ISOLATION, CHARACTERIZATION AND THE DYNAMIC STRUCTURE OF THE PEARL MILLET *D2* DWARFING GENE

by

RAJIV KRISHNA PARVATHANENI

(Under the Direction of Katrien M. Devos)

ABSTRACT

The *d2* dwarfing gene is widely used in pearl millet breeding programs. The *ABCB1* gene was identified as a candidate for *d2* based on mapping data from ~1500 progeny from two mapping populations, a haplotype analysis of the *d2* region in six diverse tall and *d2* dwarf inbred lines and comparative genome analysis of the *d2* region in sorghum. The *abcb1* allele isolated from the *d2* dwarf inbred differed from the tall allele (*ABCB1*) by the presence of two high-copy long terminal repeat (LTR) retrotransposons, one in the coding region and one 665 bp upstream of the start codon. All reported independent *d2* mutants tested contained the transposable element suggesting a single origin of all *d2* plants. Expression profiles of the *ABCB1* gene in different tissues and at different stages of plant development showed that 1) *ABCB1* was expressed in all analyzed tissues and 2) *ABCB1* was expressed at much higher levels in tall inbreds in comparison to *d2* dwarf plants.

The *ABCB1* gene showed a dynamic structure in angiosperms with the number of introns ranging from one to nine. A comprehensive analysis of the structure of *ABCB1* across angiosperms

revealed that seven out of the nine introns were lost independently in two or more species while two introns underwent a single loss. To investigate whether intron splicing efficiency played a role in the evolutionary loss of introns, nascent RNA (pre-mRNA) was isolated and sequenced from rice and sorghum and intron read coverage in a subset of lost and conserved genes was studied. We observed the following: 1) The genes with lost introns are expressed at significantly higher levels than the conserved set of genes 2) The length of an intron or the relative position of the intron in the gene did not affect the splicing efficiency and 3) The lost introns did not differ in splicing efficiency from conserved introns disproving our hypothesis. Through my dissertation work, we have also developed markers for the easy screening of the *d2* dwarf trait in pearl millet which will save time, money and resources in pearl millet breeding programs.

INDEX WORDS: Pearl millet, Gene mapping, Intron loss, Splicing, *D2* gene, Comparative genome analysis, RNA-seq, Nuclei isolation, *ABCB1* gene, Dynamic intron

ISOLATION, CHARACTERIZATION AND THE DYNAMIC STRUCTURE OF THE PEARL MILLET *D2* DWARFING GENE

by

RAJIV KRISHNA PARVATHANENI

B.Tech., Tamil Nadu Agricultural University, India, 2008

A Dissertation to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2015

ISOLATION, CHARACTERIZATION AND THE DYNAMIC STRUCTURE OF THE PEARL MILLET *D2* DWARFING GENE

by

Rajiv Krishna Parvathaneni

Major Professor: Katrien M. Devos

Committee: Peggy Ozias-Akins
Richard Meagher
John M. Burke
Xiaoyu Zhang

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2015

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Katrien M. Devos, for her continued guidance and support throughout my entire period of graduate study. I would also like to thank each of my advisors for periodic advice on my dissertation research. I would also like to thank past and current members of the Devos lab for suggestions on bench work and professional advice. The Plant Biology greenhouse staff, Michael Boyd and Kevin Tarner have been tremendously helpful in teaching me how to grow and maintain pearl millet and rice. I would also like to thank my family and close friends for their continued support to help me keep my morale high during this entire period. My friends in Athens: Douglas Eudy, Victoria Eudy, Guillaume Daverdin, Bochra Bahri, Johnathan Corbi, Trudi Thomas, Ganesh Bonde, Pankaj Sethi, Vani Hariharan, Suchitra Lagisetty, Judy Gray, Mike and Connie Wilson have made my enriched my cultural experience in the US. My friends in Atlanta: Ganesh Srinivasamoorthy, Boby Anthony, Parag Dedhia, Arun Ramachandran, Rocky D'Souza, Azim Rehmatullah and Darly Ezequiel made my weekends very special with their company. I specially thank Dr. Bochra Bahri on her help with the statistical analyses in Chapter 5. I also specially thank Himanshu Mishra who has been my pillar of support in the last 2 years of my PhD studies.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I:

INTRODUCTION AND LITERATURE REVIEW

Pearl millet

Pearl millet (*Cenchrus americanus* (L.) Morrone) (2n=2x=14) is an important cereal crop in many parts of the developing world. It is grown on an estimated 27 million hectares in Asia (10 million ha.) and sub-Saharan Africa (17 million ha.) (Rai et al., 2012). Pearl millet is a key crop in harsh dryland environmental conditions where other crops fail to be economically viable (Yadav et al. 2002; Devos et al. 2006). Pearl millet is also adapted to acidic and sandy soils which generally have poor fertility. Nutritionally, pearl millet seed has a high content of oil, protein, energy, calcium and iron (reviewed in Devos et al. 2006). It has a better amino acid profile compared to maize or sorghum (Ejeta et al. 1987). It is the main source of calorie intake in the West African Sudano-Sahelian belt (Devos et al. 2006; Haussmann et al. 2012). This region is home to some of the poorest people living on under 2$ a day (http://povertydata.worldbank.org/). Pearl millet is used as a dual purpose crop in many developing countries (food and forage), but it is almost exclusively grown for forage in the United States, Australia, and South America. It is well adapted to the Southeast Coastal Plains of the United States where soils are well-drained sandy or light loams (http://plants.usda.gov/). Pearl millet grain performed equally well or better to corn-based poultry feed (Singh and Barsaul 1976; Sharma et al. 1979). It has potential in cattle feed as well provided the higher protein content in pearl millet is efficiently utilized in feed formulations (Hill and Hanna 1990).

The International Crops Research Institute for Semi-Arid Tropics (ICRISAT), Patancheru, India and the Coastal Plain Experiment Station, Tifton, USA have had successful pearl millet breeding programs for many decades. The work of Dr. Glenn Burton is especially noteworthy in the improvement of the pearl millet crop. To mention a few, he developed the first cytoplasmic male sterile line Tift 23A (Burton 1958; Burton 1965), the *d2* dwarf germplasm (Burton and Fortson 1966), early maturity and photoperiod insensitivity germplasm (Burton 1981) and the trichomeless (*tr*) mutants (Powell and Burton 1971) which have reduced leaf transpiration and increased insect resistance (reviewed in Kumar and Andrews 1993). All the above traits are controlled by single genes and one or more of these genes are extensively used in current pearl millet breeding programs. Over the years, ICRISAT has been instrumental in developing and disseminating improved varieties to marginal farmers in developing countries (http://www.icrisat.org/). In India, the land under pearl millet cultivation has shown a small decrease since the 1950s, but the yield per hectare has increased more than four-fold (Prey and Nagarajan 2009)

Genetic resources for pearl millet

Pearl millet (*Cenchrus americanus* (L.) Morrone) was previously known as *Pennisetum glaucum* (Morrone et al. 2012). It belongs to one of the largest tribes, the Paniceae (over 2000 members), in the grass sub-family Panicoidae (Morrone et al. 2012). The basic chromosome number of these *Cenchrus sp.* varies with x = 5, 7, 8 and 9. These species also have a wide range of ploidy levels from diploid to octaploid (Martel et al. 1997; Devos et al. 2006). Many of these polyploid members (*e.g. Cenchrus squamulatus*) are facultative or obligate apomicts (Ozias-Akins et al. 1998; Devos et al. 2006). Pearl millet is a diploid with a basic chromosome number of seven (x=7). It is an annual crop which undergoes sexual reproduction. Pearl millet produces

protogynous (stigma develops and emerges before the anthers) flowers and is self-compatible. The nuclear DNA content of pearl millet was reported to be in the range of 4 pg to 4.8 pg/2C (4 pg/2C, (Bennett and Smith 1976); 4.71 pg/2C, (Martel et al. 1997); 4.8 pg/2C, (Bennett and Leitch 1995). This is similar to the haploid size of the maize genome (2,500 Mbp) (Chandler and Brendel 2002).

Genomic research in pearl millet started with the development of the first restriction fragment length polymorphism (RFLP)-based genetic map by Liu and colleagues in 1994 (Liu et al. 1994) on which 200 loci had been mapped using *Pst*I-based DNA probes. Highlights of this first study included: 1) that markers were arranged into seven linkage groups corresponding to the seven chromosomes 2) that the *Pst*I DNA probes detected high polymorphism levels within the set of 19 tested pearl millet genotypes and 3) that the maps were also extremely short. The total length of the map was estimated at 303 cM (Liu et al. 1994) which is much shorter on a per chromosome basis than maps of other genomes such as rice (1389 cM) (McCouch et al. 1988). Subsequent pearl millet genetic maps were also short suggesting that the maps were still incomplete. Subsequently, breeder friendly simple sequence repeats (SSRs) were added to the pearl millet genetic map (Qi et al. 2004). A comparison of maps generated in different crosses showed that high recombination in the distal chromosome regions was typical for pearl millet and resulted in loci that were closely linked physically, but were located tens of centimorgans apart on the genetic map (Qi et al. 2004). The fact that targeted marker development for one of these 'gaps' only yielded markers closely linked to the flanking markers supported the hypothesis that at least some of these gaps were due to high recombination rather than marker paucity. It was therefore concluded that recombination was not reduced in pearl millet in comparison to other species but rather was highly skewed towards the chromosome ends. With the advancement in genotyping technologies, more SSR and Diversity Arrays Technology (DArT) markers were added to the

genetic maps which now span a total length of 1148 cM (Supriya et al. 2011). More recently, 314 single nucleotide polymorphism (SNP) markers were mapped in a small F_2 population using genotyping-by-sequencing (GBS) (Moumouni et al. 2015).

The pearl millet linkage groups (LG) have been incorporated into the grass 'Crop Circle' where the colinear relationship of each pearl millet LG to other grass chromosomes has been determined (Gale and Devos 1998; Devos 2005). The pearl millet genome is highly rearranged in comparison to rice (Devos et al. 2000) which shared a common ancestor with pearl millet ~50 MYA. In contrast, the genome of the closest panicoid cereal to pearl millet, foxtail millet, is largely colinear to that of rice at the gross chromosomal level (Devos et al. 2000; Bennetzen et al. 2012). Low density mapping of a few simple traits controlled by single genes such as the *d1* dwarfing gene and the purple foliage color locus *P* has been reported (Azhaguvel et al. 2003). The maps have also been used to conduct QTL studies for yield-related traits (Vengadessan et al. 2013), drought tolerance (Yadav et al. 2002; Yadav et al. 2011) and downy mildew resistance (Jones et al. 1995; Jones et al. 2002). One QTL for downy mildew resistance was successfully introduced into the variety 'HHB-67' via marker-assisted breeding and the resulting variety, 'HHB 67-improved', has been deployed commercially (Hash *et al*., 2006). However, no gene candidates have as yet been reported for any of the simple traits or for the QTL. The high-density mapping and gene isolation of the *d2* dwarfing gene conducted in our lab is the first report of isolation of a gene underlying a phenotype in pearl millet (Parvathaneni et al. 2013; Chapter 2).

In addition to the genetic maps, a bacterial artificial chromosome (BAC) library of 5.6 genome equivalents was generated with an average insert size of 90 kb for the pearl millet *d2* dwarf inbred 'Tift 23DB' (Allouis et al. 2001). Another BAC library with an average insert size of 82 kb and 3.7 genome equivalents was developed from an apomictic polyhaploid resulting from

a cross between *Cenchrus squamulatus* and pearl millet (Roche et al. 2002). Additionally, two fosmid libraries with insert sizes in the range 35-40 kb are also available, one of which was generated in our lab from the tall inbred line ICMP 451 (2.2 genome equivalents; RK Parvathaneni and KM. Devos, unpublished) and one from the inbred line IP18296 with a fourfold genome coverage (Sujeeth et al. 2012).

Currently there are 5916 nucleotide sequences, 3577 expressed sequence tags (EST) and 4105 genome survey sequences (GSS) available for '*Cenchrus americanus*' in NCBI (http://www.ncbi.nlm.nih.gov/). Shotgun sequencing and assembly of the pearl millet genome is currently underway by the International Pearl Millet Genome Sequencing Consortium (IPMGSC)(http://www.ceg.icrisat.org/ipgsc.html). A comprehensive transcriptome assembly with RNA-seq datasets is also a part of the above sequencing strategy. Lack of an efficient pearl millet transformation system has been a limiting factor in both basic and applied research. Pearl millet transformation has been achieved using biolistic and *Agrobacterium*-mediated methods (Goldman et al. 2003; O'Kennedy et al. 2011; Ramineni et al. 2014). However, the transformation efficiency is low and no labs offer it as a service.

Current breeding efforts are focused on improving pearl millet for major biotic and abiotic stresses such as downy mildew caused by the oomycete *Sclerospora graminicola* (Hash et. al,2006)), leaf blast caused by *Pyricularia grisea* (Rai et al., 2012), infestation by the parasitic weed striga (*Striga hermonthica* (Del.) Benth) (Kountche et al. 2013) and tolerance to drought (Vadez et al. 2012). Efforts to improve the micro nutrient content in pearl millet are happening in parallel. Anemia (Fe deficit) is the most common nutritional deficiency in the developing world affecting some 2 billion people (World Health Organization, 2007)). Recent deployment of biofortified High-Fe pearl millet varieties has shown tremendous promise to combat this

deficiency (Tako et al. 2015). Pearl millet is a very promising food crop for the future, in particular as drought-prone regions are predicted to expand due to climate change. Development of genomic resources will help accelerate the process of crop improvement.

Dwarfing genes in agriculture

The Green Revolution in the 1960-70s helped South Asian countries achieve food security. One of the main reasons for its success was the deployment of early maturing dwarf rice (*sd-1*) and wheat (*Rht-1*) varieties that showed lodging resistance and a better response to fertilizer application (Hedden 2003). These improved varieties coupled with more intense farming practices increased cereal yield/acre and production by >99% and reduced the percentage poverty by 28% in developing Asia between 1970 and 1995 (Harzell 2009)

Dwarf mutants have been broadly classified as gibberrelic acid (GA)-sensitive or GA-insensitive based on their response to external gibberellin hormone application (Milach and Federizzi 2001). Some of these dwarf mutants have been extensively used in the breeding of cereal crops and genes underlying several dwarf phenotypes have been cloned. The *sd-1* rice mutant encodes a defective GA-20 oxidase, an enzyme in the GA biosynthetic pathway, leading to reduced levels of gibberrelic acid (GA-20) within the plant (Monna et al. 2002; Sasaki et al. 2002). The GA-sensitive *sdw/denso* gene in barley is used in many feed and malt cultivars and is potentially orthologous to the *sd-1* gene in rice (Jia et al. 2009; Jia et al. 2011). The *Ddw1* gene in rye is used in Eastern European breeding programs (Milach and Fedderizi 2001). The *Ddw1* rye mutant is classified as a GA-sensitive mutant but the candidate gene underlying the trait is unknown (Börner et al. 1996). Other cereal GA-sensitive dwarfing genes have had very limited commercial success including the *Dw6*, *Dw7* and *Dw8* dwarfing genes in oats. Although environment-specific yield

increases have been noted for *Dw6* cultivars (Anderson and McLean 1989), they generally have a reduced seed size, quality and yield (reviewed in Milach et al. 2002). An example of a GA-insensitive mutant is the widely used wheat *Rht-1* gene. This gene, which is homologous to the *gai* gene in Arabidopsis, contains an N-terminal domain called DELLA. Mutations in the DELLA domain of *Rht-1*/*gai* gene constitutively repress GA-signaling which leads to GA insensitive semi-dwarf phenotypes (Dill et al. 2001; Hedden 2003).

Another category of dwarf mutants are those that affect the transport of auxin hormones. This type of dwarf has also been used commercially. A combination of three dwarfing genes (3-dwarf system) is used in commercial varieties of sorghum, with *dw3* being present in most combinations (Schertz et al. 1974). The gene underlying *dw3*, *ABCB1*, has been cloned and was shown to be involved in re-export of the auxin hormone from the intercalary meristems of the stem (Multani et al. 2003; Knoller et al. 2010). Several mutant alleles of the *ABCB1* gene have also been reported in maize (*br2*) but they were not agriculturally useful because they yielded extreme dwarf phenotypes. However, a rare allele of *br2* has recently been uncovered which reduces plant height by only 20% and increases the yield potential in maize (Xing et al. 2015).

In pearl millet, Burton and colleagues identified several sources of dwarfing labelled *D1* to *D5* (Burton and Fortson 1966). Only *D1* and *D2* were shown to be controlled by single genes that were subsequently designated as *d1* and *d2*. Two additional dwarf mutants (*d3* and *d4*) controlled by a single gene were identified in a germplasm screen (Rao et al. 1986). However, both showed an extreme dwarf phenotype that has no commercial utility. Most of the commercially used pearl millet dwarf cultivars across the United States and Australia contain the *d2* dwarfing gene. Presence of the *d2* mutation has a small yield penalty but this can be mitigated by choosing an appropriate genetic background (Bidinger and Raju 1990; Rai and Rao 1991).

Because of a higher leaf to stem ratio, *d2* semi-dwarfs produce a higher quality of forage which has better digestive indices for the animals that feed on them (Johnson et al. 1968). Also, dwarf plants are more amenable for mechanical harvest due to their uniformity. Low-density mapping of *d2* located the gene on linkage group 4 in pearl millet (Azhaguvel et al. 2003). One aim of my dissertation was to isolate the gene underlying the *d2* phenotype. The candidate gene for *d2*, *ABCB1*, was identified using a combination of high-density mapping, haplotype analysis of the chromosomal region in diverse tall and *d2* dwarfs and comparative analysis with the sequenced sorghum genome (Chapter 2: Parvathaneni et al. 2013). We went on to determine the structure and expression of *ABCB1* in tall and dwarf lines (Chapter 3).

Structure and function of the *ABCB1* gene

The *ABCB1* gene encodes a P-glycoprotein (PGP) that modulates long distance transport of the natural auxin, indole 3-acetic acid, in monocots (Multani et al. 2003). This gene belongs to the ATP-binding cassette (ABC) superfamily, which is not only one of the largest families of proteins but is also present in all organisms (Henikoff et al. 1997). The "full" ABC protein comprises two hydrophobic transmembrane domains (TMDs) and two cytosolic nucleotide-binding domains (NBDs). The majority of eukaryotic ABCs are full transporters but prokaryotic and some eukaryotic transporters are "half" transporters containing just one NBD and one TMD (reviewed in Jasinski et al. 2003).

More than 100 ABC-type proteins including 53 full transporters are encoded in the Arabidopsis genome (Martinoia et al. 2002; Kang et al. 2011). Of these, 22 transporters belong to the PGP sub-family. In plants, ABC proteins transport different compounds required for a range of cellular processes (reviewed in Kang et al. 2011). Some of these processes include cellular

detoxification, plant growth and development, and response to pathogen and environmental stimuli. The *ABCB1* protein is involved in plant growth by the efflux of cellular indole-3-acetic acid (IAA) in *A. thaliana* (Geisler et al. 2005). The exact mode of auxin efflux is still under investigation, but it is known that a TWISTED DWARF 1 (TWD1) protein, which is a FK506 binding protein 42, physically interacts with the ABCB1 protein and aids its movement to the plasma membrane where the efflux of auxin occurs (Wang et al. 2013). The TWD1 protein was also shown to interact with the AGC kinase PINOID (PID) which phosphorylates S634 in the linker domain of ABCB1 but the outcome of this interaction is unclear (Henrichs et al. 2012). In Arabidopsis, *ABCB1* mutants show a zero or small reduction in plant height depending on the position of the mutation in the gene (Noh et al. 2001; Ye et al. 2013). In contrast, *ABCB1* mutants in grasses show a drastic reduction in height (Multani et al. 2003; Knoller et al. 2010). It remains to be tested if ABCB1 proteins fulfil the same functional roles in monocots as in dicots.

Isolation of pearl millet *ABCB1* (*CaABCB1)* as the gene underlying the *d2* phenotype led to the discovery of differences in the gene structure of *ABCB1* across angiosperms. The number of introns present in the *ABCB1* gene in dicots is different than what is seen in most monocots. For example, the *ABCB1* gene in Arabidopsis comprises 9 introns (Noh et al. 2001) but comprises just 4 introns in sorghum and maize (Multani et al. 2003). This observation led to a study of intron loss in *ABCB1* (Chapter 4) and an investigation into the mechanism of intron loss (Chapter 5).

Characteristics and mechanisms of intron loss

Introns are a characteristic of eukaryotic genomes but the number of intron containing genes, number of introns per gene and size of the introns vastly differ between eukaryotes (Fedorova and Fedorov 2003). A number of functions have been associated with introns. To list

a few: 1) The expression level of a gene can be governed by the relative position of introns (intron-mediated enhancement) (Callis et al. 1987; Palmiter et al. 1991); 2) Intron-retention is the predominant mechanism of alternate splicing in plants and a way to create more transcriptome diversity within a cell (Ner-Gaon et al. 2004); 3) Long introns can contain nested genes which may have important physiological roles (Henikoff et al. 1986; Yu et al. 2005); 4) Functional non-coding RNA sequences such as small nucleolar RNAs and microRNAs present in introns may be involved in *cis* and *trans* regulation of genes (Filipowicz and Pogacic 2002; Bartel 2004); and 5) Intron sequences themselves can have key regulatory roles such as transcription initiation, transcription termination and cytoplasmic localization. Some sequence motifs in promoter-proximal introns enhance gene expression while others harbor alternate promoter sequences giving rise to new gene isoforms (Rose et al. 2008; Chorev and Carmel 2012).

Genomic loss of introns in species or lineages occurs at a rate that is an order of magnitude higher than intron gain (Coulombe-Huntington and Majewski 2007; Roy and Penny 2007; Fawcett et al. 2012). Single loss of introns was also shown to occur much more frequently than single gain in a genome-wide comparison of 5 sequenced grass genomes (Wang et al. 2014). The frequency of intron loss in humans was estimated at 4.28×10^{-13} per intron per year (Coulombe-Huntington and Majewski 2007). Higher frequencies of intron loss were observed in plants such as 1.64×10^{-10} per intron per year for *A. thaliana* (Fawcett et al. 2012), 2.73×10^{-11} for *A. lyrata* (Fawcett et al. 2012), and 3.3×10^{-10} for rice (Lin et al. 2006). Parallel (independent) loss of the same intron across multiple species is presumably rare as only a few cases have been reported in the literature. The mammalian glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene which has 10 introns shows recurrent loss of intron 9 in the opossum, dog and the primate/rodent lineages and also single loss of introns 2, 3, 4, 5 and 7 across the different mammalian lineages (Coulombe-

Huntington and Majewski 2007). The *white* gene (ABC transporter), which has 14 introns, has lost introns 10, 11 and 13 independently at least 3 times in the distantly related dipteran lineages (Krzywinski and Besansky 2002). In another ABC transporter, *MRP1*, convergent loss of one intron, Z, was observed in *Drosophila* and mosquito species (Zhan et al. 2012). In plants, the isochorismate synthase (ICS) gene which consists of 16 introns shows recurrent loss of intron 2 in the *Glycine max*/*Medicago truncatula* lineage and in the grass clade. It also shows single loss of introns 3 and 5 in the Arabidopsis ICS1 homolog (Yuan et al. 2009). In a genome wide study of intron loss in Angiosperms, only 93 cases of recurrent intron loss were observed, further confirming that this phenomenon is relatively rare (Wang et al. 2014).

For an intron loss to be fixed in a species, it needs to occur in the cells that lead to the germline. Whether selection or genetic drift plays a role in the fixation of gene copies that have undergone intron loss is still unknown. There are two proposed mechanisms for the evolutionary loss of introns: 1) Reverse transcriptase (RT)-mediated intron loss caused by recombination between a genomic copy and its fully or partially processed cDNA and 2) genomic deletion of introns by double stranded DNA breakage and repair. RT-mediated intron loss will result in the precise removal of introns, will typically remove adjacent introns and will display a bias towards removal of introns at the 3' end of genes (Roy and Gilbert 2006; Coulombe-Huntington and Majewski 2007). Non-adjacent removal of introns can be explained by recombination with semi-processed or fragmented cDNA (Coulombe-Huntington and Majewski 2007). Genomic deletions may or may not result in precise deletion of introns. It is a more error prone process that is driven by small-scale sequence homology (Fawcett et al. 2012). Non-homologous end joining (NHEJ), a DNA repair mechanism, requires sequence homology of 2-4 bp for the repair to take place (Fawcett et al. 2012). When the splice junctions contain these homologous sequences, a precise

removal of introns may occur as has been seen in many cases of intron loss in *A. thaliana* (Fawcett et al. 2012). However, neither RT-mediated intron removal nor removal by NHEJ has direct experimental evidence that supports these proposed mechanisms. Due to the precision of intron removal, RT-mediated intron loss is considered the most likely mechanism leading to intron loss (Roy and Gilbert 2005; Coulombe-Huntington and Majewski 2007). The splicing efficiency of an intron determines the probability of its occurrence in a semi-processed transcript. In Chapter 5, we investigate whether splicing efficiency is correlated with evolutionary loss of an intron in a set of genes that have undergone parallel (or independent) loss of introns across multiple lineages.

REFERENCES

Allouis S, Qi X, Lindup S, Gale MD, Devos KM. 2001. Construction of a BAC library of pearl millet, *Pennisetum glaucum*. *Theoret. Appl. Genetics* **102**(8): 1200-1205.

Anderson W, McLean R. 1989. Increased responsiveness of short oat cultivars to early sowing, nitrogen fertilizer and seed rate. *Australian Journal of Agricultural Research* **40**(4): 729-744.

Azhaguvel P, Hash CT, Rangasamy P, Sharma A. 2003. Mapping the *d1* and *d2* dwarfing genes and the purple foliage color locus *P* in pearl millet. *Journal of Heredity* **94**(2): 155-159.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**(2): 281-297.

Baurhoo N, Baurhoo B, Mustafa AF, Zhao X. 2011. Comparison of corn-based and Canadian pearl millet-based diets on performance, digestibility, villus morphology, and digestive microbial populations in broiler chickens. *Poultry science* **90**(3): 579-586.

Bennett MD, Leitch IJ. 1995. Nuclear DNA Amounts in Angiosperms. *Annals of Botany* **76**(2): 113-176.

Bennett MD, Smith JB. 1976. Nuclear dna amounts in angiosperms. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **274**(933): 227-274.

Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J et al. 2012. Reference genome sequence of the model plant Setaria. *Nat. Biotechnol.* **30**(6): 555-561.

Bidinger FR, Raju DS. 1990. Effects of the D2 Dwarfing Gene in Pearl-Millet. *Theoret. Appl. Genetics* **79**(4): 521-524.

Börner A, Plaschke J, Korzun V, Worland AJ. 1996. The relationships between the dwarfing genes of wheat and rye. *Euphytica* **89**(1): 69-75.

Burton GW. 1958. Cytoplasmic male-sterility in pearl millet (*Pennisetum glaucum* (L.) R. Br.). *Agronomy Journal* **50**:230.

Burton GW. 1965. Pearl Millet Tift 23A released. *Crops Soils* **17:**19.

Burton GW. 1981. A gene for early maturity and photoperiod insensitivity in pearl millet. *Crop Sci.* **21**: 317-318.

Burton GW, Fortson JC. 1966. Inheritance and utilization of five dwarfs in pearl millet (*Pennisetum typhoides*) Breeding. *Crop Sci* **6**(1): 69-72.

Callis J, Fromm M, Walbot V. 1987. Introns increase gene expression in cultured maize cells. *Genes Dev* **1**(10): 1183-1200.

Chandler VL, Brendel V. 2002. The Maize Genome Sequencing Project. *Plant physiology* **130**(4): 1594-1597.

Chorev M, Carmel L. 2012. The function of introns. *Front Genet.* **3**: 55.

Coulombe-Huntington J, Majewski J. 2007. Characterization of intron loss events in mammals. *Genome Res.* **17**(1): 23-32.

Devos KM. 2005. Updating the 'crop circle'. *Current opinion in plant biology* **8**(2): 155-162.

Devos KM, Hanna WW, Ozias-Akins P. 2006. Pearl Millet. **In:** The Genomes, Vol. 1: Cereals and Millets. Pp. 478-506, Kole, C., Ed. Indus Intl., New Delhi.

Devos KM, Pittaway TS, Reynolds A, Gale MD. 2000. Comparative mapping reveals a complex relationship between the pearl millet genome and those of foxtail millet and rice. *Theoret. Appl. Genetics* **100**(2): 190-198.

Dill A, Jung H-S, Sun T-p. 2001. The DELLA Motif is Essential for Gibberellin-Induced Degradation of RGA. *Proceedings of the National Academy of Sciences of the United States of America* **98**(24): 14162-14167.

Ejeta G, Hassen MM, Mertz ET. 1987. In vitro digestibility and amino acid composition of pearl millet (*Pennisetum typhoides*) and other cereals. *Proceedings of the National Academy of Sciences of the United States of America* **84**(17): 6016-6019.

Fawcett JA, Rouze P, Van de Peer Y. 2012. Higher intron loss rate in Arabidopsis thaliana than A. lyrata is consistent with stronger selection for a smaller genome. *Mol Biol Evol* **29**(2): 849-859.

Fedorova L, Fedorov A. 2003. Introns in gene evolution. *Genetica* **118**(2-3): 123-131.

Filipowicz W, Pogacic V. 2002. Biogenesis of small nucleolar ribonucleoproteins. *Curr Opin Cell Biol* **14**(3): 319-327.

Gale MD, Devos KM. 1998. Comparative genetics in the grasses. *Proceedings of the National Academy of Sciences* **95**(5): 1971-1974.

Geisler M, Blakeslee JJ, Bouchard R, Lee OR, Vincenzetti V, Bandyopadhyay A, Titapiwatanakun B, Peer WA, Bailly A, Richards EL et al. 2005. Cellular efflux of auxin catalyzed by the Arabidopsis MDR/PGP transporter AtPGP1. *Plant J* **44**(2): 179-194.

Goldman JJ, Hanna WW, Fleming G, Ozias-Akins P. 2003. Fertile transgenic pearl millet [Pennisetum glaucum (L.) R. Br.] plants recovered through microprojectile bombardment and phosphinothricin selection of apical meristem-, inflorescence-, and immature embryo-derived embryogenic tissues. *Plant cell reports* **21**(10): 999-1009.

Hash CT, Sharma A, Kolesnikova-Allen MA, Singh SD, Thakur RP, Bhasker Raj AG, Ratnaji Rao MNV, Nijhawan DC, Beniwal CR, Sagar P, Yadav HP, Yadav YP, Srikant , Bhatnagar SK, Khairwal IS, Howarth CJ, Cavan GP, Gale MD, Liu C, Devos KM, Breese WA, Witcombe JR. 2006. Teamwork delivers biotechnology products to Indian small-holder crop-livestock producers: Pearl millet hybrid "HHB 67 Improved" enters seed delivery pipeline. J. SAT. Agric. Res. 2(1): http://www.icrisat.org/journal/bioinformatics/v2i1/v2i1teamwork.pdf

Haussmann BIG, Fred Rattunde H, Weltzien-Rattunde E, Traoré PSC, vom Brocke K, Parzies HK. 2012. Breeding Strategies for Adaptation of Pearl Millet and Sorghum to Climate

Variability and Change in West Africa. *Journal of Agronomy and Crop Science* **198**(5): 327-339.

Harzell PB. 2009. The Asian Green Revolution. IFPRI Discussion Paper 00911. This paper has been prepared for the project on Millions Fed: Proven Successes in Agricultural Development (URL:www.ifpri.org/millionsfed)

Hedden P. 2003. The genes of the Green Revolution. *Trends Genet* **19**(1): 5-9.

Hill GM, Hanna WW. 1990. Nutritive characteristics of pearl millet grain in beef cattle diets. Journal of animal science **68**(7): 2061-2066.

Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L. 1997. Gene Families: The Taxonomy of Protein Paralogs and Chimeras. *Science* **278**(5338): 609-614.

Henikoff S, Keene MA, Fechtel K, Fristrom JW. 1986. Gene within a gene: Nested Drosophila genes encode unrelated proteins on opposite DNA strands. *Cell* **44**(1): 33-42.

Henrichs S, Wang B, Fukao Y, Zhu J, Charrier L, Bailly A, Oehring SC, Linnert M, Weiwad M, Endler A et al. 2012. Regulation of ABCB1/PGP1-catalysed auxin transport by linker phosphorylation. *The EMBO journal* **31**(13): 2965-2980.

Jasinski M, Ducos E, Martinoia E, Boutry M. 2003. The ATP-Binding Cassette Transporters: Structure, Function, and Gene Family Comparison between Rice and Arabidopsis. *Plant Physiology* **131**(3): 1169-1177.

Jia Q, Zhang J, Westcott S, Zhang XQ, Bellgard M, Lance R, Li C. 2009. GA-20 oxidase as a candidate for the semidwarf gene sdw1/denso in barley. *Funct Integr Genomics* **9**(2): 255-262.

Jia Q, Zhang XQ, Westcott S, Broughton S, Cakir M, Yang J, Lance R, Li C. 2011. Expression level of a gibberellin 20-oxidase gene is associated with multiple agronomic and quality traits in barley. *Theoret. Appl. Genetics* **122**(8): 1451-1460.

Johnson JC, Lowrey RS, Monson WG, Burton GW. 1968. Influence of Dwarf Characteristic on Composition and Feeding Value of near-Isogenic Pearl Millets. *Journal of Dairy Science* **51**(9): 1423-&.

Jones ES, Breese WA, Liu CJ, Singh SD, Shaw DS, Witcombe JR. 2002. Mapping Quantitative Trait Loci for Resistance to Downy Mildew in Pearl Millet. *Crop Science* **42**(4): 1316-1323.

Jones ES, Liu CJ, Gale MD, Hash CT, Witcombe JR. 1995. Mapping quantitative trait loci for downy mildew resistance in pearl millet. *Theoret. Appl. Genetics* **91**(3): 448-456.

Kang J, Park J, Choi H, Burla B, Kretzschmar T, Lee Y, Martinoia E. 2011. Plant ABC Transporters. *The Arabidopsis Book / American Society of Plant Biologists* **9**: e0153.

Knoller AS, Blakeslee JJ, Richards EL, Peer WA, Murphy AS. 2010. Brachytic2/ZmABCB1 functions in IAA export from intercalary meristems. *J. Exp. Bot.* **61**(13): 3689-3696.

Kountche BA, Hash CT, Dodo H, Laoualy O, Sanogo MD, Timbeli A, Vigouroux Y, This D, Nijkamp R, Haussmann BIG. 2013. Development of a pearl millet Striga-resistant genepool: Response to five cycles of recurrent selection under Striga-infested field conditions in West Africa. *Field Crops Research* **154**(0): 82-90.

Krzywinski J, Besansky NJ. 2002. Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. *Mol Biol. Evol.* **19**(3): 362-366.

Kumar KA, Andrews DJ. 1993. Genetics of qualitative traits in pearl millet: a review. *Crop Sci.* **33**:1-20.

Lin H, Zhu W, Silva JC, Gu X, Buell CR. 2006. Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol.* **7**(5): R41.

Liu CJ, Witcombe JR, Pittaway TS, Nash M, Hash CT, Busso CS, Gale MD. 1994. An RFLP-based genetic map of pearl millet (*Pennisetum glaucum*). *Theoret. Appl. Genetics* **89**(4): 481-487.

Martel E, De Nay D, Siljak-Yakoviev S, Brown S, Sarr A. 1997. Genome Size Variation and Basic Chromosome Number in Pearl Millet and Fourteen Related Pennisetum Species. *Journal of Heredity* **88**(2): 139-143.

Martinoia E, Klein M, Geisler M, Bovet L, Forestier C, Kolukisaoglu Ü, Müller-Röber B, Schulz B. 2002. Multifunctionality of plant ABC transporters – more than just detoxifiers. *Planta* **214**(3): 345-355.

McCouch SR, Kochert G, Yu ZH, Wang ZY, Khush GS, Coffman WR, Tanksley SD. 1988. Molecular mapping of rice chromosomes. *Theoret. Appl. Genetics* **76**(6): 815-829.

Milach SCK, Federizzi LC. 2001. Dwarfing genes in plant improvement. **In** *Advances in Agronomy*, Vol 73. pp. 35-63. Academic Press.

Milach SCK, Rines HW, Phillips RL. 2002. Plant height components and gibberellic acid response of oat dwarf lines. *Crop Sci* **42**(4): 1147-1154.

Monna L, Kitazawa N, Yoshino R, Suzuki J, Masuda H, Maehara Y, Tanji M, Sato M, Nasu S, Minobe Y. 2002. Positional cloning of rice semidwarfing gene, sd-1: rice "green

revolution gene" encodes a mutant enzyme involved in gibberellin synthesis. *DNA Res.* **9**(1): 11-17.

Morrone O, Aagesen L, Scataglini MA, Salariato DL, Denham SS, Chemisquy MA, Sede SM, Giussani LM, Kellogg EA, Zuloaga FO. 2012. Phylogeny of the Paniceae (Poaceae: Panicoideae): integrating plastid DNA sequences and morphology into a new classification. *Cladistics* **28**(4): 333-356.

Moumouni KH, Kountche BA, Jean M, Hash CT, Vigouroux Y, Haussmann BIG, Belzile F. 2015. Construction of a genetic map for pearl millet, Pennisetum glaucum (L.) R. Br., using a genotyping-by-sequencing (GBS) approach. *Mol Breeding* **35**(1): 1-10.

Multani DS, Briggs SP, Chamberlin MA, Blakeslee JJ, Murphy AS, Johal GS. 2003. Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. *Science* **302**(5642): 81-84.

Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, Fluhr R. 2004. Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J* **39**(6): 877-885.

Noh B, Murphy AS, Spalding EP. 2001. Multidrug Resistance-Like Genes of Arabidopsis Required for Auxin Transport and Auxin-Mediated Development. *The Plant Cell* **13**(11): 2441-2454.

O'Kennedy MM, Stark HC, Dube N. 2011. Biolistic-mediated transformation protocols for maize and pearl millet using pre-cultured immature zygotic embryos and embryogenic tissue. *Methods in molecular biology (Clifton, NJ)* **710**: 343-354.

Ozias-Akins P, Roche D, Hanna WW. 1998. Tight clustering and hemizygosity of apomixis-linked molecular markers in Pennisetum squamulatum implies genetic control of apospory by a divergent locus that may have no allelic form in sexual genotypes. *Proceedings of the National Academy of Sciences of the United States of America* **95**(9): 5127-5132.

Palmiter RD, Sandgren EP, Avarbock MR, Allen DD, Brinster RL. 1991. Heterologous introns can enhance expression of transgenes in mice. *Proceedings of the National Academy of Sciences of the United States of America* **88**(2): 478-482.

Parvathaneni RK, Jakkula V, Padi FK, Faure S, Nagarajappa N, Pontaroli AC, Wu X, Bennetzen JL, Devos KM. 2013. Fine-mapping and identification of a candidate gene underlying the d2 dwarfing phenotype in pearl millet, *Cenchrus americanus* (L.) Morrone. *G3* **3**(3): 563-572.

Pray CE and Nagarajan L. 2009. Pearl millet and sorghum improvement in India. IFPRI discussion paper 00919. This paper has been prepared for the project on Millions Fed: Proven Successes in Agricultural Development (URL:www.ifpri.org/millionsfed)

Powell JB, Burton GW, 1971. Genetic suppression of shoot-trichomes in pearl millet, Pennisetum typhoides. *Crop Sci*. 11:763-765.

Qi X, Pittaway TS, Lindup S, Liu H, Waterman E, Padi FK, Hash CT, Zhu J, Gale MD, Devos KM. 2004. An integrated genetic map and a new set of simple sequence repeat markers for pearl millet, Pennisetum glaucum. *Theoret Appl Genetics* **109**(7): 1485-1493.

Rai KN, Rao AS. 1991. Effect of *d2* dwarfing gene on grain-yield and yield components in pearl millet near-Isogenic Lines. *Euphytica* **52**(1): 25-31.

Rai KN, Yadav OP, Gupta SK, Mahala RS, Gupta SK. 2012. Emerging research prorities in pearl millet. J. SAT Agric. Res. **10**:1-5

Ramineni R, Sadumpati V, Khareedu VR, Vudem DR. 2014. Transgenic pearl millet male mertility restorer Line (ICMP451) and hybrid (ICMH451) expressing *Brassica juncea* nonexpressor of pathogenesis related genes 1 (*BjNPR1*) exhibit resistance to downy mildew disease. *PLoS ONE* **9**(3): e90839.

Rao SA, Mengesha MH, Reddy CR. 1986. New Sources of Dwarfing Genes in Pearl-Millet (Pennisetum-Americanum). *Theoret Appl Genetics* **73**(2): 170-174.

Roche D, Conner A, Budiman A, Frisch D, Wing R, Hanna W, Ozias-Akins P. 2002. Construction of BAC libraries from two apomictic grasses to study the microcolinearity of their apospory-specific genomic regions. *Theoret Appl Genetics* **104**(5): 804-812.

Rose AB, Elfersi T, Parra G, Korf I. 2008. Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression. *The Plant cell* **20**(3): 543-551.

Roy SW, Gilbert W. 2005. The pattern of intron loss. *Proceedings of the National Academy of Sciences of the United States of America* **102**(3): 713-718.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* **7**(3): 211-221.

Roy SW, Penny D. 2007. Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of O. sativa and *A. thaliana*. *Mol Biol Evol* **24**(1): 171-181.

Sasaki A, Ashikari M, Ueguchi-Tanaka M, Itoh H, Nishimura A, Swapan D, Ishiyama K, Saito T, Kobayashi M, Khush GS et al. 2002. Green revolution: a mutant gibberellin-synthesis gene in rice. *Nature* **416**(6882): 701-702.

Schertz KF, Rosenow DT, Johnson JW, Gibson PT. 1974. Single *dw3* height gene effects in 4- and 3-dwarf hybrids of *Sorghum bicolor* (L.) Moench. *Crop Sci* **14**(6): 875-877.

Sharma B, Sadagopan V, Reddy V. 1979. Utilization of different cereals in broiler diets. British Poultry Science **20**(4): 371-378.

Singh S, Barsaul C. 1976. Replacement of maize by coarse grains for growth production in White Leghorn and Rhode Island Red birds [India]. Indian Journal of Animal Sciences.

Sujeeth N, Kini R, Shailasree S, Wallaart E, Shetty S, Hille J. 2012. Characterization of a hydroxyproline-rich glycoprotein in pearl millet and its differential expression in response to the downy mildew pathogen Sclerospora graminicola. *Acta Physiol Plant* **34**(2): 779-791.

Supriya A, Senthilvel S, Nepolean T, Eshwar K, Rajaram V, Shaw R, Hash CT, Kilian A, Yadav RC, Narasu ML. 2011. Development of a molecular linkage map of pearl millet integrating DArT and SSR markers. *Theoret Appl Genetics* **123**(2): 239-250.

Tako E, Reed S, Budiman J, Hart J, Glahn R. 2015. Higher iron pearl millet (*Pennisetum glaucum* L.) provides more absorbable iron that is limited by increased polyphenolic content. *Nutrition Journal* **14**(1): 11.

Vadez V, Hash T, Bidinger FR, Kholova J. 2012. II.1.5 Phenotyping pearl millet for adaptation to drought. *Frontiers in Physiology* **3**: 386.

Vengadessan V, Rai KN, Kannan Bapu JR, Hash CT, Bhattacharjee R, Senthilvel S, Vinayan MT, Nepolean T. 2013. Construction of Genetic Linkage Map and QTL Analysis of Sink-Size Traits in Pearl Millet (*Pennisetum glaucum*). *ISRN Genetics* **2013**: 14.

Wang B, Bailly A, Zwiewka M, Henrichs S, Azzarello E, Mancuso S, Maeshima M, Friml J, Schulz A, Geisler M. 2013. Arabidopsis TWISTED DWARF1 Functionally Interacts with Auxin Exporter ABCB1 on the Root Plasma Membrane. *The Plant Cell* **25**(1): 202-214.

Wang H, Devos KM, Bennetzen JL. 2014. Recurrent Loss of Specific Introns during Angiosperm Evolution. *PLoS Genet* **10**(12): e1004843.

World Health Organization, Centre for Disease Control and Prevention. 2007. Assessing the iron status of populations. Second edition. (URL:http://www.who.int/nutrition/publications/micronutrients/anaemia_iron_deficiency/9789241596107/en/)

Xing A, Gao Y, Ye L, Zhang W, Cai L, Ching A, Llaca V, Johnson B, Liu L, Yang X et al. 2015. A rare SNP mutation in Brachytic2 moderately reduces plant height and increases yield potential in maize. *Journal of Experimental Botany* **66**(13): 3791-802.

Yadav RS, Hash CT, Bidinger FR, Cavan GP, Howarth CJ. 2002. Quantitative trait loci associated with traits determining grain and stover yield in pearl millet under terminal drought-stress conditions. *Theoret Appl Genetics* **104**(1): 67-83.

Yadav RS, Sehgal D, Vadez V. 2011. Using genetic mapping and genomics approaches in understanding and improving drought tolerance in pearl millet. *Journal of Experimental Botany* **62**(2): 397-408.

Ye L, Liu L, Xing A, Kang D. 2013. Characterization of a dwarf mutant allele of Arabidopsis MDR-like ABC transporter AtPGP1 gene. *Biochemical and Biophysical Research Communications* **441**(4): 782-786.

Yu P, Ma D, Xu M. 2005. Nested genes in the human genome. *Genomics* **86**(4): 414-422.

Yuan Y, Chung J-D, Fu X, Johnson VE, Ranjan P, Booth SL, Harding SA, Tsai C-J. 2009. Alternative splicing and gene duplication differentially shaped the regulation of isochorismate synthase in Populus and Arabidopsis. *Proceedings of the National Academy of Sciences* **106**(51): 22020-22025.

Zhan L-L, Ding Z, Qian Y-H, Zeng Q-T. 2012. Convergent Intron Loss of MRP1 in Drosophila and Mosquito Species. *Journal of Heredity* **103**(1): 147-151.

CHAPTER II:

FINE-MAPPING AND IDENTIFICATION OF A CANDIDATE GENE UNDERLYING THE *D2* DWARFING PHENOTYPE IN PEARL MILLET, *CENCHRUS AMERICANUS* (L.) MORRONE[1]

[1] Parvathaneni, R.K., Jakkula V., Padi, F. K., Faure S., Nagarajappa, N., Pontaroli A. C., Wu X., Bennetzen, J. L., and Devos, K. M. 2013. Genes, Genomes and Genetics (G3). 3(3):562-572. Reprinted here with permission of publisher.

ABSTRACT

Pearl millet is one of the most important subsistence crops grown in India and sub-Saharan Africa. In many cereal crops, reduced height is a key trait for enhancing yield and dwarf mutants have been extensively used in breeding to reduce yield loss due to lodging under intense management. In pearl millet, the recessive *d2* dwarfing gene has been deployed widely in commercial germplasm grown in India, the US and Australia. Despite its importance, very little research has gone into determining the identity of the *d2* gene. We used comparative information, genetic mapping in two F_2 populations representing a total of some 1500 progeny, and haplotype analysis of three tall and three dwarf inbred lines to delineate the *d2* region by two genetic markers that, in sorghum, define a region of 410 kb with 40 annotated genes. One of the sorghum genes annotated within this region is *ABCB1* which encodes a P-glycoprotein involved in auxin transport. This gene had previously been shown to underlie the economically important *dw3* dwarf mutation in sorghum. The cosegregation of *ABCB1* with the *d2* phenotype, its differential expression in the tall inbred ICMP 451 and the dwarf inbred Tift 23DB, and the similar phenotype of stacked lower internodes in the sorghum *dw3* and pearl millet *d2* mutants suggest that *ABCB1* is a likely candidate for *d2*.

INTRODUCTION

Pearl millet (*Cenchrus americanus* (L.) Morrone, previously *Pennisetum glaucum* (L.) R. Br.)(Chemisquy *et al.* 2010), is an important cereal crop grown on an estimated 27 million hectares in Asia and sub-Saharan Africa (FAOSTAT; faostat.fao.org). Being highly drought tolerant, the crop is well adapted to the arid and semi-arid tropical environments in these countries. Pearl millet is also grown as forage crop in regions of the US, Australia and South America. Most commercial

pearl millet hybrids, whether grown for grain or forage, carry the recessive height-reducing gene *d2* (e.g. Burton 1980; Hanna *et al.* 1997; Gulia *et al.* 2007; Rai *et al.* 2009). The *d2* gene does not affect the coleoptile and mesocotyl length (Soman *et al.* 1989), but reduces overall plant height by some 50% through a shortening of all internodes except the peduncle (Burton and Fortson 1966; Burton *et al.* 1969; Rai and Hanna 1990a). Dwarfs tend to yield less grain than their tall isolines, but this negative effect can be mitigated by manipulating the genetic background (Bidinger and Raju 1990; Rai and Rao 1991). Forage quantity is also reduced, but forage quality is higher in the dwarfs than in the talls due to a higher leaf to stem ratio (Johnson *et al.* 1968). The higher digestibility of the leaves compared to the stems results in higher animal yields in feed trials (Burton *et al.* 1969; Hanna *et al.* 1979).

