY design .................................................. 154

Appendix 8- Sequenom outputs categorised by gene ................................... 155
Appendix 1 - Alignment of rice 18 rice domestication related genes to its respective blastn results showing primer positions

On CD: Appendix 1 – Supplementary data of assemblies of the MPS data to the rice target. Folder name: Appendix 1 - Supplementary 3.3.1. To view the file, download Sequencher® from http://www.genecodes.com/download/external-tools-download.
Appendices

Appendix 2 - Assemblies of the MPS data to the rice target genes

Appendices

Appendix 3  - Calculation on pooled PCR amplicons

On CD: Appendix 3 – Supplementary data of calculation of pooling strategy for PCR amplicons prior to MPS. File name: Appendix 3 - Pooling strategy.pdf.
Appendix 4  - De novo alignment of MPS reads

On CD: Appendix 4 – Supplementary data of de novo assembly of the MPS reads. Contigs were formed according to the PCR amplicons generated. Folder name: Appendix 4 - Pooling strategy. To view the file, download Sequencher® from http://www.genecodes.com/download/external-tools-download.
Appendices

Appendix 5  -  Genbank files (files with .gbk extensions)


List of files:
1. badh2x1-5allele1.gbk – BADH2 amplicon of exons 1-5 putative allele 1
2. badh2x1-5allele2.gbk – BADH2 amplicon of exons 1-5 putative allele 2
3. badh2x5-9allele1.gbk - BADH2 amplicon of exons 5-9 putative allele 1
4. badh2x5-9allele2.gbk - BADH2 amplicon of exons 5-9 putative allele 2
5. badh2x8-15allele1.gbk - BADH2 amplicon of exons 8-15 putative allele 1
6. badh2x8-15allele2.gbk - BADH2 amplicon of exons 8-15 putative allele 2
7. gw2x1-5allele1.gbk – GW2 amplicon of exons 1-5 putative allele 1
8. gw2x1-5allele2.gbk – GW2 amplicon of exons 1-5 putative allele 2
9. gw2x4-8allele1.gbk - GW2 amplicon of exons 4-8 putative allele 1
10. gw2x4-8allele2.gbk - GW2 amplicon of exons 4-8 putative allele 2
11. hd6x1-5allele1.gbk – Hd6 amplicon of exons 1-5 putative allele 1
12. hd6x1-5allele2.gbk – Hd6 amplicon of exons 1-5 putative allele 2
13. hd6x2-5allele3.gbk – Hd6 amplicon of exons 2-5 putative allele 3
14. hd6x2-5allele4.gbk – Hd6 amplicon of exons 2-5 putative allele 4
15. hd6x4-10allele3.gbk – Hd6 amplicon of exons 4-10 putative allele 3
16. hd6x4-7allele4.gbk – Hd6 amplicon of exons 4-7 putative allele 4
17. amplicons_summary.pdf – a summary showing the structure of all the amplicons
Appendices

Appendix 6 - Results of MPS SNP detection using CLC Genomics Workbench

Appendix 7  - SNPs for MassARRAY design

On CD: Appendix 7 – Tables of SNPs for MassARRAY design. File name: Appendix 7 – SNPs for MAssARRAY design.pdf
Appendix 8 - Sequenom outputs categorised by gene

On CD: Appendix 8 – Sequenom outputs showing individuals carrying the allele(s) are shown on Excel spreadsheets. File name: Appendix 8 - Sequenom results categorised by gene.