The precise origin of the *d2* dwarf mutation is unknown. In the US, Burton and colleagues discovered in 1939 an extremely leafy pearl millet plant with short internodes among the progeny of a plant obtained through mass selection from five introductions of pearl millet acquired a few years earlier from the Vavilov Institute of Plant Industry, Russia (Burton and Devane 1951; Hein 1953). Based on information in the Germplasm Resource Information Network (GRIN) database, the five introductions originated from Tunisia (PI 115055), Eritrea (PI 115056, PI 115058), Arabia (PI 115057) and India (PI 115059). The dwarf line was true-breeding and used in crosses with an adapted pearl millet line to form the highly successful synthetic variety 'Starr' (Hein 1953). Although there are no records confirming that Starr millet carried the *d2* gene, the described morphology makes this a plausible hypothesis (Kumar and Andrews 1993). Around the same time, in India, Kadam *et al.* (1940) obtained dwarf phenotypes after inbreeding local pearl millet lines. The dwarfs had shortened internodes, overlapping leaf sheaths and shortened peduncles and were attributed to a recessive mutation. Again, it is unclear whether any of these represented *d2*.

In 1966, Burton and Fortson (1966) reported identification of five non-allelic dwarf mutants (*D1* to *D5*). Two of those, *D1* and *D2*, were shown to be controlled by different single recessive genes and were assigned the gene symbols *d1* and *d2*, respectively. The *d2* gene was subsequently incorporated into Indian cultivars through backcross breeding using seed stocks provided by Dr. G.W. Burton (Bakshi *et al.* 1966) and is now widely used in commercial hybrids in the US, India and Australia (reviewed by Kumar and Andrews 1993; Gulia *et al.* 2007; Rai *et al.* 2009). The *d2* gene has been mapped on pearl millet linkage group 4 to a 23.2 cM interval flanked by RFLP markers PSM84 and PSM413.2 (Azhaguvel *et al*. 2003).

Height-reducing genes have played key roles in enhancing yield in a range of cereals. The best known examples are the gibberellic acid (GA)-insensitive *Rht-1* and GA-sensitive *sd-1* dwarfing genes that were essential to the Green Revolutions in wheat and rice, respectively (Peng *et al.* 1999; Monna *et al.* 2002; Sasaki *et al.* 2002), but height mutants have also been widely employed in other cereals. For example, in barley, the GA-sensitive *sdw/denso* gene located on chromosome 3H (Laurie *et al.* 1993) has been used extensively in feed and malt cultivars in the Western United States, Canada, Europe and Australia (reviewed by Mickelson and Rasmusson 1994). Most commercial sorghum lines are "3-dwarf", which indicates that they carry mutations in three of the four dwarfing genes that have been identified in this species (Schertz *et al.* 1974). The height-reducing gene *Ddw1* on rye chromosome 5R has been deployed in many Eastern-European and Finnish rye breeding programs to develop short-straw cultivars (Milach and Federizzi 2001). A number of these dwarfing genes have been isolated and characterized. The *Rht-1* genes encode DELLA proteins that act as repressors of plant growth (Peng *et al.* 1999). Mutations in the DELLA domain inhibit GA-induced degradation of the DELLA proteins, which results in a GA-insensitive dwarfing phenotype (Peng *et al.* 1997; Dill *et al.* 2001; reviewed by

Hedden 2003). The rice *sd-1* gene is a GA 20-oxidase, which catalyzes multiple steps in the GA biosynthetic pathway (Monna *et al.* 2002; Sasaki *et al.* 2002; Spielmeyer *et al.* 2002), and it has recently been shown that the *sdw/denso* gene in barley is likely an ortholog of *sd-1* in rice (Jia *et al.* 2009; Jia *et al.* 2011). The *dw3* dwarf phenotype in sorghum is caused by an 882 bp tandem duplication in the fifth exon of the *ABCB1* gene. This rearrangement results in the loss of the encoded P-glycoprotein, which modulates polar auxin transport in the stalk (Multani *et al.* 2003).

The aim of our research was to fine-map the *d2* gene in pearl millet, and to use comparative information to identify putative candidate genes for the locus. Genetic mapping of an identified candidate gene and preliminary expression analysis provide support for a model that the pearl millet *d2* gene is the ortholog of sorghum *dw3.*

MATERIALS AND METHODS

Mapping populations

An F_2 mapping population of a few thousand seed was generated by selfing a single F_1 hybrid from a cross between the *d2* dwarf inbred Tift 23DB (female; Figure 2.1A) and the tall inbred ICMP 451 (male; Figure 2.1B). Tift 23DB was obtained from Wayne Hanna, University of Georgia, Tifton, USA. ICMP 451 was obtained from the International Crop Research Institute for Semi-Arid Tropics (ICRISAT), Patancheru, India.

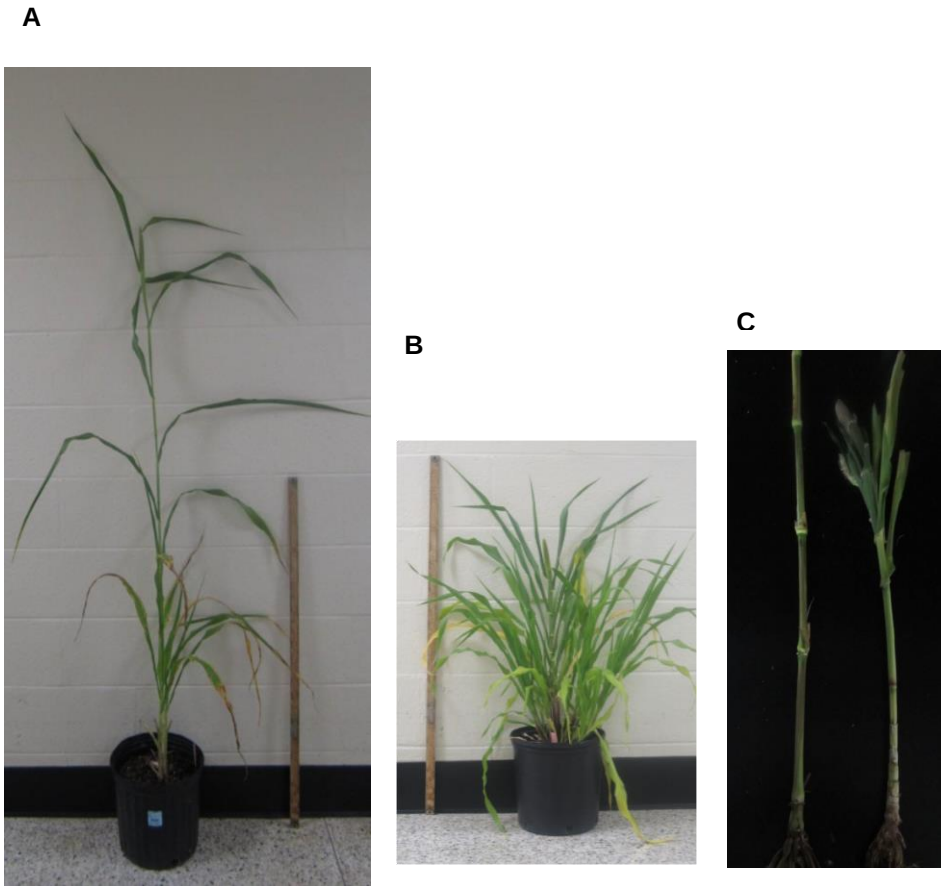


Figure 2.1 Architecture of A) inbred ICMP 451 (*D2D2*) and B) inbred Tift 23DB (*d2d2*), the parents of the fine-mapping population at flowering time (panicle on main tiller 50% exerted). A 1 m ruler is shown for height comparison. Tift 23DB is ~50% shorter and has a higher leaf to stem ratio compared to ICMP 451. C) Phenotype of the stem of ICMP 451 (left) and Tift 23DB (right) after the leaves were removed from the plants shown in A) and B) showing stacking of, in particular, the lower internodes in Tift 23DB compared to ICMP 451.

A second pearl millet mapping population, originally developed to segregate for a downy mildew resistance gene on linkage group 4, also segregated for *d2*. To construct this population, an F_2 individual was identified from among the progeny of a genotyped F_2 population derived from the cross PT 732B (*d2d2*) x P1449-2 (*D2D2*) (Qi *et al.* 2004) that was heterozygous at most marker loci on linkage group 4, including the loci that spanned the region carrying both the resistance gene

and the *d2* gene. Twenty-two $F_{2:3}$ plants were grown and analyzed with markers for the region of interest on linkage group 4 and a heterozygous F_3 plant was selfed to produce a 552 progeny population. In a genetic context, this population, at least in the *d2* region, behaves as an F_2 population and will be referred to as such. This population was phenotyped for *d2* and mapped with restriction fragment length polymorphism markers in early 2000 (Padi 2002). The Padi (2002) study located the *d2* gene to a 2.8 cM interval between marker PSMP344 and the cosegregating markers B224C4P2 and RGR1963. Because *d2* was scored as a dominant trait (dwarf – *d2d2* and tall – *D2D2* or *D2d2*), the precise position of *d2* could not be determined. However, of the 19 recombination events that could be allocated, 18 events occurred between PSMP344 and *d2*, and 1 occurred between B224C4P2/RGR1963 and *d2*, indicating a tight linkage of *d2* with B224C4P2/RGR1963.

Bacterial artificial chromosome (BAC) sequencing and sequence analysis

Rice RFLP marker RGR1963 was used to screen a pearl millet BAC library (Allouis *et al.* 2001). Eleven positive clones were identified of which BAC 293B22 was selected for sequencing. BAC DNA was isolated from clone 293B22 and shotgun libraries prepared as described (Dubcovosky *et al.* 2001). A total of 1152 subclones were sequenced from both ends using Sanger technology. PHRED, PHRAP and CONSED were used with default parameters for base calling and quality control, sequence assembly and contig ordering, respectively. The sequence of BAC 293B22 has been deposited in GenBank under accession number KC463796. Gene prediction was done using FGENESH with the monocot training set (www.softberry.com). Repetitive DNA was identified by BLASTN searches against the Gramineae repeat database (http://plantrepeats.plantbiology.msu.edu/search.html).

Markers

B224C4P2 is a PCR-based marker derived from an end-sequence of pearl millet BAC 224C4, which was one of the 11 clones identified after screening a pearl millet BAC library with the rice RFLP marker RGR1963 (Padi 2002). PSMP344 and PSMP305 are sequence-tagged-site (STS) markers derived from RFLP probes PSM305 and PSM344, respectively (Money *et al.* 1994). Three PCR-based markers, Ca_Sb07g023840, Ca_Sb07g023850 and Sb07g023860 were derived from genes identified on BAC 293B22. The prefix Ca stands for *Cenchrus americanus* and is followed by the designation of the sorghum ortholog. In addition, primer sets were developed against 40 genes located in the regions 28.17 Mb – 28.44 Mb (chromosome end) on rice chromosome 8 and 58.37 Mb – 59.04 Mb on sorghum chromosome 7. Based on grass comparative data, these rice and sorghum chromosome segments were expected to be syntenic to the pearl millet *d2* region (Devos *et al.* 2000; Devos 2005). Genes selected from the rice and sorghum genomic sequences were used in BLASTN searches to find corresponding expressed sequence tags from other grass species. The genomic and expressed sequence tags sequences were aligned using the program Multalin (Corpet 1988) and primer sets were designed against conserved exon regions flanking an intron. For polymorphism screening and mapping, amplification products were separated on MDE™ (Mutation Detection Enhancement) acrylamide gels to display single strand conformation polymorphisms (Martins-Lopes *et al.* 2001). The primer sequences, annealing temperature and location in the sorghum, rice and *Setaria italica* genomes for all markers are given in Supplementary Table 1. To facilitate interpretation of the data, all markers, irrespective of whether they were developed from rice or sorghum, were named after their sorghum ortholog.

Genotyping

A total of 915 F_2 individuals from the cross Tift 23DB x ICMP 451 were grown in the glasshouse in batches of 100 to 200 plants. Genomic DNA was extracted from F_2 individuals using an ultra-quick DNA extraction protocol (Steiner *et al.* 1995) and genotyped with markers B224C4P2 and PSMP305 on MDE™ gels to identify recombinants in the *d2* region. Plants carrying a recombination event in the *d2* region are referred to as 'informative plants'. Informative plants were selfed to produce F_3 seed. High quality DNA for further genotyping of the informative plants was obtained either from leaves of the F_2 plants using a standard CTAB protocol (Murray and Thompson 1980) or from 25 bulked F_3 seeds using the protocol described by Busso *et al.* (2000). All genotyping was done on MDE™ gels. DNA fragments were visualized by silver staining (Beidler 1982).

PCR amplifications were performed in 20 µl reaction volumes containing 1X PCR buffer (Promega), 1.5 mM $MgCl_2$, 0.25 mM of each dNTP, 0.5 µM of forward and reverse primers, 100 ng of template DNA, and 0.8 U of Taq DNA Polymerase (Promega). Amplification conditions consisted of an initial denaturation at 94 °C for 3 minutes, 38 cycles of denaturation at 94 °C for 30 sec, primer annealing for 30 sec (see Supplementary Table 1 for primer-specific annealing temperatures), and extension at 72 °C for 1 minute followed by a final extension at 72 °C for 5 minutes. For touch-down PCR (primers indicated with a temperature range in Supplementary Table 1), the annealing temperature was decreased by 1 °C every two cycles until the target temperature was reached and 35 PCR cycles were done at the target temperature.

Phenotyping

To determine the allelic composition at the *d2* locus, 13 to 25 F_3 plants were analyzed for each of the informative F_2 plants. Because of space limitations in the glasshouse, phenotyping was done in multiple batches. Plant height was measured at maturity from the basal node to the top of the panicle (Supplementary Table 2). Families with a median height <90 cm consisted of plants with the dwarf phenotype and the corresponding F_2 genotype was scored as homozygous dwarf (Table 2.1). Because of the broad range of heights observed for the tall plants, which was caused both by the segregation of other height-affecting genes in the population and environmental effects, we were very conservative in converting the phenotypic measurements to genotypic scores for the *d2* locus. Families with a median height >135 cm were considered to be derived from an F_2 plant heterozygous at the *d2* locus if the number of plants shorter than 110 cm was not significantly different from 25%. The height of 110 cm was chosen as threshold because less than 2% of dwarf plants were ≥ 110 cm. F_2 genotypes giving rise to F_3 families with a median height between 90 cm and 130 cm were considered to be either heterozygous or homozygous for the tall allele and were not further classified (Table 2.1).

Comparative analyses

The protein corresponding to the primary transcript and its location in the genome was retrieved for all annotated genes in the region between 54.31 Mb and 64.31 Mb on sorghum chromosome 7 (JGI v.1.0; Paterson *et al.* 2009). Similarly, we retrieved the protein corresponding to the primary transcript for all annotated genes in *S. italica* (35,158 loci in 9 chromosomes, JGI v2.1; Bennetzen *et al.* 2012), *Oryza sativa* (34,781 representative gene loci in 12 chromosomes, IRGSP build 5; Matsumoto *et al.* 2005), and *Brachypodium distachyon* (23,558 loci in 5 chromosomes, JGI v.1.0;

Vogel *et al.* 2010). In the first instance, a BLASTP search was carried out with the sorghum proteins as queries against the *S. italica* proteins. The top hit was recorded if the E-value was less than $1e^{-5}$ and the maximum number of hits at the threshold value was four. Homologous pairs were used to detect syntenic blocks by MCscan (multiple collinearity scan) (Tang *et al.* 2008) and colinear segments were identified using the empirical scoring scheme *min* $\{-\log_{10}E, 50\}$ for one gene pair and -1 gap penalty for each 10 kb distance between any two consecutive gene pairs. Syntenic blocks with scores >300 and an E-value $<1e^{-10}$ were retrieved. *S. italica* proteins located within these syntenic blocks were then used to identify the syntenic regions in rice and *B. distachyon*. The positions of orthologous gene pairs were plotted using a script in R (R Development Core Team 2005), and the dot plots were used to identify rearrangements. The precise breakpoints of the rearrangements were determined manually.

Expression analyses

The mapping parents ICMP 451 and Tift 23DB were grown under greenhouse conditions for 7 to 9 weeks. The top-most internode and its corresponding nodal leaf were harvested when 50% of the panicle had emerged from the flag leaf sheath. The harvested material was immediately frozen in liquid nitrogen and, if needed, stored at -80 ºC. Total RNA was extracted from leaves and internodes with TRIzol reagent (Chomczynski and Sacchi 1987) and approximately 5-10 µg of isolated RNA was treated with DNase using the Ambion TURBO DNA-*free*™ kit. cDNA synthesis was performed on ~1 µg of DNase treated RNA using the Roche Transcriptor First Strand cDNA Synthesis kit. The manufacturer's protocol was used for all experimental steps involving kits. PCR conditions for semi-quantitative PCR consisted of an initial denaturation at 95 ºC for 3 minutes followed by 29 cycles of denaturation at 95 ºC for 30 seconds, annealing at 59 ºC for 30 seconds, extension at 72 ºC for 1 minute, and a final extension at 72 ºC for 3 minutes.

Primers Ca_Sb07g023730F10 (5'-GCAGGTTCTCCTTGATGCTC-3') and Ca_Sb07g07g23730R10 (5'-CTCGGAGGCACCTACTTCAC-3'), designed against gene Sb07g023730 (*dw3*), were used to study expression of the pearl millet *dw3* ortholog, while actin primers ActinF (5'-ACCGAAGCCCCTCTTAACCC-3') and ActinR (5'-GTATGGCTGACACCATCACC-3') were used as internal controls (Van Den Berg *et al.* 2004). The expression analysis was conducted twice with samples collected from plants grown at different times of the year.

RESULTS

Identifying recombinants in the *d2* region

Mapping of plant height in the PT 732B x P1449-2 population had shown the *d2* gene to be located between the markers B224C4P2 and PSMP344 (Padi 2002). The STS marker PSMP344 (primer set PSMP344F/R) derived from RFLP marker PSM344 did not amplify from ICMP 451 and thus could only be scored as a dominant marker in the Tift 23DB x ICMP 451 population. Hence, PSMP305 which cosegregates with PSMP344 in most pearl millet maps (Qi *et al.* 2004) was used in combination with B224C4P2 to identify F_2 plants that carried a recombination event in the *d2* region. Genotyping of 915 F_2 plants from the cross Tift 23DB x ICMP 451 yielded 29 recombinants providing an estimate of 1.6 cM for the genetic distance between B224C4P2 and PSMP305. Five plants did not survive the seedling stage, and one plant was removed because of the presence of non-parental alleles, so the fine-mapping and phenotyping was carried out on 23 recombinants.

Determining the genotype at the *d2* locus

The height of the $F_{2:3}$ plants varied both within and between F_3 families and ranged from 37 cm to 270 cm (Supplementary Table 2). Five F_3 families comprised only dwarf plants and the corresponding F_2 plants were genotyped as *d2d2* (Table 2.1). Five F_3 families were segregating for plant height in a 3:1 (tall:dwarf) ratio, and the corresponding F_2 plants were genotyped as *D2d2*. Nine F_3 families contained either no plants <110 cm (7 families) or a single plant <110 cm (2 families) and the corresponding F_2 plants were genotyped as *D2D2*. Integrating this data with the genotypic data for markers B224C4P2 and PSMP305 confirmed the location of *d2* in the B224C4P2 – PSMP305 interval.

Table 2.1 Median height, mean height and standard deviation of F_3 families derived from informative F_2 plants, the number of F_3 plants per family with height <110 cm and F_2 genotypic score

F_2 plant ID	No. of F_3 progeny (batch)[1]	Median plant height (cm)	Mean plant height (cm)	Standard deviation (cm)	Number (percentage) of plants <110 cm[2]	F_2 genotype at *d2* locus
1	23 (1)	85	87	13.99	23 (100%)	*d2d2*
55	23 (1)	139	137.6	22.63	3 (13%)	*D2d2*
177	23 (1)	62	63	11.82	23 (100%)	*d2d2*
263	25 (1)	104	109.1	16.94	ND[3]	*D2D2 or D2d2*
310	25 (1)	160	155.4	22.66	0 (0%)	*D2D2*
320	24 (1)	150.5	150.5	21.57	0 (0%)	*D2D2*
344	23 (1)	85	86	17.20	21 (91.3%)	*d2d2*
349	-	-	-	-	-	-
374	23 (2)	89	82.3	18.16		*d2d2*
477	25 (1)	61	64.8	16.04	25 (100%)	*d2d2*
479	22 (1)	155	153	29.27	1 (4.6%)*	*D2D2*
486	24 (1)	107	109.6	30.09	ND	*D2D2 or D2d2*
486	15 (3)	130	121.8	23.21	ND	*D2D2 or D2d2*
496	23 (1)	142	132.5	28.17	4 (17.4%)	*D2d2*
514	21 (2)	121	129.9	34.66	ND	*D2D2 or D2d2*
612	23 (2)	154	150.3	24.35	2 (8.7%)	*D2d2*
701	15 (2)	181	179.8	23.63	0 (0%)	*D2D2*
778	21 (2)	145	145.1	26.81	1 (4.8%)*	*D2d2*
787	17 (3)	157	150.7	26.51	2 (11.8%)	*D2d2*
812	17 (2)	210	203.2	22.54	0 (0%)	*D2D2*

900	13 (2)	189	182	23.88	0 (0%)	*D2D2*
914	14 (2)	192.5	185.9	47.83	2 (14.3%)	*D2d2*
924	22 (2)	186	186.4	24.49	0 (0%)	*D2D2*
930	18 (2)	206	208.4	28.34	0 (0%)	*D2D2*

[1] F3 families with the same batch number were grown concurrently.

[2] Except where the percentage of plants <110 cm is 0% or 100%, significant deviation from 25% at $P \leq 0.05$ is indicated with *

[3] Not determined in F3 families with a median height between 90 and 135 cm

Marker development and fine-mapping

Marker RGR1963, which corresponds to sorghum gene Sb07g023850, the two additional markers developed from BAC clone 293B22 which was identified with RGR1963, and the primers designed against selected genes on rice chromosome 8 and sorghum chromosome 7 were tested on Tift 23DB, ICMP 451 and 4 recombinant progeny for their ability to amplify and to detect variation. This initial screen also allowed us to identify markers that would likely map to the *d2* region based on their segregation pattern in the four recombinant lines. Seventy-five percent of the primer sets amplified well in pearl millet and 30% were polymorphic in the mapping population. Four markers, Ca_Sb07g023460, Ca_Sb07g023470, Ca_Sb07g023500 and Ca_Sb07g023740 had segregation patterns in the four recombinant progeny that were inconsistent with their location in the *d2* region and were not further analyzed. Eight markers corresponding to sorghum genes Sb07g023430, Sb07g023440, Sb07g023520, Sb07g023630, Sb07g023810, Sb07g023840, Sb07g023850 and Sb07g023910 were mapped in the full set of 23 informative progeny. All markers cosegregated (Figure 2.2A, Supplementary Table 3). Following the development of a new primer set for PSMP344 (PSMP344F2/R2; Supplementary Table 1) which amplified in both ICMP 451 and Tift 23DB, PSM344 was mapped between PSMP305 and the cluster of cosegregating markers. Placement of *d2* relative to the fine-mapped markers showed, with the exception of one double recombination event in the *d2* score, complete cosegregation of

the *d2* phenotype with the 8-marker cluster (Supplementary Table 3). The F_3 family derived from the F_2 line showing the double recombination event (F_2 plant with ID 612, Table 2.1) consisted of 23 plants, 2 of which were 2 and 4 cm shorter than the threshold of 110 cm. Furthermore, while 21:2 did not deviate significantly from a 3:1 ratio, the P-value was close to the 5% significance threshold (P=0.071). It seems therefore likely that the genotype for this plant was *D2D2* rather than *D2d2*. When this hypothesis was taken into consideration, the *d2* phenotype cosegregated with the 8-marker cluster (Figure 2.2A, Supplementary Table 3).

In an attempt to order the newly generated markers in the pearl millet genome, we selected 16 out of the 29 individuals from the PT 732B x P1449-2 population with a recombination event in the interval B224C4P2 – PSMP344 for which F_3 seed was available. Bulked F_3 seed was grown and seedlings were used for DNA extraction. Mapping of RGR1963, Ca_Sb07g023430, Ca_Sb07g023520, Ca_Sb07g023630, Ca_Sb07g023810 and Ca_Sb07g023910, and one additional sorghum-derived marker, Ca_Sb07g024020, in the 16 informative plants identified two marker clusters (Fig. 2.2B, Supplementary Table 4). One cluster cosegregated with B224C4P2 and comprised the markers RGR1963, Ca_Sb07g023910, Ca_Sb07g024020 and Ca_Sb07g023810 which, in sorghum, are located in the interval 58.78 Mb – 59.04 Mb on chromosome 7. The second cluster comprised *d2*, and markers Ca_Sb07g023630, Ca_Sb07g023520, and Ca_Sb07g023430 which were derived from region 58.37 Mb – 58.54 Mb on sorghum chromosome 7 (Figure 2.2B, Supplementary Table 4).

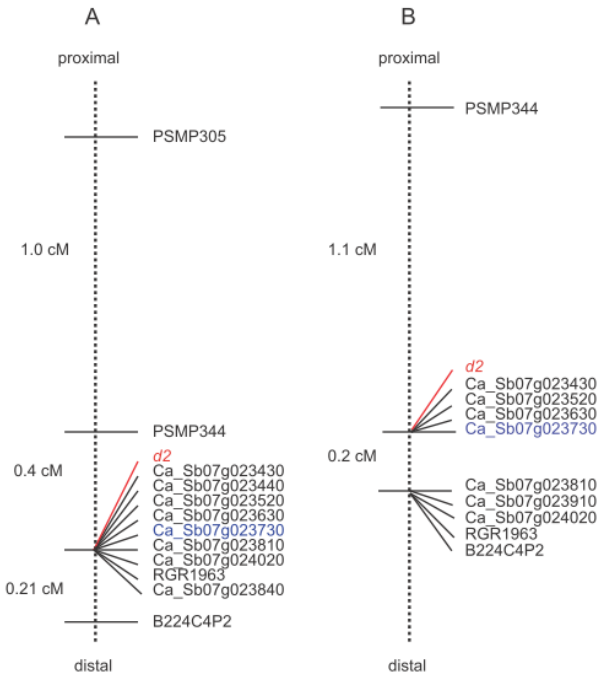


Figure 2.2 Genetic map of the *d2* region on linkage group 4 of pearl millet generated in A) the Tift 23DB x ICMP 451 mapping population and B) the PT 732B x P1449-2 mapping population. The map position of the *d2* phenotype is indicated in red. The map position of the sorghum *dw3* ortholog (Ca_Sb07g023730) is indicated in blue.

Comparative analysis of the *d2* region

If we assume that the *d2* region is completely colinear in the pearl millet and sorghum genomes, the ortholog of *d2* should be present in the sorghum genome proximal to location 58.78 Mb (distal boundary). However, our mapping data did not allow us to determine the proximal boundary of the *d2* region. A number of RFLP probes, including PSM344 and PSM305, had previously been end-sequenced (Money *et al.* 1994). The two end-sequences of PSM344, which is a 2 kb probe,

mapped 24.7 kb apart on sorghum chromosome 7 (locations 12.59 Mb and 12.61 Mb). For PSM305, one end does not have homology in sorghum at an e-value threshold $\leq 1e^{-05}$ while the other end identifies sequences on all 10 sorghum chromosomes. We also assessed the location in sorghum of PSM364, which had previously been mapped, depending on the cross, 1.8 – 6.1 cM distal of PSM305 (Qi *et al.* 2004) but no BLASTN hits were identified. A BLASTN analysis of these same markers in the foxtail millet genome identified hits for the two PSM344 end-sequences 2.7 kb apart at location 9.10 Mb on foxtail millet chromosome VI. One end of PSM305 and both ends of PSM364 identified homologous sequences on foxtail millet chromosome VI at locations 1.36 Mb and 4.18 Mb, respectively. The locations of PSM344, PSM305 and PSM364 in foxtail millet suggest that the region distal to PSM344 is rearranged in pearl millet compared to foxtail millet.

Lack of recombination in the pearl millet *d2* region precluded precise ordering of the developed markers. However, three of the mapped markers Ca_Sb07023860, Ca_Sb07g023840 and RGR1963, were derived from/present on BAC 293B22, which was sequenced to a depth of approximately 12X. The sequence of BAC 293B22 assembled into a single scaffold consisting of three contigs. *De novo* gene identification as well as homology-based annotation identified three genes in the order Ca_Sb07g023860 - 1.7 kb - Ca_Sb07g023850 (which corresponds to RGR1963) - 60.1 kb - Ca_Sb07g023840. Marker B224C4P2 was located 3.1 kb from Ca_Sb07g023840 and both markers were separated by a minimum of one and a maximum of three recombination event in the Tift 23DB x ICMP 451 map (Ca_Sb07g023840 was scored as a dominant marker hence not all recombination events could be identified). Combining the genetic mapping data with the gene order information from BAC 293B22 indicated that part of the *d2* region was inverted in pearl millet compared to sorghum (Figure 2.3).

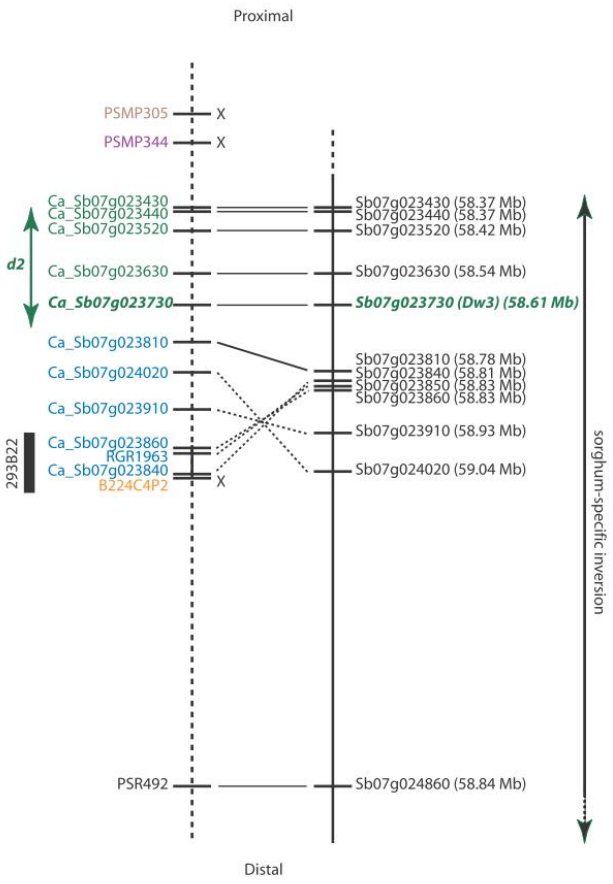


Figure 2.3 Comparative relationship between the *d2* region in pearl millet (left) and the orthologous region in sorghum (right). Orthologous markers in pearl millet and sorghum are connected with solid lines. or, for markers that are inverted in pearl millet relative to sorghum, with dotted lines. Pearl millet markers for which no ortholog could be identified in the depicted sorghum region are indicated with X. Ca_Sb07g023860, RGR1963, Ca_Sb07g023840 and B224C4P2 were located on pearl millet BAC clones 293B22 and distances between those markers are drawn to scale. Distances between other markers in pearl millet are taken from sorghum. Markers shown in the d2 region in pearl millet in the same color could not be separated by recombination events based on data from both the Tift 23DB x ICMP 451 and PT 732B x P1449-2 mapping populations. Marker Ca_Sb07g023730 (indicated in bold italic) represents the gene

underlying the dw3 phenotype in sorghum. The genome location in sorghum is given in parentheses after the marker name.

In order to better understand the evolution of the *d2* region in grasses, we conducted a comparative analysis at the genome level of the distal 10 Mb of sorghum chromosome 7 (54.31 Mb – 64.31 Mb). This region was largely colinear between sorghum chromosome 7, foxtail millet chromosome VI, rice chromosome 8 and *B. distachyon* chromosome 3, but a number of species-specific inversions were observed. The distal region of sorghum chromosome 7 (from 58.36 Mb – end) is inverted relative to the other three species (Figure 2.4 and Supplementary Table 5). This places Si012129m, the foxtail millet ortholog of marker Ca_Sb07g023430, as the most distal marker on foxtail millet chromosome VI for which an ortholog is present on sorghum chromosome 7. In rice, the ortholog of Ca_Sb07g023430 is present on rice chromosome 12 and the rice ortholog to Si015189m, the proximal neighbor of Si013129m, is the last marker on rice chromosome 8 with an ortholog on sorghum chromosome 7. In *B. distachyon*, the ortholog of Ca_Sb07g023430 marks the breakpoint of an ancestral chromosome fusion event. Other rearrangements include an inversion of the region 35.19 Mb – 35.50 Mb in foxtail millet, two inversions comprising the regions 23.82 Mb – 23.97 Mb and 25.72 Mb – 26.07 Mb in rice, and two inversions spanning the regions 40.52 Mb – 40.80 Mb and 41.65 Mb – 41.74 Mb in *B. distachyon.* None of these inversions correspond to the inversion that differentiates pearl millet from sorghum.

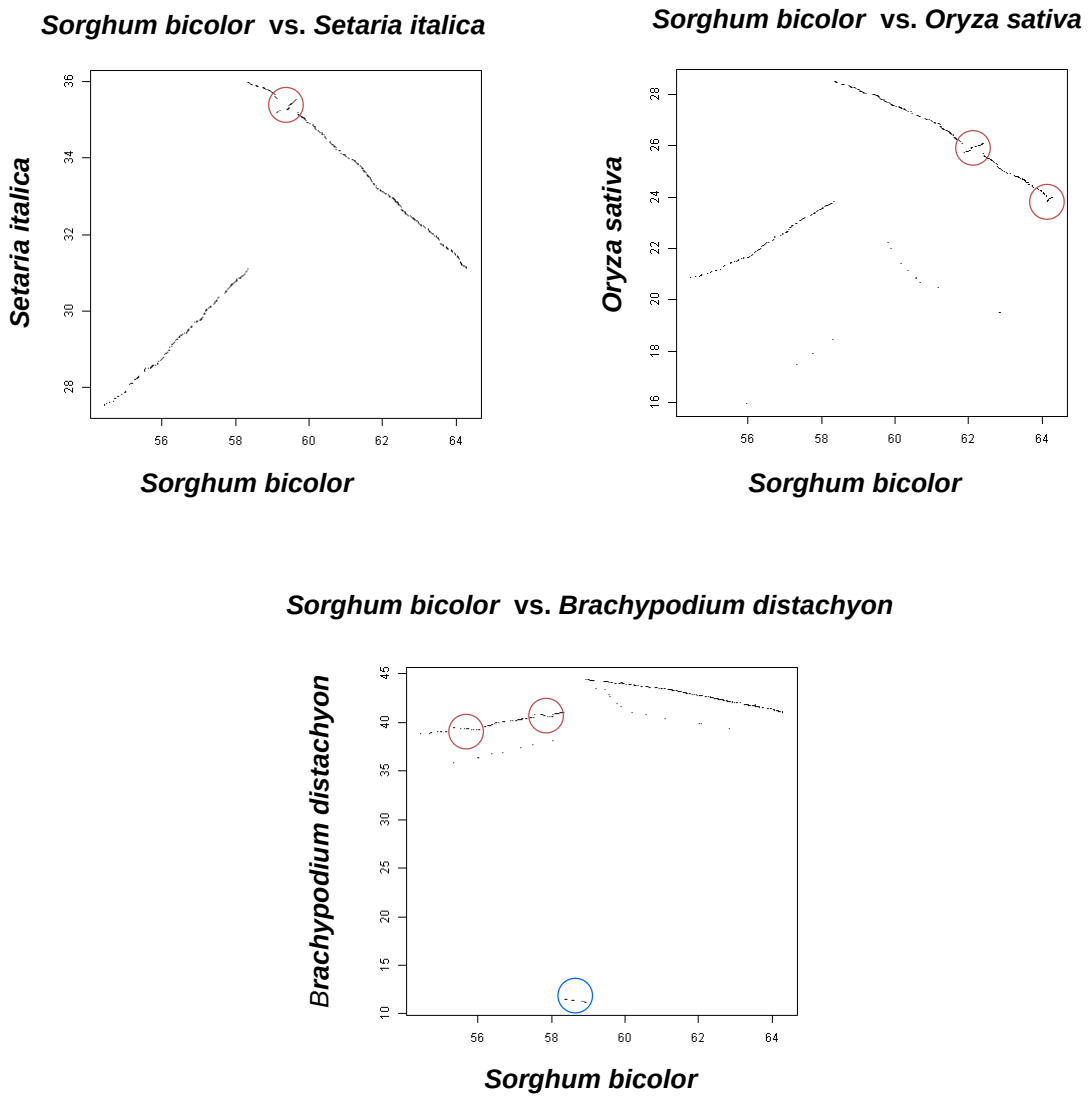


Figure 2.4 Dot plots showing comparisons at the genome level between region 54.31 Mb - 64.31 Mb in sorghum and the orthologous regions in *Setaria italica*, *Oryza sativa* and *Brachypodium distachyon*. Inversions in *S. italica*, *O. sativa* and *B. distachyon* relative to sorghum are circled in red and other rearrangements are circled in blue. Comparison of the three dot plots also shows that the region 58.36 Mb – 64.31 Mb in sorghum is inverted relative to the other three species.

Haplotype of the *d2* region in three tall and three dwarf lines

We used the mapped markers to determine the allelic configuration at each of the loci in three dwarf lines Tift 23DB, PT 732B and 81B, and three tall lines, ICMP 451, P1449-2 and Tift red (Table 2.2). At all loci distal to Ca_Sb07g023430, the three dwarf lines carried the same allele, referred to as 'a' while the tall lines carried alleles that were different (either 'b' or 'c'; Table 2.2) than those observed in the dwarfs. The only exception was at locus Ca_Sb07g023910 where the tall Tift red appeared to have the same allele as the dwarfs. The differentiation between dwarf and tall haplotypes was lost at the three most proximal markers analyzed, Ca_Sb07g023430, PSMP344 and PSMP305.

Table 2.2 Allele composition at 12 loci in three tall and three dwarf inbred lines

	Tall inbreds			Dwarf inbreds		
Marker	ICMP 451	P1449-2	Tift red	Tift 23DB	81B	PT 732B
B224C4P2	b	b	b	a	a	a
Ca_Sb07g023840	b	b	b	a	a	a
RGR1963	b	b	b	a	a	a
Ca_Sb07g023910	b	b	a	a	a	a
Ca_Sb07g024020	ND[1]	b	b	ND	ND	a
Ca_Sb07g023810	b	c	b	a	a	a
Ca_Sb07g023630	b	c	b	a	a	a
Ca_Sb07g023520	b	b	b	a	a	a
Ca_Sb07g023440	b	b	c	a	a	a
Ca_Sb07g023430	b	b	a	a	a	a
PSMP344	b	b	a	a	b	a
PSMP305	b	b	a	a	b	a

[1] ND = No data

Using comparative information to identify a putative candidate gene for *d2*

Combining the mapping information with the haplotype data yielded Ca_Sb07g023810 and Ca_Sb07g023430 as the distal and proximal boundary, respectively, of the *d2* region. These two markers defined a 410 kb region in sorghum in which 40 genes had been annotated

(www.phytozome.net; Sbi1.4 gene set). The genes, together with their location and functional annotation, are given in Supplementary Table 6. Sixty-three percent of the 40 genes had no functional annotation. Focusing on the remaining 15 that had homology to characterized proteins, Sb07g023730 became a likely candidate for *d2* as it had previously been identified as the gene underlying the *dw3* and *br2* dwarf phenotypes in sorghum and maize, respectively (Multani *et al.* 2003). Sb07g023730, an ABC transporter of the B subfamily (member 1), encodes a P-glycoprotein that modulates auxin transport in the stalk (Multani *et al.* 2003; Knoller *et al.* 2010).

Preliminary validation of *ABCB1* as a candidate for *d2*

Several forward and reverse primers were designed against the sequence of gene Sb07g023730. One primer combination, Ca_Sb07g023730F1/R5 (Supplementary Table 1), yielded a strong amplification product in the tall inbreds ICMP 451, P1449-2 and Tift red but did not amplify in the dwarf inbreds Tift 23DB, PT 732B and 81B. Sanger sequencing of the fragment amplified from the tall inbred ICMP 451 (Supplementary File 1) and BLASTX analysis of the resulting sequence to the 'nr' section of GenBank confirmed that the amplified fragment was derived from gene *ABCB1* (96% identity with both the sorghum and maize ABCB1 protein). Mapping of this product in the informative progeny of the Tift 23DB x ICMP 451 and PT 732B x P1449-2 F_2 populations showed Ca_Sb07g023730 to fall within the cluster of markers that cosegregated with *d2* (Figs. 2A and 2B). This demonstrated that the pearl millet ortholog of Sb07g023730 was located within the *d2* interval.

A second primer set Ca_Sb07g023730F10/R10 gave strong amplification products in both ICMP 451 and Tift 23DB. Sequence analysis showed that the ICMP 451 and Tift 23DB fragments differed by a single synonymous SNP (Supplementary File 2). Because this SNP could not be

visualized by single strand conformation polymorphism gel electrophoresis, the amplicon obtained with primer set Ca_Sb07g023730F10/R10 was not mapped. A BLASTX analysis, however, confirmed that the Ca_Sb07g023730F10/R10 amplification product corresponded to *ABCB1* (92% identity with both the sorghum and maize ABCB1 protein). This primer set, which flanked introns 2 and 3 in Sb07g023730, was used in a semi-quantitative reverse transcriptase (RT)-PCR experiment and showed that Ca_Sb07g023730 was differentially expressed in both the top internode and corresponding leaf between ICMP 451 (*D2D2*) and Tift 23DB *(d2d2)* (Figure 2.5). The expression data provided support for our hypothesis that Sb07g023730 is the *d2* gene.

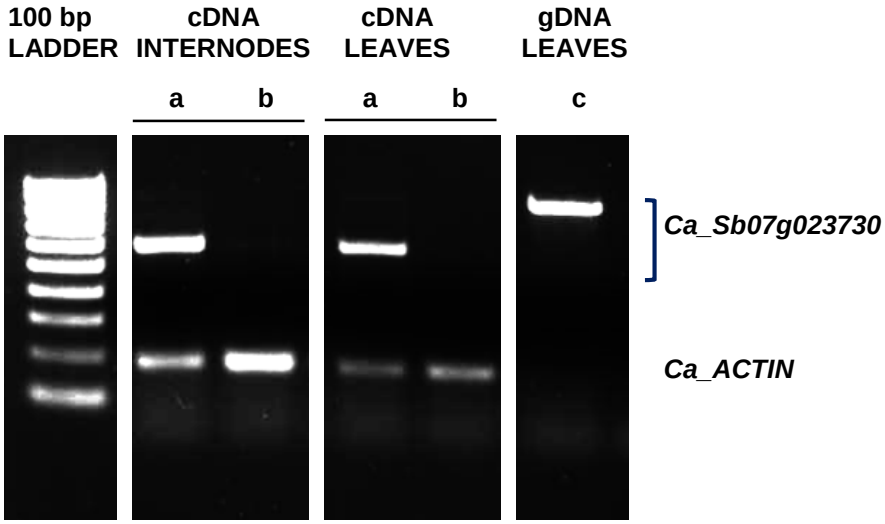


Figure 2.5 Semi-quantitative RT-PCR with primers designed against sorghum gene Sb07g023730 showing a 564 bp fragment in cDNA extracted from leaves and internodes of ICMP 451 (*D2D2*) (lane a) and no/a weak fragment of the same size in cDNA extracted from leaves and internodes of Tift 23DB (*d2d2*) (lane b). Lane c shows the 841 bp fragment obtained in pearl millet genomic DNA. Primers homologous to an actin gene were used as an internal control.

DISCUSSION

Organization of the *d2* region

To our knowledge, no traits have been fine-mapped in pearl millet. The *d2* dwarfing gene, despite its widespread incorporation into commercial germplasm, had previously only been located to a 23.2 cM interval on pearl millet linkage group 4 (Azhaguvel *et al.* 2003). We initially mapped *d2* to a 1.6 cM region, but attempts to further narrow the interval in an approximately 1000 progeny F_2 population largely failed due to a lack of recombination. Markers developed from a 670 kb region on sorghum chromosome 7 (58.37 Mb – 59.04 Mb) all cosegregated with the *d2* phenotype. Recombination in pearl millet is very unevenly distributed (Liu *et al.* 1994; Qi *et al.* 2004) and it might be that the *d2* region has inherently low recombination rates. An alternative explanation is that the two mapping parents, ICMP 451 and Tift 23DB, differ by an inversion in this region. Combining sequence information from a BAC clone originating from the *d2* region with recombination data on the four markers that were identified on this BAC indicated the presence of an inversion in the *d2* region in Tift 23DB relative to sorghum (Figure 2.3). Although we could not precisely determine the boundaries of the inversion, the fact that markers Ca_Sb07g023810 and PSR492 (which is orthologous to Sb07g024860) were located outside the inversion suggests that the inversion likely encompasses less than 100 genes. As the pearl millet genome is characterized by a large number of chromosomal rearrangements relative to other grass genomes (Devos *et al.* 2000), it was not particularly surprising to observe this inversion at the *d2* locus. However, it is unclear if the inversion is present in all pearl millet lines or if it is limited to the *d2* dwarfs or, possibly, even the inbred Tift 23DB. We therefore cannot exclude the possibility that the lack of recombination seen in this region between Tift 23DB and ICMP 451 is the result of the differential presence of this rearrangement in the two parental lines. The pattern of recombination

seen in the PT 732B x P1449-2 mapping population was consistent with both overall reduced recombination and the presence of an inversion in one of the parents, and hence did not provide further insights into the specific organization of the *d2* region.

Comparative analysis of the *d2* region

Comparative information has been used to develop markers for specific chromosome regions and, in some cases, to identify candidate genes underlying traits (Kilian *et al.* 1997; Yan *et al.* 2003; Yan *et al.* 2004). In pearl millet, comparative relationships are complicated by the extensive chromosomal rearrangements that have taken place in the pearl millet genome since its divergence from a common ancestor with foxtail millet some 8.3 million years ago (MYA) (Devos *et al.* 2000; Bennetzen *et al.* 2012). While there is little data on comparative relationships at the DNA sequence level for pearl millet, a comparative study involving *Aegilops tauschii*, rice, sorghum and *B. distachyon* suggests that the relative frequency with which gross chromosomal rearrangements and small-scale rearrangements, mainly the insertion of duplicated gene copies, occur is significantly correlated (Massa *et al.* 2011). Therefore, disruption of colinearity at the gene level might be higher in pearl millet relative to the other grasses. However, even if the larger number of chromosomal rearrangements in pearl millet relative to other grasses means a higher number of gene insertions, the expectation is still that gene orders will have remained sufficiently conserved to exploit comparative relationships for marker development. Nine of the 13 markers developed from the orthologous sorghum region that were polymorphic in the mapping population mapped to the *d2* region in pearl millet. The other four primer sets generated segregation patterns in the preliminary screen, which consisted of the parents and four recombinant progeny that indicated that the polymorphic fragments were located outside the *d2* region. Since we did not attempt by sequence analysis to establish orthology between the scored pearl millet fragments and the sorghum genes

used for primer design, we cannot state with certainty that those four genes are located in non-colinear positions in pearl millet and sorghum. Two of those, Sb07g023460 and Sb07g023470, are found in colinear positions in sorghum, foxtail millet, rice and *B. distachyon*. The other two are either located in non-colinear positions or have duplicated gene copies in non-colinear positions in at least some of the sequenced grass genomes (www.gramene.org).

While gene orders were overall highly conserved between the regions orthologous to *d2* in foxtail, sorghum, rice and *B. distachyon*, species-specific inversions were identified in all four species. The entire distal region of sorghum chromosome 7 had undergone an inversion with a breakpoint between 58.34 Mb and 58.36 Mb. This meant that the genes that were located immediately distal to the inversion breakpoint region in sorghum had been located near the telomere in the ancestral grass genome. The ancestral distal position was maintained only in foxtail millet and rice. In sorghum, *B. distachyon*, and pearl millet, chromosomal rearrangements had moved the ancestral telomere to an interstitial position. This almost certainly had been accompanied by a reduction in recombination, in particular in pearl millet, where a very strong recombination gradient exists along the chromosomes from the centromere to the telomere (Qi *et al.* 2004). Although all the inversions observed were species-specific, it is interesting to note that the distal inversion in sorghum and the 23.82 Mb – 23.97 Mb inversion in rice have one of their breakpoints in common, suggesting that the breakpoint might represent a region on the ancestral grass chromosome that is prone to breakage.

ABCB1 as a candidate for *d2*

Our mapping data had indicated that the distal boundary for the *d2* region in sorghum was at location 58.78 Mb on sorghum chromosome 7, but we had not been able to establish a proximal

boundary due to cosegregation of the markers with the *d2* phenotype. However, haplotype analysis of three tall and three dwarf inbred lines with the mapped markers suggested that the *d2* gene was located distal to marker Ca_Sb07g023430, whose ortholog is located at position 58.37 Mb in sorghum. In the region spanned by the markers B224C4P2 and Ca_Sb07g023440, the allelic composition of the dwarf inbreds was identical and different from that of the tall inbreds, except at locus Ca_Sb07g023910 in line Tift red (Table 2.2). This suggests that the three dwarfs analyzed (Tift 23DB, PT 732B and 81B) are derived from the same *d2* source. As expected, we find different haplotypes in this region in the tall inbreds ICMP 454, P1449-2 and Tift red (Table 2.2). The inbred 81B is a downy mildew resistant selection from gamma-irradiation treated Tift 23DB (Rai and Hanna 1990b). While this line maintains the dwarf haplotype in the *d2* region, it likely underwent a recombination event with a tall line between markers Sb07g023430 and PSMP344 (Table 2.2). Tift red is a backcross line produced by the late Glen Burton that carries a gene for purple plant color and the tall *D2* allele in a Tift 23 background. Considering that no recombination was identified between Ca_Sb07g023430 and Ca_Sb07g023440 in the ~1500 progeny we analyzed from two crosses, it was surprising to see that Tift 23DB and Tift red, which are near-isogenic lines, differ by a recombination event between those two markers. The unexpected haplotype of Tift red was crucial to determining the proximal boundary of the *d2* region.

The region delineated in sorghum as being orthologous to the *d2* region in pearl millet contained the adenosine triphosphate (ATP)-binding cassette (ABC) subfamily B1 gene, an obvious candidate for *d2* because a mutation in this gene was shown to underlie the recessive sorghum *dw3* dwarfing phenotype (Multani *et al.* 2003). A mutation in the same gene is also responsible for the brachytic2 (*br2*) phenotype in maize. The ABCB1 protein belongs to the multidrug resistant (MDR) class of P-glycoproteins and plays a role in auxin transport in the

nodal/intercalary meristem regions (Knoller *et al.* 2010). Consequently, in the *br2/Zmpgp1* and *dw3/Sbpgp1* mutants, auxin accumulates in the vicinity of the nodes. Because auxin is synthesized mainly in the shoot apex and young leaves, and then transported basipetally, the lowermost internodes in the sorghum *dw3* and maize *br2* mutants are affected the most by the modulation of auxin transport caused by a knockout of the *ABCB1* gene (Multani *et al.* 2003; Knoller *et al.* 2010). The phenotype of stacked lower internodes in the pearl millet *d2* dwarf is very similar to that observed in sorghum *dw3* and maize *br2* mutants (Fig 2.1C).

In an attempt to validate *ABCB1* as a candidate for *d2*, we designed multiple primer sets against the sorghum *ABCB1* gene and tested them in three dwarf and three tall lines. One primer set, which spanned intron 1, yielded an amplification product in all tall lines tested and in none of the dwarfs, and this polymorphism cosegregated with the height phenotype in the set of informative F_2 plants in both the Tift 23DB x ICMP 451, and PT 732B x P1449-2 populations. The most likely cause for the lack of amplification in the dwarf lines is either a single nucleotide polymorphism (SNP) or deletion at the primer site(s) that prohibits primer extension or the presence of an insertion in the region between the two primer binding sites that extends the fragment to be amplified beyond the length limit of a typical PCR reaction. Further work is needed to determine whether the observed variation could be the underlying cause of the dwarf phenotype.

We also analyzed the expression of the *ABCB1* gene in the topmost internode and the corresponding leaf in flowering plants of one *d2* dwarf plant, Tift 23DB, and one tall plant, ICMP 451, using semi-quantitative RT-PCR. The RT-PCR yielded a strong amplification product in both tissues in ICMP 451 and a weak product in Tift 23DB. This suggests that *ABCB1* is differentially expressed in the tall and dwarf inbreds in both tissues or that the stability of the *ABCB1* mRNA is reduced in the dwarf mutant (Figure 2.5). A reduced transcript level in the dwarf

mutant is in agreement with the recessive nature of the *d2* mutation. In maize, *ABCB1* is expressed in nodal tissue, and possibly in internodes although reports on the latter are not consistent (Multani *et al.* 2003; Knoller *et al.* 2010). No expression was detected in maize leaves (Multani *et al.* 2003). Nodes were not included in our preliminary expression analysis because of the difficulty of extracting RNA from the hard node tissue. In Arabidopsis, *ABCB1* expression is highest in nodes, but the gene is also expressed in a range of other tissues (Titapiwatanakun and Murphy 2009). More detailed expression analyses are needed in pearl millet to determine precisely where the *ABCB1* gene is expressed and at what levels. The higher expression in the tall compared to the dwarf inbred, however, provided some support for our hypothesis that *ABCB1* is a reasonable candidate for *d2*.

Conclusion

Using a combination of genetic mapping, haplotype analysis and comparative genomics, we have fine-mapped the pearl millet *d2* dwarf phenotype to a region which, in sorghum, spans 410 kb and contains 40 annotated genes. A candidate gene, *ABCB1*, was identified as putatively underlying *d2*. Work is currently underway to isolate full length copies of the *ABCB1* gene from both a tall and a dwarf inbred to further test our candidate gene hypothesis. In the meantime, our work provides breeders with a set of markers that can be used to identify the presence of the recessive *d2* gene in heterozygous condition and at the seedling stage. Phenotypically, the dwarf phenotype can only be scored in homozygous condition and at the booting stage, so the markers will enhance the efficiency of breeding programs that use tall lines as sources of novel genes for the improvement of dwarf inbreds.

Acknowledgements

Early work on the PT 732B x P1449-2 cross was carried out by FKP, SF and KMD at the Comparative Genomics Unit, John Innes Centre, Norwich, UK. The research was funded by fellowships from the Pioneer Hibred – Generation Challenge Program (to VJ), the Kirkhouse Trust (to NN), and by a National Institute of Food and Agriculture (NIFA) 1890_CSREES Capacity Building Grants (grant 2008-38814-04740; PI: B. Singh, Fort Valley State University, GA, USA). We thank W. Hanna for providing seeds of Tift 23DB and Tift red, and generating the Tift 23DB x ICMP 451 F_2 mapping population

REFERENCES

Allouis, S., X. Qi, S. Lindup, M. D. Gale and K. M. Devos, 2001 Construction of a BAC library of pearl millet, *Pennisetum glaucum*. Theor. Appl. Genet. **102:** 1200-1205.

Azhaguvel, P., C. T. Hash, P. Rangasamy and A. Sharma, 2003 Mapping the d_1 and d_2 dwarfing genes and the purple foliage color locus *P* in pearl millet. J. Hered. **94:** 155-159.

Bakshi, J. S., K. O. Rachie and S. Amarjit, 1966 Development of dwarf strains of pearl millet and an assessment of their yield potential. Curr. Sci. **35:** 355-356.

Beidler, J. L., P.R. Hillard and R.L. Rill, 1982 Ultrasensitive staining of nucleic acids with silver. Anal.Biochem. **126:** 374-380

Bennetzen, J. L., J. Schmutz, H. Wang, R. Percifield, J. Hawkins *et al.*, 2012 Reference genome sequence of the model plant Setaria. Nat. Biotechnol. **30:** 555-561.

Bidinger, F. R., and D. S. Raju, 1990 Effects of the d_2 dwarfing gene in pearl millet. Theor. and Appl. Genet. **79:** 521-524.

Burton, G. W., 1980 Registration of pearl millet inbred Tift 383 and Tifleaf 1 pearl millet (Reg. PL 8 and Reg. No. 60). Crop Sci. **20:** 293.

Burton, G. W., and E. H. Devane, 1951 Starr millet: Synthetic cattail lasts longer, produces more beef per acre. Southern Seedsman, March issue.

Burton, G. W., and J. C. Fortson, 1966 Inheritance and utilization of five dwarfs in pearl millet. (*Pennisetum typhoides*) breeding. Crop Sci. **6:** 69-72.

Burton, G. W., W. G. Monson, J. C. Johnson, R. S. Lowrey, H. D. Chapman *et al.*, 1969 Effect of the d_2 dwarf gene on the forage yield and quality of pearl millet. Agron. J. 61: 607-612

Busso, C. S., K. M. Devos, G. Ross, M. Mortimore, W. M. Adams *et al.*, 2000 Genetic diversity within and among landraces of pearl millet (*Pennisetum glaucum*) under farmer management in West Africa. Genet. Resour. and Crop Evol. **47:** 561-568.

Chemisquy, M. A., L. M. Giussani, M. A. Scataglini, E. A. Kellogg and O. Morrone, 2010 Phylogenetic studies favour the unification of *Pennisetum*, *Cenchrus* and *Odontelytrum* (*Poaceae*): a combined nuclear, plastid and morphological analysis, and nomenclatural combinations in Cenchrus. Ann. Bot. **106:** 107-130.

Chomczynski, P., and N. Sacchi, 1987 Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. **162:** 156-159.

Corpet, F., 1988 Multiple sequence alignment with hierarchical-clustering. Nucleic Acids Res. **16:** 10881-10890.

Devos, K. M., 2005 Updating the 'Crop circle'. Curr. Opin. in Plant Biol. **8:** 155-162.

Devos, K. M., T. S. Pittaway, A. Reynolds and M. D. Gale, 2000 Comparative mapping reveals a complex relationship between the pearl millet genome and those of foxtail millet and rice. Theor. and Appl. Genet. **100:** 190-198.

Dill, A., H. S. Jung and T. P. Sun, 2001 The DELLA motif is essential for gibberellin-induced degradation of RGA. Proc. Natl. Acad. Sci. USA **98:** 14162-14167.

Dubcovsky, J., W. Ramakrishna, P. J. SanMiguel, C. S. Busso, L. Yan *et al.*, 2001 Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. Plant Physiol. **125:** 1342-1353.

Gulia, S. K., J. P. Wilson, J. Carter and B. P. Singh, 2007 Progress in grain pearl millet research and market development, pp. 196-203 in *Issues in new crops and new uses*, edited by J. Janick and A. Whipkey. AHES Press, Alexandria.

Hanna, W. W., T. P. Gaines and W. G. Monson, 1979 Seed set and d_2 gene effects on pearl millet forage quality. Agron. J. **71:** 1027-1029.

Hanna, W. W., G. M. Hill, R. N. Gates, J. P. Wilson and G. W. Burton, 1997 Registration of 'Tifleaf 3' pearl millet. Crop Sci. **37:** 1388-1388.

Hedden, P., 2003 The genes of the Green Revolution. Trends Genet. **19:** 5-9.

Hein, M. A., 1953 Registration of varieties and strains of pearl millet (*Pennisetum glaucum* (L.) R. Br.). Agron. J. **45:** 573-574.

Jia, Q., J. Zhang, S. Westcott, X. Q. Zhang, M. Bellgard *et al.*, 2009 GA-20 oxidase as a candidate for the semidwarf gene *sdw1/denso* in barley. Funct. Integr. Genom. **9:** 255-262.

Jia, Q., X. Q. Zhang, S. Westcott, S. Broughton, M. Cakir *et al.*, 2011 Expression level of a gibberellin 20-oxidase gene is associated with multiple agronomic and quality traits in barley. Theor. Appl. Genet. **122:** 1451-1460.

Johnson, J. C., R. S. Lowrey, W. G. Monson and G. W. Burton, 1968 Influence of dwarf characteristic on composition and feeding value of near-isogenic pearl millets. J. of Dairy Sci. **51:** 1423-1424.

Kadam, B.S., S. M. Patel and R. K. Kulkarni, 1940 Consequences of inbreeding in bajri. J. Hered.31: 201-207

Kilian, A., J. Chen, F. Han, B. Steffenson and A. Kleinhofs, 1997 Towards map-based cloning of the barley stem rust resistance genes *Rpgl* and *rpg4* using rice as an intergenomic cloning vehicle. Plant Mol. Biol. **35:** 187-195.

Knoller, A. S., J. J. Blakeslee, E. L. Richards, W. A. Peer and A. S. Murphy, 2010 *Brachytic2/ZmABCB1* functions in IAA export from intercalary meristems. J. Exp. Bot. **61:** 3689-3696.

Kumar, K. A., and D. J. Andrews, 1993 Genetics of qualitative traits in pearl millet: a review. Crop Sci. **33:** 1-20.

Laurie, D. A., N. Pratchett, C. Romero, E. Simpson and J. W. Snape, 1993 Assignment of the *denso* dwarfing gene to the long arm of chromosome 3 (3H) of barley by use of RFLP markers.Plant Breed. **111**: 198-203.

Liu, C. J., J. R. Witcombe, T. S. Pittaway, M. Nash, C. T. Hash *et al.*, 1994 An RFLP-based genetic map of pearl millet (*Pennisetum glaucum*). Theor. and Appl. Genet. **89:** 481-487.

Martins-Lopes, P., H. Zhang and R. Koebner, 2001 Detection of single nucleotide mutations in wheat using single strand conformation polymorphism gels. Plant Mol. Biol. Rep. **19:** 159-162.

Massa, A. N., H. Wanjugi, K. R. Deal, K. O'Brien, F. M. You *et al.*, 2011 Gene space dynamics during the evolution of *Aegilops tauschii*, *Brachypodium distachyon*, *Oryza sativa*, and *Sorghum bicolor* genomes. Mol. Biol. and Evol. **28:** 2537-2547.

Matsumoto, T., J. Z. Wu, H. Kanamori, Y. Katayose, M. Fujisawa *et al.*, 2005 The map-based sequence of the rice genome. Nature **436**: 793-800.

Mickelson, H. R and D.C. Rasmusson, 1994 Genes for short stature in barley. Crop Sci. **34:** 1180-1183.

Milach, S. C. K., and L. C. Federizzi, 2001 Dwarfing genes in plant improvement. Adv. in Agron. **73:** 35-63.

Money, T. A., C. J. Liu and M. D. Gale, 1994 Conversion of RFLP markers for downy mildew resistance in pearl millet to sequence-tagged-sites, pp. 65-68 in *Use of molecular markers in sorghum and pearl millet breeding for developing countries*, edited by J.R.Witcombe and R.R. Duncan. Overseas Development Administration, London.

Monna, L., N. Kitazawa, R. Yoshino, J. Suzuki, H. Masuda *et al.*, 2002 Positional cloning of rice semidwarfing gene, *sd-1*: rice "green revolution gene" encodes a mutant enzyme involved in gibberellin synthesis. DNA Res. **9:** 11-17.

Multani, D. S., S. P. Briggs, M. A. Chamberlin, J. J. Blakeslee, A. S. Murphy *et al.*, 2003 Loss of an MDR transporter in compact stalks of maize *br2* and sorghum *dw3* mutants. Science **302:** 81-84.

Murray, M. G., and W. F. Thompson, 1980 Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. **8:** 4321-4325.

Padi, F.K., 2002 Genetic analyses of adaptive traits in pearl millet (*Pennisetum glaucum* (L.) R. Br. PhD Thesis, University of East Anglia, Norwich, UK. 245pp.

Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood *et al.*, 2009 The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**: 551-556.

Peng, J., P. Carol, D. E. Richards, K. E. King, R. J. Cowling *et al.*, 1997 The Arabidopsis GAI gene defines a signaling pathway that negatively regulates gibberellin responses. Genes Dev. **11:** 3194-3205.

Peng, J., D. E. Richards, N. M. Hartley, G. P. Murphy, K. M. Devos *et al.*, 1999 'Green revolution' genes encode mutant gibberellin response modulators. Nature **400:** 256-261.

Qi, X., T. S. Pittaway, S. Lindup, H. Liu, E. Waterman *et al.*, 2004 An integrated genetic map and a new set of simple sequence repeat markers for pearl millet, *Pennisetum glaucum*. Theor. And Appl. Genet. **109:** 1485-1493.

R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.15.1. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rai, K. N., S. K. Gupta, R. Bhattacharjee, V. N. Kulkarni, A. K. Singh *et al.*, 2009 Morphological characteristics of ICRISAT-bred pearl millet hybrid seed parents. J. SAT Agric. Res. 7.

Rai, K. N., and W. W. Hanna, 1990a Morphological characteristics of tall and dwarf pearl millet isolines. Crop Sci. **30:** 23-25.

Rai, K. N., and W. W. Hanna, 1990b Morphological changes in an inbred line of pearl millet selected for downy mildew resistance. J. of Genet. and Breed. **44:** 199-202.

Rai, K. N., and A. S. Rao, 1991 Effect of d_2 dwarfing gene on grain yield and yield components in pearl millet near-isogenic lines. Euphytica **52:** 25-31.

Sasaki, A., M. Ashikari, M. Ueguchi-Tanaka, H. Itoh, A. Nishimura *et al.*, 2002 Green revolution: a mutant gibberellin-synthesis gene in rice. Nature **416:** 701-702.

Schertz, K. F., D. T. Rosenow, J. W. Johnson and P. T. Gibson, 1974 Single *Dw3* height-gene effects in 4- and 3-dwarf hybrids of *Sorghum bicolor* (L.) Moench. Crop Sci. **14:** 875-877.

Soman, P., K. N. Rai and F. R. Bidinger, 1989 Dwarfing gene effects on coleoptile length in pearl millet. Crop Sci. **29:** 956-958.

Spielmeyer, W., M. H. Ellis and P. M. Chandler, 2002 Semidwarf (*sd-1*), "green revolution" rice, contains a defective gibberellin 20-oxidase gene. Proc. Natl. Acad. Sci. USA **99:** 9043-9048.

Steiner, J. J., C. J. Poklemba, R. G. Fjellstrom and L. F. Elliott, 1995 A rapid one-tube genomic DNA extraction process for PCR and RAPD Analyses. Nucleic Acids Res. **23:** 2569-2570.

Tang, H., X. Wang, J. E. Bowers, R. Ming, M. Alam *et al.*, 2008 Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. **18:** 1944-1954.

Titapiwatanakun, B., and A. S. Murphy, 2009 Post-transcriptional regulation of auxin transport proteins: cellular trafficking, protein phosphorylation, protein maturation, ubiquitination, and membrane composition. J. of Expt. Bot. **60:** 1093-1107.

Van den Berg, N., B. G. Crampton, I. Hein, P. R. Birch and D. K. Berger, 2004 High-throughput screening of suppression subtractive hybridization cDNA libraries using DNA microarray analysis. Biotechniques **37:** 818-824.

Vogel, J. P., D. F. Garvin, T. C. Mockler, J. Schmutz, D. Rokhsar *et al.*, 2010 Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature **463**: 763-768.

Yan, L., A. Loukoianov, A. Blechl, G. Tranquilli, W. Ramakrishna *et al.*, 2004 The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. Science **303:** 1640-1644.

Yan, L., A. Loukoianov, G. Tranquilli, M. Helguera, T. Fahima *et al.*, 2003 Positional cloning of the wheat vernalization gene *VRN1*. Proc. of the Natl. Acad. Sci. USA **100:** 6263-6268.

The supplementary tables for this published report can be found at

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583462/bin/supp_3_3_563__index.html

CHAPTER III:

STRUCTURAL CHARACTERIZATION OF *ABCB1*, THE GENE UNDERLYING THE *D2* DWARF PHENOTYPE IN PEARL MILLET, *CENCHRUS AMERICANUS* (L.) MORRONE[2]

[2] Parvathaneni, R.K., Wu, X., and K.M. Devos. To be submitted to Plant Genome

ABSTRACT

Pearl millet is an important food crop in arid and semi-arid regions of South Asia and sub-Saharan Africa. In Australia and the United States, it is grown to a limited extent as a summer fodder crop. The *d2* dwarf germplasm has been widely used in the last half-century to develop high-performing pearl millet hybrids. We previously mapped the *d2* trait to a 1.6 cM region in LG4 and identified the *ABCB1* gene as a candidate underlying the trait. Here, we report the sequence, structure and expression of *ABCB1* in tall (*D2D2*) and *d2* dwarf (*d2d2*) germplasm. The *ABCB1* gene from *d2* dwarfs differs from the tall copy by the presence of two different high copy transposable elements, one in the coding region and the second located 665 bp upstream of the ATG start codon. These transposons were present in all *d2* dwarfs tested that were reported to be of independent origin and absent in the analyzed wild-type tall germplasm. We also compared the expression profile of this gene in different organs of multiple tall and *d2* dwarf inbreds including the isogenic inbreds at the *d2* locus, Tift 23B (*D2D2*) and Tift 23DB (*d2d2*). Heterologous transformation studies of the tall (*Ca_ABCB1*) and the *d2* dwarf (*Ca_abcb1*) pearl millet alleles in Arabidopsis *abcb1* mutants are also described.

INTRODUCTION

Modern agriculture has benefited tremendously from the improvements that plant breeders have made to plant architecture. In the 1960s, the use of dwarf mutants in wheat (*Rht1*) and rice (*sd-1*) allowed application of higher rates of fertilizer which increased yield/acre and production by >99% (Harzell 2009). This was dubbed the "Green Revolution" (reviewed by (Hedden 2003)). Height-reducing genes have also been extensively used in other cereals such as sorghum (*e.g. dw3*; (Schertz et al. 1974)) , barley (*e.g. denso*; (Jia et al. 2009)), rye (*e.g. Ddw1;*(Milach and Federizzi 2001)) and pearl millet (*e.g. d2*; (Burton and Fortson 1966)). Millets, of which pearl millet

(*Cenchrus americanus* (L.) Morrone) is the most widely grown, are cultivated on more than 27 million hectares in regions of Africa and the Indian sub-continent as a dual purpose food and fodder crop (Rai et al., 2012). Pearl millet is also grown as a minor forage crop in the United States and Australia. In the developed world, pearl millet is grown exclusively as *d2* dwarf hybrids (Gulia et al. 2007). Tall pearl millet varieties are preferred in sub-Saharan Africa because the long stems are used for forage, fencing and roofing. In India, both tall and *d2* dwarf hybrids are grown depending on the region and season. However, in recent years, farmers' preference seems to have shifted towards tall hybrids as they have "good yield" (grain and fodder), "good taste", and a "good return value" (Asare-Marfo et al. 2011). *d2* dwarfs yield less forage compared to tall pearl millet lines but produce better quality forage due to a higher leaf to stem ratio (Johnson et al. 1968). Grain yield of *d2* dwarfs is lower in some backgrounds, but unaffected in others (Bidinger and Raju 1990; Rai and Rao 1991). Because *d2* dwarfs are more amenable to dense planting and machine harvesting, it is expected that the *d2* dwarfing trait will gain in importance as commercial agriculture in developing countries is moving towards mechanization.

A candidate gene for *d2* was recently identified using a combination of high-density gene mapping, haplotype analysis of the *d2* region in tall and dwarf lines, and comparative genome analysis with rice and sorghum (Parvathaneni et al. 2013). The gene, *ABCB1*, encodes a P-glycoprotein (PGP) which facilitates cell to cell polar auxin transport. *ABCB1* belongs to the multidrug resistance (MDR)/P-glycoprotein (PGP) sub-family of the large super-family ATP-binding cassette (ABC) transporters (reviewed in Dean et al. 2001; Sanchez-Fernandez et al. 2001). Complete ABC transporter proteins contain four domains, including two transmembrane domains (TMDs) that bind specific substrates and two nuclear binding domains (NBDs) that hydrolyze ATP and use the energy from the hydrolysis of ATP to transport the substrate. Half

transporters contain one TMD and one NBD, and function either as homo- or heterodimers (Vasiliou et al. 2009). ABC transporters transport a wide range of compounds (reviewed in Kang et al. 2011). In plants, transported compounds include but are not limited to phytohormones, heavy metals, lipids, antibiotics, and glucosylated compounds (reviewed in Kang et al. 2011). ABC proteins may also play a role in plant-pathogen interactions and in the modulation of ion channels (Rea 2007). There are eight ABC subfamilies, labeled ABCA to ABCH (Verrier et al. 2008). Proteins of the ABCB (PGP) subfamily in plants function in auxin transport, and are hypothesized to play an active role in the efflux of auxin from meristematic cells where the auxin concentration is very high (Blakeslee et al. 2005).

Reduced height as a result of loss-of-function mutations in ABCB1 is well characterized in the panicoid grasses sorghum (*Sorghum bicolor* – *dw3* mutation) and maize (*Zea mays* – *br2* mutation) (Multani et al. 2003). An 881bp tandem duplication in the fifth exon of *ABCB1* reduces internode length leading to a dwarf phenotype in sorghum (*dw3*) (Multani et al. 2003). This phenotype is unstable because recombination between the direct repeats can give rise to dwarf-to-tall revertants. Interestingly, an unequal crossover event also produced a stable *dw3* dwarf mutant that lacked the duplicated region and differed by the presence of several SNPs and one CG microsatellite in comparison to the *Dw3* allele. These mutations disrupt the reading frame and create a truncated protein lacking 200 amino acids (Multani et al. 2003). The *dw3* gene is widely deployed in sorghum breeding programs. Several recessive T-DNA insertional mutants of *ABCB1* causing dwarf phenotypes have been reported in maize (*br2*) (Multani et al. 2003). These insertion mutants have an extreme dwarf phenotype that cannot be exploited commercially. More recently, however, a *br2* allele has been identified that results in a 20 percent reduction in height and improves yield (Xing et al. 2015). This *br2* allele contains 4 synonymous SNPs and one non-

synonymous SNP in the last exon of the *ABCB1* gene (Xing et al. 2015). The potential of mutating *ABCB1* genes to create agriculturally useful dwarfs is being explored using EMS mutagenesis in the orphan crop tef (*Ergrostis tef*) to enhance lodging tolerance (Zhu et al. 2012).

In this manuscript, we describe the structural changes in *ABCB1* that led to its loss of function in the pearl millet *d2* dwarf. Expression analyses support the identity of *ABCB1* as *d2*. Transformation of the pearl millet functional and mutant alleles in an Arabidopsis abcb1 mutant however did not yield conclusive results.

MATERIALS AND METHODS

Plant material and DNA isolation

The following pearl millet inbred lines were used: ICMP 451 (*D2D2*), Tift red (*D2D2*), P-1449 (*D2D2*), Tift 23B (*D2D2*), 81B (*d2d2*), pT 732B (*d2d2*), Tift 23DB (*d2d2*), IP 8227 (*d2d2*), IP 8208 (*d2d2*), IP 8288 (*d2d2*), IP 8157 (*d2d2*), IP 8112 (*d2d2*), IP 8008 (*d2d2*), IP 8058 (*d2d2*), IP 10399 (*d2d2*) and Starr (allele composition at the *D2* locus unknown). Inbreds with allelic composition *D2D2* are tall, those with allelic composition *d2d2* are dwarf. Inbreds ICMP 451, P-1449, 81B and pT 732B, and inbreds with the prefix 'IP' were obtained from ICRISAT. Inbreds with the designation 'Tift' were obtained from Wayne Hanna, University of Georgia, Tifton, and Starr (NSL 4716) was obtained from the USDA National Plant Germplasm System (NPGS). Plants were grown in a growth chamber under a temperature of 26 ^{0}C and a 15/9 hours day/night cycle. DNA was isolated from leaves of three week old seedlings using either a CTAB extraction protocol (Murray and Thompson 1980) or a Qiagen DNA miniprep kit (Qiagen, Valencia, CA).

Heterologous transformation studies in Arabidopsis were carried out in an *abcb1* single mutant (*atpgp1-2*; CS863226). The single mutants *abcb1* and *abcb19* (*atmdr1-101*; SALK_033455) were obtained from the Arabidopsis Biological Resource Centre (ABRC). To

generate an *abcb1abcb19* double mutant, the *abcb19* mutant (*atmdr1-101*) was used as the female plant and the *abcb1* mutant (*atpgp1-2*) was used as the pollen donor. F_1 plants were selfed to generate F_2 seed. F_2 seed was germinated on half strength Murashige and Skoog (MS) (Murashige and Skoog 1962) agar plates and *abcb1abcb19* double mutants were identified phenotypically and will be confirmed by PCR in the near future.

Isolation of a BAC clone carrying the *ABCB1* allele from the dwarf inbred Tift 23DB

An available BAC library from the dwarf inbred line Tift 23DB covers 5.6 genome equivalents and has an average insert size of ~90 kb (Allouis et al. 2001). The 159,100 clones that constitute the library were pooled per 384-well plate (single-plate pool) and DNA was extracted from each pool. DNA from 10 single-plate pools was combined into a super pool. PCR-based screening of super pools and single plate-pools was performed using the primer sets Ca_Sb07g023730_F1 and Ca_Sb07g023730_R1 (Table 3.1). PCR amplifications were performed in 20 µl reaction volumes containing 1X PCR buffer (5X GoTaq® Flexi Buffer), 1.5 mM $MgCl_2$, 0.25 mM of each dNTP, 0.5 µM forward and reverse primers, 1U of GoTaq DNA polymerase (Promega) and 25 ng of DNA template. Amplification conditions consisted of an initial denaturation step at 95 ^{0}C for 5 minutes followed by 34 cycles of 95 ^{0}C for 30 sec, 59 ^{0}C for 30 sec and 72 ^{0}C for 1 min, and a final extension of 72 ^{0}C for 5 min. Amplification products were run on 1 % agarose gels. Once the plate address for a positive clone was identified, the clones in the corresponding 384-well plate were double-gridded on an Amersham HyBond N+ nylon membrane (GE Life Sciences) using a hand-held "colony-plaque lift" tool (V&P Scientific). Colonies were grown on the filters placed on LB medium with 25 µM chloramphenicol at 37 ^{0}C overnight. Lysis of the bacteria and denaturation of the DNA was performed according to established protocols (The Molecular Cloning Lab Manual; (Sambrook et al.,1989). Colony

hybridization with the 841 bp amplicon obtained with primer set Ca_Sb07g023730_F1/R1 and visualization of the hybridization sites were performed using the Amersham Gene Images AlkPhos Direct labelling and Detection System using the manufacturer's recommendations with the following modifications. To decrease non-specific hybridization, membranes were pre-hybridized with hybridization buffer at 65 ^{0}C for 45 minutes. The hybridization and 2X primary wash were also performed at 65 ^{0}C. Fluorescent signals were recorded on a high performance chemiluminescence film (Amersham Hyperfilm^{TM} ECL).

Isolation of a fosmid clone carrying the *ABCB1* allele from the tall inbred ICMP 451

To isolate the *ABCB1* allele from a tall inbred, a fosmid library was constructed of ICMP 451. High molecular weight DNA was isolated from nuclei using a modified protocol from Peterson and colleagues (Peterson *et al*., 2000). To shear the DNA to an average size of 35 kb, 10 µg of high quality nuclear DNA (50 to 100 ng/µl in 10 mM TE) was added to 1.5 volumes of AP3 buffer provided in the DNeasy plant mini kit (Qiagen), and passed through a DNeasy mini spin column (Qiagen). The manufacturer's protocol was then followed to recover the nuclear DNA. The DNA was separated for 16 hours on a 1% low melting point agarose gel in 0.5X TBE buffer by field inversion gel electrophoresis (FIGE) (BIO-RAD FIGE Mapper Electrophoresis System) with forward pulses of 180V, reverse pulses of 120V, and a linear switch time ramp of 0.1-2.0 seconds. A 5-50 kb pulsed field molecular weight ladder (Lambda DNA monocot mix, NEB) and a 42 kb control DNA sample from the Copy Control Library production kit (Epicentre Biotechnologies, Madison, USA) were used as standards. Fragments in the size range 30-43 kb were eluted from the gel using the components of the Copy Control Fosmid Library production kit (Epicentre Biotechnologies, Madison, USA) and used to construct a library following the manufacturer's recommendations. Fosmids were plated at a density of ~800 clones per plate.

Colonies from a single plate were combined using LB broth to constitute a single pool and stored as a 20 percent glycerol stock. Ten µl of bacterial culture from eight pools were combined to constitute super pools for a total of 24 super pools. The total coverage of the library was estimated at 2.2 genome equivalents. The same primer sets and conditions as used in the BAC library screening were used to screen the fosmid library using 0.5 µl of super pool culture as template. Once the address of the pool was known, the culture from the pool was titrated and plated to give ~100 colonies per plate. Twelve sub-pools, each consisting of ~100 colonies, were screened by PCR for each positive pool. The positive sub-pools were plated again and individual colonies were screened by PCR to identify the fosmid that contained the *ABCB1* gene.

Sequencing of the BAC and fosmid clones

Plasmid DNA was isolated from the *ABCB1*-positive fosmid (fosmid-19) and BAC (156A12) clones using the Qiagen Large Construct kit (Qiagen) with minor modifications such as 1) The *E. coli* cells were pelleted at a speed of 8,500 rcf. 2) Isopropanol precipitation was conducted at 13,000xg for 45 min at 4 ^{0}C and 3) Ethanol precipitation was conducted at 13,000xg for 30 min at 4 ^{0}C. DNA of fosmid-19 and BAC 156A12 was fragmented to an average size of 900 bp by mechanical shearing (nebulization), cleaned using the AMpure beads (Agencourt Bioscience, Beverly, MA), after which fragments were end repaired and ligated to adapters at the Georgia Genomic Facility (GGF), UGA, using in-house protocols. Libraries were paired-end sequenced using the Roche 454 GS FLX sequencing platform with Titanium chemistry. BAC 156A12 was sequenced to a depth of 67X and fosmid-19 to a depth of 23X. The read statistics are provided in Table 3.2. Sequences were assembled *de novo* using MIRA v. 3.2.0 (Li et al. 2008) using both the 'normal' and 'accurate' mode for the assembly of BAC 156A12 and 'accurate' settings for assembly of the fosmid-19 clone.

Classification of the transposable elements

Sequences of the transposable elements were translated in 6 frames and compared to the hidden Markov model (HMM) profiles deposited in the GYDBv2.0 (Llorens et al. 2011) to find their reverse transcriptase (RT) domains. The closest RT clade and superfamily in the database was identified by the hmmsearch program from HMMER (version 3.0) package. The closest transposable element was identified by the BLASTN analysis of the DNA sequences of the element to the MIPS repeat element database (Nussbaumer et al. 2013) and Repbase (Jurka et al. 2005). The family name of the element were identified by an in-house script.

PCR testing for the presence of the 'Juriah' transposon

The primer set Ca_Sb07g023730 F1/R5 flanks the Juriah transposable element in the coding region while the primer set RB2F1- Ca_Sb07g023730R5 spans the 3' gene-LTR boundary (Table 3.1). These primer sets were used to test the tall and *d2* dwarf inbred plants for the presence of the Juriah element. PCR amplifications were performed in 20 µl reaction volumes containing 1X PCR buffer (5X GoTaq® Flexi Buffer), 1.5 mM $MgCl_2$, 0.25 mM of each dNTP, 0.5 µM forward and reverse primers, 1U of GoTaq DNA polymerase (Promega) and 25 ng of DNA template. Amplification conditions consisted of an initial denaturation step at 95 ^{0}C for 5 minutes followed by 34 cycles of 95 ^{0}C for 30 sec, 61 ^{0}C for 30 sec and 72 ^{0}C for 1 min, and a final extension at 72 ^{0}C for 5 min. Amplification products were run on 1 % agarose gels.

Expression analyses

For expression analyses, three tall inbred lines (ICMP 451, P-1449-2 and Tift 23B) and three dwarf inbred lines (81B, pT 732B and Tift 23DB) were grown in 6 inch pots (analysis at the seedling stage) or 12 inch pots (analysis of adult tissues) in the greenhouse under 14 hour day lengths and a day/night temperature of approximately 27 ^{0}C/21 ^{0}C. When seedlings reached the

five-leaf stage (Figure 3.1C), the oldest leaf and the stem were collected for expression analyses. The panicle, top node, top internode and root were collected when 50% of the stigma on the head had emerged (Figure 3.1A; 3.1B). Samples were flash frozen in liquid nitrogen and stored at -80 ^{0}C until the time of RNA extraction. RNA was extracted using a standard protocol using TRIzol reagent (Chomczynski and Sacchi 1987). The RNA quality was checked on a 1% agarose gel. Up to 5 µg of RNA was treated with DNase using the Ambion TURBO DNA-*free* kit after which the RNA was quantified using a nanodrop (NanoDrop technologies). 800 ng (seedling tissues) or 500 ng (adult tissues) of DNase-treated total RNA was used to conduct cDNA synthesis with the SuperScript III Reverse Transcriptase system (Life Technologies) using an OligodT (20) primer according to the manufacturer's recommendations. At least three biological replicates were analyzed for each sample.

Quantitative RT-PCR was conducted as detailed in Dash and Malladi (2012). In brief, the Veriquest SYBR Green qPCR (Affymetrix, Santa Clara,CA) master mix was used with the Stratagene Mx3005P real-time PCR system (Agilent Technologies, Santa Clara, CA). PCR conditions were as follows: 50 ^{0}C for 2 min, 95 ^{0}C for 10 min, and 40 cycles of 95 ^{0}C for 30 s and 60 ^{0}C for 1 min. A melt curve analysis was performed at the end of the cycles to check for single peaks, which indicate amplification of a single fragment. The primers Ca_ABCB1_F20 and Ca_ABCB1_R20 were designed to quantify expression of the *Ca_ABCB1* gene, while primers designed against the pearl millet actin (Ca_Act_F1/R1) and glyceraldehyde 3-phophate dehydrogenase (GAPDH) (Ca_GAPDH_F1/R1) genes were used to control for variation in the amount of input RNA (Table 3.1). One complete set of replicates was analyzed per PCR run. The primer efficiency (PE) was calculated using the program LinRegPCR for each run using the amplification plots (Ruijter et al. 2009). The relative quantity (RQ) of each amplicon was

calculated by the formula (1/(PE*Ct)) with Ct the number of PCR cycles required to cross a fluorescence threshold of 0.1. The normalized relative quantity (NRQ) of the gene-specific amplicons was calculated by dividing their relative quantity (RQ) by the normalization factor (geometric mean of RQ for actin and GAPDH). Expression levels of *ABCB1* in the different genotypes were determined relative to the average NRQ value for the same tissue and developmental stage in Tift 23B. Statistics were calculated on the $\log_2$ (NRQ) values.

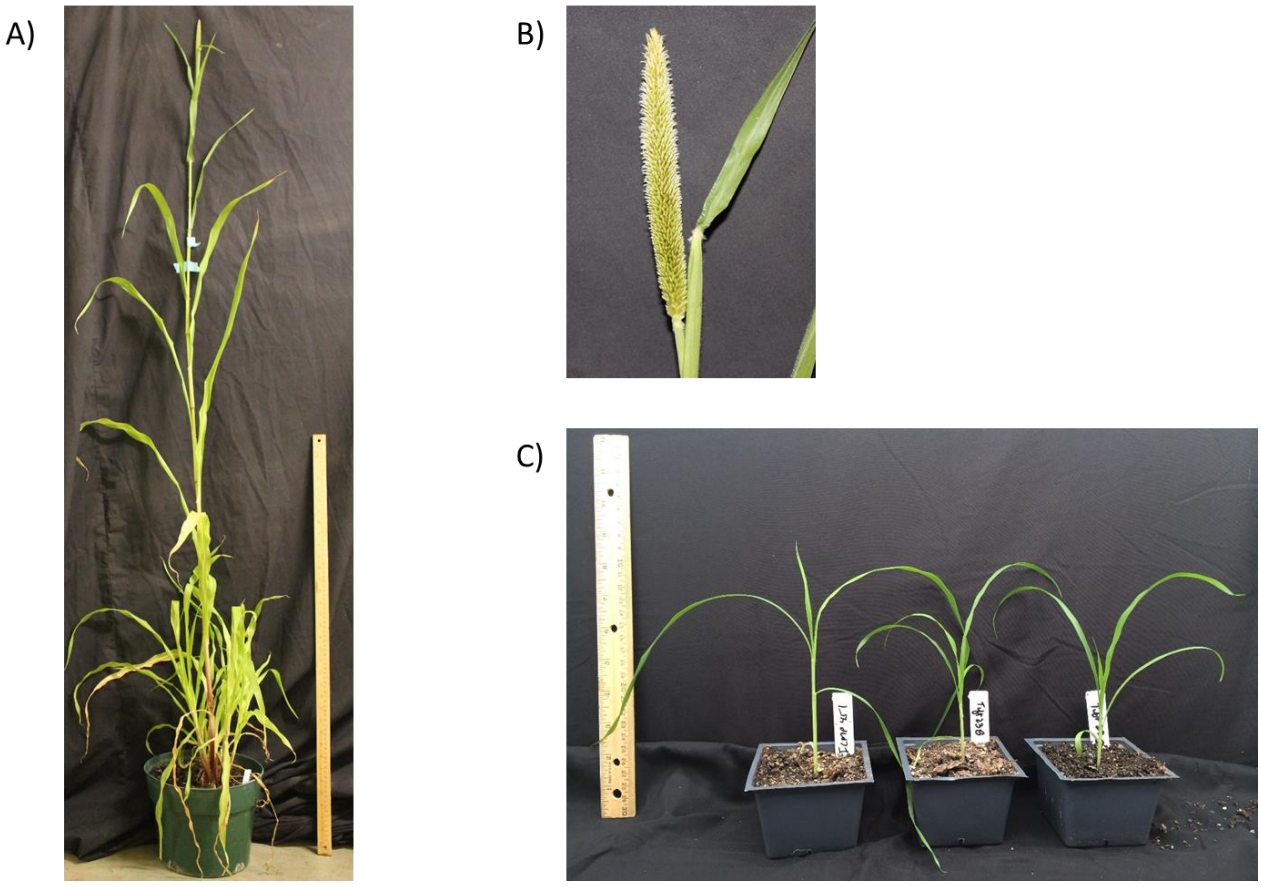


Fig 3.1 Pearl millet stages used for expression analyses. A) Tift 23B (*D2D2*) at 50% stigma emergence. 1 m ruler used to scale B) Panicle of Tift 23B (*D2D2*) at the 50% stigma emergence

and C) Representative images of the inbreds 1: ICMP 451 (*D2D2*), 2: Tift 23B (*D2D2*) and 3: Tift 23DB (*d2d2*) at the 5-leaf stage

Generation of the constructs for transformation

Three constructs driven by a 35S cauliflower mosaic virus (CaMV) promoter were tested for their ability to complement an Arabidopsis *abcb1* mutant: 1) The *D2* construct which carries the coding region of *Ca_ABCB1* (functional *ABCB1* allele); 2) The *d2* construct which carries an *ABCB1* allele (*Ca_abcb1*) in which the long terminal repeat (LTR) of a Juriah element has been engineered into the coding region (non-functional *ABCB1* allele); and 3) The coding region of the Arabidopsis *ABCB1* gene (*At_ABCB1*) as a control.

To generate the *D2* construct, primers were designed 162 bp upstream of the start codon in the 5'UTR (primer D2TF7) and 108 bp downstream of the stop codon in the 3'UTR region (primer D2TR7) of the pearl millet *ABCB1* gene. Primers D2TF7 and D2TR7 (Table 3.1) were designed to carry a *Hin*dIII and *Bam*HI restriction site, respectively. PCR amplification was conducted using the high-fidelity DNA polymerase Q5 (NEB) in a 25 µl reacting volume consisting of 1X reaction buffer, 1X enhancer buffer, 0.2 mM of each dNTP, 0.5 µM forward and reverse primers, 0.5U of Q5 polymerase (NEB), and 1 µl (50 ng/µl total RNA) of ICMP 451 cDNA. PCR reactions were assembled on ice, and then placed in a PCR block pre-heated to 98 °C. PCR conditions were as follow: an initial denaturation at 98 °C for 30 sec; 30 cycles of 98 °C for 10 sec and 72° C for 3 min; a final extension at 72 °C for 8 min. The PCR product was purified using the Qiagen PCR purification kit, double digested with *Hin*dIII and *Bam*HI, and cloned in the engineered T-DNA pCambia vector developed in our lab, pRajC, which confers kanamycin resistance in *E. coli* and hygromycin resistance in plants. The multiple cloning site (MCS) of the pCambia 1300 vector was replaced with the CaMV 35S promoter: multiple cloning site: nopaline synthase (NOS

terminator) from PCCN3_S_OX. The pRajC construct and inserts were verified by Sanger sequencing. Approximately 100 ng of the plasmid was then transformed into *Agrobacterium tumefaciens* strain GV3101.

To develop the *d2* construct, overlap PCR was conducted as follows: The D2TF7 primer was used in conjunction with the D2TR16 primer on DNA of Tift 23DB BAC clone 156A12 to generate fragment 1 which captured 162 bp of the 5'UTR, the 5' region of exon 1 in pearl millet *ABCB1* gene and part of the 5' LTR of the Juriah TE. Similarly, primer sets D2TF15 and D2TR10 were used on Tift 23DB BAC clone 156A12 to generate fragment 2 which captured the 3' LTR of the Juriah TE and most of the 3' region of *ABCB1* exon1. Primer set D2TF9 and D2TR7 was used on the *D2* construct to capture the remaining coding region of *ABCB1* and generate fragment 3. The three PCR fragments were cleaned using the Qiagen PCR purification kit (Qiagen, Valencia, CA). Because fragments 1 and 2, and 2 and 3 have partial overlaps, a PCR with primer sets D2TF7 and D2TR7 using the three fragments as template was used to construct a *d2* coding region that was essentially identical to the *D2* coding except for the presence of a Juriah solo LTR in exon 1. The primer sets and corresponding annealing temperatures are given in table 3.1. All PCR amplifications were conducted using the Q5 high-fidelity polymerase.

Primers AtABCB1_F1 and AtABCB1_R1 were used to amplify and isolate the full length coding region of *ABCB1* from Arabidopsis *Col-o* cDNA (Table 3.1). All PCR products were cloned into pRajC and transformed into *Agrobacterium* using the procedure described for the *D2* construct above and verified by Sanger sequencing.

Table 3.1: List of primers and their corresponding annealing temperatures used

Marker		Sequence	Annealing temp. (^{0}C)
Ca_Sb07g023730	F10	GCAGGTTCTCCTTGATGCTC	59
	R10	CTCGGAGGCACCTACTTCAC	
Ca_Sb07g023730	F1	TACGCCTTCTACTTCCTCGTC	61
	R5	AGCAGCAGAAGACGGTGAAGTAG	
RB2	F1[1]	ACTTGCCCCACTACAAGCAC	61
Ca_ABCB1	F20	CATGACCTCAAGAGCCTGAA	60
	R20	GAGCTTGATGATGAAGGAGTG	
Ca_GAPDH	F1	CTGTCGGTAAGGTTCTTCCTGAAT	60
	R1	CTAACAGTGAGGTCAACAACTGACAC	
Ca_Act	F1	AGATCATGTTTGAGACCTTTGAATG	60
	R1	ATCACCAGAGTCCAGCACAATAC	
D2T	F7	CCGTTGAAGCTTCAGACGCCATTCCAAATCCCCATCTT	72
	R7	CTGGTAAGGATCCGCTGCTGGTTGCTTCTTT	
	R16[2]	GCGTTCCTTCCCAATGAGCTCTGC	
	F9[3]	CCAAGATCTTCCGCATCATCGACCAC	
D2T	F15	CGTACTTGGGCATTTAACGCCCTGAC	72
	R10	CGTAGAACCTCTCGATGAGCGACACGA	
AtABCB1	F1	TCATAACACCAACAACAACTCACGAAGC	61
	R1	GTACTAAGCATCATCTTCCTTAACCC	

[1] Primer used in combination with Ca_Sb07g023730 R5

[2] Primer used in combination with D2TF7

[3] Primer used in combination with D2TR7

Transformation studies in the Arabidopsis *abcb1abcb19* mutant

The *Agrobacterium tumefaciens* strain GV3101 containing the pRajC vector with different inserts was used to transform the Arabidopsis *abcb1* single mutant using the floral dip method (Clough and Bent 1998). T1 seeds were collected and germinated on a ½ strength MS agar plate with 25 µg/ml hygromycin. The plates were placed at 4 ^{0}C for 3 days and then transferred to a growth chamber with continuous light. After 13 days in the growth chamber, the surviving plants

were transplanted into small pots and grown under long days (16 hour day length). The plants were grown to maturity and height was measured from the base to top most inflorescence. The plants were genotyped by insert specific markers to confirm their transgenic identity. LS means Student t-test was used to compare the mean height between transgenic events and controls

RESULTS

The pearl millet *ABCB1* gene

Fosmid-19 containing the wild-type *ABCB1* gene was sequenced to a depth of 23X and assembled into one large contig (C1; 31,571 bp) and one smaller contig (C2; 6952 bp) (Genbank accession number XXX). The summary statistics of the sequencing reads and assembly are provided in Table 3.2. The complete *ABCB1* gene was comprised within the larger contig C1. Contigs C1 and C2 have homology to adjacent locations on *Setaria italica* scaffold 6 (C1= scaffold_6:35858687-35876908; C2= scaffold_6:35876088-35884870), which is syntenic to pearl millet linkage group 4.

Table 3.2: Read statistics of the sequenced BAC and fosmid clones

	BAC_156A12	**Fosmid_19**
Genotype	Tift 23DB	ICMP 451
Sequencing technology	Roche 454 GLS FX	Roche 454 GLS FX
Insert size	90 kb	35 kb
Average read length	302.9 bp	181.0 bp
Read coverage (X-fold)	67.93	23
Number of contigs in assembly	15	2
Largest contig	47,002 bp	31,571 bp
Smallest contig	364 bp	6952 bp

Sequence analysis of the *ABCB1* gene from ICMP 451, hereafter referred to as *Ca_ABCB1*, revealed that it contains three exons and two introns (Figure 3.2A). In contrast, the maize and

sorghum *ABCB1* orthologs, *Zm_ABCB1* (*Br2*) and *Sb_ABCB1* (*dw3*), respectively, consist of five exons and four introns (Figure 4.1). Despite the differences in gene structure, a neighbor joining tree of ABCB1 and ABCB19 proteins (ABCB19 is the closest homolog to ABCB1) shows that pearl millet ABCB1 falls in its expected phylogenetic position within the ABCB1 clade (Figure 3.3), indicating that the gene we isolated from ICMP 451 is indeed *ABCB1*. A BLASTP analysis of pearl millet ABCB1 amino acid sequence has 89% homology with sorghum ABCB1 and 88% with maize ABCB1 at the protein level with a query coverage of 99% in both species. In comparison, aBLASTN analysis of pearl millet coding region of *ABCB1* shows 91% homology to sorghum *ABCB1* with a query coverage of 94% and 90% homology to maize *ABCB1* with a query coverage of 95%. The fact that we can see greater homology in cDNA level than in protein level can possibly be attributed to the lower query coverage in case of cDNA.

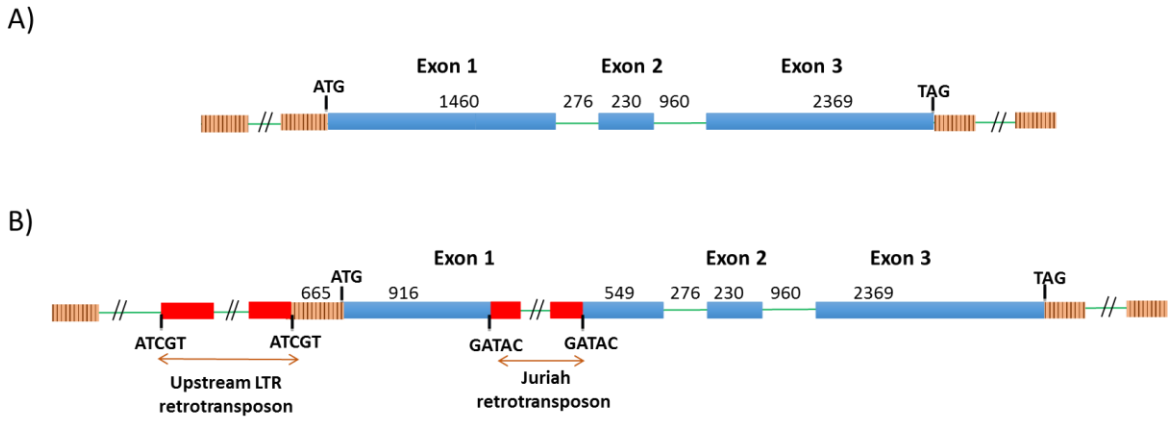


Figure 3.2 Structure of pearl millet *ABCB1* alleles. A) Structure of *Ca_ABCB1* isolated from the tall inbred line ICMP 451. B) Structure of *Ca_abcb1* isolated from the dwarf inbred line Tift 23DB. The size of the 5'UTR and 3'UTR regions is unknown (represented in yellow shaded boxes). Exons are represented by blue boxes and introns are represented as lines between them. The

transposable elements are represented by red boxes. The 5'UTR, 3'UTR and the promoter regions is indicated by shaded boxes as the length is unknown presently. The figure is not drawn to scale.

The *ABCB1* allele in the *d2* dwarf Tift 23DB

The Tift 23DB BAC (156A12) which contained the *d2* allele was sequenced to a depth of 67X and was assembled into 12 contigs using 'accurate' *de novo* genome assembly parameters in MIRA v. 3.2.0. The contigs varied in size from 457 bp to 48,634 bp and contig C2 (12,245 bp) carried the complete *ABCB1* gene. The Tift 23DB *ABCB1* allele, hereafter referred to as *Ca_abcb1*, differed from that of ICMP 451 (*Ca_ABCB1*) by the presence of a synonymous SNP in exon 1 and a solo long terminal repeat (LTR) element, also in exon 1 (Figure 3.2B). The solo LTR was 1408 bp long and flanked by the 5-bp host target site duplication GATAC (bases 912 – 916 in *Ca_ABCB1*). However, a second assembly of this region using the 'normal' assembly parameters in MIRA v. 3.2.0 had the *ABCB1* gene distributed across two contigs. The region 1188 bp – 2103 bp of contig C6 (total length of C6 is 2586 bp) corresponded to the 5' end of exon 1 of *ABCB1* (bases 1 – 916), and the region 14,571 bp – 18,961 of contig C2 (total length of C2 is 23,307 bp) corresponded to the remainder of the *ABCB1* gene (bases 910 – 5299). This assembly showed that the element inserted into *ABCB1* was not a solo LTR but a full length LTR retrotransposon with the region 2104 – 2586 bp on contig C6 corresponding to the 5' end of the 5' LTR, and regions 1 – 440 bp, 441 – 13,164 bp and 13,165 – 14,572 bp on C2 corresponding to the 3' region of the 5' LTR, the internal region and the 3' LTR, respectively. Although only a partial assembly was obtained for the 5' LTR (923 bp out of 1408 bp), the fact that (1) these regions showed 100% homology to the 3' LTR (a single base length variation in a poly (G) tract was likely a 454 sequencing error) and (2) the 5' and 3' LTRs had been collapsed into a single assembly using the 'accurate' assembly parameters in MIRA, suggest that the 5' and 3' LTRs are identical. The

LTR retrotransposon has 94% homology to an annotated 'Juriah' element in *Cenchrus americanus* BAC 311G2 (Genbank accession AF488414). The Juriah element is classified under the 'ofiek' family of the 'gypsy' superfamily with the reverse transcriptase most identical to the 'tat' clade. At the protein level, the Juriah element disrupts the first transmembrane domain in ABCB1 (Figure 3.3).

Figure 3.3 Structure of the ABCB1 protein of pearl millet. The 'Juriah' element is inserted 925 bp downstream from the start codon (codon 309) and interrupts the first transmembrane domain.

In addition to the variation in the coding region, *Ca_abcb1* also differed from *Ca_ABCB1* by the presence of one SNP 274 bp upstream of the ATG start codon, the length of a (CT) SSR 622 bp upstream of the start codon and the presence of an LTR transposable element (TE) 665 bp upstream of the start codon (Figure 3.2B). This TE was spread over contigs C6, C11, C3 and C5. Further analysis showed that contigs C6 and C11, and C11 and C3 overlapped so that the 3' LTR, the internal region and the 3' end of the 5' LTR were located on the joined contig C6rev-C11-C3rev, and the 5' end of the 5' LTR was located on contig C5. The host site duplication at the TE insertion site was ATCGT. This element has not been previously described. The element belongs to the 'owuut' family of the 'gypsy' superfamily. The reverse transcriptase domain of this element has the best hit to the 'del' clade. The closest match to this element is the 'LTR_AC186622.3_14040' element in *Zea mays* in the MIPS database.

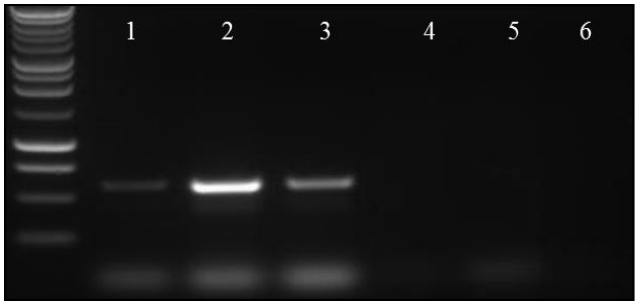
The 14 contigs resulting from the 'normal' assembly will be submitted to Genbank (accession number XXX). The contigs were ordered based on the structure of the Juriah and upstream transposable elements, the structure of the *ABCB1* gene and BLASTN analysis of the genes flanking *ABCB1* in Setaria and sorghum. In Setaria, the 3' flanking gene (Si014659m) had a BLAST hit on contig C2 while the 5' flanking gene (Si014865m.g) did not show homology to any of the 14 pearl millet contigs. In contrast, the 5' flanking gene in sorghum (Sobic.007G164000) had BLAST hits on contigs C1 and C5. Similar to the results obtained with the 3' flanking gene in Setaria, the 3' flanking gene in sorghum, Sobic.007G163700, had a BLAST hit on contig C2. Closer analysis of the *ABCB1* region in sorghum and Setaria showed that the 5' flanking gene of *ABCB1* in *Setaria* (Si014865m.g) was orthologous to sorghum Sobic.007G164400 (e-value = 0). The five annotated genes present in sorghum immediately upstream of *ABCB1* (Sobic.007G163900 (probably not a true gene); Sobic.007G164000; Sobic.007G164100; Sobic.007G164200 and Sobic.007G164300) were absent from the *ABCB1* region in *Setaria*. Hence, although *Setaria* is phylogenetically closer to pearl millet than sorghum, the gene-containing contigs were oriented and ordered (C1rev-C5-C3-C11rev-C6-C2rev) to match the gene order and orientation in sorghum. The five Tift 23DB BAC contigs constitute a total of 90,231 bp. .

Presence of the 'Juriah' element in dwarf and tall pearl millet inbreds

Three tall inbreds (ICMP 451, Tift red and P-1449-2) and 3 *d2* dwarf inbreds (Tift 23DB, 81B and pT 732B were tested for the presence of the Juriah TE using both a primer set that flanked the element and a primer set that spanned one of the gene-repeat boundaries. Inbred pT 732 was classified as a spontaneous dwarf mutant identified in the field in India (CT Hash, personal communication). In addition, the published independent *d2* mutants IP 8227, IP 8208, IP 8288, IP

8157, IP 8112 , IP 8008, IP 8058 and IP 10399 (Rao et al. 1986), the tall inbred Tift 23B and the variety 'Starr' were tested for the presence of the Juriah TE using only the primer set that spanned one of the gene-repeat boundaries. The Juriah element was absent from the four tall inbreds tested and present in all *d2* dwarf inbred lines tested. The PCR results of a subset of tall and *d2* dwarfs is shown in Figures 3.4A and 3.4B.

A)



B)

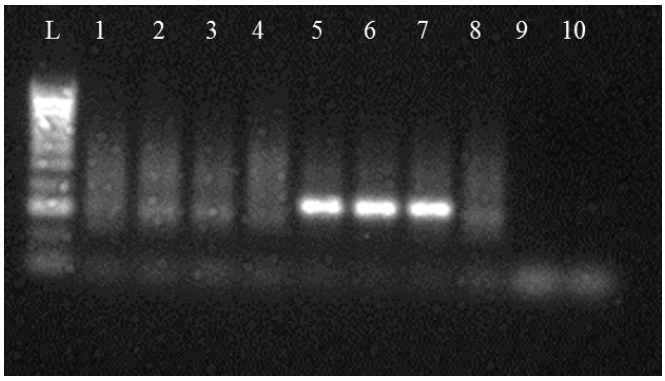


Figure 3.4 Results of PCR amplification of diverse pearl millet genotypes using (A) primers which flank the transposable element. L: 1 kbp ladder; 1) ICMP 451 (*D2D2*), 2) Tift red (*D2D2*) 3) P-1449-2 (*D2D2*), 4) Tift 23DB (*d2d2*), 5) 81B (*d2d2*) 6) Water control); and (B) a transposon specific forward primer in conjuction with a gene-specific reverse primer. L: 100 bp ladder; 1)

ICMP 451 (*D2D2*), 2) Tift 23B (*D2D2*), 3) Tift red (*D2D2*), 4) P-1449-2 (*D2D2*); 5) Tift 23DB (*d2d2*); 6) 81B (*d2d2*); 7) pT732B (*d2d2*); 8) 'Starr' Millet; 9 & 10) Water controls

Expression analysis of *Ca_ABCB1* and *Ca_abcb1*

The range of primer efficiencies in the different replicates is as follows: Ca_ABCB1F20/R20 (1.835 - 1.88); Actin (1.836 - 1.85); and GAPDH (1.828 – 1.884). No statistically significant differences in height were observed between the tall and *d2* dwarf near-isogenic lines Tift 23B and Tift 23DB at the 5-leaf stage (p value = 0.2665) but their height was significantly different at the time of 50% stigma emergence. The expression of *ABCB1* was tested at the 5-leaf stage in the *d2* dwarfs Tift 23DB, pT 732B and 81B, and in the *D2* tall accessions Tift 23B, P-1449-2 and ICMP 451, and at 50% stigma emergence in Tift 23DB (*d2d2*), Tift 23B (*D2D2*) and ICMP 451 (*D2D2*). At both growth stages, expression was significantly higher in the tall compared to the dwarf inbred lines (Figures 3.5 A-B and 3.6 A-D). The expression fold change between the isogenic inbred lines Tift 23B (*D2D2*), Tift 23DB (*d2d2*) and the tall mapping parent ICMP 451 (*D2D2*) in different organs at different developmental stages is recorded in Table 3.3. Differences in *ABCB1* expression level in different tissues within tall and dwarfs inbreds are listed in Table 3.4.

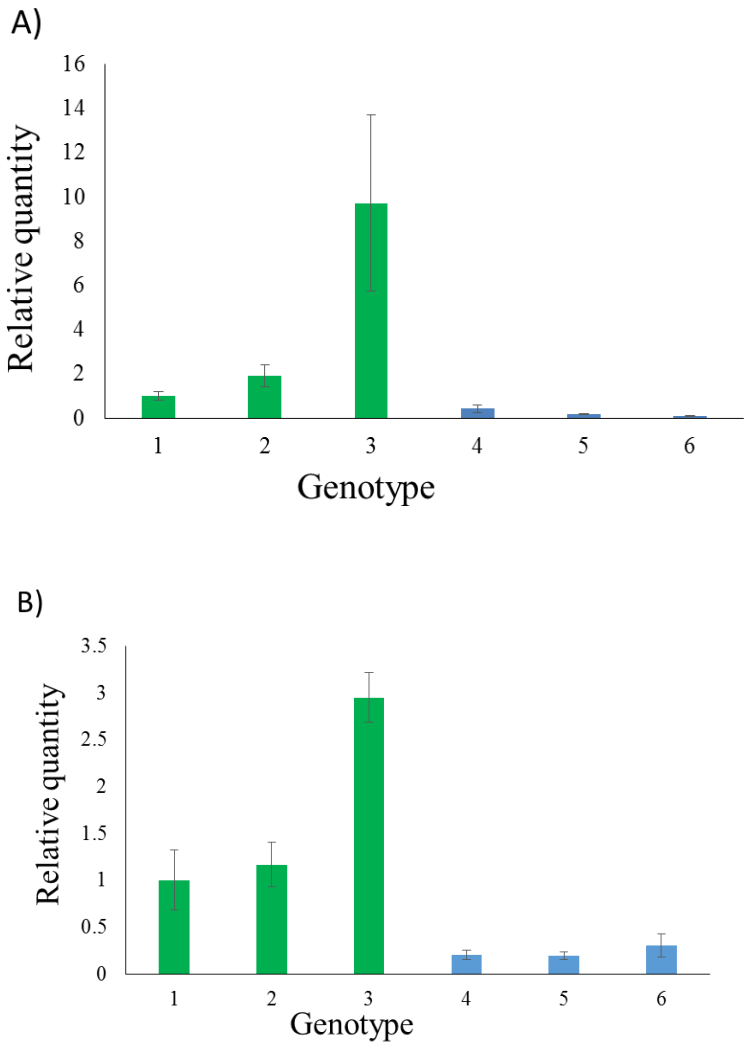


Figure 3.5 Relative expression of the *ABCB1* gene in (A) leaf tissue and (B) stem tissue at 5 leaf stage. 1: Tift 23B (*D2D2*); 2: P-1449-2 (*D2D2*); 3: ICMP 451 (*D2D2*); 4: Tift 23DB (*d2d2*); 5: 81B (*d2d2*) and 6: pT 732B (*d2d2*). The expression was normalized using pearl millet Actin-2 and GAPDH (n=3) expression levels.

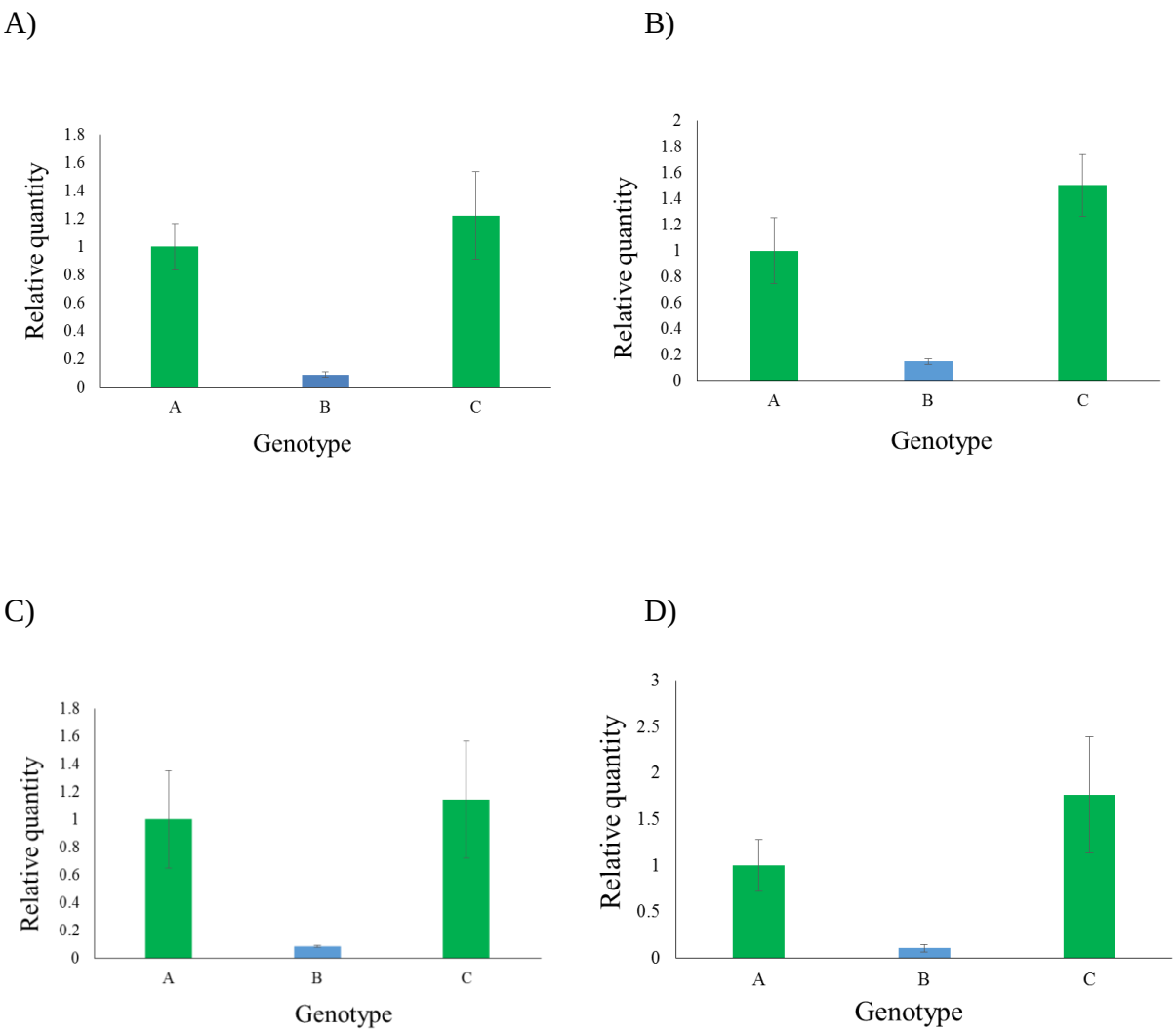


Figure 3.6 Expression of *ABCB1* in different organs in the pearl millet tall and *d2* dwarf at 50 percent stigma emergence. A) Panicle; B) Node; C) Internode; and D) Root. Genotypes: A) Tift 23B (*D2D2*); B) Tift 23DB (*d2d2*); and C) ICMP 451 (*D2D2*).

Table 3.3 Expression level comparison between the tall and the *d2* dwarf genotypes in various organs

Comparison	**Stage**	**Organ type**	**Fold difference**	**p-value (t-test)[1]**
Tift 23DB (*d2d2*) vs. Tift 23B (*D2D2*)	5-leaf	Leaf	2.28	0.2478
Tift 23DB (*d2d2*) vs. Tift 23B (*D2D2*)	5-leaf	Stem	4.48	0.0093**
Tift 23DB (*d2d2*) vs. Tift 23B (*D2D2*)	50 stigma emergence	Panicle	11.29	0.0009***
Tift 23DB (*d2d2*) vs. Tift 23B (*D2D2*)	50 stigma emergence	first node	6.85	0.0038**
Tift 23DB (*d2d2*) vs. Tift 23B (*D2D2*)	50 stigma emergence	first internode	11.8	0.0093**
Tift 23DB (*d2d2*) vs. Tift 23B (*D2D2*)	50 stigma emergence	root	9.54	0.0163*
ICMP 451 (*D2D2*) vs. Tift 23B (*D2D2*)	5-leaf	Leaf	9.71	0.0441*
ICMP 451 (*D2D2*) vs. Tift 23B (*D2D2*)	5-leaf	Stem	2.94	0.0787
ICMP 451 (*D2D2*) vs. Tift 23B (*D2D2*)	50 stigma emergence	Panicle	1.22	0.6938
ICMP 451 (*D2D2*) vs. Tift 23B (*D2D2*)	50 stigma emergence	first node	1.5	0.2409
ICMP 451 (*D2D2*) vs. Tift 23B (*D2D2*)	50 stigma emergence	first internode	1.1	0.8606
ICMP 451 (*D2D2*) vs. Tift 23B (*D2D2*)	50 stigma emergence	root	1.76	0.4021
ICMP 451 (*D2D2*) vs. Tift 23DB (*d2d2*)	5-leaf	Leaf	22.22	0.0063**
ICMP 451 (*D2D2*) vs. Tift 23DB (*d2d2*)	5-leaf	Stem	14.4	0.0004***
ICMP 451 (*D2D2*) vs. Tift 23DB (*d2d2*)	50 stigma emergence	Panicle	13.81	0.0022**
ICMP 451 (*D2D2*) vs. Tift 23DB (*d2d2*)	50 stigma emergence	first node	10.31	0.0005***
ICMP 451 (*D2D2*) vs. Tift 23DB (*d2d2*)	50 stigma emergence	first internode	13.47	0.007**
ICMP 451 (*D2D2*) vs. Tift 23DB (*d2d2*)	50 stigma emergence	root	16.72	0.014*

[1]The expression difference is statistically significant at $\alpha = 0.05$ (*), 0.01 (**) and 0.001 (***)

Table 3.4: Expression level comparison between the different organs in tall and *d2* dwarf plants

Genotype	**Comparison**	**Stage**	**p-value (t-test)[1]**
Tift 23B	leaf vs stem	5-leaf	0.048*
Tift 23B	Panicle vs node	50% stigma emergence	0.085
Tift 23B	Panicle vs internode	50% stigma emergence	0.46
Tift 23B	Panicle vs root	50% stigma emergence	0.93
Tift 23B	node vs internode	50% stigma emergence	0.57
Tift 23B	node vs root	50% stigma emergence	0.11
Tift 23B	internode vs root	50% stigma emergence	0.46
Tift 23DB	leaf vs stem	5-leaf	0.035*
Tift 23DB	Panicle vs node	50% stigma emergence	0.006**
Tift 23DB	Panicle vs internode	50% stigma emergence	0.0648
Tift 23DB	Panicle vs root	50% stigma emergence	0.98
Tift 23DB	node vs internode	50% stigma emergence	0.017*
Tift 23DB	node vs root	50% stigma emergence	0.076
Tift 23DB	internode vs root	50% stigma emergence	0.35
ICMP 451	leaf vs stem	5-leaf	0.168
ICMP 451	Panicle vs node	50% stigma emergence	0.0397*
ICMP 451	Panicle vs internode	50% stigma emergence	0.56
ICMP 451	Panicle vs root	50% stigma emergence	0.59
ICMP 451	node vs internode	50% stigma emergence	0.26
ICMP 451	node vs root	50% stigma emergence	0.18
ICMP 451	internode vs root	50% stigma emergence	0.89

[1] The expression difference is statistically significant at $\alpha = 0.05$ (*), 0.01 (**) and 0.001 (***)

Heterologous transformation studies in Arabidopsis

Eight or more independent transformation events were analyzed per transgenic construct. Plant height was measured at maturity, when the siliques had senesced. No significant height differences were observed between the any of the transgenic plants (Figure 3.7). Insect damage was noted in some plants that performed poorly.

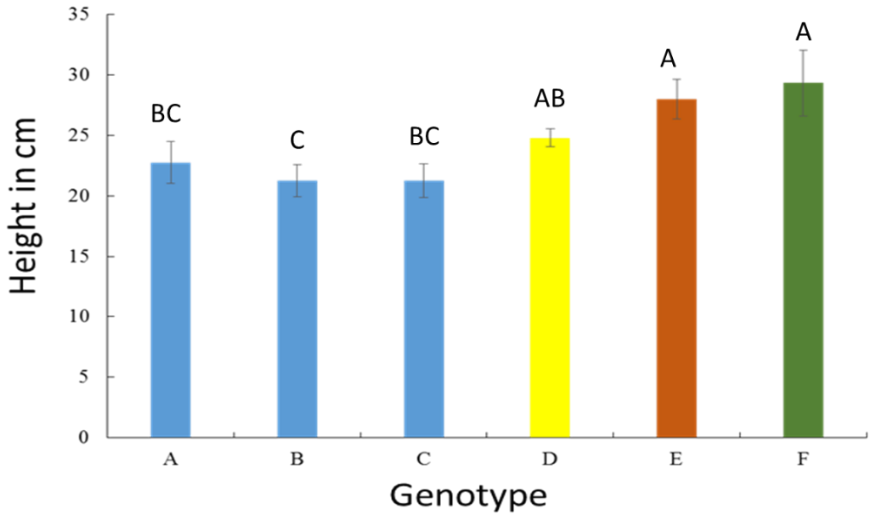


Figure 3.7 Height measurements (in cm) of the T1 transgenic plants in an Arabidopsis *abcb1-2* background. A) *D2* construct; B) *d2* construct; C) *At* native construct; D) *abcb1-2* single mutant; E) *abcb19* single mutant; and F) Col-o plants. n ≥ 8 plants. Means with the same letter are not statistically significant at the 5% level.

DISCUSSION

ABCB1 underlies the *d2* phenotype

ABCB1 was suggested as a candidate for *d2* based on comparative map information (Parvathaneni et al. 2013) and this was the impetus for isolation of the full length *ABCB1* alleles from tall (*D2*) and dwarf *(d2*) inbred pearl millet lines. To ensure that we indeed isolated *ABCB1*, a neighbor-joining tree was constructed of the *in-silico* translated *D2* sequence obtained, and the protein sequences of ABCB1 and its closest homolog ABCB19 from *Sorghum bicolor*, *Zea mays*, *Oryza sativa*, *Setaria italica*, and *Arabidopsis*. The phylogenetic placement of the *D2* protein showed that the isolated sequence was indeed *Ca_ABCB1* (Figure 3.8).

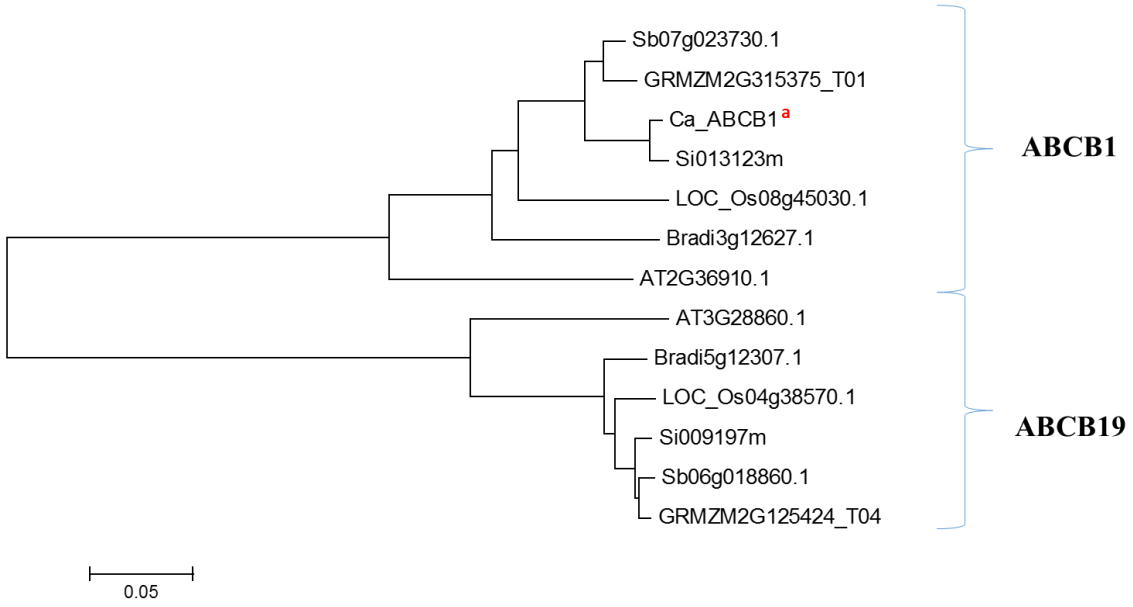


Figure 3.8 Neighbor joining tree of the ABCB1 and ABCB19 protein orthologs in grasses.

[a] The *in-silico* translated pearl millet ABCB1 homolog.

The *Ca_ABCB1* and *Ca_abcb1* alleles differed only by the presence of a synonymous SNP and of a ~15 kb LTR retrotransposon in exon 1. *ABCB1* transcript levels were significantly reduced in the *d2* dwarf compared to tall lines. Furthermore, since the LTR-retrotransposon insertion disrupts the first transmembrane domain of ABCB1, any *Ca_abcb1* protein that is formed in the *d2* dwarf is almost certainly inactive. Confirmation that the *Ca_abcb1* allele is inactive was attempted using heterologous transformation in an Arabidopsis *abcb1* background. Heterologous transformation experiments in an Arabidopsis *abcb1abcb19* double mutant are ongoing.

Heterologous transformations of *Ca_ABCB1* in Arabidopsis

Because no independent *d2* mutants could be identified, and virus-induced gene silencing has not been achieved in pearl millet, we aimed to demonstrate that the presence of the Juriah element indeed inactivated the pearl millet *ABCB1* gene by transforming *Ca_ABCB1* in an *abcb1*

background. However, transformation efficiency in cereals is still largely genotype-dependent (Ji et al. 2013) and we therefore decided to perform transformation studies in *Arabidopsis thaliana*. Several *abcb1* mutants are available in Arabidopsis. T-DNA insertions in the 8th exon of *ABCB1* (e.g. *atpgp1-100* and *atpgp1-101*) have no obvious phenotype, while T-DNA insertions in the 3rd exon (e.g. *atpgp1-2*) result in a small reduction in height under long days (Ye et al. 2013). It has been hypothesized that *atpgp1-100* and *atpgp1-101* mutants can still produce functional half-transporters while mutations in the third exon produce non-functional proteins (Ye et al. 2013). In contrast, double mutants of *abcb1* and its close relative *abcb19* display an extreme dwarf phenotype (Noh et al. 2001). Because a verified *abcb1abcb19* double mutant was not available at the time, we transformed the single mutant *atpgp1-2* with the coding region of pearl millet *ABCB1*, the coding region of a pearl millet *abcb1* allele that carried the Juriah LTR and a control Arabidopsis *ABCB1* construct (coding region of Arabidopsis *ABCB1*). In the T1 generation, no significant differences in height were observed between transformants carrying a functional *ABCB1* alleles and those that carried a non-functional *abcb1* allele on one hand, and between the transformants and the non-transformed *atpgp1-2* mutant on the other hand. Because selection on hygromycin medium may have affected growth in the T1 generation (Wolfgang Lukowitz; personal communication), transformants will be selfed, and progeny will be genotyped for the presence of the transgene. A comparison of the height of plants with and without the transgene will be conducted. However, because the phenotype of the *atpgp1-2* mutant is subtle, it is possible that the results may be inconclusive. Recently, we have successfully generated an *abcb1abcb19* double mutant which has an extreme dwarf phenotype and are in the process of repeating the transformation experiments conducted in the single mutant in the double mutant.

Origin of the *d2* dwarf

The exact source of the *d2* dwarf is unknown. In 1966, Burton (1966) reported five dwarf pearl millet inbreds, two of which carried a single recessive gene for reduced height and received the designations *d1* and *d2*. We had earlier hypothesized that the *d2* mutation had been present in heterozygous condition in one of five pearl millet accessions that Burton had acquired from the Vavilov Institute of Plant Industry in the mid-1930s (Parvathaneni et al. 2013). A leafy dwarf was identified among progeny from a plant obtained through mass selection from those introductions and was subsequently used in crosses with an adapted pearl millet line to generate the synthetic cultivar 'Starr'. The morphological characteristics of Starr fit the description of a *d2* dwarf (Burton and Devane 1951). Seeds of Starr millet (NSL 4716) were obtained from the USDA National Plant Germplasm System (NPGS). However, the Starr millet obtained was tall and thus unlikely represents the Starr millet described by Burton and Devane (1951). As expected for a tall line, both the transposable elements in the upstream and coding region of *ABCB1* were absent from Starr.

Considering that the LTRs of the Juriah element appear to be identical, the Juriah retrotransposon likely inserted into the *ABCB1* gene in the past 41,670 years. This time frame is based on the 923 bp of assembled LTR sequence, a k-value of 1/923/site (k=substitution rate/site), an r-value of 1.3×10^{-8} substitutions/site/year (Ma and Bennetzen 2004) and the formula t (divergence time) = k/2r (Jing et al. 2005). Once the LTR retrotransposon insertion inactivated the *ABCB1* gene, presumably *Ca_abcb1* was no longer under any selective constraints and free to accumulate mutations. However, only a single synonymous SNP differentiates the open reading frame of *Ca_abcb1* and *Ca_ABCB1*, suggesting that gene inactivation occurred in the past ~5076 years.

Although we do not know the precise copy number of Juriah in the pearl millet genome, the fact that it was found in both pearl millet BACs that are available in GenBank (AF488414 and KF704368) suggests that this is a high copy number element. Furthermore, BLASTN analysis of Juriah against the *Cenchrus americanus* (taxid: 4543) genomic survey sequences (GSS) deposited in GenBank shows that 6.25% of the total pearl millet GSS sequences show homology to the transposable element. According to Estep (Estep et al. 2013), Juriah is the third most abundant LTR retrotransposon in *Cenchrus americanus* covering a total of 86.8 Mb. A BLASTN analysis using the upstream transposable element in the *ABCB1* upstream region as query against the *Cenchrus* GSS sequences revealed that 3.3% of the total pearl millet GSS sequence have homology to this element. Typically, high copy LTR retrotransposons insert into one another and are inactivated through methylation (SanMiguel et al. 1998). A number of genes that show loss of function through the insertion of LTR retrotransposons have been reported. Independent insertions of *TS1-7* (gypsy-like) and *TSI-9* (copia-like) elements in exon 3 and exon 9 of the *GBSS1* gene in *Setaria* caused the conversion of non-waxy grain to waxy grain (Kawase et al. 2005). The *TSI-7* TE has a higher copy number in domesticated foxtail (*Setaria italica*) than its wild progenitor, green foxtail (*Setaria viridis*) and has been classified as "recently active" (Hirano et al. 2011). The terminal repeat retrotransposon in miniature (TRIM) insertion in the coding region of *HaCYC2c* in sunflower led to tubular ray florets (*HaCYC2c-tub* allele) (Chapman et al. 2012). The insertion of the *Gret1* LTR retrotransposon in the promoter of the grape *VvmybA2* gene caused loss of red pigmentation (Kobayashi et al. 2004). This insertion completely knocked out the *VvmybA2* gene expression (Pereira et al. 2005). In most cases, the inserting elements have low to moderate copy numbers. An exception is the potentially active high copy transposon *Rider* which inserts into or near genes in tomato. Transposition of the *Rider* element resulted in duplication and translocation

of a 24.7 kb region from chromosome 10 to chromosome 7 which included the *SUN* gene leading to an oval-shaped fruit. These genomic translocations of the *Rider* element have been shown to be important in tomato evolution and domestication (Jiang et al. 2009). Once the pearl millet genome sequence becomes available, it will be interesting to study the insertion preference of the Juriah retrotransposon.

The fact that two high copy LTR retrotransposons were inserted into *ABCB1,* one in the coding region and one in the upstream region, in relatively recent times led us to speculate that the *d2* mutation arose following transposon activation, possibly caused by treatment of pearl millet seed with Ethyl methanesulfonate (EMS) and thermal neutrons. Mutagenesis as a way to create genetic variation in crop plants was actively pursued during the period 1950 to 1970 in many breeding programs, including the pearl millet breeding program at Tifton, GA (Burton and Powell 1966; Burton and Hanna 1976; Hanna et al. 1978). Even though no published record is available that links *d2* to a mutation experiment, the time period of the discovery of the *d2* dwarf together with the characteristics of the *d2* mutation make this a plausible hypothesis. Stress induced transposon activation has been documented in a number of species. Ionizing gamma-radiation activates Ty1 copia retrotransposable elements in *Saccharomyces cerevisiae* and the DNA transposon mPing in rice (Nakazaki et al. 2003; Sacerdot et al. 2005). Activation by tissue culture is documented for the transposable elements (TE) Tto1-Tto3 in tobacco (Hirochika 1993) and nDaiZ in rice (Huang et al. 2009). Even pathogen attacks have been shown to lead to retrotransposon amplification as is the case for the Tnt1 retrotransposon in tobacco (Melayah et al. 2001). Barbara McClintock noted that cell stress caused by chromosome breakage-fusion-bridge mitosis cycles activates transposable elements (McClintock 1984). Treatment of cells with thermal neutrons results in uniform chromosome fragmentation (Deschner and Sparrow 1955; Yagyu and

Morris 1957; Rakhmatullina and Sanamyan 2007) and may provide the necessary stress for TE activation. Interestingly, a BLASTN analysis in NCBI of the Juriah element to a pearl millet EST database (*Cenchrus*, taxid: 4583) identifies two clones (CD726520 and CD726379) with >83% homology from a cDNA library from pearl millet seedling drought stress experiment which suggests that this element might be activated as response to diverse stresses.

The quest for independent *d2* mutations

The identification of independent *d2* alleles was one way to confirm that *ABCB1* underlies the *d2* phenotype. We consulted with pearl millet breeders and scoured the literature for reports on spontaneous and thus, presumably, independent *d2* dwarfs (India: pT 732B (CT Hash; personal communication), IP 8008, IP 8210; Niger: IP 8058; Senegal: IP8112; Cameroun: IP 8157:; Uganda: IP 8208; ICRISAT: IP 8227, IP 8228)(Rao et al. 1986). Several *d2* mutants had been recovered from mutagenesis experiments with gamma radiation and salicylic acid (SA) (Sukhadev et al.,1986), however, seed for these *d2* mutants was no longer available (MV Subbarao; personal communication). Analysis of the reported spontaneous *d2* dwarfs with the transposon-gene boundary primers demonstrated that the Juriah element was present in the coding region of all presumed independent pearl millet *d2* dwarf mutants tested. Except for pT 732B, which was discovered as a spontaneous mutant in the field (CT Hash; personal communication) and IP 8210, whose origin is not described, all other independent *d2* mutants were identified as 'segregants' from tall landrace populations of Indian and African origin (Rao et al. 1986). Because pearl millet is an outcrossing species and the field experiments were conducted in India where *d2* gene is widely used, most likely, the *d2* gene was introduced into these landraces via hybridization with Indian *d2* germplasm. The presence of the upstream TE was tested only in Tift 23DB, 81B and pT 732B, and was present in all three lines. 'Starr' millet did not the carry either of the transposable

elements. It is highly likely due to the close linkage of the two transposable elements that all other reported *d2* dwarfs will carry this transposable element as well.

The expression profile of *ABCB1* in pearl millet

Expression of *ABCB1* was observed in all the analyzed organs in pearl millet including leaves, nodes, internodes, panicle, peduncle and root. At the five leaf stage, levels of expression were higher in the stem compared to the leaf in isogenic inbreds Tift 23B (*D2D2*) and Tift 23DB (*d2d2*). Due to a higher expression of *ABCB1* in leaf in the tall mapping parent, ICMP 451 (*D2D2*), no statistical difference in expression between leaf and stem was observed in this genotype. Some significant differences in expression levels were observed between organs at 50% stigma emergence (Table 3.4). In the tall inbred ICMP 451 (*D2D2*) and dwarf inbred Tift 23DB (*d2d2*), the nodal tissue showed significantly higher expression than the panicle (p value = 0.039). In Tift 23DB, the nodal tissue also shows a higher expression than internodal tissue (p value = 0.017). Overall, average expression levels were higher in the nodes compared to the other organs but these differences were only statistically significant in the *d2* dwarf Tift 23DB due to the relatively high variance between replicates. The variance might be caused by environmental effects as sampling was done on different days and different times of the day based on the developmental stage of the plant.

The expression profile of *ABCB1* is well documented in maize and Arabidopsis. In Arabidopsis, the gene is expressed at high levels in a range of tissues including leaves, roots and flower although the highest expression was observed in nodal tissue (Titapiwatanakun and Murphy 2009). In maize, Multani and colleagues (2003) performed northern blot analyses in three organs (internode, root and leaf) and reported expression only in the internodes. In contrast, Knoller and

colleagues studied the expression by semi-quantitative PCR in nodes and internodes and reported expression only in the nodes of maize. Forestan and colleagues (2012) semi-quantitatively tested the expression in a wide range of vegetative and reproductive organs including seedlings, leaf, root, male and female inflorescence, different nodes and stages of kernel development (Forestan et al. 2012). The gene was expressed in all the tested tissues but expressed at lower levels in kernels and elongating roots. Similar to maize and Arabidopsis, pearl millet also exhibits expression in inflorescence, leaf, root and nodes with a relatively higher expression in nodes. Although the vascular architecture and development of shoot apical meristems differ between monocots and dicots, the role of *ABCB1* has been shown to be conserved between maize and Arabidopsis (Knoller et al. 2010). The similar gene expression patterns in pearl millet suggest functional conservation of *ABCB1* in pearl millet as well.

ABCB1 transcript levels were significantly lower in *d2* dwarfs compared to tall inbreds (Table 3.3, Figures 3.6 A-B, 3.7 A-D). Since *ABCB1* expression was analyzed with a primer pair (Ca_ABCB1F20/R20) designed to amplify a region downstream of the transposable element insertion, our data show that the presence of the LTR retrotransposon does not inhibit transcription. Although transcription through the 15 kb Juriah element might reduce the rate at which transcripts accumulate, we consider it more likely that the presence of the retrotransposon in exon 1, in addition to disrupting the *ABCB1* coding region, reduces the stability of *ABCB1* transcripts. Destabilization of transcripts caused by the insertion of LINE retrotransposons in introns has been shown in humans (Chen et al. 2006). Alternatively, it is possible that the measured transcripts initiated from promoter sequences within the Juriah element. The plant promoter prediction program TSSP (Solovyev and Salamov 1997) predicted the presence of one TATA box and one enhancer at the 3’ end of the LTR of the Juriah transposon with a linear discriminant function

(LDF) weight of 0.11 and 0.09 respectively (threshold LDF for identifying TATA boxes = 0.02; Enhancer = 0.04). However, if transcription of *ABCB1* was initiated from the Juriah 3' LTR, we would expect transcript levels to be similar across different organs in dwarf lines. The fact that transcript levels in different organs in dwarf lines are correlated with those observed in the tall lines (r = 0.99) is congruent with the hypothesis that reduced transcript levels are caused by transcript instability.

In conclusion, the tall and the *d2* dwarf plants differ by the presence of two transposable elements, one in the coding region and the second at 665 bp upstream of the start codon of the *ABCB1* gene. Only a single source of *d2* exists as no independent mutations in *ABCB1* have been identified. The *ABCB1* gene is expressed at higher levels in all analyzed tissues in the tall genotypes in comparison to *d2* dwarfs. Heterologous transformation of pearl millet tall allele (*ABCB1*) and dwarf allele (*abcb1*) in Arabidopsis is currently underway to confirm the candidate gene identity.

REFERENCES

Allouis S, Qi X, Lindup S, Gale MD, Devos KM. 2001. Construction of a BAC library of pearl millet, Pennisetum glaucum. *Theoret Appl Genetics* **102**(8): 1200-1205.

Asare-Marfo D, Birol E, Roy D. 2010. Investigating Farmers' Choice of Pearl Millet Varieties in India to Inform Targeted Biofortification Interventions: Modalities of Multi-Stakeholder Data Collection. Environmental Economy and Policy Research Discussion Paper Series, No. 51.2010. Cambridge, UK: University of Cambridge, Department of Land Economy.

Bidinger FR, Raju DS. 1990. Effects of the D2 Dwarfing Gene in Pearl-Millet. *Theoret Appl Genetics* **79**(4): 521-524.

Blakeslee JJ, Peer WA, Murphy AS. 2005. Auxin transport. *Current opinion in plant biology* **8**(5): 494-500.

Burton GW, Devane EH. 1951. *In ascendency: Starr millet: synthetic cattail lasts longer, produces more beef per acre*.

Burton GW, Fortson JC. 1966. Inheritance and Utilization of 5 Dwarfs in Pearl Millet (Pennisetum Typhoides) Breeding. *Crop Sci* **6**(1): 69-&.

Burton GW, Hanna WW. 1976. Ethidium bromide induced cytoplasmic male sterility in pearl millet. *Crop Sci* **16**(5).

Burton GW, Powell J. 1966. Morphological and Cytological Response of Pear Miller, Pennisetum typhoides to Thermal Neutron and Ethyl Methane Sulfonate Seed Treatments1. *Crop Sci* **6**(2): 180-182.

Chapman MA, Tang S, Draeger D, Nambeesan S, Shaffer H, Barb JG, Knapp SJ, Burke JM. 2012. Genetic analysis of floral symmetry in Van Gogh's sunflowers reveals independent recruitment of CYCLOIDEA genes in the Asteraceae. *PLoS Genet* **8**(3): e1002628.

Chen J-M, Férec C, Cooper DN. 2006. LINE-1 Endonuclease-Dependent Retrotranspositional Events Causing Human Genetic Disease: Mutation Detection Bias and Multiple Mechanisms of Target Gene Disruption. *Journal of Biomedicine and Biotechnology* **2006**: 56182.

Chomczynski P, Sacchi N. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* **162**(1): 156-159.

Clough SJ, Bent AF. 1998. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *The Plant journal : for cell and molecular biology* **16**(6): 735-743.

Dash M, Malladi A. 2012. The AINTEGUMENTA genes, MdANT1 and MdANT2, are associated with the regulation of cell production during fruit growth in apple (Malus x domestica Borkh.). *BMC Plant Biology* **12**(1): 98.

Dean M, Rzhetsky A, Allikmets R. 2001. The human ATP-binding cassette (ABC) transporter superfamily. *Genome research* **11**(7): 1156-1166.

Deschner E, Sparrow AH. 1955. Chromosome Rejoining Capacity with Respect to Breakage Sensitivity to X-Rays and Thermal Neutrons. *Genetics* **40**(4): 460-475.

Estep MC, DeBarry JD, Bennetzen JL. 2013. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity* **110**(2): 194-204.

Forestan C, Farinati S, Varotto S. 2012. The Maize PIN Gene Family of Auxin Transporters. *Frontiers in plant science* **3**: 16.

Gulia SK, Wilson JP, Carter J, Singh BP. 2007. Progress in grain pearl millet research and market development. In *Issues in New Crops and New Uses B2 - Issues in New Crops and New Uses*. ASHS Press, Proceedings of the Sixth National Symposium Creating Markets

Hazell PBR. 2009. The Asian green revolution. IFPRI discussion paper 911. This paper has been prepared for the project on Millions Fed: Proven Successes in Agricultural Development (URL:www.ifpri.org/millionsfed)

HANNA WW, BURTON GW, POWELL JB. 1978. Genetics of mutagen induced non-lethal chlorophyll mutants in pearl millet. *Journal of Heredity* **69**(4): 273-274.

Hedden P. 2003. The genes of the Green Revolution. *Trends Genet* **19**(1): 5-9.

Hirano R, Naito K, Fukunaga K, Watanabe KN, Ohsawa R, Kawase M. 2011. Genetic structure of landraces in foxtail millet (Setaria italica (L.) P. Beauv.) revealed with transposon display and interpretation to crop evolution of foxtail millet. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada* **54**(6): 498-506.

Hirochika H. 1993. Activation of tobacco retrotransposons during tissue culture. *The EMBO journal* **12**(6): 2521-2528.

Huang J, Zhang K, Shen Y, Huang Z, Li M, Tang D, Gu M, Cheng Z. 2009. Identification of a high frequency transposon induced by tissue culture, nDaiZ, a member of the hAT family in rice. *Genomics* **93**(3): 274-281.

Ji Q, Xu X, Wang K. 2013. Genetic transformation of major cereal crops. *The International journal of developmental biology* **57**(6-8): 495-508.

Jia Q, Zhang J, Westcott S, Zhang XQ, Bellgard M, Lance R, Li C. 2009. GA-20 oxidase as a candidate for the semidwarf gene sdw1/denso in barley. *Funct Integr Genomics* **9**(2): 255-262.

Jiang N, Gao D, Xiao H, van der Knaap E. 2009. Genome organization of the tomato sun locus and characterization of the unusual retrotransposon Rider. *The Plant journal : for cell and molecular biology* **60**(1): 181-193.

Jing R, Knox MR, Lee JM, Vershinin AV, Ambrose M, Ellis THN, Flavell AJ. 2005. Insertional Polymorphism and Antiquity of PDR1 Retrotransposon Insertions in Pisum Species. *Genetics* **171**(2): 741-752.

Johnson JC, Lowrey RS, Monson WG, Burton GW. 1968. Influence of Dwarf Characteristic on Composition and Feeding Value of near-Isogenic Pearl Millets. *Journal of Dairy Science* **51**(9): 1423-&.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**(1-4): 462-467.

Kang J, Park J, Choi H, Burla B, Kretzschmar T, Lee Y, Martinoia E. 2011. Plant ABC Transporters. *The Arabidopsis book / American Society of Plant Biologists* **9**: e0153.

Kawase M, Fukunaga K, Kato K. 2005. Diverse origins of waxy foxtail millet crops in East and Southeast Asia mediated by multiple transposable element insertions. *Mol Genet Genomics* **274**(2): 131-140.

Knoller AS, Blakeslee JJ, Richards EL, Peer WA, Murphy AS. 2010. Brachytic2/ZmABCB1 functions in IAA export from intercalary meristems. *J Exp Bot* **61**(13): 3689-3696.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. *Science* **304**(5673): 982.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**(11): 1851-1858.

Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D, Aguilar-Rodriguez J, Vicente-Ripolles M, Fuster G, Bernet GP et al. 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic acids research* **39**(Database issue): D70-74.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**(34): 12404-12410.

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* **226**(4676): 792-801.

Melayah D, Bonnivard E, Chalhoub B, Audeon C, Grandbastien M-A. 2001. The mobility of the tobacco Tnt1 retrotransposon correlates with its transcriptional activation by fungal factors. *The Plant Journal* **28**(2): 159-168.

Milach SCK, Federizzi LC. 2001. Dwarfing genes in plant improvement. *Advances in Agronomy* **73**: 35-63.

Multani DS, Briggs SP, Chamberlin MA, Blakeslee JJ, Murphy AS, Johal GS. 2003. Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. *Science* **302**(5642): 81-84.

Murashige T, Skoog F. 1962. A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. *Physiologia Plantarum* **15**(3): 473-497.

Murray MG, Thompson WF. 1980. Rapid isolation of high molecular weight plant DNA. *Nucleic acids research* **8**(19): 4321-4325.

Nakazaki T, Okumoto Y, Horibata A, Yamahira S, Teraishi M, Nishida H, Inoue H, Tanisaka T. 2003. Mobilization of a transposon in the rice genome. *Nature* **421**(6919): 170-172.

Noh B, Murphy AS, Spalding EP. 2001. Multidrug Resistance-Like Genes of Arabidopsis Required for Auxin Transport and Auxin-Mediated Development. *The Plant Cell* **13**(11): 2441-2454.

Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M. 2013. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic acids research* **41**(Database issue): D1144-D1151.

Parvathaneni RK, Jakkula V, Padi FK, Faure S, Nagarajappa N, Pontaroli AC, Wu X, Bennetzen JL, Devos KM. 2013. Fine-mapping and identification of a candidate gene underlying the

d2 dwarfing phenotype in pearl millet, Cenchrus americanus (L.) Morrone. *G3* **3**(3): 563-572.

Pereira HS, Barao A, Delgado M, Morais-Cecilio L, Viegas W. 2005. Genomic analysis of Grapevine Retrotransposon 1 (Gret 1) in Vitis vinifera. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* **111**(5): 871-878.

Peterson DG, Tomkins JP, Frisch DA, Wing RA, and Paterson AH. 2000. Construction of plant bacterial artificial chromosome (BAC) libraries: An illustrated guide. J. Agric. Genomics 5, (www.ncgr.org/research/jag).

Rai KN, Rao AS. 1991. Effect of D2 Dwarfing Gene on Grain-Yield and Yield Components in Pearl-Millet near-Isogenic Lines. *Euphytica* **52**(1): 25-31.

Rakhmatullina EM, Sanamyan MF. 2007. Estimation of the efficiency of seed irradiation by thermal neutrons for inducing chromosomal aberrations in M1 of cotton Gossypium hirsutum L. In *Russian Journal of Genetics*, Vol 43, pp. 396-403. Springer SBM, Dordrecht; Netherlands.

Rao SA, Mengesha MH, Reddy CR. 1986. New Sources of Dwarfing Genes in Pearl-Millet (Pennisetum-Americanum). *Theoret Appl Genetics* **73**(2): 170-174.

Rea PA. 2007. Plant ATP-binding cassette transporters. *Annual review of plant biology* **58**: 347-375.

Ruijter JM, Ramakers C, Hoogaars WM, Karlen Y, Bakker O, van den Hoff MJ, Moorman AF. 2009. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic acids research* **37**(6): e45.

Sacerdot C, Mercier G, Todeschini AL, Dutreix M, Springer M, Lesage P. 2005. Impact of ionizing radiation on the life cycle of Saccharomyces cerevisiae Ty1 retrotransposon. *Yeast (Chichester, England)* **22**(6): 441-455.

Sanchez-Fernandez R, Davies TG, Coleman JO, Rea PA. 2001. The Arabidopsis thaliana ABC protein superfamily, a complete inventory. *The Journal of biological chemistry* **276**(32): 30231-30244.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nature genetics* **20**(1): 43-45.

Sambrook J, Fritsch EF, and Maniatis T. 1989. Molecular cloning: a laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y

Schertz KF, Rosenow DT, Johnson JW, Gibson PT. 1974. Single Dw3 Height-Gene Effects in 4- and 3-Dwarf Hybrids of Sorghum bicolor (L.) Moench1. *Crop Sci* **14**(6): 875-877.

Solovyev V, Salamov A. 1997. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* **5**: 294-302.

Sukhadev P, Rao YS, Rao MVS, Manga V. 1987. Genetics of induced dwarf mutants in pearl millet,Pennisetum americanum (L.)Leeke. *Euphytica* **36**(1): 181-185.

Titapiwatanakun B, Murphy AS. 2009. Post-transcriptional regulation of auxin transport proteins: cellular trafficking, protein phosphorylation, protein maturation, ubiquitination, and membrane composition. *Journal of Experimental Botany* **60**(4): 1093-1107.

Vasiliou V, Vasiliou K, Nebert DW. 2009. Human ATP-binding cassette (ABC) transporter family. *Human Genomics* **3**(3): 281-290.

Verrier PJ, Bird D, Burla B, Dassa E, Forestier C, Geisler M, Klein M, Kolukisaoglu U, Lee Y, Martinoia E et al. 2008. Plant ABC proteins--a unified nomenclature and updated inventory. *Trends in plant science* **13**(4): 151-159.

Xing A, Gao Y, Ye L, Zhang W, Cai L, Ching A, Llaca V, Johnson B, Liu L, Yang X et al. 2015. A rare SNP mutation in Brachytic2 moderately reduces plant height and increases yield potential in maize. *Journal of Experimental Botany*.

Yagyu P, Morris R. 1957. Cytogenetic Effects of X-Rays and Thermal Neutrons on Dormant Tomato Seeds. *Genetics* **42**(3): 222-238.

Ye L, Liu L, Xing A, Kang D. 2013a. Characterization of a dwarf mutant allele of Arabidopsis MDR-like ABC transporter AtPGP1 gene. *Biochemical and biophysical research communications* **441**(4): 782-786.

Zhu Q, Smith SM, Ayele M, Yang L, Jogi A, Chaluvadi SR, Bennetzen JL. 2012. High-throughput discovery of mutations in tef semi-dwarfing genes by next-generation sequencing analysis. *Genetics* **192**(3): 819-829.

CHAPTER IV:

PARALLEL LOSS OF INTRONS IN THE *ABCB1* GENE IN THE ANGIOSPERMS[3]

[3] Parvathaneni, R.K., DeLeo, V.L., Chakraborty, D., and K.M. Devos. To be submitted to Molecular Biology and Evolution

ABSTRACT

The presence of non-coding introns is a characteristic feature of most eukaryotic genes. However, the size of the introns, number of introns per gene and the number of intron containing genes can vary greatly between sequenced eukaryotic genomes. Nevertheless, the structure of a gene with reference to intron presence and positions is typically conserved in closely related species such as species belonging to the grass family which diverged ~70 MYA. The *ABCB1* gene which underlies dwarfing traits in the panicoid grasses, maize (*br2*), sorghum (*dw3*) and pearl millet (*d2*) does not follow this rule as intron number in the three species ranged from 2 to 4. We therefore analyzed the *ABCB1* gene structure in 49 sequenced angiosperm genomes and in an additional 26 non-sequenced monocots by PCR and sequencing. Several interesting features were observed including 1) that the number of introns in *ABCB1* ranged from 1 to 9 within the analyzed Angiosperms; 2) that all introns of this gene underwent either a one-time loss (single loss in 1 lineage/species only) or multiple independent losses (parallel loss in two or more lineages/species); and 3) the majority of the losses occurred within the grass family. In contrast, the structure of the closest homolog to *ABCB1*, *ABCB19*, remained constant in the majority of angiosperms analyzed. Using existing phylogenetic relationships within the grasses, we determined the ancestral branch-points where the loss occurred. Expression of *ABCB1* and *ABCB19* in the germline tissues and an in-depth intron sequence analysis of the only conserved intron in the grasses, intron 7, is also reported.

INTRODUCTION

Introns are a characteristic and common feature in eukaryotic genomes and likely accumulated very early in eukaryotic evolution (Fedorov et al. 2002; Rogozin et al. 2003; Csuros et al. 2008). The conservation of some intron positions across the different kingdoms for a period

close to 2 billion years suggests that introns are functionally important (Rogozin et al. 2003). Alternatively, introns may have inserted in the same location in orthologous genes, although this hypothesis is not well supported. Simulation studies and estimations of the probability of parallel gains for individual introns have indicated that the majority of shared introns likely have a common origin (Fedorov et al. 2002; Rogozin et al. 2003; Sverdlov et al. 2005).

Evolutionary loss and gain of introns in genomic sequence data may provide a mechanism by which organisms diversify gene expression and gene function. The rate of gain and loss of introns varies with the lineage but in most cases intron loss is higher, sometimes by a few orders of magnitude, than intron gain (Roy and Gilbert 2005b; Coulombe-Huntington and Majewski 2007b; Roy and Penny 2007; Fawcett et al. 2012; Wang et al. 2014a). Rates of intron loss have been calibrated in humans at 4-5 x 10^{-10} per intron per year by Roy and Gilbert (2005) and 4.28 x 10^{-13} by Coulombe-Huntington and Majewski (2007). Similar rates of intron loss have been observed in plants, including *Arabidopsis thaliana* (1-3 x 10^{-10}, (Fawcett et al. 2012)), *A. lyrata* (2.73 x 10^{-11}, (Fawcett et al. 2012)), *Oryza sativa* (3.3 x 10^{-10}, (Lin et al. 2006); 8.1 x 10^{-11},(Wang et al. 2014a)), and the grasses *Setaria italica*, *Brachypodium distachyon*, *Sorghum bicolor* and *Zea mays* (1.1 – 1.8 x 10^{10}, Wang et al. 2014).

Two main mechanisms for intron loss have been proposed, that is reverse-transcriptase (RT)-mediated intron loss and intron deletion triggered by repair of double strand breaks via non-homologous end joining (NHEJ). RT-mediated loss occurs when a cDNA recombines with its genomic copy. Characteristics of RT-mediated loss include precise removal of introns, preferential removal of small introns, deletion of adjacent introns and a 3' bias of intron removal (Frugoli et al. 1998; Roy and Gilbert 2006; Coulombe-Huntington and Majewski 2007a). A signature of genomic deletion of introns via NHEJ is the presence of 2-8 bp of micro-homology.

The presence of 3 or 4 bp direct repeats flanking deleted introns has been observed in *A. thaliana* and *A. lyrata* (Fawcett et al. 2012). Overall, however, RT-meditated removal is considered the dominant mechanism of intron loss (Roy and Gilbert 2005a; Coulombe-Huntington and Majewski 2007a; Wang et al. 2014a).

Intron loss has been documented in whole genome studies of multiple species. Comparison of the 10 fully sequenced genomes of the genus *Drosophila* uncovered a total of 1754 intron loss events (Coulombe-Huntington and Majewski 2007b). A comparison between the genome sequences of *A. thaliana* and *A. lyrata* revealed a combined loss of 105 introns (Fawcett et al. 2012) and a similar study across five sequenced grass genomes (maize, sorghum, rice, foxtail millet and *Brachypodium*) using *Arabidopsis* and banana as outgroups revealed a total of 745 intron loss events, including 93 cases of parallel intron loss whereby the same intron was lost independently in multiple lineages (Wang et al. 2014a). Until the latter study, only a few cases of parallel intron loss had been described in the literature. The glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene which has a 10 introns shows parallel loss of intron 9 in opossum, dog and the primate/rodent lineages (Coulombe-Huntington and Majewski 2007a). The *white* gene in the distantly related dipteran (true fly) lineages contains 14 introns, three of which (introns 10, 11 and 13) underwent parallel loss 3, 4 and 5 times, respectively (Krzywinski and Besansky 2002). In the *Drosophila* and mosquito species, convergent loss of one intron, intron Z, was observed in the *MRP1* gene (Zhan et al. 2012). Incidentally, two of the three genes for which parallel intron loss has been observed in animals, *white* and *MRP1,* are ABC transporters. In Angiosperms, the isochorismate synthase gene which comprises 15 introns underwent parallel intron loss of intron 2 in the grass lineage and the *Medicago truncatula/Glycine max* lineage (Yuan et al. 2009). The

overall characteristics of introns that have been lost once and those that have been lost multiple times are very similar (Wang et al. 2014).

In the current study, we conducted a detailed analysis of intron loss in the *ABCB1* gene across plant genomes. Intron loss in *ABCB1* was first observed when pearl millet *ABCB1* was isolated as a candidate for the *d2* dwarfing trait (Parvathaneni et al. 2013) and compared with sorghum *ABCB1*. We subsequently expanded the structural analysis of the *ABCB1* gene to other sequenced and non-sequenced members of the grasses and beyond, including selected non-grass monocots, dicots and the basal species *Amborella trichopoda*.

MATERIALS AND METHODS

Retrieval of protein, cDNA and gene sequences

Protein homologs (primary proteins only) to *Arabidopsis* ABCB1 (locus AT2G36910) (similarity >82%) and ABCB19 (locus AT3G28860) (similarity >85%) were retrieved from Phytozome 10.2 (phytozome.jgi.doe.gov) (Table 4.1). The corresponding gene and cDNA sequences were also obtained from Phytozome. The sequence of *Brassica oleraceae ABCB19* was obtained from Gramene (www.gramene.org). *ABCB1* and *ABCB19* gene sequences from additional species for which whole or partial genome sequence was available (Table 4.1) were identified by reciprocal BLASTN analysis using the rice gene sequences as queries. Top hits in reciprocal blast searches were considered true orthologs to the query sequence. To obtain the corresponding protein sequences, the genomic sequences were aligned with *ABCB1* and *ABCB19* genomic and mRNA sequences of rice and Arabidopsis to delineate introns. Introns were manually removed to yield cDNA sequences which were translated using the NCBI ORF finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). For the gene sequence analysis, a total of 63

ABCB1 sequences from 24 monocot species and 26 dicot species and 64 *ABCB19* sequences from 13 monocot and 31 dicot species was retrieved from databases or obtained from collaborators and analyzed.

Table 4.1 Information of the *ABCB* genes used including gene IDs, source of the genes and identification method

	Database locus ID	***Gene Alias***	**Protein Similarity (%)**	**Database**	**Identification method**	**intron count**
Prunus persica	ppa000269m.g	*Pp_ABCB1*	90.6	Phytozome 10.1	Protein homolog database	9
Phaseolus vulgaris	Phvul.001G179300	*Ph_ABCB1a*	91.5	Phytozome 10.1	Protein homolog database	9
Phaseolus vulgaris	Phvul.007G147400	*Ph_ABCB1b*	90.2	Phytozome 10.1	Protein homolog database	9
Medicago truncatula	Medtr7g102070	*Mt_ABCB1a*	89.7	Phytozome 10.1	Protein homolog database	9
Medicago truncatula	Medtr1g063170	*Mt_ABCB1b*	88.7	Phytozome 10.1	Protein homolog database	9
Malus domestica	MDP0000231597	*Md_ABCB1a*	87.8	Phytozome 10.1	Protein homolog database	9
Malus domestica	MDP0000183294	*Md_ABCB1b*	87.8	Phytozome 10.1	Protein homolog database	9
Glycine max	Glyma.19G184300	*Gm_ABCB1a*	93.3	Phytozome 10.1	Protein homolog database	9
Glycine max	Glyma.03G183600	*Gm_ABCB1b*	92.2	Phytozome 10.1	Protein homolog database	9
Glycine max	Glyma.10G055000	*Gm_ABCB1c*	92.5	Phytozome 10.1	Protein homolog database	9
Glycine max	Glyma.13G142100	*Gm_ABCB1d*	91.4	Phytozome 10.1	Protein homolog database	9
Fragaria vesca	gene09601-v1.0-hybrid	*Fv_ABCB1*	91.4	Phytozome 10.1	Protein homolog database	9
Citrus clementina	Ciclev10010916m.g	*Cc_ABCB1*	91.2	Phytozome 10.1	Protein homolog database	9
Citrus sinensis	orange1.1g000687m.g	*Cs_ABCB1*	91.2	Phytozome 10.1	Protein homolog database	9
Eutrema salsugineum	Thhalv10016150m.g	*Es_ABCB1*	97	Phytozome 10.1	Protein homolog database	8
Capsella grandiflora	Cagra.0558s0007	*Cg_ABCB1*	99.7	Phytozome 10.1	Protein homolog database	9
Boechera stricta	Bostr.23794s0400	*Bs_ABCB1*	99.2	Phytozome 10.1	Protein homolog database	9

Arabidopsis lyrata	934330	*Al_ABCB1*	99.8	Phytozome 10.1	Protein homolog database	9
Arabidopsis thaliana	AT2G36910	*At_ABCB1*	100	Phytozome 10.1	Protein homolog database	9
Theobroma cacao	Thecc1EG022244	*Tc_ABCB1*	92.5	Phytozome 10.1	Protein homolog database	9
Gossypium raimondii	Gorai.002G246800	*Gr_ABCB1a*	90.7	Phytozome 10.1	Protein homolog database	9
Gossypium raimondii	Gorai.006G163000	*Gr_ABCB1b*	90	Phytozome 10.1	Protein homolog database	9
Ricinus communis	30078.t000079	*Rc_ABCB1*	90	Phytozome 10.1	Protein homolog database	9
Manihot esculenta	cassava4.1_000306m.g	*Me_ABCB1*	89.3	Phytozome 10.1	Protein homolog database	9
Linum usitatissimum	Lus10023929.g	*Lu_ABCB1*	87.7	Phytozome 10.1	Protein homolog database	8
Salix purpurea	SapurV1A.1230s0070	*Sp_ABCB1a*	90.4	Phytozome 10.1	Protein homolog database	9
Salix purpurea	SapurV1A.0584s0030	*Sp_ABCB1b*	88.8	Phytozome 10.1	Protein homolog database	9
Populus trichocarpa	Potri.006G123900	*Pt_ABCB1a*	90.6	Phytozome 10.1	Protein homolog database	9
Populus trichocarpa	Potri.016G093600	*Pt_ABCB1b*	89.7	Phytozome 10.1	Protein homolog database	9
Eucalyptus grandis	Eucgr.K02930	*Eg_ABCB1*	90.7	Phytozome 10.1	Protein homolog database	9
Solanum lycopersicum	Solyc09g008240.2	*Sl_ABCB1*	89.9	Phytozome 10.1	Protein homolog database	9
Solanum tuberosum	PGSC0003DMT400009924	*St_ABCB1*	90.1	Phytozome 10.1	Protein homolog database	9
Mimulus guttatus	Migut.J00652	*Mg_ABCB1a*	90.3	Phytozome 10.1	Protein homolog database	9
Mimulus guttatus	Migut.L01707	*Mg_ABCB1b*	86.1	Phytozome 10.1	Protein homolog database	8
Aquilegia coerulea Goldsmith	Aquca_030_00072	*Aq_ABCB1a*	88.8	Phytozome 10.1	Protein homolog database	9
Aquilegia coerulea Goldsmith	Aquca_062_00022	*Aq_ABCB1b*	87.9	Phytozome 10.1	Protein homolog database	9
Spirodela polyrhiza	Not assigned	*Sp_ABCB1*	ND	Dr. Eric Lam	Reciprocal Blastn	
Phoenix dactylifera	Not assigned	*Pd_ABCB1*	ND	Source:	Reciprocal Blastn	9
Musa acuminata	GSMUA_Achr4P04430_001	*Ma_ABCB1*	ND	http://banana-genome.cirad.fr/	Reciprocal Blastn	9
Ananas comosus	Not assigned	*Ac_ABCB1*	ND	Source: Dr. Ray Ming	Reciprocal Blastn	9

Oryza sativa ssp. Indica	BGIOSGA029208	*Osi_ABCB1*	ND	http://plants.ensembl.org	Reciprocal Blastn	2
Oryza sativa ssp. japonica	LOC_Os08g45030	*Osj_ABCB1*	83.4	Phytozome 10.1	Protein homolog database	2
Phyllostachys edulis	Not assigned	*Pe_ABCB1a*	ND	http://www.bamboogdb.org	Reciprocal Blastn	4
Phyllostachys edulis	Not assigned	*Pe_ABCB1b*	ND	http://www.bamboogdb.org	Reciprocal Blastn	4
Brachypodium distachyon	Bradi3g12627	*Bd_ABCB1*	82.3	Phytozome 10.1	Protein homolog database	2
Hordeum vulgare	MLOC_5438	*Hv_ABCB1*	ND	http://plants.ensembl.org	Reciprocal Blastn	3
Aegilops tauschii	F775_52675	*Ae_ABCB1*	ND	http://plants.ensembl.org	Reciprocal Blastn	3
Triticum urartu	TRIUR3_22724	*Tu_ABCB1*	ND	http://plants.ensembl.org	Reciprocal Blastn	2
Eragrostis tef	JN672669.1	*Et_ABCB1a*	ND	NCBI/Shavannor smith	Reciprocal Blastn	2
Eragrostis tef	JN672670.1	*Et_ABCB1b*	ND	NCBI/Shavannor smith	Reciprocal Blastn	2
Chasmanthium laxum	Not assigned	*Cl_ABCB1*	ND	James Schnable	Reciprocal Blastn	4
Paspalum vaginatum	Not assigned	*Pas_ABCB1*	ND	James Schnable	Reciprocal Blastn	3
Zea mays	GRMZM2G315375	*Zm_ABCB1*	83	Phytozome 10.1	Protein homolog database	4
Sorghum bicolor	AY372819.1	*Sb_ABCB1*	83.1	NCBI	Protein homolog database	4
Andropogon virginicus	Not assigned	*Av_ABCB1*	ND	James Schnable	Reciprocal Blastn	4
Dichanthelium oligosanthes	Not assigned	*Do_ABCB1*	ND	James Schnable	Reciprocal Blastn	4
Panicum virgatum	Pavir.Fb02214	*Pv_ABCB1a*	83.4	Phytozome 10.1	Protein homolog database	3
Panicum virgatum	Pavir.Fa00055	*Pv_ABCB1b*	82.3	Phytozome 10.1	Protein homolog database	3
Panicum hallii	Pahal.0383s0025	*Pah_ABCB1*	ND	Phytozome 10.1	Reciprocal Blastn	3
Urochloa abyss	Not assigned	*Ua_ABCB1*	ND	James Schnable	Reciprocal Blastn	2
Setaria italica	Si013123m	*Si_ABCB1*	83.1	Phytozome 10.1	Protein homolog database	2
Cenchrus americanus	Not assigned	*Ca_ABCB1*	ND	Isolated in Devos lab	Blastn	2
Amborella trichopoda	evm_27.TU.AmTr_v1.0_scaffold00076.86	*Am_ABCB1*	80.4	Phytozome 10.2	Reciprocal Blastn	9

Prunus persica	ppa000359m.g	*Pp_ABCB19*	94.2	Phytozome 10.1	Protein homolog database	9
Phaseolus vulgaris	Phvul.004G027800	*Ph_ABCB19a*	93.3	Phytozome 10.1	Protein homolog database	9
Phaseolus vulgaris	Phvul.004G027800	*Ph_ABCB19b*	93.3	Phytozome 10.1	Protein homolog database	9
Medicago truncatula	Medtr6g011680	*Mt_ABCB19*	93.7	Phytozome 10.1	Protein homolog database	9
Malus Domestica	MDP0000265271	*Md_ABCB19*	90.5	Phytozome 10.1	Protein homolog database	9
Glycine max	Glyma.13G063700	*Gm_ABCB19 a*	94.8	Phytozome 10.1	Protein homolog database	9
Glycine max	Glyma.19G021500	*Gm_ABCB19 b*	94.6	Phytozome 10.1	Protein homolog database	9
Cucumis sativus	Cucsa.158380	*Cu_ABCB19*	92	Phytozome 10.1	Protein homolog database	9
Fragaria vesca	gene03853-v1.0-hybrid	*Fv_ABCB19*	94.1	Phytozome 10.1	Protein homolog database	9
Citrus clementina	Ciclev10010931m.g	*Cc_ABCB19*	93.5	Phytozome 10.1	Protein homolog database	9
Citrus sinensis	orange1.1g000856m.g	*Cs_ABCB19*	93.5	Phytozome 10.1	Protein homolog database	9
Eutrema salsugineum	Thhalv10003528m.g	*Es_ABCB19*	97.7	Phytozome 10.1	Protein homolog database	9
Capsella grandiflora	Cagra.14450s0001	*Cg_ABCB19*	91.1	Phytozome 10.1	Protein homolog database	9
Capsella rubella	Carubv10016588m.g	*Cr_ABCB19*	99.6	Phytozome 10.1	Protein homolog database	9
Boechera stricta	Bostr.0556s0536	*Bs_ABCB19*	99.8	Phytozome 10.1	Protein homolog database	9
Brassica rapa	Brara.F03123	*Br_ABCB19a*	96.4	Phytozome 10.1	Protein homolog database	8
Brassica rapa	Brara.I00328	*Br_ABCB19b*	92.7	Phytozome 10.1	Protein homolog database	8
Brassica oleracea	Bo7g087020	*Bo_ABCB19a*	ND	www.gramene.org	Protein homolog database	9
Brassica oleracea	Bo9g008680	*Bo_ABCB19b*	ND	www.gramene.org	Protein homolog database	8
Arabidopsis lyrata	484657	*Al_ABCB19*	99.8	Phytozome 10.1	Protein homolog database	9
Arabidopsis thaliana	AT3G28860	*At_ABCB19*	ND	Phytozome 10.1	Protein homolog database	9
Theobroma cacao	Thecc1EG018479t	*Tc_ABCB19*	93.7	Phytozome 10.1	Protein homolog database	9
Carica papaya	evm.TU.supercontig_14.14	*Cp_ABCB19*	93.8	Phytozome 10.1	Protein homolog database	9
Gossypium raimondii	Gorai.006G021600	*Gr_ABCB19a*	94	Phytozome 10.1	Protein homolog database	9
Gossypium raimondii	Gorai.001G256100	*Gr_ABCB19b*	93.1	Phytozome 10.1	Protein homolog database	9
Ricinus communis	29822.t000171	*Rc_ABCB19*	93.7	Phytozome 10.1	Protein homolog database	9
Manihot esculenta	cassava4.1_000385m.g	*Me_ABCB19 a*	93.5	Phytozome 10.1	Protein homolog database	9

Manihot esculenta	cassava4.1_000386m	*Me_ABCB19b*	93.6	Phytozome 10.1	Protein homolog database	9
Linum usitatissimum	Lus10030674.g	*Lu_ABCB19a*	95	Phytozome 10.1	Protein homolog database	9
Linum usitatissimum	Lus10005249.g	*Lu_ABCB19b*	93.5	Phytozome 10.1	Protein homolog database	9
Salix purpurea	SapurV1A.0334s0130	*Sp_ABCB19*	93.4	Phytozome 10.1	Protein homolog database	9
Populus trichocarpa	Potri.017G081100	*Pt_ABCB19*	93.5	Phytozome 10.1	Protein homolog database	9
Eucalyptus grandis	Eucgr.J01214	*Eg_ABCB19*	94.9	Phytozome 10.1	Protein homolog database	9
Solanum lycopersicum	Solyc02g087870.2	*Sl_ABCB19*	93.7	Phytozome 10.1	Protein homolog database	9
Solanum tuberosum	PGSC0003DMG400001419	*St_ABCB19*	93.7	Phytozome 10.1	Protein homolog database	9
Mimulus guttatus	Migut.H00909	*Mg_ABCB19*	93.6	Phytozome 10.1	Protein homolog database	9
Aquilegia coerulea Goldsmith	Aquca_002_00550	*Aq_ABCB19a*	91.9	Phytozome 10.1	Protein homolog database	9
Aquilegia coerulea Goldsmith	Aquca_074_00020	*Aq_ABCB19b*	93.5	Phytozome 10.1	Protein homolog database	9
Amborella trichopoda	evm_27.TU.AmTr_v1.0_scaffold00010.381	*Am_ABCB19*	ND	Phytozome 10.2	Reciprocal Blastn	10
Musa acuminata	GSMUA_Achr8G24160_001	*Ma_ABCB19a*	ND	Phytozome 10.2	Reciprocal Blastn	9
Musa acuminata	GSMUA_Achr4G09980_001	*Ma_ABCB19b*	ND	Phytozome 10.2	Reciprocal Blastn	9
Oryza sativa ssp. Indica	BGIOSGA014966	*Osi_ABCB19a*	ND	Gramene	Protein homolog database	9
Oryza sativa ssp. Indica	BGIOSGA014245	*Osi_ABCB19b*	ND	Gramene	Protein homolog database	8
Oryza sativa ssp. japonica	LOC_Os04g38570	*Osj_ABCB19a*	91.1	Phytozome 10.1	Protein homolog database	9
Oryza sativa ssp. japonica	LOC_Os04g54930	*Osj_ABCB19b*	ND	Phytozome 10.1	Reciprocal Blastn	8
Phyllostachys edulis	PH01000947	*Pe_ABCB19*	ND	http://www.bamboogdb.org	Protein homolog database	9
Brachypodium distachyon	Bradi5g12307	*Bd_ABCB19a*	90.7	Phytozome 10.1	Protein homolog database	9
Brachypodium distachyon	Bradi5g23600	*Bd_ABCB19b*	86.1	Phytozome 10.1	Protein homolog database	8
Hordeum Vulgare	MLOC_64400	*Hv_ABCB19a*		http://plants.ensembl.org	Gramene protein homolog database	9
Hordeum Vulgare	MLOC_56096	*Hv_ABCB19b*	ND	http://plants.ensembl.org	Reciprocal Blastn	8

Aegilops tauschii	F775_21781	*Ae_ABCB19a*	ND	http://plants.ensembl.org	Reciprocal Blastn	9
Aegilops tauschii	F775_17083	*Ae_ABCB19b*	ND	http://plants.ensembl.org	Reciprocal Blastn	8
Triticum urartu	TRIUR3_16643	*Tu_ABCB19a*	ND	http://plants.ensembl.org	Reciprocal Blastn	9
Triticum urartu	TRIUR3_16789	*Tu_ABCB19b*	ND	http://plants.ensembl.org	Reciprocal Blastn	8
Zea mays	GRMZM2G125424	*Zm_ABCB19a*	88.6	Phytozome 10.1	Protein homolog database	9
Zea mays	GRMZM2G072850_T01	*Zm_ABCB19b*	86.2	Phytozome 10.1	Protein homolog database	8
Zea mays	GRMZM2G085236_T01	*Zm_ABCB19c*	85.3	Phytozome 10.1	Protein homolog database	8
Sorghum bicolor	Sobic.006G105600	*Sb_ABCB19a*	88.6	Phytozome 10.1	Protein homolog database	9
Sorghum bicolor	Sobic.006G236700	*Sb_ABCB19b*	86.7	Phytozome 10.1	Protein homolog database	8
Panicum virgatum	Pavir.Ga01466	*Pv_ABCB19a*	86.5	Phytozome 10.1	Protein homolog database	9
Panicum virgatum	Pavir.J34236	*Pv_ABCB19b*	ND	Phytozome 10.1	Protein homolog database	8
Panicum hallii	Pahal.G02507	*Pah_ABCB19*	ND	Phytozome 10.1	Reciprocal Blastn	9
Setaria italica	Si009197m.g	*Si_ABCB19a*	88.7	Phytozome 10.1	Protein homolog database	9
Setaria italica	Si009196m	*Si_ABCB19b*	86.5	Phytozome 10.1	Protein homolog database	8
Amborella trichopoda	evm_27.TU.AmTr_v1.0_scaffold00010.381	*Am_ABCB19*	ND	Phytozome 10.2	Reciprocal Blastn	10

Sequence alignment and determination of intron sites

The retrieved *ABCB1* and *ABCB19* gene sequences were aligned using the online multiple sequence alignment (MSA) program MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/) (Edgar 2004). Because of the large number of entries, sequence alignments were done separately for monocot *ABCB1*, dicot *ABCB1*, monocot *ABCB19* and dicot *ABCB19*. Where necessary, sequence alignments were edited manually using Jalview (Waterhouse et al. 2009). A MSA between genomic DNA and cDNA of *ABCB1* and *ABCB19* of Arabidopsis, rice, banana and sorghum

identified the corresponding introns between monocots and dicots in each of the genes and across the two genes, allowing conclusions to be drawn from the separate alignments on the differential presence of introns across species and genes.

Phylogenies of the ABCB1 and ABCB19 proteins

The retrieved and *in-silico* translated ABCB1 and ABCB19 protein sequences were aligned using the MUSCLE (Edgar, 2004) aligner in the program Mega 6.0 (Tamura et al. 2013). Only ABCB genes for which the entire coding sequence was available were included in the analysis. A total of 58 ABCB1 protein sequences from 19 monocot species and 26 dicot species and 62 ABCB19 sequences from 12 monocot and 30 dicot species was analyzed. A neighbor joining tree was constructed using the Poisson model with uniform substitution rates in Mega 6.0. The bootstrap method with 1000 iterations was used to test the robustness of the tree. The tree was rooted to Arabidopsis protein AT2G47000 (ABCB4).

Plant material and DNA extractions

Seeds of the grass species *Danthoniopsis dinteri, Steinchisma laxum, Sacciolepis myosuroides, Arundinella hirta,* and leaf material from *Brachiaria spp., Andropogon gerardii, Bambusa* spp. and *Acroceras macrum* were obtained from Melanie L. Harrison-Dunn at the Plant Genetic Resources Conservation unit (PGRCU), Griffin, GA. The seeds were planted in 3 inch pots and grown in a greenhouse under natural light and day/night temperatures of 32 ^{0}C/28 ^{0}C. *Paspalum vaginatum* cultivar 'sea-dwarf' was provided as a clonal propagate by Paul Raymer (UGA). Leaves from the Zoysia grass cultivar 'Emerald' (*Zoysia japonica* X *Zoysia tenuifolia*) were collected from the lawn outside the UGA Miller Plant Sciences building. Leaves from *Phyllostachys* spp. were collected from a bamboo stand on the UGA North Campus. Leaves from

Typha latifolia, *Acorus americanus*, *Tillandsia usneoides*, *Centrolepis* spp. and *Xyris* spp. were collected from the UGA Plant Biology greenhouse teaching collection. Leaves from *Pharus mezhii* and *Streptochaeta* spp. and DNA from *Anomochloa* spp., *Phragmites australis* and *Micraria subulifolia* were provided by Elizabeth Kellogg (Donald Danforth Center, St. Louis, Missouri). DNA was extracted from leaf samples using the Qiagen DNA extraction kit. DNA of *Mayaca* spp. and *Ecdieocolea* spp. was provided by Jim Leebens-Mack (UGA), of *Zea mays* accessions NSL 2840 and B73 by Kelly Dawe (UGA) and of *Oryza sativa var. Japonica* (Nipponbare) by Jeff Bennetzen (UGA).

Primer design and PCR amplification across introns

Using the MUSCLE alignment of the monocot *ABCB1* gene sequences, primer sets were designed against conserved exon regions to amplify the introns in a range of monocot species. The primer pairs were developed to amplify across introns 1+2, 3+4, 5, 6, 7, 8 and 9 (Table 4.2). PCR amplifications were performed in 20 µl reaction volumes containing 1X PCR buffer (GoTaq buffer), 1.5 mM $MgCl_2$, 0.25 mM of each dNTP, 0.5 µM forward and reverse primers, 10-25 ng of DNA template and 1 U of Taq DNA Polymerase (Promega,Madison, WI). Amplification conditions consisted of initial denaturation at 95 ^{0}C for 5 minutes followed by 34 cycles of 95 ^{0}C for 30 sec, 61 ^{0}C for 30 sec and 72 ^{0}C for 2 min, and a final extension of 72 ^{0}C for 10 min. PCR products were separated on 0.8% agarose gels. Amplicon sizes were used as a proxy for the presence and absence of introns. DNA from the sequenced genomes rice, sorghum and maize was used as positive control.

Table 4.2 List of primers used to amplify across different introns

Primers	Sequence	Annealing temp. (^{0}C)	Intron
Forward			
ABCB1_F5	GACCTCTTCCGCTTCGCCGACG	61	1+2
ABCB1_F8	GACCTCTTCCGCTTCGCCGACG	61	1+2
ABCB1_F6	TTCCTCCGCTTCTTCGCCGACCTCG	61	1+2
ABCB1_F22	GACGTCATCTACGCCATCAACGC	61	3+4
ABCB1_F16	ACCTACTTCACCGTCTTCTGCTG	61	5
ABCB1_F17-1	TCCGGGTCAGGGAAGAGCAC	61	6
ABCB1_F27	CAGGAGCCGACGCTGTTCGC	61	7
ABCB1_F26	CATCAAGGAGAACCTGCTGCTGGG	61	7
ABCB1_F25	GGCGCTGGACCGCTTCATGAT	61	8
ABCB1_F24	GCCACCAGCGCGCTGGAC	61	8
ABCB1_F29	CTTCAGCGCCATCTTCGCCTAC	61	9
Reverse			
ABCB1_R9	GCTTCTCGCTGATGGCGTCCTG	61	1+2
ABCB1_R20	TAGCAGCAGAAGACGGTGAAGTAGGT	61	3+4
ABCB1_R16-2	GCCATGCTCGGAGCCGACT	61	5
ABCB1_F19-1	TCCCCAGCAGCAGGTTCTC	61	6
ABCB1_R25	GCGCTGGTGGCCTCGTCCA	61	7
ABCB1_R26	CGGTCCAGCGCCTCCTGCA	61	7
ABCB1_R29	CCGATGAGGAGTAGCAGTA	61	8
ABCB1_R27	ACGTAGGCGAAGATGGCGCT	61	8
ABCB1_R1	GCGTASGASGCGTASAGSAGGAACTG	61	9

Long-range PCR to isolate and validate the gene structure

Full-length *ABCB1* gene isolation and sequencing were performed in selected species, including *Ecdeiocolea* spp., *Mayaca* spp., *Typha latifolia*, *Streptochaeta* spp., *Pharus mezhii*, *Tillansia usneoides*, *Phyllostachys* spp., *Zoysia* spp., *Paspalum vaginatum*, *Brachiaria* spp., and *Carex* spp., to verify that PCR results had been correctly interpreted. The 11 species were selected because of their phylogenetic position and/or their unusual intron composition. Several sets of primers were designed against conserved regions, identified based on the MSA, in the first and

last exons of *ABCB1* (Table 4.3). The primer combination ABCB1F1/R1 could amplify *ABCB1* from the majority of the selected species (Table 4.3). Other species required the use of different forward/reverse primers (Table 4.3). Long range PCR was performed in 12.5 µl reaction volumes comprising 0.4 µM forward primer, 0.4 µM reverse primer, 0.3 mM of each dNTP, 1X Long Amp buffer and 1.25 U of Long Amp Taq DNA polymerase (New England Biolabs, Ipswich, MA). The PCR conditions were as follows: Pre-heat the PCR block to 95^0C; 95 ^{0}C for 30 s for initial denaturation followed by 35 cycles of 95 ^{0}C for 30 sec, 61 ^{0}C for 30 sec and 68 ^{0}C for 6 min, and a final extension at 68 ^{0}C for 20 min. Amplification products were checked on a 0.8% agarose gel. Products with sizes greater than 3 kb were either cleaned with a QIAquick PCR purification kit (Qiagen, Valencia, CA) or gel extracted using a GeneJet gel extraction kit (Life Technologies, Grand Island, NY). PCR products were cloned in the pGEM-T vector (Promega, Madison, WI) and sequenced using the BigDye Terminator v3.1 kit (Life Technologies, Grand Island, NY) using the manufacturer's protocol except that reaction volumes were halved and 1/8 the recommended volume of Big dye was used. Primers used for sequencing included the vector primers T7 (5'-GTAATACGACTCACTATAGGGC-3') and SP6 (5'-ATTTAGGTGACACTATAGAATACTC-3', and the intron-flanking primers (Table 2.1). Sequencing products were cleaned on Sephadex^{TM} G-50 fine (GE Healthcare Bio-Sciences, Pittsburgh, PA) using an in-house protocol and run on an ABI 3730xl (Life Technologies, Grand Island, NY).

Table 4.3 List of primers used for long range PCR across all *ABCB1* introns

Primers	**Sequence**	**Annealing temp. (^{0}C)**	**Species in which primers were successfully used**
Forward			
ABCB1_F5	GACCTCTTCCGCTTCGCCGACG	61	*Zoysia, Pharus, Tillandsia*
ABCB1_F1	CGCTTCTTYGCSGABCTBGTSGACTC	61	*Phyllostachys, Brachiaria, Typha, Ecdieocolea,*

			Mayaca, Paspalum, Carex, Streptochaeta
Reverse			
ABCB1_R4	CCGTGATCTTGCGCTCCGCGTTG	61	*Zoysia*
ABCB1_R1	GCGTASGASGCGTASAGSAGGAACTG	61	*Phyllostachys, Brachiaria, Typha, Ecdieocolea, Paspalum, Carex, Streptochaeta*
ABCB1_R22	GACACCATCAGCACCATGAA	61	*Mayaca*
ABCB1_R23	CACCACCTCCGCCTCCGT	61	*Pharus, Tillandsia*

Motif finding

The intron 7 sequences of 58 *ABCB1* orthologs from 23 monocots and 26 dicots were used to identify motifs using the 'Multiple EM for Motif Elicitation' (MEME) version 4.10.1 (http://meme-suite.org/tools/meme) (Bailey et al. 2009). The maximum number of characters allowed in the webserver limited us to use just 58 ABCB1 sequences. Different settings of maximum width were used to identify motifs. The settings were as follows: number of motifs = 10, minimum and maximum widths = 12, 15, 18, 20, 30 and 50. Default settings were used for all other parameters.

Expression of *ABCB1* and *ABCB19* genes in germline tissues

Microarray expression profiles of *ABCB1* and *ABCB19* in germline tissues were studied in *O. sativa* and *A. thaliana* relative to stably expressed housekeeping genes. Expression heat maps of *ABCB1* and *ABCB19* were generated using the RiceXPro browser (http://ricexpro.dna.affrc.go.jp) against the housekeeping genes ubiquitin 5 (*UBQ5*) and eukaryotic elongation factor 1-alpha (*EF1-α*) using the dataset 'Spatio-temporal gene expression of various tissues/organs throughout entire growth in the field'. UBQ5 and EF1-α are the most stably expressed control genes across tissues used in qPCR experiments for rice (Jain et al. 2006). The gene nomenclature from the rice annotation project database (RAP-DB) was used in the

RiceXPro browser which is different from the gene nomenclature used in Phytozome 10.2. The gene ID's are *ABCB1*: Os08g0564300; *ABCB19*: Os04g0459000; *EF1-α*: Os03g0177400 and *UBQ5*: Os01g0328400. Only the unique probes for each gene (*ABCB1*: S-2532, feature number 3137; S-12091, feature number 14315; S-19963, feature number 24050; and S-22319, feature number 27017; *ABCB19*: S-16663, feature number 19945; *UBQ5*: S-18048, feature number 21680; EF1-α: S-4018, feature number 4831; S-4227, feature number 5069; and S-10066, feature number 11921) were used to analyze expression levels. In Arabidopsis, the AtGenExpress visualization tool (AVT) browser (http://jsp.weigelworld.org/expviz/expviz.jsp) was used to compare absolute expression values of Arabidopsis *ABCB1* (AT2G36910), *ABCB19* (AT3G28860), *Act-2* (AT3G18780) and Ubiquitin (AT5G25760) reported in the AtGE development dataset (Schmid et al. 2005). This data set contains expression data in multiple organs but as we were interested in expression in germline tissue, we only considered expression in the inflorescence after bolting and in various embryo stages (mid globular to early heart embryo, early to late heart embryo, late heart to mid torpedo embryo and mid to late torpedo embryos).

RESULTS

Structural analysis of *ABCB1* and *ABCB19* across sequenced genomes

Investigation into the structure of *ABCB1* genes across grasses started with the isolation and sequencing of pearl millet *ABCB1*, the gene underlying the mutation in *d2* dwarfs (Parvathaneni et al. 2013). Comparison of pearl millet *ABCB1* (*Ca_ABCB1*) with *ABCB1* in the panicoid species sorghum (*Sb_ABCB1*; AY372819.1), maize (*Zm_ABCB1*; GRMZM2G315375) and foxtail millet (*Si_ABCB1*; Si013123m) revealed that *ABCB1* in pearl millet and foxtail millet carried two introns while *ABCB1* in maize and sorghum carried four introns (Figure 4.1). Analysis

into the structure of *ABCB1* in other grass sub-family members such as *Brachypodium* and rice revealed further variation in the *ABCB1* gene structure (Figure 4.1). Interestingly, while the number of introns in grasses appeared to be limited to four, the Arabidopsis *ABCB1* gene carried 9 introns. Expanding the *ABCB1* comparison to all sequenced genomes available at the time demonstrated that 23 out of the 26 dicots analyzed and the four non-grass monocots *Spirodela polyrhiza*, *Musa acuminata*, *Ananas comosus* and *Phoenix dactylifera* had 9 introns (Figure 4.2). *ABCB1* in the dicots *Eutrema salsugineum* (Thhalv10016150m) and *Linum usitatissimum* (Lus10023929) and one of the two *ABCB1* copies in *Mimulus guttatus* (Migut.L01707) had 8 introns and lacked introns 8, 2 and 7, respectively (Table 4.4). In contrast, the number of introns in *ABCB1* in grasses varied from 2 to 4. All sequenced grass species carried intron 7, lacked introns 1, 3, 4, 8 and 9, and had varying combinations of introns 2, 5 and 6 (Table 4.4).

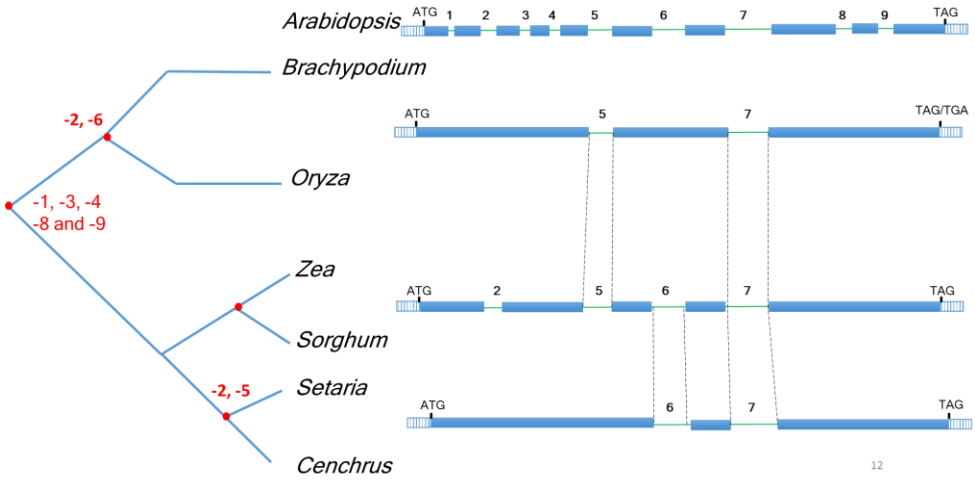


Figure 4.1 Gene structure variation of the *ABCB1* gene in members of the grass family. Arabidopsis *ABCB1* carries 9 introns. The 5' and 3' UTR regions are shown in shaded blue boxes and their sizes are unknown. The figure is not drawn to scale.

Table 4.4 Presence/absence polymorphisms of the *ABCB1* introns in all the sequenced Angiosperms. P = present; A = absent

Species	**Locus ID**	**I[1] 1**	**I 2**	**I 3**	**I 4**	**I 5**	**I 6**	**I 7**	**I 8-ABCB1**	**I 9**
Dicots										
Prunus persica	ppa000269m.g	P	P	P	P	P	P	P	P	P
Phaseolus vulgaris	Phvul.001G179300	P	P	P	P	P	P	P	P	P
Phaseolus vulgaris	Phvul.007G147400	P	P	P	P	P	P	P	P	P
Medicago truncatula	Medtr7g102070	P	P	P	P	P	P	P	P	P
Medicago truncatula	Medtr1g063170	P	P	P	P	P	P	P	P	P
Malus domestica	MDP0000231597	P	P	P	P	P	P	P	P	P
Malus domestica	MDP0000183294	P	P	P	P	P	P	P	P	P
Glycine max	Glyma.19G184300	P	P	P	P	P	P	P	P	P
Glycine max	Glyma.03G183600	P	P	P	P	P	P	P	P	P
Glycine max	Glyma.10G055000	P	P	P	P	P	P	P	P	P
Glycine max	Glyma.13G142100	P	P	P	P	P	P	P	P	P
Fragaria vesca	gene09601-v1.0-hybrid	P	P	P	P	P	P	P	P	P
Citrus clementina	Ciclev10010916m.g	P	P	P	P	P	P	P	P	P
Citrus sinensis	orange1.1g000687m.g	P	P	P	P	P	P	P	P	P
Eutrema salsugineum	Thhalv10016150m.g	P	P	P	P	P	P	P	A	P
Capsella grandiflora	Cagra.0558s0007	P	P	P	P	P	P	P	P	P
Boechera stricta	Bostr.23794s0400	P	P	P	P	P	P	P	P	P
Arabidopsis lyrata	934330	P	P	P	P	P	P	P	P	P
Arabidopsis thaliana	AT2G36910	P	P	P	P	P	P	P	P	P
Theobroma cacao	Thecc1EG022244	P	P	P	P	P	P	P	P	P
Theobroma cacao	Thecc1EG022244	P	P	P	P	P	P	P	P	P
Gossypium raimondii	Gorai.002G246800	P	P	P	P	P	P	P	P	P
Gossypium raimondii	Gorai.006G163000	P	P	P	P	P	P	P	P	P
Ricinus communis	30078.t000079	P	P	P	P	P	P	P	P	P

Manihot esculenta	cassava4.1_000306m.g	P	P	P	P	P	P	P	P	P
Linum usitatissimum	Lus10023929.g	P	A	P	P	P	P	P	P	P
Salix purpurea	SapurV1A.1230s0070	P	P	P	P	P	P	P	P	P
Salix purpurea	SapurV1A.0584s0030	P	P	P	P	P	P	P	P	P
Populus trichocarpa	Potri.006G123900	P	P	P	P	P	P	P	P	P
Populus trichocarpa	Potri.016G093600	P	P	P	P	P	P	P	P	P
Eucalyptus grandis	Eucgr.K02930	P	P	P	P	P	P	P	P	P
Solanum lycopersicum	Solyc09g008240.2	P	P	P	P	P	P	P	P	P
Solanum tuberosum	PGSC0003DMT400009924	P	P	P	P	P	P	P	P	P
Mimulus guttatus	Migut.J00652	P	P	P	P	P	P	P	P	P
Mimulus guttatus	Migut.L01707	P	P	P	P	P	P	A	P	P
Aquilegia coerulea Goldsmith	Aquca_030_00072	P	P	P	P	P	P	P	P	P
Aquilegia coerulea Goldsmith	Aquca_062_00022	P	P	P	P	P	P	P	P	P
Basal dicot										
Amborella trichopoda	evm_27.TU.AmTr_v1.0_scaffold00076.86	P	P	P	P	P	P	P	P	P
Non-grass monocots										
Phoenix dactylifera	Not assigned	P	P	P	P	P	P	P	P	P
Spirodela polyrhiza	Not assigned	P	P	P	P	P	P	P	P	P
Musa acuminata	GSMUA_Achr4P04430_001	P	P	P	P	P	P	P	P	P
Ananas comosus	Not assigned	P	P	P	P	P	P	P	P	P
Grasses										
Oryza sativa ssp. indica	BGIOSGA029208	A	A	A	A	P	A	P	A	A
Oryza sativa ssp. japonica	LOC_Os08g45030	A	A	A	A	P	A	P	A	A
Phyllostachys edulis	Not assigned	A	P	A	A	P	P	P	A	A
Phyllostachys edulis	Not assigned	A	P	A	A	P	P	P	A	A
Brachypodium distachyon	Bradi3g12627	A	A	A	A	P	A	P	A	A

Hordeum vulgare	MLOC_5438	A	A	A	A	P	P	P	A	A
Aegilops tauschii	F775_52675	A	A	A	A	P	P	P	A	A
Triticum urartu	TRIUR3_22724	A	A	A	A	P	P	P	A	A
Eragrostis tef	JN672669.1	A	A	A	A	A	P	P	A	A
Eragrostis tef	JN672670.1	A	A	A	A	A	P	P	A	A
Chasmanthium laxum	Not assigned	A	P	A	A	P	P	P	A	A
Paspalum vaginatum	Not assigned	A	P	A	A	A	P	P	A	A
Zea mays	GRMZM2G315375_T01	A	P	A	A	P	P	P	A	A
Sorghum bicolor	Sobic.007G163800	A	P	A	A	P	P	P	A	A
Andropogon virginicus	Not assigned	A	P	A	A	P	P	P	A	A
Dichanthelium oligosanthes	Not assigned	A	P	A	A	A	P	P	A	A
Panicum virgatum	Pavir.Fb02214	A	P	A	A	A	P	P	A	A
Panicum virgatum	Pavir.Fa00055	A	P	A	A	A	P	P	A	A
Panicum hallii	Pahal.0383s0025	A	P	A	A	A	P	P	A	A
Urochloa abyss	Not assigned	A	A	A	A	A	P	P	A	A
Setaria italica	Si013123m	A	A	A	A	A	P	P	A	A
Cenchrus americanus	Not assigned	A	A	A	A	A	P	P	A	A

[1] I = intron

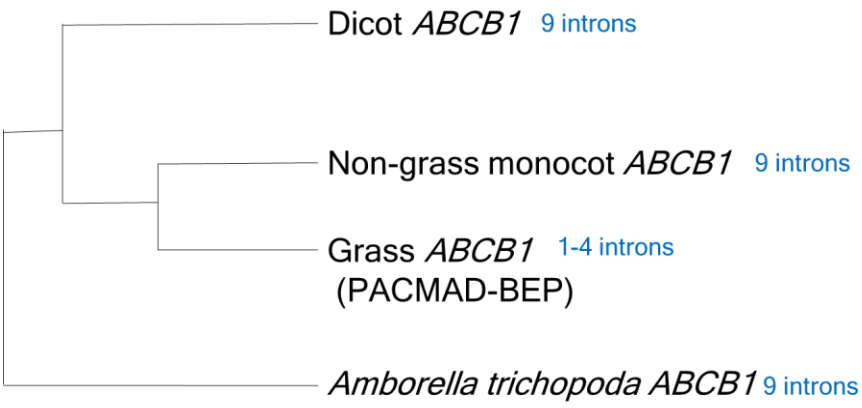


Figure 4.2: An overview of plant phylogeny showing intron loss in *ABCB1* in Angiosperms. Most species in dicots and non-grass monocots contain 9 introns but there are a few species which show single intron loss.

Because *ABCB1* is a member of a multigene family, we also examined the structure of *ABCB19*, the closest extant paralog of *ABCB1*. *ABCB19* contained 9 introns in all sequenced monocot and dicot species, except in *Brassica rapa* where both *ABCB19* gene copies (Brara.F03123; Brara.I00328) lacked intron 7 (Table 4.5). To verify if intron 7 was present in *ABCB19* in other sequenced *Brassica* species, the *ABCB19* gene sequences (2 copies) from *Brassica oleraceae* were retrieved from Gramene and their intron number analyzed. One copy, Bo7g087020, carried all 9 introns while intron 7 was absent from the second copy, Bo7g008680. Interestingly, all intron positions were conserved between *ABCB1* and *ABCB19* except for the position of intron 8. *ABCB1* intron 8 was located ~500 bp upstream of *ABCB19* intron 8. To avoid confusion, *ABCB1* intron 8 will be hereafter referred to as intron 8-ABCB1 and *ABCB19* intron 8 as intron 8-ABCB19 (Table 4.6). The basal dicot *Amborella trichopod*a contains an extra intron in *ABCB19*, Amborella-Intron 10, after the 9th intron which is absent in all other dicots and monocots analyzed.

Table 4.5 The presence/absence polymorphism of the *ABCB19* introns in all sequenced Angiosperms. P = present; A = absent; ND = not determined

Species	**Locus name**	**I[1] 1**	**I 2**	**I 3**	**I 4**	**I 5**	**I 6**	**I 7**	**I 8**	**I 9**	**Am-I 10**
Prunus persica	ppa000359m.g	P	P	P	P	P	P	P	P	P	A
Phaseolus vulgaris	Phvul.004G027800	P	P	P	P	P	P	P	P	P	A
Phaseolus vulgaris	Phvul.004G027800	P	P	P	P	P	P	P	P	P	A
Medicago truncatula	Medtr6g011680	P	P	P	P	P	P	P	P	P	A
Malus Domestica	MDP0000265271	P	P	P	P	P	P	P	P	P	A
Glycine max	Glyma.13G063700	P	P	P	P	P	P	P	P	P	A
Glycine max	Glyma.19G021500	P	P	P	P	P	P	P	P	P	A
Cucumis sativus	Cucsa.158380	P	P	P	P	P	P	P	P	P	A
Fragaria vesca	gene03853-v1.0-hybrid	P	P	P	P	P	P	P	P	P	A
Citrus clementina	Ciclev10010931m.g	P	P	P	P	P	P	P	P	P	A

Citrus sinensis	orange1.1g000856m.g	P	P	P	P	P	P	P	P	P	A
Eutrema salsugineum	Thhalv10003528m.g	P	P	P	P	P	P	P	P	P	A
Capsella grandiflora	Cagra.14450s0001	P	P	P	P	P	P	P	P	P	A
Capsella rubella	Carubv10016588m.g	P	P	P	P	P	P	P	P	P	A
Boechera stricta	Bostr.0556s0536	P	P	P	P	P	P	P	P	P	A
Brassica rapa	Brara.F03123	P	P	P	P	P	P	A	P	P	A
Brassica rapa	Brara.I00328	P	P	P	P	P	P	A	P	P	A
Brassica oleracea	Bo7g087020	P	P	P	P	P	P	P	P	P	A
Brassica oleracea	Bo9g008680	P	P	P	P	P	P	A	P	P	A
Arabidopsis lyrata	484657	P	P	P	P	P	P	P	P	P	A
Arabidopsis thaliana	AT3G28860	P	P	P	P	P	P	P	P	P	A
Theobroma cacao	Thecc1EG018479t	P	P	P	P	P	P	P	P	P	A
Carica papaya	evm.TU.supercontig_14.14	P	P	P	P	P	P	P	P	P	A
Gossypium raimondii	Gorai.006G021600	P	P	P	P	P	P	P	P	P	A
Gossypium raimondii	Gorai.001G256100	P	P	P	P	P	P	P	P	P	A
Ricinus communis	29822.t000171	P	P	P	P	P	P	P	P	P	A
Manihot esculenta	cassava4.1_000385m.g	P	P	P	P	P	P	P	P	P	A
Manihot esculenta	cassava4.1_000386m	P	P	P	P	P	P	P	P	P	A
Linum usitatissimum	Lus10030674.g	P	P	P	P	P	P	P	P	P	A
Linum usitatissimum	Lus10005249.g	P	P	P	P	P	P	P	P	P	A
Salix purpurea	SapurV1A.0334s0130	P	P	P	P	P	P	P	P	P	A
Populus trichocarpa	Potri.017G081100	P	P	P	P	P	P	P	P	P	A
Eucalyptus grandis	Eucgr.J01214	P	P	P	P	P	P	P	P	P	A
Solanum lycopersicum	Solyc02g087870.2	P	P	P	P	P	P	P	P	P	A
Solanum tuberosum	PGSC0003DMG400001419	P	P	P	P	P	P	P	P	P	A
Mimulus guttatus	Migut.H00909	P	P	P	P	P	P	P	P	P	A
Aquilegia coerulea Goldsmith	Aquca_002_00550	P	P	P	P	P	P	P	P	P	A
Aquilegia coerulea Goldsmith	Aquca_074_00020	P	P	P	P	P	P	P	P	P	A
Basal dicot											
Amborella trichopoda	evm_27.TU.AmTr_v1.0_scaffold00076.86	P	P	P	P	P	P	P	P	P	P
Non-Grass Monocots											
Musa acuminata	GSMUA_Achr8G24160_001	P	P	P	P	P	P	P	P	P	A
Musa acuminata	GSMUA_Achr4G09980_001	P	P	P	P	P	P	P	P	P	A

Grasses											
Oryza sativa ssp. Indica	BGIOSGA014966	P	P	P	P	P	P	P	P	P	A
Oryza sativa ssp. Indica	BGIOSGA014245	P	P	A	P	P	P	P	P	P	A
Oryza sativa ssp. japonica	LOC_Os04g38570	P	P	P	P	P	P	P	P	P	A
Oryza sativa ssp. japonica	LOC_Os04g54930	P	P	A	P	P	P	P	P	P	A
Phyllostachys edulis	PH01000947	P	P	P	P	P	P	P	P	P	A
Brachypodium distachyon	Bradi5g12307	P	P	P	P	P	P	P	P	P	A
Brachypodium distachyon	Bradi5g23600	P	P	A	P	P	P	P	P	P	A
Hordeum Vulgare	MLOC_64400	P	P	P	P	P	P	P	P	P	A
Hordeum Vulgare	MLOC_56096	P	P	A	P	P	P	P	P	P	A
Aegilops tauschii	F775_21781	P	P	P	P	P	P	P	P	P	A
Aegilops tauschii	F775_17083	P	P	A	P	P	P	P	P	P	A
Triticum urartu	TRIUR3_16643	P	P	P	P	P	P	P	P	P	A
Triticum urartu	TRIUR3_16789	P	P	A	P	P	P	P	P	P	A
Zea mays	GRMZM2G125424	P	P	P	P	P	P	P	P	P	A
Zea mays	GRMZM2G072850_T01	P	P	A	P	P	P	P	P	P	A
Zea mays	GRMZM2G085236_T01	P	P	A	P	P	P	P	P	P	A
Sorghum bicolor	Sobic.006G105600	P	P	P	P	P	P	P	P	P	A
Sorghum bicolor	Sobic.006G236700	P	P	A	P	P	P	P	P	P	A
Panicum virgatum	Pavir.Ga01466	P	P	P	P	P	P	P	P	P	A
Panicum virgatum	Pavir.J34236	P	P	P	P	P	P	P	P	ND	A
Panicum hallii	Pahal.G02507	P	P	A	P	P	P	P	P	P	A
Setaria italica	Si009197m.g	P	P	P	P	P	P	P	P	P	A
Setaria italica	Si009196m	P	P	A	P	P	P	P	P	P	A

[1] I = Intron

Table 4.6 Intron structure of the *ABCB1* and *ABCB19* genes from *Arabidopsis* and rice

Species	**Locus ID**	**Gene**	**I[1] 1**	**I 2**	**I 3**	**I 4**	**I 5**	**I 6**	**I 7**	**I 8-ABCB1**	**I 8-ABCB19**	**I 9**
Oryza sativa	LOC_Os08g45030	*ABCB1*	A	A	A	A	P	A	P	A	A	A
Oryza sativa	LOC_Os04g38570	*ABCB19*	P	P	P	P	P	P	P	A	P	P
Arabidopsis thaliana	AT2G36910	*ABCB1*	P	P	P	P	P	P	P	P	A	P
Arabidopsis thaliana	AT3G28860	*ABCB19*	P	P	P	P	P	P	P	A	P	P

[1] I = Intron

PCR analysis of intron presence in *ABCB1*:

A total of 26 species were analyzed by PCR of which 22 species belonged to the order Poales, one to the Bromeliales, two to the Commelinales and one to the order Acorales (Table 4.7). Of the 22 Poales species, 19 belonged to the family *Poaceae*, one to the *Centrolepidaceae*, one to the *Cyperaceae* and one to the *Typhaceae*. *Cenchrus americanus*, *Oryza sativa* and *Zea mays*, three species for which the *ABCB1* gene structure was known, were used as controls.

For most introns, a single amplification product was obtained in the analyzed species, allowing unambiguous determination of the presence/absence of that intron. Some primer pairs, however, generated two amplicons of different sizes with one fragment corresponding in size to an intron-less amplicon and the other to an amplicon containing an intron. This was observed especially, although not exclusively, in species outside the *Poaceae* (*Carex*, *Mayaca*, *Phyllostachys*, *Tillandsia*, *Typha* and *Xyris*). Multiple fragments might have been due to allopolyploidy of the analyzed species (the ploidy level of many non-sequenced species is unknown) whereby different subgenomes carried *ABCB1* genes with a different intron composition. However, especially for the non-grass species, multiple amplicons might have been

due to non-specific amplification of other ABC genes. As the identity of these amplicons was not confirmed by sequencing, we conservatively scored intron presence/absence in species that displayed two PCR fragments as 'not determined' (ND) (Table 4.7). Exon 2 and exon 4 were small, and because introns 1, 3 and 4 were absent in all sequenced grasses initially analyzed, primers were designed to simultaneously amplify introns 1 and 2, and introns 3 and 4. Fragments of varying sizes were obtained and the size of the amplification product were used as guidance to determine the presence/absence of the individual introns. For introns 1 and 2, amplicons with a size similar to those of rice and pearl millet were considered to lack both introns 1 and 2, with a size similar to that in maize were considered to lack intron 1 but carry intron 2, and with sizes larger than that obtained in maize were scored as carrying both introns 1 and 2. We also developed a primer (ABCB1_F8; Table 4.2) that spanned the exon 1 – exon 2 boundary, and hence amplification from this primer should occur only if intron 1 was absent. Results obtained with this primer suggested that several of the larger fragments, which were assumed to carry both introns 1 and 2, likely lacked intron 1 and potentially carried a larger intron 2. For introns 3 and 4, small, intermediate and larger amplicons were obtained. Small fragments lacked both introns 3 and 4 (e.g. rice, pearl millet and maize), larger fragments were predicted to contain both introns 3 and 4, while intermediate fragments likely carried only intron 3 or intron 4. Because introns 5 and 6 were amplified individually, the presence/absence of these introns could be determined unambiguously. Several primer pairs were designed to amplify across the largest individual intron, intron 7, but all primer pairs tested generated either no amplification product or non-specific amplification products. A BLAST analysis of the 3' region of exon 7 and the 5' region of exon 8, which were used for primer design, in *Zea mays* (GRMZM2G315375) indicated high homology of these regions to other ABC gene family members. Intron 7 is positioned in an ATP-binding cassette

transporter nucleotide-binding domain which is common to all members of the ABC super-family. The primer pairs that flanked introns 8 and 9 did not amplify in the majority of non-grass species. Interpretations of the PCR results for all nine introns across the 29 species analyzed are given in Table 4.7.

Table 4.7 Intron presence/absence in *ABCB1* of the 26 monocot species determined by PCR and/or sequencing. P = presence; A = absence; ND = not determined and N/A is not analyzed.

Species	**I 1**	**I 2**	**I 3**	**I 4**	**I 5**	**I 6**	**I 7**	**I 8**	**I 9**	**Method used**
Acorus spp.	A	A	ND	ND	ND	P	ND	P	ND	PCR
Mayaca spp.	P	P	P	A	P	P	P	P	P	Sequencing
Mayaca spp.	ND	ND	P/A[2]	P/A[2]	ND	ND	ND	ND	ND	PCR
Xyris spp.	ND	ND	P	P	ND	A	ND	ND	ND	PCR
Tillandsia usneoides	A	A	P	P	ND	A	ND	ND	ND	PCR
Typha latifolia	P	P	P	P	P	P	P	P	P	Sequencing
Typha latifolia	ND	ND	P	P	ND	P	ND	ND	ND	PCR
Centrolepis spp.	ND	ND	P	P	P	A	ND	ND	ND	PCR
Ecdeiocolea spp.	A	P	P	P	P	P	P	P	P	Sequencing
Ecdeiocolea spp.	A	P	P/A[2]	P/A[2]	P	P	ND	P	ND	PCR
Pharus mezhii	A	P	P	P	P	P	ND	ND	ND	PCR
Streptochaeta spp.	A	P	P	P	P	P	P	P	P	Sequencing
Streptochaeta spp.	A	P	P	P	ND	P	N/A	P	ND	PCR
Anomochloa spp.	N/A	N/A	P/A[2]	P/A[2]	ND	N/A	ND	P	A	PCR
Oryza sativa	A	A	A	A	P	A	P	A	A	PCR
Oryza sativa	A	A	A	A	P	A	P	A	A	WGS
Aulonemia spp.	N/A	N/A	A	A	P	N/A	ND	A	ND	PCR
Bambusa spp.	A	P	A	A	P	P	ND	A	A	PCR
Phyllostachys edulis	A	P	A	A	P	P	P	A	A	WGS
Eleusine coracana	A	A	A	A	A	A	ND	A	A	WGS[1]
Eleusine coracana	N/A	N/A	A	A	ND	N/A	ND	ND	A	PCR

Zoysia spp.	A	A	A	A	A	A	P	A	A	Sequencing
Zoysia spp.	A	A	A	A	A	A	ND	A	A	PCR
Micraira spp.	ND	ND	A	A	P	P	ND	A	A	PCR
Phragmites australis	A	P	A	A	P	P	ND	A	A	PCR
Andropogon virginicus	A	P	A	A	P	P	P	A	A	WGS
Andropogon gerardii	A	P	A	A	P	P	ND	A	ND	PCR
Zea mays	A	P	A	A	P	P	P	A	A	WGS
Zea mays	A	P	A	A	P	P	ND	A	A	PCR
Arundinella hirta	A	P	A	A	P	P	ND	A	A	PCR
Paspalum vaginatum	A	P	A	A	A	P	P	A	A	Sequencing/WGS
Paspalum vaginatum	A	P	A	A	A	P	ND	A	A	PCR
Steinchisma laxum	ND	ND	A	A	P	P	ND	A	A	PCR
Brachiaria spp.	A	A	A	A	A	P	ND	A	ND	PCR
Cenchrus americanus	A	A	A	A	A	P	P	A	A	Sequencing
Cenchrus americanus	A	A	A	A	A	P	ND	A	A	PCR
Sacciolepis myosuroides	A	P	A	A	P	P	ND	A	A	PCR
Acroceras macrum	A	P	A	A	A	P	ND	A	ND	PCR
Danthoniopsis dinteri	A	P	A	A	P	P	ND	A	A	PCR

[1] Based on a limited number of 454 reads.

[2] P/A: Fragment size indicates that either intron 3 is absent and intron 4 present or *vice versa*

Validation of the PCR-predicted gene structure by sequence analysis

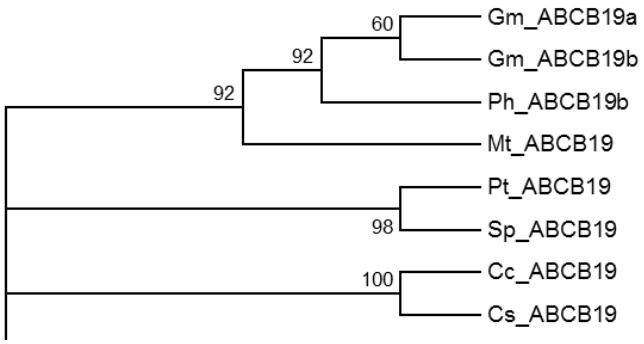
Long range amplification across all *ABCB1* introns was successful for 7 out of the 11 species selected. None of the primer pairs tested generated amplification products for *Tillandsia* and *Pharus*, while non-specific amplification products were obtained for *Brachiaria* and *Carex*. *ABCB1* amplification products generated from the species *Typha*, *Ecdieocolea*, *Streptochaeta*, *Mayaca*, *Paspalum*, *Zoysia* and *Phyllostachys* were sequenced and the sequences have been

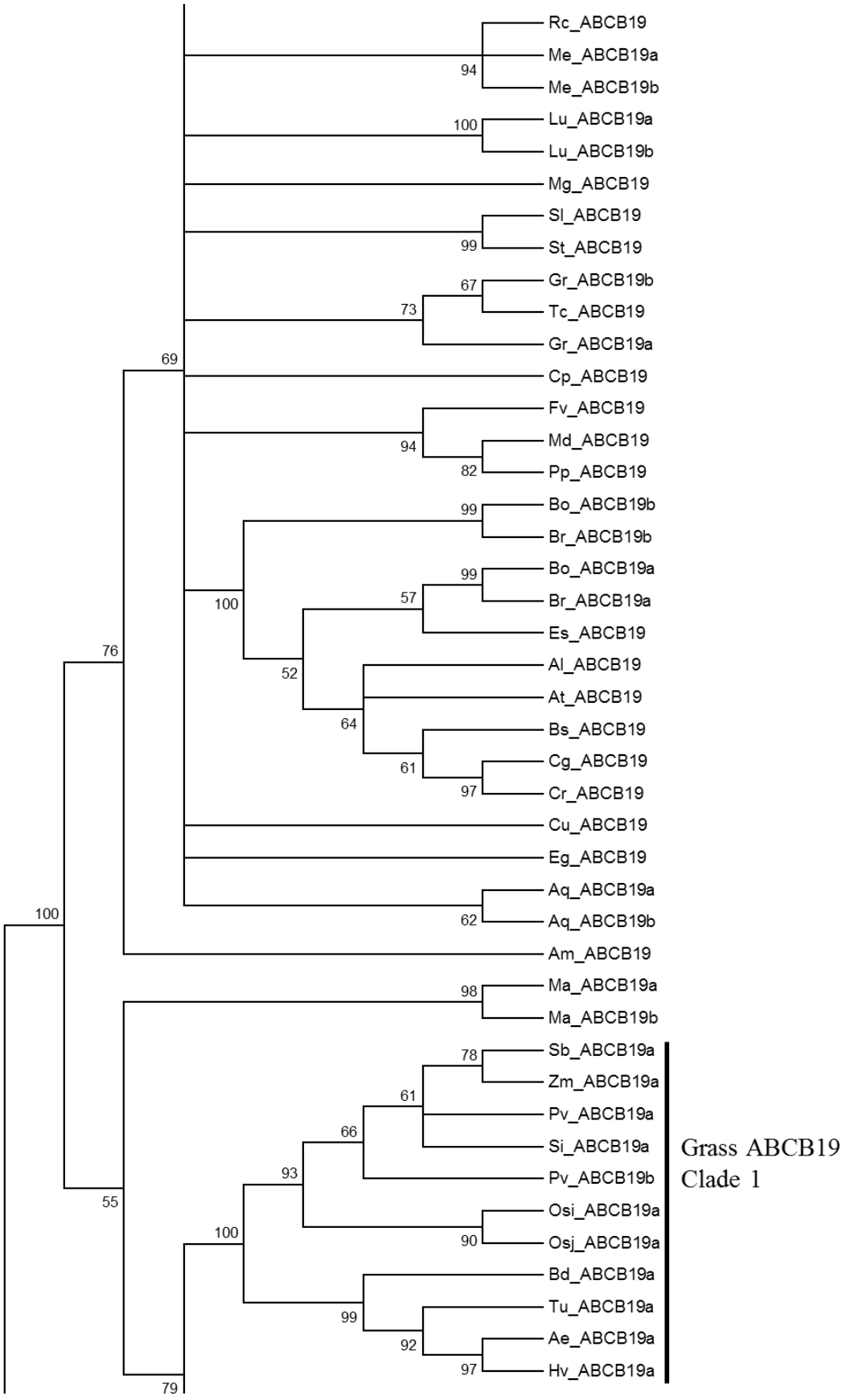
deposited in Genbank (IDs XXXXXX). The intron structure of these *ABCB1* orthologs is given in Table 4. Sequencing confirmed the PCR results for the grass species *Paspalum, Zoysia, Phyllostachys,* and *Streptochaeta*. The PCR results suggested the presence of either intron 3 or intron 4 in *Mayaca* and *Ecdeiocolea*. Sequencing confirmed that intron 3 was present and intron 4 was absent in *Mayaca* but also showed that both introns were present in *Ecdeiocolea*. Upon closer analysis, introns 3 and 4 in *Ecdeiocolea ABCB1* have a combined size of 140 bp while the size of intron 3 in *Mayaca* was 97 bp. In comparison, the combined size of introns 3 and 4 in *Streptochaeta ABCB1* was 208 bp. The intermediate size of the intron 3+4 amplicon in *Ecdeiocolea* was interpreted as the presence of a single intron but, in fact, this amplicon comprised two smaller introns. This suggests that amplicon size is not a reliable estimator for intron presence when two introns are amplified simultaneously. Overall, we observed variation for the presence of all introns except intron 7 across the sample of tested monocot species (Table 4.7).

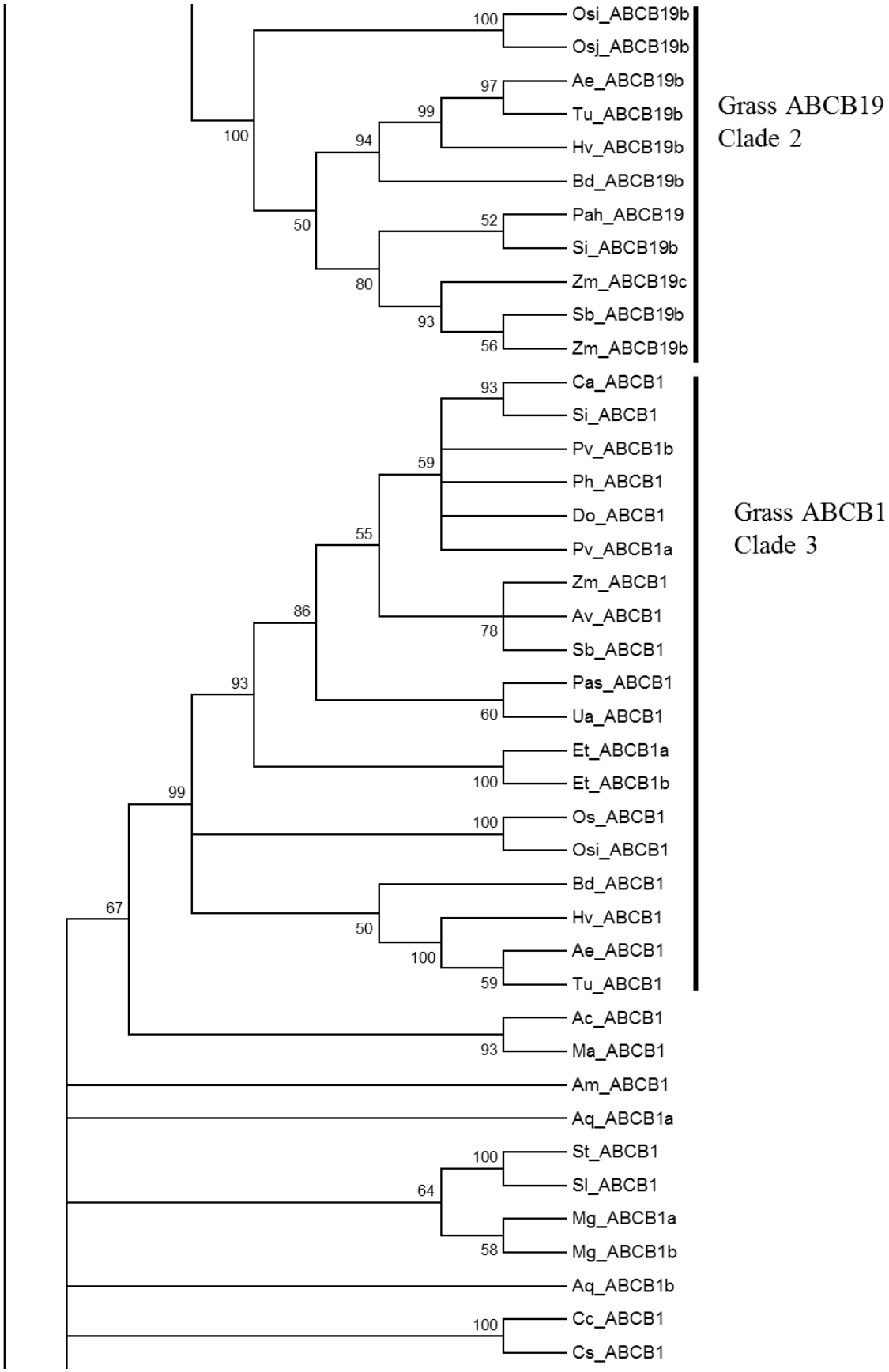
Phylogeny of ABCB1 and ABCB19

To ensure that all genes analyzed for intron loss were indeed orthologous, a neighbor-joining tree was constructed based on the corresponding protein sequences. The ABCB1 and ABCB19 proteins formed two distinct clusters (Figure 4.3). The relative position of some ABCB1 and ABCB19 proteins within each of the two clades such as the *ABCB1* and *ABCB19* genes in the grass family largely agreed with known phylogenetic relationships, thereby confirming the identity of the *ABCB1* and *ABCB19* genes. However, due to low bootstrap support, the phylogenetic placements of other proteins could not be established. *ABCB1* was present in a single copy per monoploid genome in all monocots and dicots with the exception of *Glycine max, Phaseolus vulgaris, Medicago truncatula, Mimulus guttatus and Malus domestica*. The lineages of *Glycine Phaseolus* and *Medicago* share a duplication event in a common ancestor (Clade 4; Figure 4.3).

Most members of the grass family had two copies of *ABCB19* per monoploid genome (Clades 1 and 2; Figure 4.3), indicating that *ABCB19* or a chromosome segment comprising *ABCB19* duplicated before the radiation of the grasses. One of the *ABCB19* copies in *O. sativa ssp. Japonica* (LOC_Os04g54930) was annotated to produce a truncated protein of 649 amino acids in Phytozome 10.2. Manual inspection of this gene, however, indicated that it was misannotated and represented, in fact, a full-length gene that encoded a 1268 amino acid protein. One set of duplicated *ABCB19* copies in the grasses (Clade 1; Figure 4.3) contained 9 introns while the other set (Clade 2; Figure 4.3) lacked intron 3 and contained 8 introns. The non-grass monocot *Musa acuminate* also carried two *ABCB19* copies (*ABCB19a* and *ABCB19b*), but since both copies were equidistant from the grass *ABCB19* copies, the *ABCB19* duplication in the grasses and in *Musa* represent independent events that occurred after the divergence of the two lineages. Both *Musa ABCB19* copies contained 9 introns.




Rc_ABCB19
Me_ABCB19a
Me_ABCB19b
Lu_ABCB19a
Lu_ABCB19b
Mg_ABCB19
Sl_ABCB19
St_ABCB19
Gr_ABCB19b
Tc_ABCB19
Gr_ABCB19a
Cp_ABCB19
Fv_ABCB19
Md_ABCB19
Pp_ABCB19
Bo_ABCB19b
Br_ABCB19b
Bo_ABCB19a
Br_ABCB19a
Es_ABCB19
Al_ABCB19
At_ABCB19
Bs_ABCB19
Cg_ABCB19
Cr_ABCB19
Cu_ABCB19
Eg_ABCB19
Aq_ABCB19a
Aq_ABCB19b
Am_ABCB19
Ma_ABCB19a
Ma_ABCB19b
Sb_ABCB19a
Zm_ABCB19a
Pv_ABCB19a
Si_ABCB19a
Pv_ABCB19b
Osi_ABCB19a
Osj_ABCB19a
Bd_ABCB19a
Tu_ABCB19a
Ae_ABCB19a
Hv_ABCB19a
Grass ABCB19
Clade 1


Osi_ABCB19b
Osj_ABCB19b
Ae_ABCB19b
Tu_ABCB19b
Hv_ABCB19b
Bd_ABCB19b
Pah_ABCB19
Si_ABCB19b
Zm_ABCB19c
Sb_ABCB19b
Zm_ABCB19b
Grass ABCB19
Clade 2
Ca_ABCB1
Si_ABCB1
Pv_ABCB1b
Ph_ABCB1
Do_ABCB1
Pv_ABCB1a
Zm_ABCB1
Av_ABCB1
Sb_ABCB1
Pas_ABCB1
Ua_ABCB1
Et_ABCB1a
Et_ABCB1b
Os_ABCB1
Osi_ABCB1
Bd_ABCB1
Hv_ABCB1
Ae_ABCB1
Tu_ABCB1
Grass ABCB1
Clade 3
Ac_ABCB1
Ma_ABCB1
Am_ABCB1
Aq_ABCB1a
St_ABCB1
Sl_ABCB1
Mg_ABCB1a
Mg_ABCB1b
Aq_ABCB1b
Cc_ABCB1
Cs_ABCB1

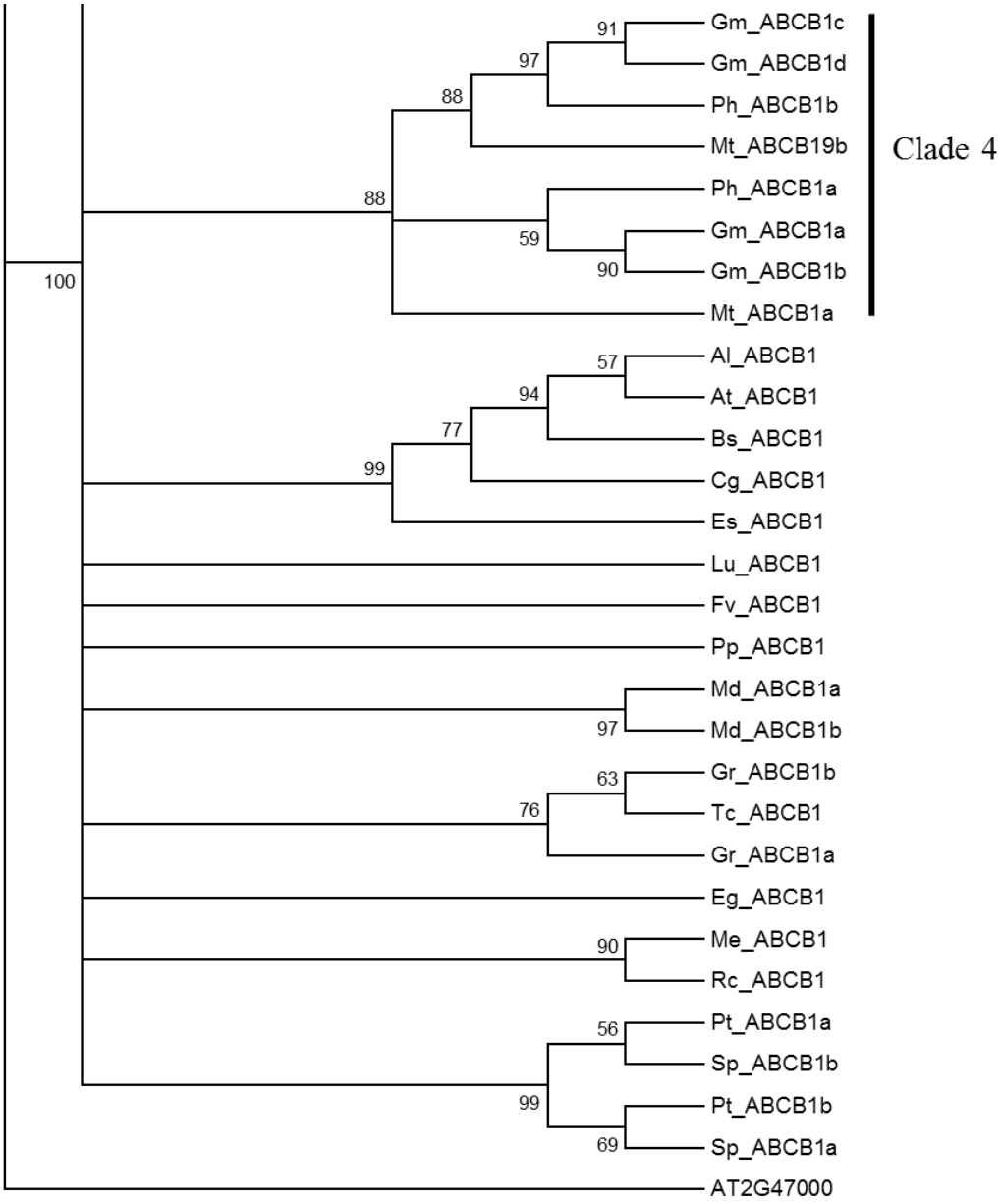


Figure 4.3 Neighbor joining tree of the ABCB1 and ABCB19 proteins in Angiosperms. Only boot-strap values higher than 50% are shown. The species abbreviations are given in Table 4.1 in materials and methods

Conserved motifs in intron 7

For the width setting of 12 letters in the MEME suite, 10 motifs were identified in intron 7 with e values < 7.7e-013 (Figure 4.4). The most common motif with a width of 12 bp is present in all 58 analyzed *ABCB1* intron 7 sequences and has an e-value of 2.1e-060. The last two base pairs of this motif are highly variable and (A/G)(G/A)GTAACATG turns out to be the 10 bp sequence, that is highly conserved across monocots and dicots. All the individual position p-values of the 58 *ABCB1* intron 7 sequences were equal to or less than 1.9e-03, which can be considered to be statistically significant. The width settings were adjusted to further analyze the sequence around this 10 bp conserved sequence and find additional motifs. For a width setting of 15 letters, these 10 bp were still contained in a 15 bp motif with the highest e-value (8.5e-075), also present in all 58 *ABCB1* intron 7 sequences. Although, the other 5 bp in this 15 bp motif did not look as conserved. When the width settings were adjusted to 18, 20, 30 and 50 letters, the sequence around the 10 bp motif is not conserved across all species. Additionally, we also a found a monocot specific motif, "TTTTGTCAGGAATGTCAAGT", present in 20 out of the 26 analyzed *ABCB1* intron 7 sequences (e-value = 2.0e-229).

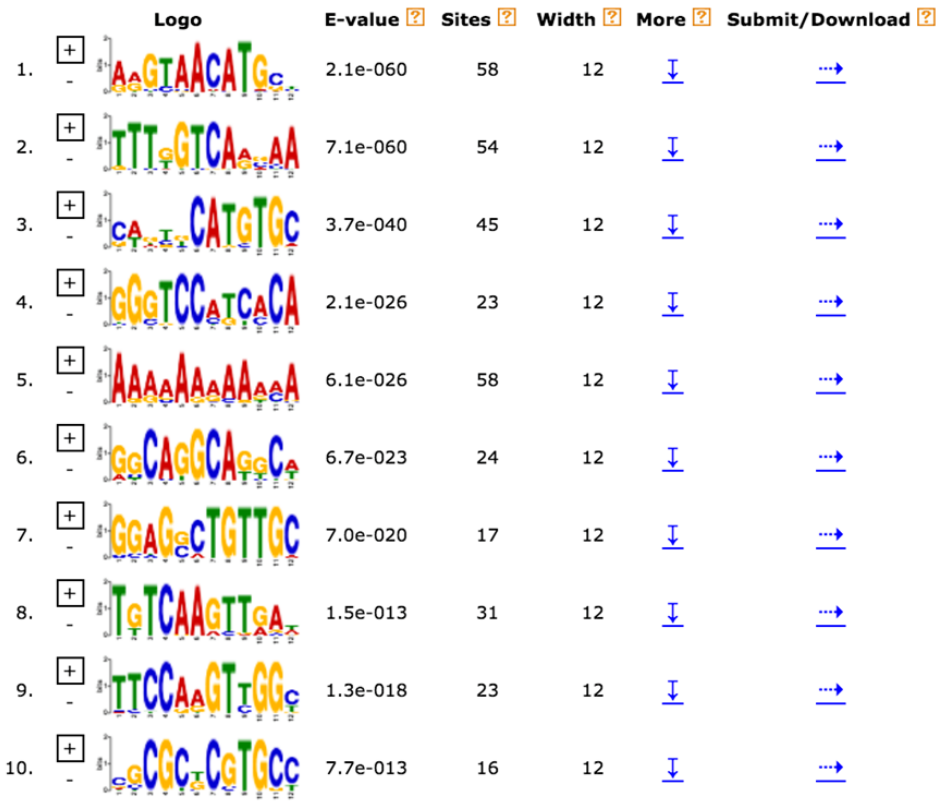


Figure 4.4 The 10 best sequence motifs identified by MEME in intron 7 using the width setting = 12.

Expression analysis of *ABCB1* and *ABCB19* in germline tissues

In rice, four and three unique probes were available for *ABCB1* and *EF1-α* respectively while only one probe each was available for *ABCB19* and *UBQ5*. However, only three (feature numbers 3137, 14315 and 27017) out of four probes showed similar expression profiles across most tissue types in *ABCB1* and two (feature numbers 5069 and 11921) out of three showed similar results in EF1-α. Across the tissues, the *ABCB1* gene is expressed at higher levels in germline tissues such as some stages of inflorescence, ovary and anther (Figure 4.5). It is also expressed at higher levels in vegetative tissues such as root, stem, and vegetative parts of a flower but shows low expression in leaf, embryo and endosperm tissues (Figure 4.5). In comparison, the *ABCB19*

gene is also expressed in germline tissues, inflorescence, anther and ovary but at much higher levels (Figure 4.4). The *ABCB19* gene is also expressed at higher levels in stages of other tissues such as embryo and root but expressed at lower levels in leaf, stem and endosperm. Both *ABCB1* and *ABCB19* are expressed at similar or higher levels to housekeeping genes (UBQ5 and EF1- α) in most stages of inflorescence, anther and ovary. In the stages of embryo, *ABCB19* is expressed at similar levels to housekeeping genes but *ABCB1* is expressed at lower levels.

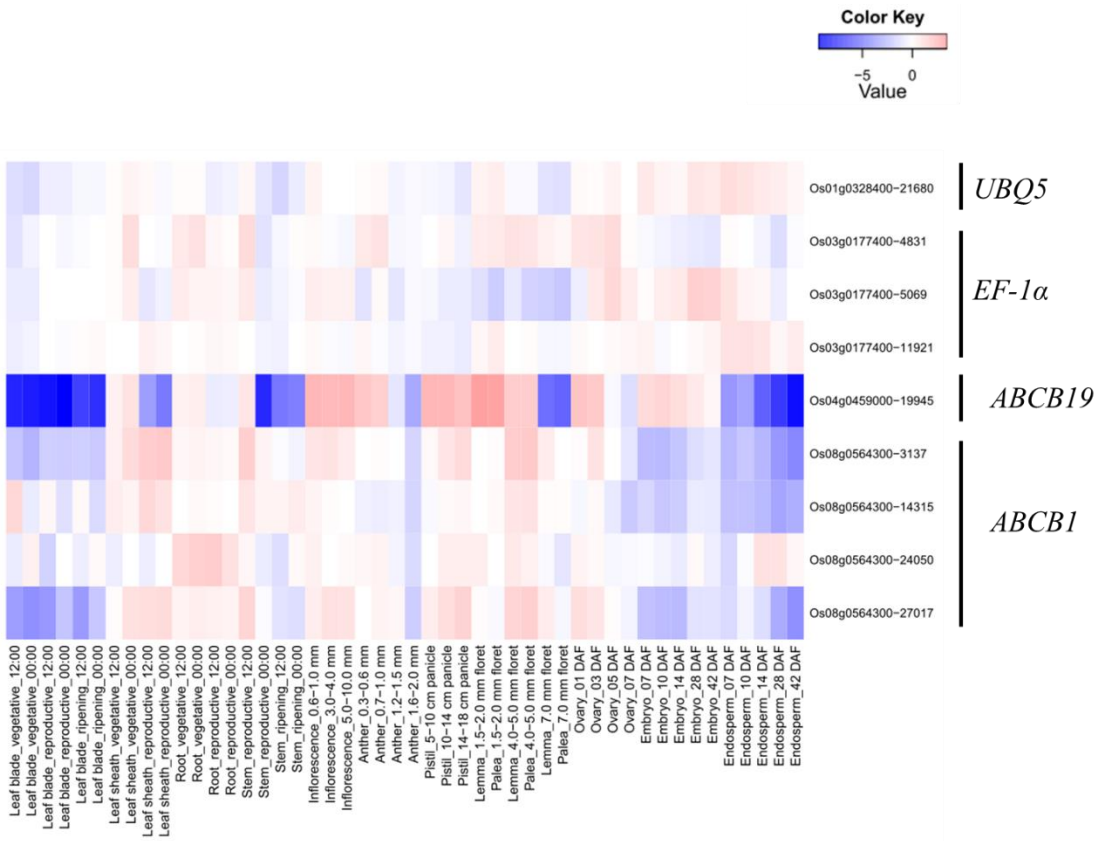


Figure 4.5 Expression profiles of *ABCB* and control genes in different rice tissues based on microarray data. Expression profiles based on unique probes are shown for ATP-binding cassette (ABC) transporter genes, *ABCB19* and *ABCB1,* and housekeeping genes, *EF-1α* and *UBQ5.*

In Arabidopsis, except in the stem tissue where *ABCB1* is expressed at much higher levels, similar expression profile was observed for *ABCB1* and *ABCB19* in other tissues. Both *ABCB1* and *ABCB19* were expressed at comparable or higher levels than the housekeeping gene ubiquitin in all tissues (Figure 4.6). The actin gene is expressed at higher levels than *ABCB1* and *ABCB19* in most tissues except apex and floral tissues. The shoot apex contains vegetative tissue and the germline inflorescence tissue.

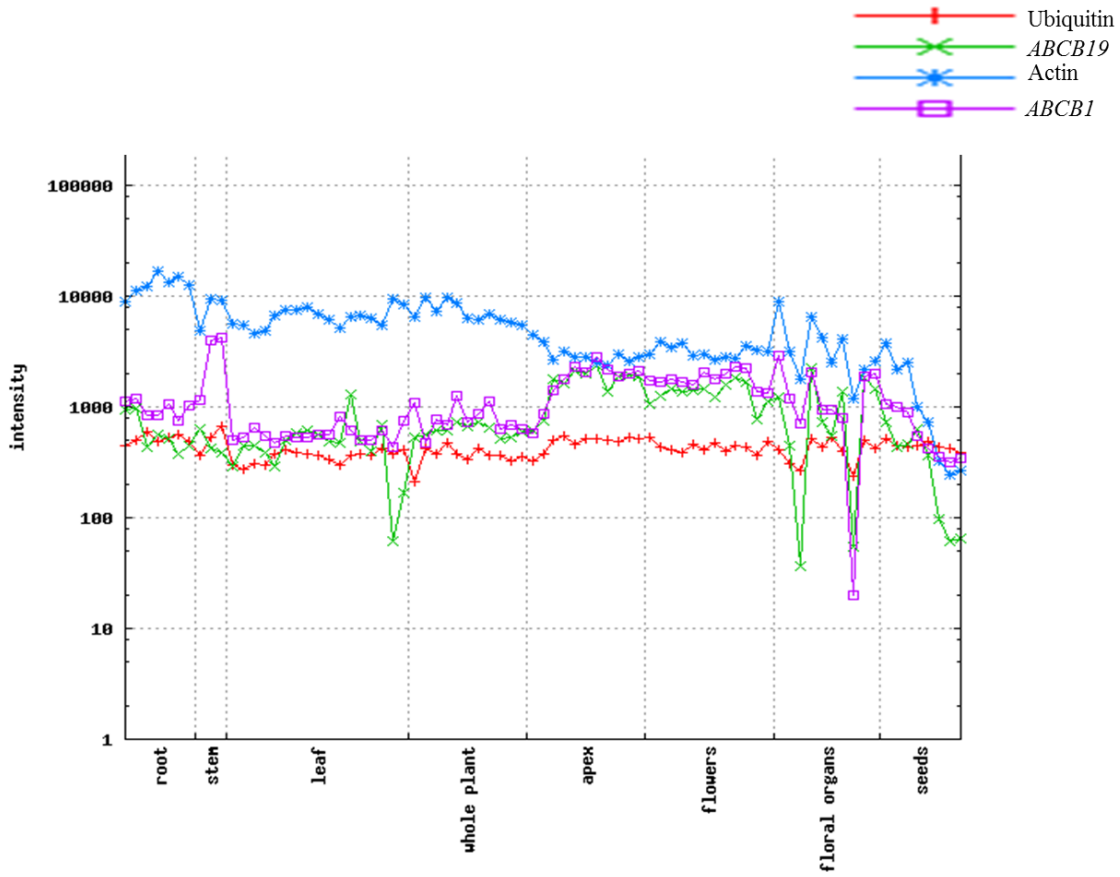


Figure 4.6 Expression profile of *ABCB* and control genes in different Arabidopsis tissues in a study based on micro array data expression. The Y axis is the absolute intensity of florescence in the microarray experiment.

DISCUSSION

The genomic structure of *ABCB1* is highly dynamic

Structural data from a range of sequenced dicot and monocot genomes (Tables 4.4 and 4.5) suggests that the ancestral *ABCB1* gene consisted of 10 exons and nine introns. In the discussion below, we will refer to introns by their ancestral intron number. The ancestral gene structure has been maintained in the majority of dicot species and non-grass monocots. In most grass species, however, the maximum number of introns in *ABCB1* is four. Although our analysis focused on only a single gene, *ABCB1*, a number of observations could be made. Firstly, the overall frequency of intron loss in *ABCB1* is 0.9% in dicots (based on 36 *ABCB1* genes in 26 dicot species) and 5.44% in monocots (based on 45 *ABCB1* genes in 42 monocot species), suggesting that intron loss is not random. Secondly, intron removal was precise in all 16 events for which sequence information was available. Thirdly, intron loss, including parallel intron loss, occurs more frequently than previously thought. An initial analysis of *ABCB1* in Arabidopsis, *Musa*, rice, *Brachypodium*, pearl millet, foxtail millet, sorghum and maize suggested that introns 1, 3, 4, 8 and 9 had been lost before the radiation of the grasses, intron 5 had been lost in the Paniceae lineage, intron 6 in the Pooideae lineage and intron 2 had been lost independently in both the Paniceae and Pooideae. However, as more genomes were added to the analysis, many additional intron loss events were observed. Intron 2 was lost independently at least 6 times in Angiosperms with the most recent loss occurring in the Melinidinae/Cenchrinae lineage after they split from the Panicinae sometime during the past 13.1 MYA (Bennetzen et al. 2012). Intron 5 was lost independently at least 4 times, with at least 3 losses having taken place within the grass sub-family Panicoidae, and intron 6 was lost at least 6 times, three times in the grass family. Intron 1 was lost 3 times, once each in the common ancestor of all grasses, in the order Acorales and in the

Bromeliales. Introns 4, 8 and 9 were each lost a total of two times in the analyzed Angiosperms. Considering the highly dynamic nature of the genomic structure of *ABCB1* in the sample of analyzed species, further expansion of the species set analyzed will likely identify additional events of intron loss.

Intron loss or gain?

Throughout our analysis, we have assumed that differential presence of introns in *ABCB1* was caused by intron loss. Although intron loss has been shown to occur more frequently than intron gain (Wang et al. 2014a), a valid concern is that some of the parallel intron loss events described for *ABCB1* could, in fact, represent single loss events that occurred early in the evolution and were followed by intron gain in specific lineages. Using known phylogenetic relationships between the species, we determined which model – one consisting solely of intron losses (Figure 4.7) or one that combined intron loss with gain – required the least number of events to explain intron evolution (Figure 4.8). This analysis was limited to the grasses because phylogenetic relationships are well established in this family, and to introns 2, 5 and 6 because these introns were differentially present in several grass lineages. Using this maximum parsimony approach, a minimum of 3 events were required to explain the presence/absence variation of intron 6 by intron loss alone but a total of 4 events (2 loss and 2 gain) were required to explain the events by a mixed model. In this particular case, the intron loss model is more likely than a mixed model of intron loss and gain. For intron 5 and intron 2, both intron loss and mixed models are equally likely as the intron presence/absence variation can be explained by 4 loss events, or 3 loss and 1 gain event. If an intron is gained independently, then the sequence of this gained intron would likely be unique. However, because introns are typically not under selective constraint, they evolve fast. Therefore, significant intron homology can only be identified between species that are closely related.

Unfortunately, the closest species to *Phyllostachys*, which putatively has gained intron 2, that have an ancestral intron 2 and for which sequence information is available are Panicoid species, which diverged ~70 MYA from the Bambusoideae lineage to which *Phyllostachys* belongs. Similarly, we currently do not have the sequence information to test if intron 5 in *Sacciolepis* could be a true gain.

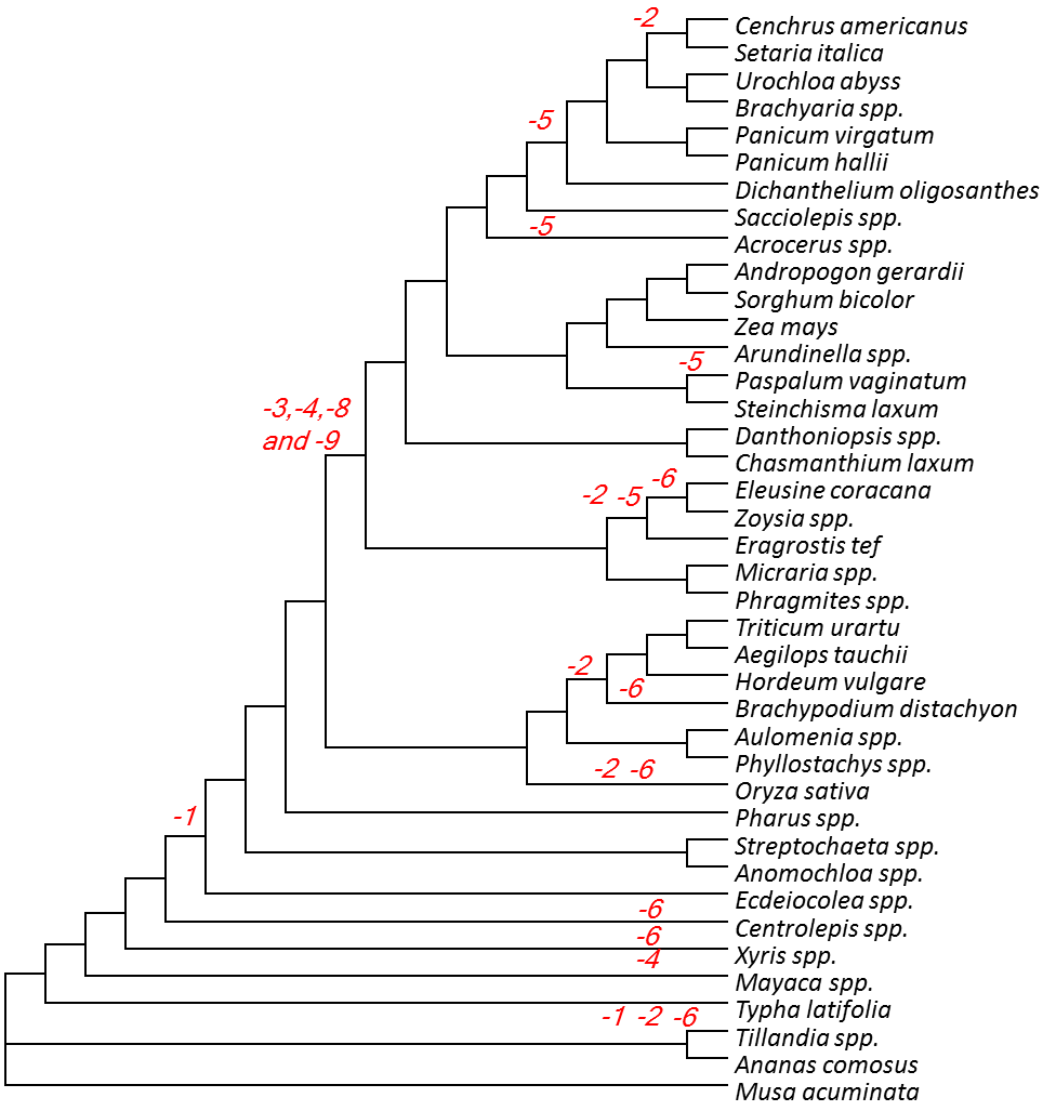


Figure 4.7 A model for *ABCB1* intron variance in the Poales explained only by intron loss.

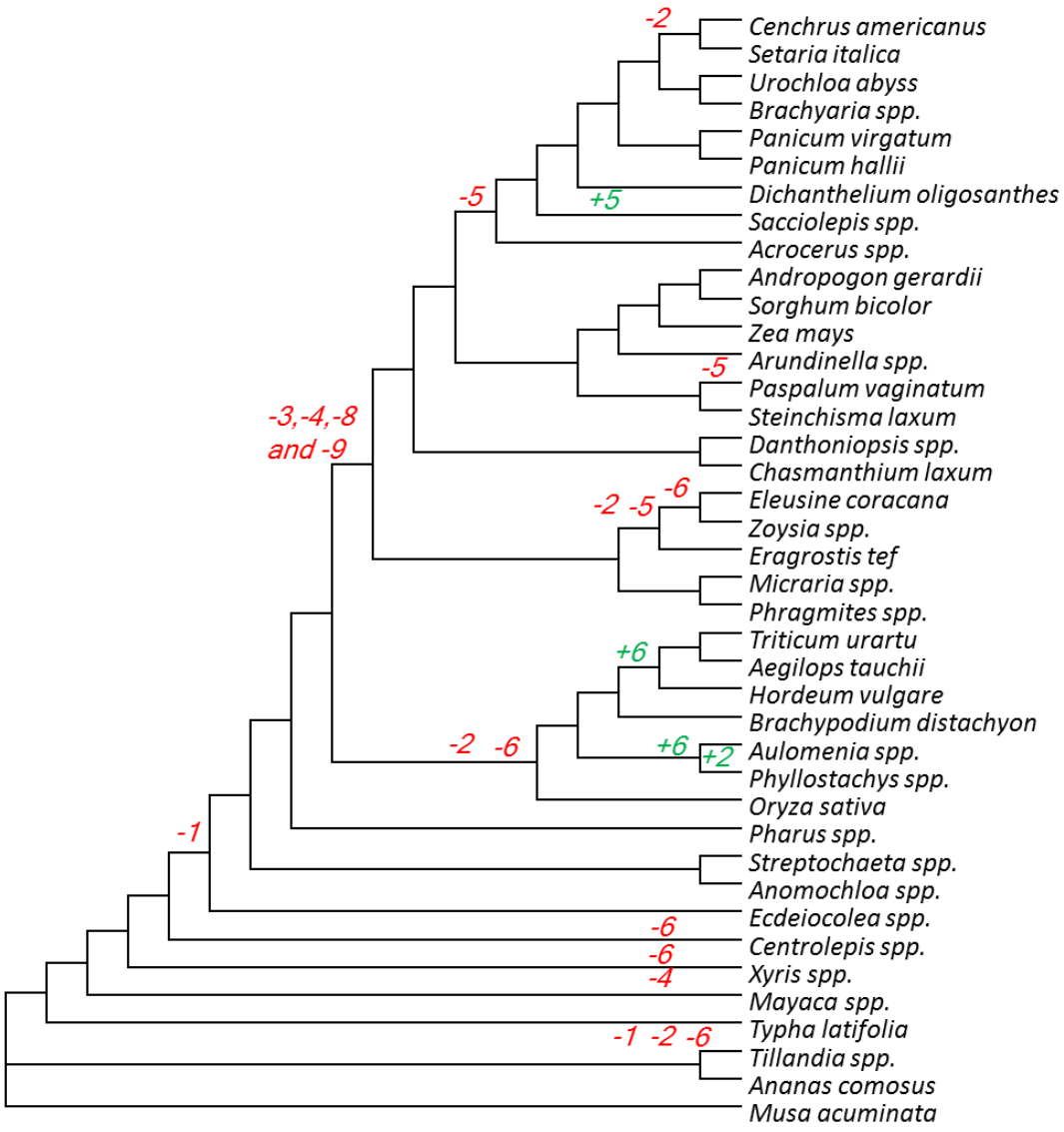


Figure 4.8 A model for *ABCB1* intron variance in the Poales explained by both intron loss and intron gain events (mixed model)

Parallel intron loss is prevalent in *ABCB1* but not in *ABCB19*

While *ABCB1* has undergone parallel loss of all introns except one (intron 7), the genomic structure of its close relative, *ABCB19*, has largely remained stable over more than 100 million years of evolution. Only two cases of intron loss were observed in *ABCB19* which, as *ABCB1*, has nine introns, and both loss events occurred after gene duplication. *ABCB19* was duplicated in the

common ancestor of the grasses, and intron 3 was lost in one of the duplicated gene copies before the radiation of the grasses. *ABCB19* was independently duplicated in the common ancestor to *Brassica rapa* and *B. oleracae*. Interestingly, both *B. rapa* copies lack intron 7, while only one of the gene copies in *B. oleraceae* lost intron 7. If intron loss occurred before the duplication of *ABCB19*, one of the *ABCB1* copies must have gained an intron in *B. oleraceae*. Alternatively, two independent losses occurred of intron 7, one in one of the duplicated *ABCB19* copies in the lineage leading to and before the divergence of *B. rapa* and *B. oleraceae*, and one in the other duplicated gene copy in the *B. rapa* lineage. While maximum parsimony cannot differentiate between the two models, the fact that intron 7 of *B. oleracae* has 82% identity to that of *Capsella rubella* indicates that both introns are related by descent and argues against the mixed intron loss/gain model. A high rate of intron loss after gene duplication has been previously recorded in *Arabidopsis* and rice (Knowles and McLysaght 2006; Lin et al. 2006).

Structure of the common ancestor to *ABCB1* and *ABCB19*

Interestingly, although both *ABCB1* and *ABCB19* carried nine introns, the positions of only eight introns were conserved across the two genes. To uncover the ancestral structure of the *ABCB1* and *ABCB19* genes, we conducted a MSA of *ABCB1* and *ABCB19* genes in rice and *Arabidopsis* with their closest known homologs. The AT1G27940 (*At_ABCB13*), AT1G28010 (*At_ABCB14*), AT1G10680 (*At_ABCB10*) and AT4G25960 (*At_ABCB2*) genes from *Arabidopsis* and LOC_Os02g46680, LOC_Os08g05690, LOC_Os08g05710 from rice are the closest homologs based on the ABCB phylogeny (Andolfo et al. 2015). As *ABCB* alias names are not assigned for the rice homologs, they were renamed as follows: LOC_Os02g46680 as *ABCB-X*, LOC_Os08g05690 as *ABCB-Y* and LOC_Os08g05710 as *ABCB-Z*. The MSA showed that the position of intron 8 in *ABCB19* (8-ABCB19) was shared by all analyzed *ABCB* genes except

ABCB-X in rice (Table 4.8). However, this homolog is also missing two neighboring introns suggesting an intron loss event that led to the removal of introns 6, 7 and 8-ABCB19 in *ABCB-X*. Sequenced grass orthologs of *ABCB-X* also lacked these introns suggesting that the loss occurred in the common grass ancestor. The presence of intron 8-ABCB19 in other *ABCB* genes indicates that this intron is ancestral.

Table 4.8 Intron structure of *ABCB* homologs in rice and *Arabidopsis*

Gene ID	**Gene Alias**	**I 1**	**I 2**	**I 3**	**I 4**	**I 5**	**I 6**	**I 7**	**8-ABCB1**	**8-ABCB 19**	**I 9**	**I 10**	**I 11**	**I 12**
AT2G36910	*At_ABCB1*	P	P	P	P	P	P	P	P	A	P	A	A	A
AT3G28860	*At_ABCB19*	P	P	P	P	P	P	P	A	P	P	A	A	A
AT1G27940	*At_ABCB13*	P	P	P	P	P	A	P	ND	P	P	A	A	A
AT1G28010	*At_ABCB14*	P	P	P	P	P	A	P	ND	P	P	A	A	A
AT1G10680	*At_ABCB10*	P	P	P	P	P	A	P	ND	P	P	P	P	P
AT4G25960	*At_ABCB2*	P	P	P	P	P	A	P	ND	P	P	P	P	P
LOC_Os08g 45030	*Os_ABCB1*	A	A	A	A	P	A	P	A	A	A	A	A	A
LOC_Os04g 38570	*Os_ABCB19*	P	P	P	P	P	P	P	A	P	P	A	A	A
LOC_Os02g 46680	*Os_ABCB-X*	P	P	P	P	P	A	A	ND	A	P	P	P	A
LOC_Os08g 05690	*Os_ABCB-Y*	P	P	P	P	P	P	P	ND	P	A	A	A	A
LOC_Os08g 05710	*Os_ABCB-Z*	P	P	P	P	P	P	P	ND	P	A	A	A	A

Exon sequences flanking intron 8-ABCB1 had very poor homology across the analyzed ABCBs. We identified that *ABCB2* and *ABCB10* in Arabidopsis carry an intron in the same location as 8-ABCB1. However, the 3' splice site positions of this intron in *At_ABCB2* and *At_ABCB10* were not conserved with that in *At_ABCB1*. Similarly, *ABCB-X*, *ABCB-Y* and *ABCB-Z* also contained an intron in this position but both 5' and 3' splice site junctions were different from those of 8-ABCB1. The *At_ABCB13* and *At_ABCB14* genes do not contain the 8-ABCB1

intron. It is possible that these introns represent novel insertions or, alternatively, that the 8-ABCB1 intron is ancestral but splice sites were mutated in some ABCB homologs during evolution. At this point, we cannot establish with certainty that the 8-ABCB1 intron is ancestral. An additional three 3' introns downstream of intron 9 were discovered in Arabidopsis homologs *ABCB10* and *ABCB2*. Based on the phylogenetic position of these homologs (Andolfo et al. 2015), it is possible that these 3' introns were a result of intron gain rather than loss. The rice homolog *ABCB-X* contains intron 10 and 11 but is missing intron 12. Also, intron 6 is only present in *Arabidopsis ABCB1* (*At_ABCB1*), rice and *Arabidopsis ABCB19* (*Os_ABCB19* and *At_ABCB19*), rice homologs *ABCB-Y* and *ABCB-Z*, and absent in all other analyzed ABCBs.

In the monocot *ABCB1* genes (32 orthologs from 27 Poales, 1 Alismatales, 1 Zingiberales and 1 Arecales species), where the majority of the loss events are reported, introns 1 to 5, 8-ABCB1 and 9 have mean and median lengths below 200 bp. Intron 6 (mean length: 382 bp and median length: 375 bp) and intron 7 (mean length: 1060 bp and median length: 1124 bp) are the larger introns in the monocot *ABCB1* gene. In monocot *ABCB19* (22 orthologs from 10 Poales and 1 Zingiberales species), introns 1, 2, 3, 5, 6, 7 and 8-ABCB19 have mean and median lengths below 200 bp while intron 4 (mean length: 812 bp and median length: 712 bp) and intron 9 (mean length: 713 bp and median length: 687 bp) are larger introns. Smaller introns have a higher tendency to undergo evolutionary loss than larger introns (Roy et al. 2003; Cho et al. 2004; Coulombe-Huntington and Majewski 2007a; Wang et al. 2014a). However, intron size clearly is not the only factor driving intron loss since most introns had comparable sizes in *ABCB1* and *ABCB19*. Within the Angiosperms, parallel loss introns were shown to have an average length of 212 bp while conserved introns had an average size of 360 bp (Wang et al. 2014a). All the parallel lost introns of *ABCB1* except intron 6 fall into the small size category. Upon closer analysis, intron 6 shows

three losses, one in the Chloridoideae lineage, one in the Ehrhartoideae lineage and one in the Pooideae lineage. The size of intron 6 in these sub-families is at most 103 bp. This suggests that a reduction in intron size likely preceded the phenomenon of evolutionary intron removal, at least in the case of intron 6 in *ABCB1*.

The largest intron, intron 7, is conserved in all the monocot species where sequence information is available. The 10 bp motif that we have identified, (A/G)(G/A)GTAACATG, does not correspond to any previously reported motif in plant species as far as we could establish. Analysis of the presence of this motif in other orthologous Angiosperm introns is currently in process to gain insights into its potential role in gene regulation. Also, we are in the process of testing the function of intron 7 using intron knock-out experiments in *Arabidopsis*.

Expression of *ABCB1* and *ABCB19* in germline tissues

At this point, it is unclear what causes the high frequency of parallel intron loss in *ABCB1*. While some intron loss was seen in *ABCB19*, it appeared to be mostly associated with gene duplication, an observation previously made by (Knowles and McLysaght 2006; Lin et al. 2006). *ABCB1*, however, is single copy in many of the species that underwent intron loss. Wang and colleagues (2014) did not identify any differences in structural characteristics between genes that underwent single intron loss and those that underwent parallel intron loss. They did, however, observe that the GC content of the 20 bp of exon sequence flanking conserved introns was some 8% (p-value < 2.2 e^{-16}) lower than that of exons flanking lost introns. The TG/CG ratio, which is a measure of historical methylation, was also significantly higher in exons flanking conserved introns compared to exons flanking lost introns (Wang et al. 2014a). The conserved introns also had a higher TG/CG ratio than the lost introns. This suggests that introns with a history of less

CG methylation are more likely to be removed. Because gene body methylation is highly correlated with expression (Wang et al. 2014b), it may simply be that more highly expressed genes have a greater chance to recombine with their cDNA leading to intron loss than lowly expressed genes. Furthermore, for intron loss to be fixed in a species, it should occur in germline tissues. Several studies have indeed reported that intron loss was more prevalent in genes representing ubiquitous housekeeping functions, including biosynthesis, metabolism, translation, transcription, and RNA processing (Coulombe-Huntington and Majewski 2007a; Zhu and Niu 2013). In Arabidopsis, genes that lost two or more introns were associated with the GO terms 'other membranes', 'transport' and 'transport activity' (Knowles and McLysaght 2006). Many members of the ABC family are associated with these functions. When we compared the expression levels of two *ABCB* genes, *ABCB1* and *ABCB19*, in *Oryza sativa*, *ABCB19* was expressed at higher levels in germline tissues than *ABCB1*. This contrasted with our expectations as *ABCB1* underwent parallel loss of multiple introns while the *ABCB19* gene structure was highly stable throughout evolution. However, both *ABCB1* and *ABCB19* are abundantly present when compared to housekeeping genes in both *Arabidopsis* and rice. It is plausible that the ancestral species that underwent a loss event in the grass lineage may have had different expression patterns of *ABCB1* and *ABCB19* than rice.

In conclusion, the *ABCB1* gene underwent dynamic structural changes in the Angiosperms, specifically in the order Poales while the structure of *ABCB19* largely remained constant. All introns in the *ABCB1* gene underwent either single (3 and 7) or parallel loss (1, 2, 4, 5, 6, 8-ABCB1 and 9) in the Angiosperms. Intron 7 is the largest intron and the only conserved intron in the grass species and testing the functional importance of this intron is currently in process. Overall, an intron loss model is more likely than a mixed model. Abundant transcript presence in germline

tissues, precise removal of introns and preferential removal of smaller introns supports an RT-based model of intron loss.

REFERENCES

Andolfo G, Ruocco M, Di Donato A, Frusciante L, Lorito M, Scala F, Ercolano MR. 2015. Genetic variability and evolutionary diversification of membrane ABC transporters in plants. *BMC Plant Biol* **15**: 51.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**(Web Server issue): W202-208.

Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J et al. 2012. Reference genome sequence of the model plant Setaria. *Nat Biotechnol* **30**(6): 555-561.

Cho S, Jin SW, Cohen A, Ellis RE. 2004. A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. *Genome research* **14**(7): 1207-1220.

Coulombe-Huntington J, Majewski J. 2007a. Characterization of intron loss events in mammals. *Genome Res* **17**(1): 23-32.

-. 2007b. Intron loss and gain in Drosophila. *Mol Biol Evol* **24**(12): 2842-2850.

Csuros M, Rogozin IB, Koonin EV. 2008. Extremely intron-rich genes in the alveolate ancestors inferred with a flexible maximum-likelihood approach. *Mol Biol Evol* **25**(5): 903-911.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**(5): 1792-1797.

Fawcett JA, Rouze P, Van de Peer Y. 2012. Higher intron loss rate in Arabidopsis thaliana than A. lyrata is consistent with stronger selection for a smaller genome. *Mol Biol Evol* **29**(2): 849-859.

Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proceedings of the National Academy of Sciences* **99**(25): 16128-16133.

Frugoli JA, McPeek MA, Thomas TL, McClung CR. 1998. Intron loss and gain during evolution of the catalase gene family in angiosperms. *Genetics* **149**(1): 355-365.

Jain M, Nijhawan A, Tyagi AK, Khurana JP. 2006. Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. *Biochem Biophys Res Commun* **345**(2): 646-651.

Knowles DG, McLysaght A. 2006. High Rate of Recent Intron Gain and Loss in Simultaneously Duplicated Arabidopsis Genes. *Molecular Biology and Evolution* **23**(8): 1548-1557.

Krzywinski J, Besansky NJ. 2002. Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. *Mol Biol Evol* **19**(3): 362-366.

Lin H, Zhu W, Silva JC, Gu X, Buell CR. 2006. Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol* **7**(5): R41.

Parvathaneni RK, Jakkula V, Padi FK, Faure S, Nagarajappa N, Pontaroli AC, Wu X, Bennetzen JL, Devos KM. 2013. Fine-mapping and identification of a candidate gene underlying the d2 dwarfing phenotype in pearl millet, Cenchrus americanus (L.) Morrone. *G3 (Bethesda)* **3**(3): 563-572.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* **13**(17): 1512-1517.

Roy SW, Fedorov A, Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proceedings of the National Academy of Sciences* **100**(12): 7158-7162.

Roy SW, Gilbert W. 2005a. The pattern of intron loss. *Proceedings of the National Academy of Sciences of the United States of America* **102**(3): 713-718.

Roy SW, Gilbert W. 2005b. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc Natl Acad Sci U S A* **102**(16): 5773-5778.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* **7**(3): 211-221.

Roy SW, Penny D. 2007. Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of O. sativa and A. thaliana. *Mol Biol Evol* **24**(1): 171-181.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of Arabidopsis thaliana development. *Nature genetics* **37**(5): 501-506.

Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. 2005. Conservation versus parallel gains in intron evolution. *Nucleic Acids Res* **33**(6): 1741-1748.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*.

Wang H, Devos KM, Bennetzen JL. 2014a. Recurrent Loss of Specific Introns during Angiosperm Evolution. *PloS Genet* **10**(12): e1004843.

Wang J, Marowsky NC, Fan C. 2014b. Divergence of Gene Body DNA Methylation and Evolution of Plant Duplicate Genes. *PloS ONE* **9**(10): e110357.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**(9): 1189-1191.

Yuan Y, Chung JD, Fu X, Johnson VE, Ranjan P, Booth SL, Harding SA, Tsai CJ. 2009. Alternative splicing and gene duplication differentially shaped the regulation of isochorismate synthase in Populus and Arabidopsis. *Proc Natl Acad Sci U S A* **106**(51): 22020-22025.

Zhan L-L, Ding Z, Qian Y-H, Zeng Q-T. 2012. Convergent Intron Loss of MRP1 in Drosophila and Mosquito Species. *Journal of Heredity* **103**(1): 147-151.

Zhu T, Niu D-K. 2013. Frequency of intron loss correlates with processed pseudogene abundance: a novel strategy to test the reverse transcriptase model of intron loss. *BMC Biology* **11**(1): 23.

CHAPTER V:

THE ROLE OF INTRON SPLICING EFFICIENCY ON THE GENOMIC LOSS OF INTRONS

ABSTRACT

The mechanism of evolutionary loss of introns is not well understood. Precise removal of introns, removal of adjacent introns and 3' bias towards intron removal support a model of gene conversion with a reverse transcribed copy of fully or partially spliced mRNA. In this study, we analyzed the pre-processed RNA (pre-mRNA) isolated from the nucleus in two grass species, rice and sorghum. As expected, the nuclear pre-mRNA has more read depth in the introns than the cytoplasmic fraction which made it possible to compare reads between the conserved and lost introns. The reads in introns and exons were studied in a set of 151 conserved genes and 142 genes which underwent intron loss in one or more of the 5 grass species rice, maize, sorghum, *Setaria* and *Brachypodium*. The genes that underwent intron loss showed higher expression than the conserved set of genes and their introns also contained higher reads. In a subset of 90 genes that contain both loss and conserved introns, the intron size and the relative position did not have an effect on the number of reads. Our findings indicate that the splicing efficiency does not play a role in evolutionary loss of introns.

INTRODUCTION

A typical eukaryotic gene is comprised of protein coding exons separated by non-coding introns. The number of intron-containing genes, introns per gene and the average intron length varies between species (Fedorova and Fedorov 2003; Hong et al. 2006; Jeffares et al. 2006). The

origin, evolution and functions of introns are still not fully understood (Rogozin et al. 2012) but it is generally accepted that introns were present at early stages of eukaryotic evolution. Interkingdom conservation of intron positions in orthologous genes provides strong support for the intron-early theory as random parallel insertions in the same position by chance are unlikely (Fedorov et al. 2002; Rogozin et al. 2003). Some lineages have undergone massive loss of introns while others have maintained ancestral intron numbers. Genome-wide studies of intron loss in mammalian lineages of mouse, rat, dog and human which shared a common ancestor 95 MYA revealed only 122 cases of intron loss and these were largely confined to the rat lineage (Roy et al. 2003a; Coulombe-Huntington and Majewski 2007a). Conversely, in a genome-wide study between the different *Drosophila* species which diverged 40 MYA ago, a high number (1,754) of intron loss events were identified (Coulombe-Huntington and Majewski 2007b).

Several studies comparing the structure of orthologous genes have been performed in plant species as well. A study comparing wheat genes with their potential orthologs in rice and *Brachypodium* showed that some 8.5% of the analyzed genes had an altered structure (Akhunov et al. 2013). However, most of these genes were pseudogenes with premature termination codons or were located in non-syntenic positions and may thus not have been true orthologs. A genome-wide comparison of the sequenced grasses *Brachypodium*, rice, maize, sorghum and *Setaria* which shared a common ancestor ~70 MYA found only 883 cases of intron presence/absence variation (Wang et al. 2014). Although the structure of orthologous genes thus appears to be mostly conserved within the grasses, other plant lineages showed a higher frequency of loss of introns. For example, a higher rate of intron loss (~25 fold) was observed in *Brassica rapa* in comparison to its close relative *Eustrema salsugineum* (Milia et al. 2015).

Gene structure evolution through intron gain or loss may be a way to enhance the functional complexity of a species with limited gene resources. Whole genome, segmental and gene duplication events accelerate this process of structural evolution as evidenced by the rampant loss/gain patterns seen in duplicated genes (Knowles and McLysaght 2006; Lin et al. 2006; Kawaguchi et al. 2010; Zhou et al. 2012). Loss of an intron in the 3' untranslated region following gene duplication led to the embryonic expression of the duplicated *Bmglv2* member of the *gloverin* family of antibacterial genes in *Bombyx mori* (Mrinal and Nagaraju 2008). Similarly, hatching enzyme genes in Teleosti that underwent intron loss express at higher levels than genes which maintained the ancestral intron structure (Kawaguchi et al. 2010). Alternately, a non-adaptive theory of evolution for complexity was proposed by Lynch and colleagues asserting that purifying selection is effective only to remove introns from populations with a large effective population size such as unicellular eukaryotes but not from vascular plants or vertebrates which have small effective population sizes (reviewed in Rogozin et al. 2012). In these more complex eukaryotes, intron loss more likely is a consequence of genetic drift rather than selection. For an altered gene structure to be fixed in a species, it is essential that the change occurs in a germline tissue.

There are two proposed mechanisms for the genomic loss of introns: 1) Gene conversion with a reverse-transcribed copy of a spliced mRNA (RT-mRNA) and 2) genomic deletions via DNA breakage and repair (reviewed in Roy and Gilbert 2006). Genomic deletions lead to precise or imprecise removal of the introns and the precision of intron removal is dependent on the micro-homology existing at the splice junctions (Farlow et al. 2011). Genomic deletions are speculated to be the major mechanism of intron removal in Arabidopsis (Fawcett et al. 2012), *Drosophila* and mammals (reviewed in Farlow et al. 2011). The signature of RT-mediated intron loss is precise removal of introns, removal of adjacent introns and bias towards the removal of 3' introns.

Overall, this mechanism is considered the predominant process through which genomic introns are removed (Roy et al. 2003b; Coulombe-Huntington and Majewski 2007a). When the same intron is lost independently at least two times evolutionarily, then the intron is defined as a "parallel loss intron". Parallel loss of introns is rarer than single loss of introns (lost only once). Only 93 cases of parallel loss have been reported in the grass family (Wang et al. 2014). The parallel loss events in *GAPDH* in mammals (Coulombe-Huntington and Majewski 2007a), *white* gene in the dipteran lineages, *MRP1* gene in *Drosophila* and mosquito species (Krzywinski and Besansky 2002) and the isochorismate synthase gene in Angiosperms (Yuan et al. 2009) are other reported examples of parallel intron loss. We conducted a detailed study of the structure of *ABCB1* genes in Angiosperms and demonstrated that *ABCB1* underwent parallel loss of 7 out of 9 introns and single loss of the other two introns while the structure of its closest homolog *ABCB19* remained largely constant over more than 100 million years of evolution (Chapter 4).

While many studies have assessed the prevalence of intron loss in eukaryotes and the characteristics of lost introns, none have investigated whether there is a link between RT-mediated intron loss and RNA processing. Since the rate with which introns are spliced out will affect the structure of semi-processed RNA, it is not inconceivable that splicing efficiency and genomic loss of introns may be correlated. To perform this type of study, it is imperative that we capture and sequence mRNA at various levels of processing. RNA-seq analysis of pre-mRNA (nascent RNA) has been conducted in recent years in several animal species including *Drosophila*, mouse and humans (Ameur et al. 2011; Khodor et al. 2011; Khodor et al. 2012). Observations of these studies include 1) that there is a vast difference in splicing efficiency between the different introns 2) that the size, position and length of an intron determine the splicing efficiency and 3) that co-transcriptional splicing is commonly seen in humans and *Drosophila* while post-transcriptional

splicing is prevalent in mouse tissue. The process of splicing *i.e.*, if it occurs co-transcriptionally or post-transcriptionally is not known in plants and may be species dependent. To gain insight into RNA processing in grasses, we sequenced nascent RNA in rice and sorghum. This data was used to test our hypothesis that introns that underwent evolutionary loss in the grasses were removed earlier during the splicing process than evolutionarily conserved introns.

MATERIALS AND METHODS

Plant material and growth conditions

Plant material of the sequenced accessions *Oryza sativa* L. *japonica* cultivar 'Nipponbare' (Goff et al. 2002) and *Sorghum bicolor* cultivar 'Btx-623' were used for this study (Paterson et al. 2009). The plants were grown in a controlled environment chamber under a day length of 16 hours, temperature of 28 ^{0}C and relative humidity of 70 percent. For sorghum, an in-house pine-bark mix was used and no fertilizer was added. Rice was sown in a mix of three parts fafard soil and one part turf-face, and supplemented with an iron micronutrient (Sprint® 330 Iron Chelates, BASF Corporation, Research triangle Park, NC) following germination. One application of a slow release fertilizer (osmocoat) 7 days after germination and weekly applications of 20-10-20 fertilizer were made. Above ground tissue was harvested of 9 day old sorghum seedlings and 28 day old rice plants and flash frozen in liquid nitrogen. Plants were grown in two batches with batch 1 providing the material for replicate 1, and batch 2 the material for replicates 2 and 3.

Nuclei isolation from rice and sorghum

A modified protocol from Luthe and Quatrano (1980) was used to isolate nuclei. The modifications include a more thorough cell disruption using liquid nitrogen, pestle and mortar and employment of fewer percol-gradient layers for nuclei recovery. In brief, crushed plant material (approximately 7 gm) was suspended in ~7 volumes (50 ml) of Honda buffer with spermine. All

the following steps were performed on ice. The suspended plant material was further crushed using a Polytron (3 pulses for 60 seconds on setting 7). The lysate was passed through 2 layers of cheese cloth, 2 layers of Miracloth and a 40 µm cell strainer (Falcon Cell Strainer™, Fisher Scientific). The flow-through was centrifuged at 3,600xg for 10 minutes at 4 ^{0}C. The supernatant was discarded and the pellet was resuspended in 10 ml of Honda buffer without spermine. The nuclei were layered on a sucrose-percol gradient consisting of a layer of 2 M sucrose, a layer of 80% percol and a layer of 40% (sorghum) or 30% (rice) percol, and centrifugation was performed at 4100xg for 30 minutes. For sorghum, the 80% percol layer was carefully removed and diluted with an equal volume of Honda buffer without spermine. For rice, both the 30% and 80% percol layers were collected. Samples were centrifuged at 3600xg for 20 minutes. Pellets were washed twice with Honda buffer without spermine and twice with nuclear resuspension buffer; centrifugation was at 5860xg for 5 minutes. Finally, pellets were resuspended and stored in 100-200 µl of nuclear resuspension buffer. Buffer compositions and concentrations of working solutions are provided in the Luthe and Quatrano (1980) protocol.

Quality control of the extracted nuclei

The quality of the nuclei and the extent of cellular contamination were checked under a microscope. Nuclei were counted using a hemacytometer (Bright-Line™ Hemacytometer, Sigma-Aldrich). About 10 µl of the extracted nuclei, which corresponds to at least 10^6 nuclei for rice and $5x10^6$ nuclei for sorghum, were added to an equal volume of loading buffer (50 mM Tris-HCL pH 6.8, 2% SDS, 10% glycerol, 1% β-mercaptoethanol, 12.5 mM EDTA, 0.02% bromophenol blue) and used for protein analysis. Control samples consisted of 50 mg of ground tissue suspended in 100 µl of loading buffer. The samples were heated at 95 ^{0}C for 5 minutes, centrifuged and the supernatant was collected. Different amounts of the total crude extracted protein (extracted from

3.5 mg, 2.5 mg and 0.5 mg of tissue), and the rice and sorghum nuclear protein samples (extracted from an estimated 10^6 nuclei for rice to $5x10^6$ nuclei for sorghum) were run on a 10% SDS-PAGE gel and transferred to a nitrocellulose membrane (Amersham Hybond™-P, GE Healthcare, Piscataway, NJ) using the Mini Trans-Blot electrophoretic transfer cell (Biorad, Hercules, CA) using the manufacturer's protocol. The membrane was probed with the nuclear antibody histone 3 (H3) (Abcam, Cambridge, MA) to check for nuclei integrity and with an antibody against the chloroplast protein, phosphoenolpyruvate carboxylase (PEPC) (Rockland antibodies and assays, Limerick, PA) to check for the extent of cytoplasmic contamination.

RNA extraction and preparation of sequencing libraries

Total RNA was extracted from ground plant tissue (~80 mg) using the Qiagen RNeasy Plant Mini Kit (Qiagen, Germantown, MD) to provide cytoplasmic RNA. The INTACT RNA isolation protocol was used for the nuclear RNA extraction (Deal and Henikoff 2011). All RNA extractions were done in triplicate. Up to 5 µg of RNA was treated with the Turbo DNA-*free*™ kit (Life technologies, Grand Island, NY) to remove DNA contamination. The RNA integrity number (RIN) was checked using the Agilent 2100 Bioanalyzer Nano chip (Agilent Technologies, Santa Clara, CA) and the concentration was estimated using the Qubit RNA HS assay kit (Life Technologies, Grand Island, NY). The manufacturer's protocol was followed for all kits used. To obtain poly(A) enriched mRNA, approximately 1 µg of high quality RNA was used to construct the Illumina sequencing libraries using the KAPA Stranded mRNA-Seq Kit, with KAPA mRNA Capture Beads (Kapa Biosystems, Wilmington, MA). Alternately, rRNA was removed from 1 µg of cytoplasmic or nuclear RNA using the Epicentre rRNA removal kit-plant leaf (Epicentre, Madison, WI) followed by preparation of libraries using the KAPA Stranded mRNA-Seq Kit (Kapa Biosystems, Wilmington, MA). Lower quantities of RNA were used for replicates 1 in both

rice and sorghum (Table 5.1). For all library preparations, RNA was fragmented to a size of 200 to 300 bp by heating in the presence of Mg^{2+} using the KAPA fragment, prime and elute buffer following the manufacturer's recommendations. The libraries were differentially barcoded with dual I5/I7 Illumina indices (Table 5.1). The libraries were checked on a fragment analyzer to confirm the insert size, and quantified via Qubit and qPCR by the Georgia Genomics Facility (GGF) at the University of Georgia (UGA). For a preliminary analysis, the rice and sorghum libraries generated from replicate 1 were pooled and sequenced on a NextSeq (150 cycles) Mid Output Flow Cell (PE 75). The libraries from the two additional replicates were pooled and sequenced on a NextSeq (150 cycles) high output flow cell (PE 75).

Processing of the reads

Barcodes and adapter contamination were removed from the RNA-Seq reads using the Trim Galore tool from the Babraham Institute, Cambridge, UK (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Sorghum genome sequence assembly and gene set v1.4 (GFF format) were downloaded from Phytozome v9. Annotated rice pseudomolecules (*Oryza sativa*, IRGSP/RAP Build 5) were downloaded from the International Rice Genome Sequencing project (IRGSP; http://rgp.dna.affrc.go.jp/). Sequences of gene loci were extracted to generate FASTA files with the sequences of all annotated genes of the rice and sorghum genomes. Coordinates for the intron and exon locations in each gene model were also extracted and written into files with suffixes .exon.gff and .intron.gff.

Default settings of Tophat 2 were used to align the cleaned RNA-Seq reads onto the extracted rice and sorghum gene sequences. The number of reads mapped to introns and exons as defined by the coordinates in the .exon.gff and .intron.gff files were counted using an in-house perl script and converted to 'reads per kilobase per million mapped reads' (RPKM) values using the

formula RPKM = (10^9xC)/(NxL) where C= the number of reads mapped to an intron or exon, N= the total number of mapped reads in the experiment and L= the intron/exon length in base-pairs.

Statistical analyses

A total of 151 genes with a conserved structure which were randomly selected from a bigger set of conserved genes and 142 genes that had undergone loss of an intron in at least one of five grass species (*Brachypodium distachyon*, rice, sorghum, maize and *Setaria italica*) were used in our analysis. Only intron sites that show high collinearity in the flanking exons between 5 grasses, banana and *Arabidopsis* were considered as true introns and included in the analysis. All other introns were removed. For example, even if an intron site shows collinearity between rice and sorghum but the flanking exons do not show high alignment score in multiple alignment with other species, it was removed. Out of the 142 genes with lost introns, 44 genes contained only a single intron and intron loss had occurred in only a single species (single loss (SL)), three genes contained only a single intron and that intron had been lost independently in at least two species/lineages (parallel loss (PL)). 81 genes carried multiple introns, some of which were conserved while others underwent single loss (SL+con). Eight genes carried multiple introns, some of which were conserved while others underwent parallel loss (PL+con). Four genes carried multiple introns including conserved, single loss and parallel loss introns (PL+SL+con), one gene with multiple introns carried single and parallel loss introns but no conserved introns, and one gene that underwent loss of the same intron in both rice and sorghum and contained one conserved intron. Three scenarios for the lost introns exist; 1) The intron loss occurred in a grass species other than rice or sorghum and intron read coverage can thus be calculated for both rice and sorghum; 2) The intron loss occurred in rice and the read count for this intron can be calculated in

sorghum or *vice versa* and 3) The intron loss occurred in both rice and sorghum in which case there is missing data for this intron in both species.

Exon and intron RPKM values of conserved genes (151 genes) and intron loss genes (142 genes) were compared for the whole dataset within RNA isolation method (ribosomal RNA depletion or polyA selection), within species (rice or sorghum) and within RNA source (cytoplasmic or nuclear RNA). The effect of RNA type (cytoplasmic rRNA depleted RNA, nuclear rRNA depleted RNA, cytoplasmic polyA RNA and nuclear polyA RNA) on read coverage across introns was performed on the whole dataset (293 genes) within each species. Effects of species, method, and type of introns (conserved introns, SL introns PL introns), intron position (introns located in the 5' half of a gene *vs.* introns located in the 3' half of a gene), and intron size on read coverage were tested on the nuclear RNA-Seq data of the 90 genes that comprised both conserved and lost introns. The effect of intron type on read coverage was also carried out on a gene-by-gene basis. Intron read coverage for different gene scenarios (genes where the lost introns had higher RPKM values than the conserved introns; genes where the conserved introns showed a higher RPKM value than the lost introns and genes where no difference was observed in read coverage between conserved and lost introns) were also compared. The effect of gene type (SL, PL, PL+con, SL+con, and PL+SL+con) on read coverage was tested on nuclear RNA-Seq data for 141 genes that underwent intron loss. For all comparisons, we performed (M)ANOVAs to test for main and interaction effects of factors on RPKM values which, when relevant, were followed by Tukey's HSD multiple comparison tests at a 5% significance level. Method, species, RNA type, gene type, intron type, gene scenario, intron position and gene structure were treated as fixed factors. Intron size was added as a covariable. Analyses were carried out using R package version 3.1.3.

RESULTS

Nuclei extraction and quality assessment

Most nuclei in sorghum were present in the 80% percol layer, whereas rice nuclei were found in both the 30% (Figure 5.1A) and 80% percol layers (Figure 5.1B). The yield of clean nuclei ranged from $5x10^5$ to $7x10^6$ per gram of tissue in sorghum and $7x10^5$ to $1.4x10^6$ nuclei per gram of tissue in rice for the three replicates. Western hybridization showed that nuclear extracts had lower levels of PEPC and higher levels of histone H3 compared to the crude extracts (Figures 5.2A and 5.2B) indicating that they were enriched in nuclei.

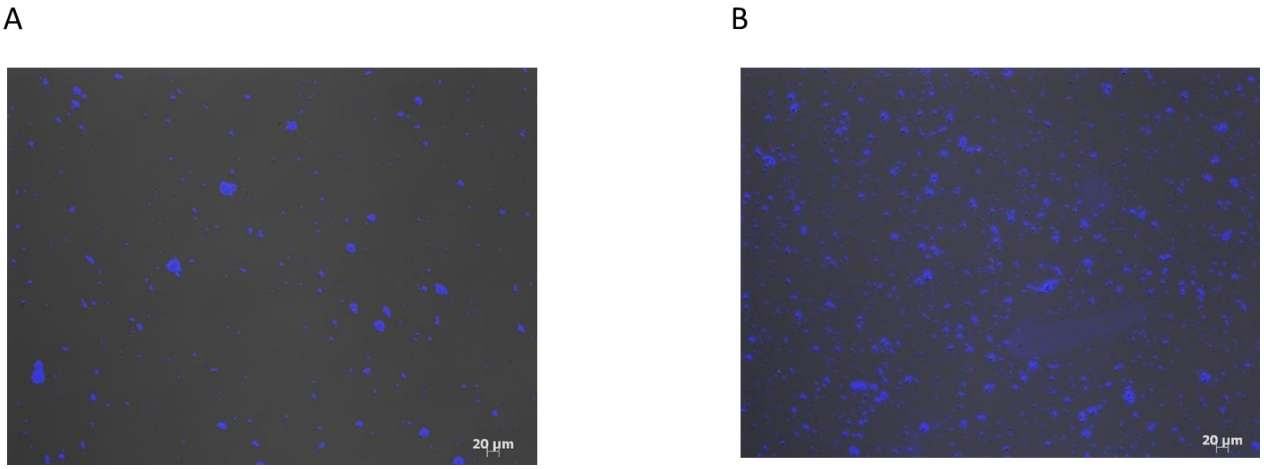


Figure 5.1 Rice nuclei in (A) the 30% percol layer and (B) the 80% percol layer. The nuclei were stained by DAPI (1:100 dilution). A 20 µm scale is presented in the bottom right hand corner of the pictures.

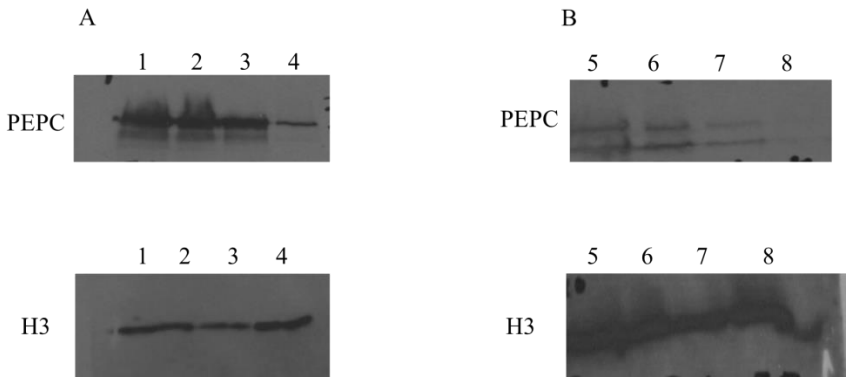


Figure 5.2 Results of western hybridization with PEPC and H3 for rice and sorghum. (A) sorghum (replicate 2) 1. Total crude extract 3.75 mg 2. Total crude extract 2.5 mg 3. Total crude extract 0.5 mg and 4. Sorghum nuclei estimated at 5 X 10^6 nuclei; and (B) rice (replicate 3) with 5. Total crude extract 3.75 mg 6. Total crude extract 2.5 mg 7. Total crude extract 0.5 mg and 8. Rice nuclei estimated at 10^6 nuclei. The total protein was not quantified in the nuclei or crude extracts.

Quality of RNA and RNA-Seq read outputs

The RIN scores for each of the RNA samples are given in Table 5.1. The nuclear RNA samples consistently had lower RIN values (range 6.4-7.1) than the cytoplasmic RNA samples (range 8.1-8.6). It is unclear whether the differences in RIN values are due to technical or biological reasons. The insert sizes of the Illumina libraries were in the range 167-538 bp for the replicate 2/replicate 3 pool with the peak at 281 bp. The number of raw reads and cleaned reads for each sample are given in Table 5.1. All but two samples (rice.total.ribozero.dup3 and sorghum.total.ribozero.dup3) yielded at least 10 million reads per sample. Replicates 2 and 3 have higher read outputs per sample as they were sequenced on a high output flow cell.

Table 5.1: Read and quality measurements of the different replicates

#	Sample	Initial quantity (ng)	RIN score	I7 Index	I5 Index	Untrimmed read count	Cleaned read count
1	rice.nuclei.polyA.dup1	300	6.75	AAGGCGTT	TTCGAAGC	13448142	13366941
2	rice.nuclei.polyA.dup2	1000	6.6	CAGAACTG	GTTCCATG	33325167	33118382
3	rice.nuclei.polyA.dup3	1000	7.1	ACACCGAT	TAGCTGAG	36171186	36013013
4	rice.nuclei.ribozero.dup1	650	6.75	GTCCACAT	AACACGCT	14968012	14935246
5	rice.nuclei.ribozero.dup2	1000	6.6	CCTCGAAT	CACAGACT	42942217	42789823
6	rice.nuclei.ribozero.dup3	1000	7.1	ACGTATGG	CGACACTT	21424091	21351205
7	rice.total.polyA.dup1	680	8.05	GAATCCGA	AGACGCTA	18713010	18634662
8	rice.total.polyA.dup2	1000	8.1	TTCACGGA	TGGATGGT	45128280	44940173
9	rice.total.polyA.dup3	1000	8.45	GTGTGTTC	TTCGAAGC	34302826	34159122
10	rice.total.ribozero.dup1	1000	8.05	ATGGCGAA	CTTCGCAA	19350502	19313952
11	rice.total.ribozero.dup2	1000	8.1	AAGCTCAC	CACTGTAG	21506784	21429236
12	rice.total.ribozero.dup3	1000	8.45	TGGCTACA	AGACGCTA	6828082	6803821
13	sorghum.nuclei.polyA.dup1	450	6.8	AGCCAAGT	CAACTCCA	13842338	13787575
14	sorghum.nuclei.polyA.dup2	1000	6.4	CCTTAGGT	GACTTGTG	41722645	41492113
15	sorghum.nuclei.polyA.dup3	1000	6.6	TCCACGTT	GTGAGACT	34528119	34315348
16	sorghum.nuclei.ribozero.dup1	930	6.8	AGACCGTA	CACTGTAG	18946699	18915645
17	sorghum.nuclei.ribozero.dup2	1000	6.4	CCACTAAG	ACCGACAA	29855451	29750244
18	sorghum.nuclei.ribozero.dup3	1000	6.6	ACCAAGCA	AGTGGCAA	43640994	43487316
19	sorghum.total.polyA.dup1	650	8.6	AGCCAAGT	CAACTCCA	17631593	17569706
20	sorghum.total.polyA.dup2	1000	8.4	ACGGTACA	CAACTCCA	25263357	25163639
21	sorghum.total.polyA.dup3	1000	8.1	ATCTCCTG	AACACGCT	28566001	28452241
22	sorghum.total.ribozero.dup1	1000	8.6	AGACCGTA	CACTGTAG	24613377	24548027
23	sorghum.total.ribozero.dup2	1000	8.4	AGCCAAGT	CTTCGCAA	12000478	11951912
24	sorghum.total.ribozero.dup3	1000	8.1	GTCCTGTT	GTGGTATG	7425303	7398057

Read coverage across exons and introns in different types of genes.

RPKM values were determined separately for exons and introns in two sets of genes, a conserved gene set and an intron loss gene set. RPKM values across exons provided a measure for gene expression. A comparison of exon RPKM values of conserved genes and intron loss genes, within RNA isolation method, within species and within RNA type, showed that expression levels were significantly higher in genes that underwent genomic loss of introns (ANOVA, p

values <0.001) (Figure 5.3). Intron RPKM values were also significantly higher in the intron loss gene set compared to the conserved gene set (p values < 0.024).

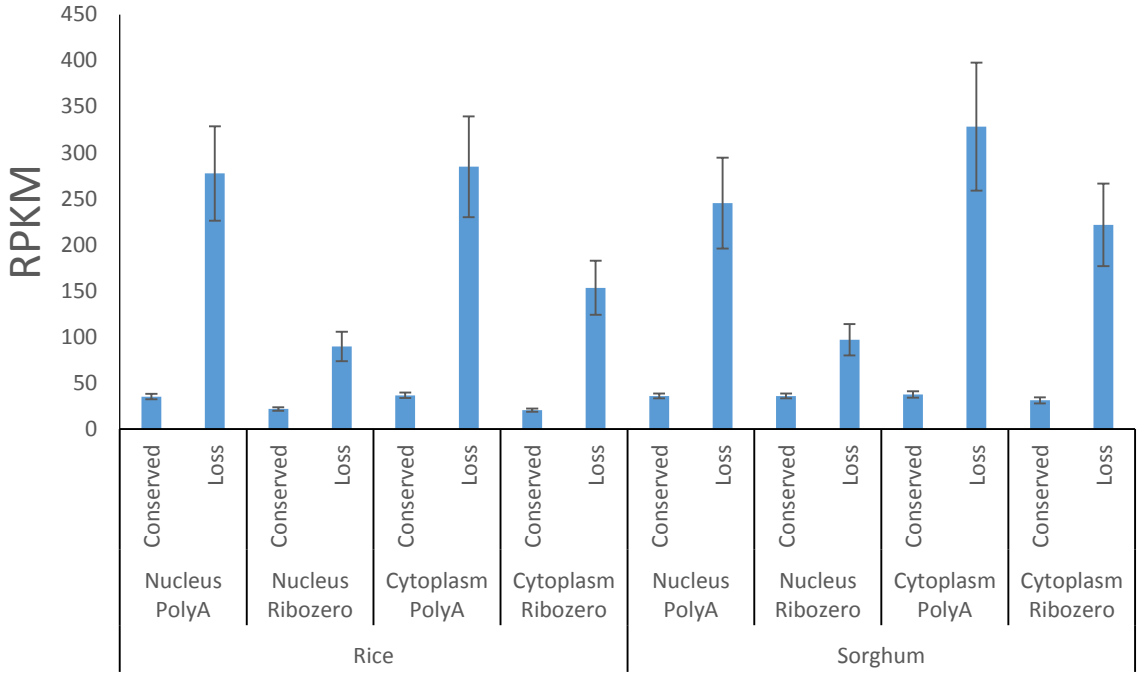


Figure 5.3 Average RPKM values across exons in the conserved gene set (n =151) and the intron loss gene set (n=142) by RNA type, RNA isolation method and species. All pairwise comparisons were statistically significant at p < 0.001.

Read coverage across introns in different types of RNA

ANOVA tests within species showed that read coverage across introns was affected by the RNA type (p values < 0.001). Tukey's HSD test was subsequently used to compare, within species, intron RPKM values across different RNA types (cytoplasmic rRNA depleted RNA, nuclear rRNA depleted RNA, cytoplasmic polyA RNA and nuclear polyA RNA) (Table 5.2). Intron RPKM values were significantly higher in nuclear compared to cytoplasmic RNA (p values ≤ 0.001), while the method used to enrich for protein-coding transcripts (rRNA depletion or polyA selection)

had no effect on intron RPKM values except in the nuclear RNA preparation in sorghum (Table 5.2).

Table 5.2 Tukey's pairwise comparison tests on read coverage within species across different RNA-types (n = 293 genes)

Pairwise comparison[a]	Organism	Average RPKM[1]	Average RPKM[2]	p value
cytoplasmic rRNA depleted[1] vs cytoplasmic poly(A) [2]	Rice	2.92	1.17	0.458
nuclear poly(A) [1] vs cytoplasmic poly(A)[2]	Rice	9.55	1.17	<0.001***
nuclear rRNA depleted[1] vs cytoplasmic poly(A) [2]	Rice	11.23	1.17	<0.001***
nuclear poly(A) [1] vs cytoplasmic rRNA depleted[2]	Rice	9.55	2.92	<0.001***
nuclear rRNA depleted[1] vs cytoplasmic rRNA depleted[2]	Rice	11.23	2.92	<0.001***
nuclear rRNA depleted[1] vs nuclear poly(A) [2]	Rice	11.23	11.23	0.49
cytoplasmic rRNA depleted[1] vs cytoplasmic poly(A) [2]	Sorghum	2.41	1.25	0.88
nuclear poly(A) [1] vs cytoplasmic poly (A) [2]	Sorghum	13.8	1.25	<0.001***
nuclear rRNA depleted[1] vs cytoplasmic poly(A) [2]	Sorghum	8.4	1.25	<0.001***
nuclear poly(A) [1] vs cytoplasmic rRNA depleted[2]	Sorghum	13.8	2.41	<0.001***
nuclear rRNA depleted[1] vs cytoplasmic rRNA depleted[2]	Sorghum	8.4	2.41	0.001***
nuclear rRNA depleted[1] vs nuclear poly(A) [2]	Sorghum	8.4	13.8	0.004**

[a] Correspondence between samples in the pairwise comparison and RPKM values is indicated with superscripts [1] and [2].

Read coverage across introns in genes with different intron composition

To test whether there was a gene structure (intron composition) effect on intron read coverage, a MANOVA followed by a Tukey's HSD test were performed on nuclear RNA-Seq

data for 141 genes that underwent intron loss. Intron RPKM values were compared for single intron genes that underwent single loss (SL) or parallel loss (PL), for multiple intron genes that carried both single loss and conserved introns (SL+con), both parallel loss and conserved introns (PL+con), or a combination of single loss, parallel loss and conserved introns (PL+SL+con). A single gene that contained only SL and PL introns was included in the latter category. One gene in which the same intron was lost from both rice and sorghum was excluded from the analysis. The MANOVA showed a significant effect of gene structure/intron composition on intron read coverage (p value < 0.001) (Table 5.3). Introns in genes that underwent only parallel loss have significantly higher RPKM values than introns of genes with other intron profiles including genes that carry both parallel loss and conserved introns (p value < 0.001) (Figure 5.4). However, no difference in RPKM values was observed between introns in genes that underwent only single loss, and genes that carry both single loss and conserved introns (p value = 0.067) (Figure 5.4).

Table 5.3: Analysis of variance of gene structure effects on intron read coverage, performed on nuclear RNA-Seq data for 141 genes that underwent intron loss.

Source of variation	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Method	1	15225	15225	3.259	0.0711
Species	1	1223	1223	0.262	0.6089
Gene structure	4	228186	57047	12.212	6.96e-10***
Method:Species	1	28585	28585	6.119	0.0134*
Method: Gene structure	4	45443	11361	2.432	0.04541*
Species: Gene structure	4	125532	31383	6.718	2.17e-05***
Method:Species: Gene structure	4	38943	9736	2.084	0.0802
Residuals	5260	24570704	4671		

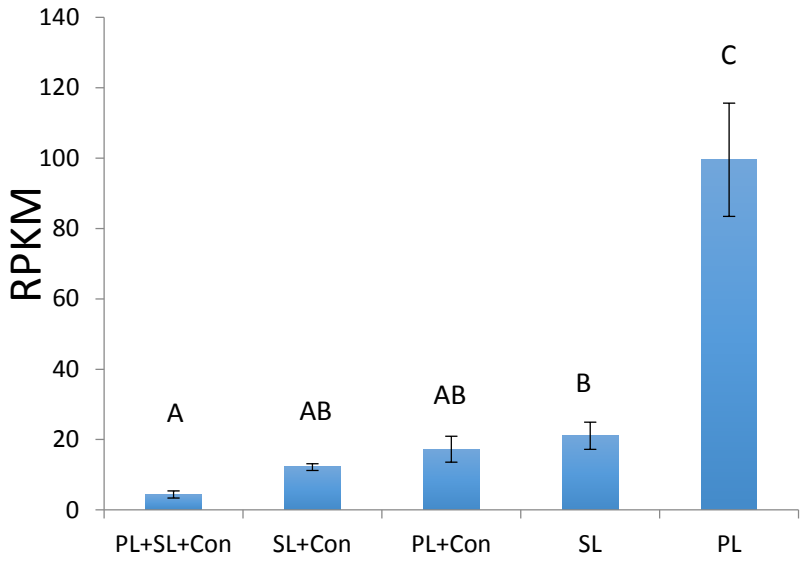


Figure 5.4 Average intron RPKM values in genes with different intron profiles. Means with the same letter are not statistically significant at the 5% level.

Read coverage across conserved and dynamic introns

We investigated read coverage across different types of introns (conserved introns, SL introns, and PL introns) in the nuclear RNA-Seq datasets only as these datasets had the highest intron read coverage. Because genes that contained lost introns had significantly higher expression levels than genes with a conserved structure, we limited our analysis to 90 genes that comprised both conserved and lost introns. A MANOVA showed that the method of mRNA enrichment (rRNA depletion or polyA selection) (p value= 0.315) and the species (rice or sorghum) (p value = 0.096) did not affect intron RPKM values which allowed us to pool datasets across species and across enrichment method for further analyses (Table 5.4). No significant difference was detected between RPKM values of conserved and dynamic introns (p value = 0.689) across these 90 genes (Table 5.4).

However, a gene by gene analysis showed different scenarios. In 64 genes, no difference was observed in read coverage of conserved and loss introns (p values > 0.051) (conserved = loss; set 1); in 13 genes, loss introns had a higher RPKM value than conserved introns (p values < 0.034) (conserved<loss; set 2); and in 13 genes the inverse occurred (p values < 0.048) (conserved >loss; set 3). A significant effect of gene scenario was detected on the intron RPKM coverage (p value < 0.001). Average intron RPKM values across all introns of genes in set 2 (conserved<loss) and set 3 (conserved>loss) were significantly lower than those of genes in set 1 (conserved = loss) (p values <0.003). No significant difference (p value = 0.948) was observed between intron RPKM values in set 2 and set 3 (Figure 5.5).

Table 5.4: Analysis of variance of method, species, intron type, intron position and intron size effects on intron read coverage, performed on the nuclear RNA-Seq data of the 90 genes that comprised both conserved and lost introns

Source of variation	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Method	1	4474	4474	1.008	0.31545	
Species	1	12319	12319	2.775	0.09581	
Intron type	2	3303	1651	0.372	0.68938	
Intron position	1	7079	7079	1.595	0.20671	
Gene scenario	2	108320	54160	12.201	5.19E-06	***
Intron size	1	2098	2098	0.473	0.49183	
Method:Species	1	44485	44485	10.021	0.00156	**
Method:Intron type	2	2481	1240	0.279	0.75623	
Species:Intron type	2	37967	18984	4.276	0.01395	*
Method:Intron position	1	4699	4699	1.058	0.30361	
Intron type:Intron position	2	5142	2571	0.579	0.56042	
Method:Gene scenario	2	1202	601	0.135	0.87338	
Species:Gene scenario	2	1268	634	0.143	0.8669	
Intron type:Gene scenario	4	9357	2339	0.527	0.71592	
Intron position:Gene scenario	2	1441	721	0.162	0.85014	
Method:Species:Intron type	2	10164	5082	1.145	0.31836	
Method:Intron type:Intron position	2	749	374	0.084	0.91911	
Method:Species:Gene scenario	2	8279	4140	0.933	0.39363	
Method:Intron type:Gene scenario	4	1151	288	0.065	0.99229	

Species:Intron type:Gene scenario	4	14510	3628	0.817	0.51397
Method:Intron position:Gene scenario	2	1389	695	0.156	0.85514
Intron type:Intron position:Gene scenario	2	1980	990	0.223	0.80007
Method:Species:Intron type:Gene scenario	4	3966	991	0.223	0.92549
Method:Intron type:Intron position:Gene scenario	2	299	149	0.034	0.9669
Residuals	4617	20495311	4439		

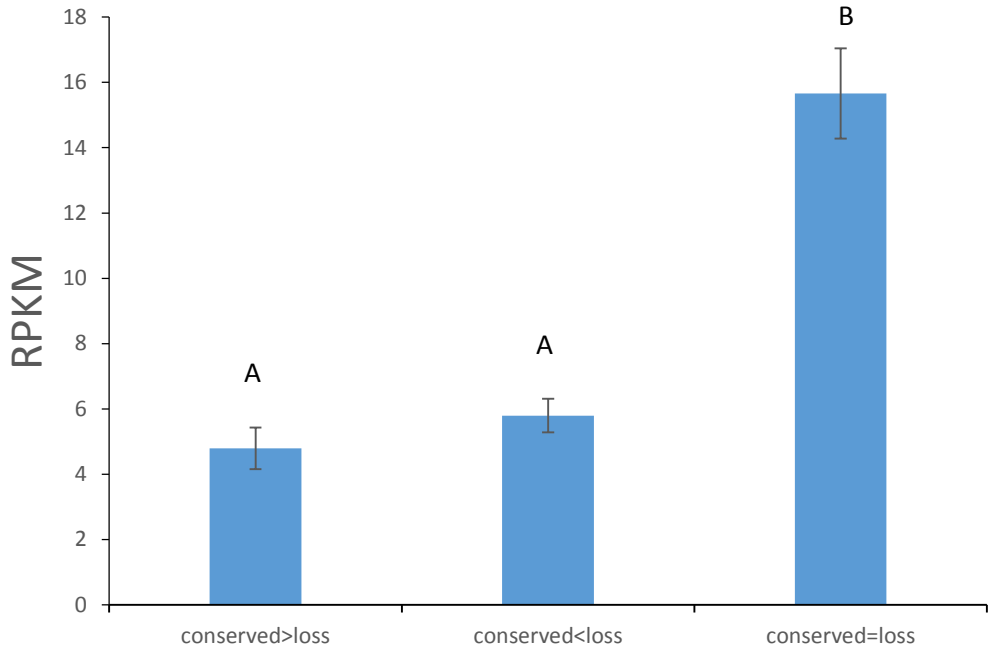


Figure 5.5 Average intron RPKM values in genes in which conserved introns have higher RPKM values than loss introns (conserved>loss; 13 genes), conserved introns have lower RPKM values than loss introns (conserved<loss; 13 genes), and both intron types have similar RPKM values (conserved=loss; 64 genes) (n=90). Means with the same letter are not statistically significant at the 5% level.

Read coverage across introns in different intron locations and intron size effect

To test whether there was a position effect on intron read coverage, intron RPKM values of introns located in the 5' half of a gene (5' introns) were compared with those in the 3' half of a gene (3' introns) in the 90 genes that carried both conserved and loss introns. This analysis was conducted only on nuclear RNA-Seq data. No significant difference between intron RPKM values in the 3' introns and 5' introns was observed (p value = 0.206). There also was no correlation between the intron size and the RPKM value (p value= 0.491) (Table 5.3). In the 13 genes where the loss introns had higher RPKM values than the conserved introns, 10 of the loss introns were present in the first half (5' part) of the gene and 16 were present in the second half (3' part) of the gene. The size of the loss introns ranged from 76 to 1493 bp in rice and 81 to 2053 bp in sorghum. In the 13 genes where the conserved introns showed a higher RPKM value than the loss introns, 13 of the loss introns were present in the 5' end of the gene and 10 were present in the 3' end of the gene. The size of these introns ranged from 89 to 472 bp in rice and 90 to 2349 bp in sorghum. We statistically tested if an effect of intron location and intron size on RPKM values was still lacking under different gene scenarios. MANOVA tests within each scenario showed no significant effect of intron location (p values ≥ 0.256) or intron size on the RPKM values (p values ≥ 0.066).

DISCUSSION

Genes that experience intron loss are expressed at higher levels

Germline tissue expression is a prerequisite for the loss of an intron to be fixed in a species. So, ideally, we should have conducted RNA-Seq on nascent and cytoplasmic RNA from germline tissues. However, we deliberately did not do this because of the technical difficulties associated with 1) isolating germline tissues (Schmidt et al. 2012) and 2) obtaining sufficient quantities of

germline tissue for nuclear isolation. Nevertheless, we observed that genes that experienced intron loss were expressed at higher levels than genes with a conserved intron structure in both rice and sorghum. As expected, intron RPKM values were also higher in these intron loss genes. Similar to our results, intron loss has been observed more frequently in highly expressed genes in mammals (Coulombe-Huntington and Majewski 2007a) and fish (Kawaguchi et al. 2010) but not in *Arabidopsis thaliana* (Yang et al. 2013). In mammals, genes that underwent intron loss are associated with cellular house-keeping functions such as biosynthesis, metabolism, translation, transcription, and RNA processing. These are genes that are 1) highly expressed and 2) expressed in germline tissues (Coulombe-Huntington and Majewski 2007a). Our set of 142 intron loss genes chosen for intron expression analysis also have a higher percentage of genes expressed in early embryogenesis and pollen development (Wang et al. 2014) which are germline tissues.

Enrichment of intron reads in the nuclear RNA

The biological process of transcription has been studied extensively in mammalian species. Wide-spread co-transcriptional splicing has been observed in *Drosophila* while a far lower frequency of co-transcriptional splicing is observed in mouse (Khodor et al. 2011; Khodor et al. 2012). *In vivo* studies in mouse have shown that the poly(A) addition occurs largely before splicing when the transcript is still attached to the chromatin (Bhatt et al. 2012). In plant systems, no experimental data currently exists on the prevalent mechanism of splicing. We observed that the nuclear mRNA fraction (both polyA and ribosome depleted preparations) yielded a higher number of intronic reads than the cytoplasmic fraction in both rice and sorghum. The presence of large numbers of intronic reads in the nuclear fractions of rice and sorghum suggests that, for many introns, post-transcriptional splicing is the major mechanism of intron removal.

Conserved introns vs. lost intron

Analysis of read coverage across a set of 141 genes that contained different types of introns showed that the RPKM values of introns in the parallel loss (PL) genes was significantly higher than in genes containing conserved or single loss introns, or a combination of conserved and loss introns. However, the PL category contained only 3 genes and when the data were scrutinized more closely, it appeared that one gene (Os07g0577600.01/ Sb02g037410.1) with a functional annotation of 'chlorophyll binding' showed very high gene and intron RPKM values which skewed the overall PL RPKM value. There was no difference in intron RPKM values between the SL and SL+con genes suggesting that overall splicing efficiency was not different in these two types of genes. Limiting our analysis to 90 genes which contained both lost and conserved introns, we established that intron retention of lost introns and conserved introns was similar. These initial analyses thus indicate that intron loss is not correlated with splicing efficiency, because intron RPKM values can differ greatly between genes, we also compared RPKM values of loss and conserved introns on a gene-by-gene basis. This analysis showed that 1) similar to the combined analysis, there was no difference between read coverage of 'lost' introns and 'conserved' introns in the majority of genes analyzed (64) and 2) thirteen genes were identified where the read coverage was significantly higher in the 'lost' introns than the conserved introns and 13 genes were identified where the reverse was true. We did not observe any common GO terms specific to each of these scenarios. Therefore, our data suggest that evolutionary loss of introns is not correlated with splicing efficiency.

One unavoidable flaw in the experiment is that the splicing efficiency of lost introns can be studied only in species in which the intron in question has not been lost. These introns may have evolved a different role than in the species in which the loss was incurred. Furthermore,

splicing efficiency of orthologous introns may vary across species. Because we only analyzed the nuclear RNA at this stage, we cannot exclude other possibilities such as 1) some transcripts in the nuclear mRNA are pseudo-transcripts which will never mature and be subjected to degradation (Pandya-Jones and Black 2009) and 2) some transcripts are lowly expressed alternately spliced transcripts as intron retention is the major form of alternate splicing in plants. As all these transcripts will also be captured in the experiment, they could bias the output. However, alternately spliced introns can be excluded in future analyses by studying the cytoplasmic mRNA.

Intron characteristics did not affect intron RPKM

In a subset of 90 genes which contained both lost and conserved introns, we observed no effect of intron size or intron position on the intron RPKM values. This observation is not in compliance with an RT-based model of genomic intron removal, where small introns are more likely to be lost than large introns because they can be more efficiently spliced (Roy et al. 2003b; Coulombe-Huntington and Majewski 2007b; Wang et al. 2014). The fact that smaller introns have similar RPKM values than large introns is consistent with our earlier conclusion that genomic intron loss is not caused by more efficient splicing. Intron length has also been shown to be negatively correlated with co-transcriptional splicing efficiency in *Drosophila* but no correlation has been observed in mouse, where co-transcriptional splicing occurs at a far lower rate than in fly (Khodor et al. 2011; Khodor et al. 2012). Again, the lack of a correlation between intron size and RPKM value is in agreement with our observation that splicing in rice and sorghum occurs mainly post-transcriptionally. Other factors may affect why smaller introns do not show lower RPKM values. For example, the structure of the chromatin determines the efficiency with which splicing factors are recruited to specific gene locations (Schor et al. 2012). Pandya-Jones and Black (2009) reported that within human HeLa cells, most introns are excised in a 5' to 3' manner in the

chromatin bound mRNA which results in 3' introns being more highly represented in RNA-Seq data than 5' introns. Again, we do not see this pattern in the sub-set of 90 genes, suggesting that the pattern of intron excision may be more random in plants.

Conclusions

In conclusion, genes that underwent genomic loss of introns are, on average, expressed at higher levels than genes with an evolutionary conserved structure. This is consistent with a model of RT-mediated gene loss, as higher expression will lead to more mRNA which, in turn, will increase chances for gene conversion between genomic and cDNA sequences to occur. We did not find any correlation between the size and position of introns intron RPKM values suggesting that splicing does not occur in a strict 5' to 3' fashion nor is splicing efficiency directed by intron size. Both global and gene-by-gene analyses indicated that lower RPKM values are not associated with introns that underwent evolutionary loss and hence that our hypothesis that splicing efficiency plays a role in intron loss is incorrect.

For future analysis, we plan to calculate intron retention by normalization to gene expression rather than based on RPKM values which are normalized to the total number of output reads for a given sample. This would make it easier to compare across genes. Pandya-Jones and Black (2009) identified that the nuclear mRNA has a different intron makeup than chromatin bound mRNA in human HeLa cells. Future work to isolate and sequence chromatin bound mRNA would provide another measure to calculate the intron splicing efficiency.

REFERENCES

Akhunov ED, Sehgal S, Liang H, Wang S, Akhunova AR, Kaur G, Li W, Forrest KL, See D, Šimková H et al. 2013. Comparative Analysis of Syntenic Genes in Grass Genomes

Reveals Accelerated Rates of Gene Structure and Coding Sequence Evolution in Polyploid Wheat. *Plant Physiology* **161**(1): 252-265.

Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L. 2011. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18**(12): 1435-1440.

Bhatt DM, Pandya-Jones A, Tong AJ, Barozzi I, Lissner MM, Natoli G, Black DL, Smale ST. 2012. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**(2): 279-290.

Coulombe-Huntington J, Majewski J. 2007a. Characterization of intron loss events in mammals. *Genome Res* **17**(1): 23-32.

-. 2007b. Intron loss and gain in Drosophila. *Molecular biology and evolution* **24**(12): 2842-2850.

Deal RB, Henikoff S. 2011. The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana. *Nat Protocols* **6**(1): 56-68.

Farlow A, Meduri E, Schlötterer C. 2011. DNA double-strand break repair and the evolution of intron density. *Trends in genetics : TIG* **27**(1): 1-6.

Fawcett JA, Rouze P, Van de Peer Y. 2012. Higher intron loss rate in Arabidopsis thaliana than A. lyrata is consistent with stronger selection for a smaller genome. *Mol Biol Evol* **29**(2): 849-859.

Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proceedings of the National Academy of Sciences* **99**(25): 16128-16133.

Fedorova L, Fedorov A. 2003. Introns in gene evolution. *Genetica* **118**(2-3): 123-131.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**(5565): 92-100.

Hong X, Scofield DG, Lynch M. 2006. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol* **23**(12): 2392-2404.

Jeffares DC, Mourier T, Penny D. 2006. The biology of intron gain and loss. *Trends Genet* **22**(1): 16-22.

Kawaguchi M, Hiroi J, Miya M, Nishida M, Iuchi I, Yasumasu S. 2010. Intron-loss evolution of hatching enzyme genes in Teleostei. *BMC Evolutionary Biology* **10**(1): 260.

Khodor YL, Menet JS, Tolan M, Rosbash M. 2012. Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. *RNA*.

Khodor YL, Rodriguez J, Abruzzi KC, Tang C-HA, Marr MT, Rosbash M. 2011. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & Development* **25**(23): 2502-2512.

Knowles DG, McLysaght A. 2006. High Rate of Recent Intron Gain and Loss in Simultaneously Duplicated Arabidopsis Genes. *Molecular Biology and Evolution* **23**(8): 1548-1557.

Krzywinski J, Besansky NJ. 2002. Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. *Mol Biol Evol* **19**(3): 362-366.

Lin H, Zhu W, Silva JC, Gu X, Buell CR. 2006. Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol* **7**(5): R41.

Luthe DS, Quatrano RS. 1980. Transcription in Isolated Wheat Nuclei: I. Isolation of nuclei and elimination of endogenous ribonuclease activity. *Plant Physiol* **65**(2): 305-308.

Milia G, Camiolo S, Avesani L, Porceddu A. 2015. The dynamic loss and gain of introns during the evolution of the Brassicaceae. *The Plant Journal* **82**(6): 915-924.

Mrinal N, Nagaraju J. 2008. Intron Loss Is Associated with Gain of Function in the Evolution of the Gloverin Family of Antibacterial Genes in Bombyx mori. *Journal of Biological Chemistry* **283**(34): 23376-23387.

Pandya-Jones A, Black DL. 2009. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15**(10): 1896-1908.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A et al. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**(7229): 551-556.

Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. *Biol Direct* **7**(1): 11.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* **13**(17): 1512-1517.

Roy SW, Fedorov A, Gilbert W. 2003a. Large-Scale Comparison of Intron Positions in Mammalian Genes Shows Intron Loss but No Gain. *Proceedings of the National Academy of Sciences of the United States of America* **100**(12): 7158-7162.

-. 2003b. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proceedings of the National Academy of Sciences* **100**(12): 7158-7162.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* **7**(3): 211-221.

Schmidt A, Schmid MW, Grossniklaus U. 2012. Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights. *The Plant Journal* **70**(1): 18-29.

Schor IE, Llères D, Risso GJ, Pawellek A, Ule J, Lamond AI, Kornblihtt AR. 2012. Perturbation of Chromatin Structure Globally Affects Localization and Recruitment of Splicing Factors. *PLoS ONE* **7**(11): e48084.

Wang H, Devos KM, Bennetzen JL. 2014. Recurrent Loss of Specific Introns during Angiosperm Evolution. *PLoS Genet* **10**(12): e1004843.

Yang YF, Zhu T, Niu DK. 2013. Association of intron loss with high mutation rate in Arabidopsis: implications for genome size evolution. *Genome biology and evolution* **5**(4): 723-733.

Yuan Y, Chung J-D, Fu X, Johnson VE, Ranjan P, Booth SL, Harding SA, Tsai C-J. 2009. Alternative splicing and gene duplication differentially shaped the regulation of isochorismate synthase in Populus and Arabidopsis. *Proceedings of the National Academy of Sciences* **106**(51): 22020-22025.

Zhou M, Yan J, Ma Z, Zhou Y, Abbood NN, Liu J, Su L, Jia H, Guo A-Y. 2012. Comparative and Evolutionary Analysis of the HES/HEY Gene Family Reveal Exon/Intron Loss and Teleost Specific Duplication Events. *PLoS ONE* **7**(7): e40649.

CHAPTER VI:

DISCUSSION AND OVERALL CONCLUSIONS

Height-reducing genes provide resistance to lodging (bending of the stalk) in cereal crops and have been successfully deployed, amongst others, in rice (*sd-1*), wheat (*Rht-1*), rye (*Ddw1*), sorghum (*dw3*) and pearl millet (*d2*) (reviewed in chapter 1). The *d2* gene has been widely used in many commercially successful cultivars in India, Australia and the United States (Gulia et al. 2007; Rai et al. 2009). The pearl millet *d2* dwarf plants are better suited to mechanical harvesting and provide better forage quality than tall plants due to their higher leaf and reduced stem mass (Johnson et al. 1968). In our efforts to identify the gene underlying the *d2* trait, we found that the *d2* region in pearl linkage group (LG) 4 is characterized by low recombination (Chapter 2). An inversion in the *d2* region is present with respect to the orthologous region in sorghum. We speculate that this inverted region is specific to the *d2* dwarfs and the cause for the reduced recombination. Hence, our mapping efforts with ~1500 progeny from two different mapping populations did not yield sufficient resolution to make a bacterial artificial chromosome (BAC) physical map of the region in a labor- and cost-efficient manner. As the pearl millet genome had not been sequenced at that time, we had to rely on the sequenced genome of the close relative sorghum to identify candidate genes for the *d2* trait. Because the *d2* region, which was interstitial on pearl millet LG4, was potentially spread over two sorghum chromosomes due to chromosomal rearrangements between the two species, it was imperative that we further narrowed the *d2* region. The mapping data along with a haplotype analysis of the *d2* region in six diverse tall and *d2* dwarf inbreds unambiguously located the marker boundaries of the *d2* region to the distal region of

sorghum chromosome 7. The sorghum *dw3* dwarfing gene a.k.a *ABCB1* present in this location was identified as a strong candidate for the *d2* trait. The *ABCB1* gene encodes an export protein that modulates auxin transport in plants (Multani et al. 2003). *ABCB1* as a likely candidate for the *d2* trait was supported by the fact that 1) The phenotype of sorghum and maize *ABCB1* mutants was similar to that of pearl millet *d2* mutants. 2) A marker developed in the pearl millet *ABCB1* coding region showed presence/absence polymorphism differentiating tall and *d2* dwarf pearl plants and 3) The *ABCB1* gene was differentially expressed between the tall and *d2* dwarf mapping parents in internodes and leaf tissue using semi-quantitative PCR. Through our work in chapter 1, we have developed several breeder friendly molecular markers that can be used for foreground selection of the *d2* trait. Currently the *d2* dwarf millet is extensively grown in the United States and Australia and to a limited extent in India. However, as more developing countries in South Asia and sub-Saharan Africa will move towards farm mechanization, the *d2* trait will become more applicable and these markers will aid the faster movement of the *d2* trait into elite varieties. This is particularly important as the *d2* trait is recessive, and hence cannot be scored for phenotypically during backcrossing.

In chapter 3, we isolated the *ABCB1* gene from the pearl millet tall mapping parent (ICMP 451) and the *d2* dwarf mapping parent (Tift 23DB). The *abcb1* allele in the *d2* dwarf differs from the tall allele (*ABCB1*) by the presence of two LTR retrotransposon insertions, the first in the coding region and the second LTR 665 bp upstream of the start codon. The transposon in the coding region has high homology to a previously reported 'Juriah' element while the upstream element has not previously been described. Both LTR retrotransposons are full elements which belong to the gypsy superfamily and are present in high copy numbers in the pearl millet genome. Gene inactivation by high copy elements is very unusual as high copy elements usually insert into

one other in heterochromatic regions (SanMiguel et al. 1998). Furthermore, the host genome typically inactivates high copy transposable elements (TE's) by small RNA mediated DNA methylation (Almeida and Allshire 2005; Lee and Kim 2014). In Arabidopsis, the older methylated TEs are present farther away from the genes as methylation of TEs in genic regions results in inactivation of neighboring genes (Hollister and Gaut 2009). Hence, the young LTR retrotransposons in genic regions are preferentially removed by purifying selection as they are deleterious to the host genome (Hollister and Gaut 2009). The left and right LTRs of the two transposable elements are, as far as we could establish, 100% identical which suggests that they are evolutionary recent insertions. In fact, our calculations indicate that the transposable element may have inserted itself as recently as 5000 years ago which coincides with the timeframe of pearl millet domestication (Clotault et al. 2011). The absence of the dwarf pearl millet in the wild germplasm also provides support to this hypothesis. The inserted transposable elements most likely are currently inactive, at least under normal environmental conditions, as *d2* plants are phenotypically stable. This hypothesis can be tested using transposon display experiments.

As identified in chapter 2, the *d2* region is characterized by reduced recombination. It should be interesting to establish whether the recombination reduction is due to a single factor or a combination of factors such as 1) the presence of an inversion in the *d2* region in dwarf plants; 2) the differential presence of transposons in the tall inbred plants in comparison to *d2* plants in this region and 3) to a differential methylation status of genes in the *d2* region in tall and *d2* dwarf plants. Suppression of recombination due to DNA methylation has been observed in the fungus *Ascobolus* (Maloisel and Rossignol 1998). In *Arabidopsis*, the loss of DNA methylation causes global redistribution of recombination events (Mirouze et al. 2012). After the pearl millet genome sequence assembly becomes publicly available, it will be easier to test which of these factors holds

true. The presence of the one or both of the TEs led to differential expression of the *ABCB1* gene in all the tested seedling and adult organs by qPCR. Small quantities of the *ABCB1* transcript were observed in *d2* dwarf plants and it will be interesting to test whether this is due to transcription initiated from the native *ABCB1* promoter across the entire 15 kb Juriah element or to re-initiation from the promoter sequence in the Juriah LTR.

Confirming the identity of the candidate gene has been the biggest road block during my doctoral study. All the reported 'independent' *d2* dwarf mutants were screened and they all showed the presence of the Juriah transposon suggesting a single source for *d2* originating from Dr. Glenn Burton's breeding program in Tifton, GA. We even tested several phenotypic dwarf plants from a Targeted Local Lesions in Genome (TILLING) population of the tall inbred P-1449-2 (ICRISAT, Patancheru, India) but did not observe mutations in *ABCB1*. Unfortunately, while a pearl millet mutant population had been generated using the mutagen ethyl methanesulfonate (EMS), the actual TILLING platform had not yet been set up, so we were unable to obtain independent *d2* mutants. No efficient protocols for pearl millet transformation and tissue culture exist and no plant transformation facilities in the United States or elsewhere in the world provide a service for pearl millet transformation, eliminating that option of candidate gene confirmation as well. Hence, we pursued several alternative lines of support such as 1) Quantify auxin levels and auxin transport efficiency in tall and *d2* dwarf inbred lines and 2) Transformation of the functional (*ABCB1*) and non-functional alleles (*abcb1*) in a easily transformable heterologous system, *Arabidopsis*. The results of the auxin experiments, carried out in collaboration with Dr. Jerry Cohen, were inconclusive due to the large variation observed between replicates. We discontinued these experiments as our lab did not have the expertise to standardize protocols for hormonal studies. In Arabidopsis, mutations in both *abcb1* and its closest homolog, *abcb19*, produce a

distinct dwarf phenotype while some single *abcb1* mutants show a mild reduction in height (Ye et al. 2013). We transformed an *abcb1* single mutant and did not observed a difference in height in the T1 generation. Further analysis in the T2 generation is underway. We were not successful in obtaining seeds of a true *abcb1abcb19* double mutant. However, I recently created a double mutant by crossing an *abcb1* mutant with an *abcb19* mutant, which will be transformed with the transgenic constructs in the near future.

Mutations in orthologous genes may or may not lead to important agricultural phenotypes. The reduced height mutants caused by mutations in the *ABCB1* are currently used in breeding programs of sorghum (*dw3*) (Multani et al. 2003) and pearl millet (*d2*) (Chapter 2 and 3) but not maize (*br2*) (Multani et al. 2003). The maize *br2* mutants reported by Multani and colleagues are knock out mutations in the *ABCB1* in maize produced severe and agriculturally non-viable dwarf mutants (Xing et al. 2015). However, recently, another allele of *br2* was discovered that not only provided a moderate reduction in height with no adverse effect on grain yield (Xing et al. 2015). As mutations in the *ABCB1* gene produced agriculturally important dwarf plants in multiple species, its potential needs to be explored in orphan crops where lodging tolerance is needed. Minor millets such as kodo millet (*Paspalum scrobiculatum*) and little millet (*Panicum sumatrense*) are also susceptible to lodging (Goron and Raizada 2015). They are the best candidates to test the utility of *ABCB1* gene in creating dwarfs because they belong to the same subfamily, *Panicoideae*, as sorghum, maize and pearl millet. Reverse genetics approaches such as mutagenize seed with EMS followed by TILLING or site-directed mutagenesis using the CRISPR-Cas system can be an efficient way to create allelic variants or knockdown of *ABCB1*. The potential of *ABCB1* gene is currently being explored for lodging resistance in another orphan crop, tef, important to Ethiopia (Zhu et al. 2012).

Through our work in chapter 3, we discovered that the structure of the *ABCB1* gene of pearl millet was different from that of the panicoid species, maize and sorghum. Pearl millet *ABCB1* contained 2 introns while maize and sorghum *ABCB1* contained 4 introns. As gene structure is typically conserved between closely related species such as species belonging to the grass family, this differential presence of introns was unexpected. A detailed analysis of the *ABCB1* gene structure in sequenced and non-sequenced angiosperms showed that the ancestral *ABCB1* likely had 9 introns (chapter 4). However, 7 out of the 9 introns were lost independently in 2 or more species (parallel loss) while 2 introns were shown to have undergone single loss. In contrast, the structure of its closest homolog, *ABCB19*, remained fairly constant over 100 million years of evolutionary divergence across monocots and dicots. Genes that show parallel loss of introns are very rare. In a genome wide study across 5 sequenced grasses, separated by 70 MY of evolutionary divergence, only a total of 93 cases of parallel intron loss are reported (Wang et al. 2014). A few well characterized examples of genes that show parallel intron loss include the *white* gene in the fly lineage, GAPDH gene in mammals, *MRP1* gene in *Drosophila* and mosquito species and isochorismate synthase gene in Angiosperms (Chapter 4). However, these genes only incur parallel loss of 1 to 3 introns. The *ABCB1* is the first unusual gene in Angiosperms that showed parallel loss of majority of its introns. This is based on the structural analysis of *ABCB1* genes from 75 Angiosperm species. In a post-genomics era, with more sequenced genomes are added every year, the phenomenon of parallel loss of introns can be studied in more detail in other genes and it may be more prevalent than previously reported.

For introns to undergo removal, the loss event should occur in germline tissue. Both *ABCB1* and *ABCB19* genes show expression levels similar to house-keeping genes in germline tissues in *Arabidopsis* and rice. But, it is unclear why the *ABCB1* gene underwent loss and the

ABCB19 structure is constant. Only, a single intron, intron 7, appears to be conserved within the grass family. Interestingly, this intron contains an 8 bp motif which is highly conserved between monocots and dicots. The role of this conserved motif is currently unknown, but intron experiments are underway to test if intron 7 is important in gene regulation.

The frequency of intron loss in *ABCB1* gene is higher in the monocots, in specific the grass family, compared to dicots suggesting that the intron loss is not random across species. A reverse transcriptase (RT) mediated intron loss model where a genomic copy undergoes recombination with an intron less cDNA copy is more widely accepted than genomic deletions as the cause for intron removal (Roy and Gilbert 2005). The *ABCB1* gene follows signatures of an RT-based model of intron loss that include characteristics such as precise removal of introns and removal of smaller introns. All introns in *ABCB1* that underwent parallel intron loss in monocots are below 200 bp. We did not observe other features of a RT-based model such as the removal of adjacent intron and preference towards removal of 3' introns. A more complex RT-model is necessary to explain the intron loss events that occurred in the *ABCB1* gene. Smaller introns can be processed more efficiently by the spliceosome (Khodor et al. 2011) and thus reduction in size may be crucial for evolutionary intron loss. As in the case of intron 6, which also underwent parallel loss, but has a mean length of 382 bp in monocots. Upon closer examination, the grass subfamilies where the loss of intron 6 occurred, the length of intron 6 is at most 103bp. We hypothesized that the more efficiently spliced introns undergo more frequent evolutionary loss.

While there have been several studies investigating the prevalence of intron loss and characteristics of lost introns, the link between the RNA processing and RT-mediated loss has not been pursued. To understand the mechanism behind the genomic loss of introns, we provide the first investigation of the role of intron splicing efficiency towards evolutionary intron removal

(Chapter 5). To perform this study, we successfully isolated mRNA (pre-mRNA) from the nucleus of two grass species, rice and sorghum. The pre-mRNA is present at various stages of processing. All introns are not processed at the same time. Co-transcriptional splicing is the prevalent mechanism in *Drosophila* (Khodor et al. 2011), while both the post-transcriptional and co-transcriptional splicing are prevalent in the mouse (Khodor et al. 2012) but co-transcriptional splicing occurs at lower levels in comparison to *Drosophila*. The intron location determines the co-transcriptional splicing efficiency in *Drosophila* and mouse while the intron length is only important in *Drosophila* and not in mouse. The process of splicing *i.e.,* whether a co-transcriptional or post-transcriptional mechanism is prevalent has not been investigated in plants. Through our study we have observed that for many introns, post-transcriptional splicing is prevalent for intron removal as reads in introns were observed in the nuclear poly(A) mRNA. However, this is based on ~290 genes and we plan to analyze the entire gene set in rice and sorghum in future work. To study if the co-transcriptional splicing occurs in plants, we need to isolate the chromatin-bound mRNA. Our attempts to isolate this fraction were currently unsuccessful because we were unable to optimize a protocol for plant tissue.

For a preliminary analysis, a dataset of 151 conserved genes and 142 genes that underwent intron loss was selected to study intron splicing efficiency. We observed that the genes that underwent intron loss are expressed at significantly higher levels than the conserved gene set in rice and sorghum. In mammals, it was noted that genes that underwent intron loss are highly expressed genes involved in housekeeping functions (Coulombe-Huntington and Majewski 2007). The set of genes that underwent intron loss that we used in our study were selected from Hao and colleagues (2014) and the 5 most significant GO terms were ‘catalytic activity’, ‘oxidoreductase activity’, ‘metabolic process’, ‘omega-3 fatty acid desaturase activity’ and ‘positive regulation of

protein modification process' which suggest involvement in housekeeping function. Overall, we identified 3 categories of genes: 1) there is no difference in read coverage between lost introns and conserved introns 2) conserved introns have a higher read coverage than lost introns and 3) conserved introns have lower read coverage than lost introns. We also observed no correlation between intron length and intron position towards the splicing efficiency in a subset of 90 genes. This suggests that intron processing does not occur sequentially from 5' to 3' and smaller introns are not processed faster than longer introns. The above results, *i.e*., 1) no difference in splicing efficiency between evolutionarily lost and conserved introns and 2) smaller introns are not processed faster than longer introns disproves our hypothesis. As this was a preliminary study, we plan to repeat the analysis changing a few analysis parameters to calculate intron splicing efficiency and check if we get the same results. In the future work, we also plan to check the gene body methylation of the lost and conserved genes. The genes that underwent loss show a lower TG/CG nucleotide ratio in lost introns and flanking exons in comparison to conserved introns and flanking exons which indicates lower historical methylation. We will check the current methylation status of these genes in several plant species using the gene methylation data housed in the Schmitz lab (http://schmitzlab.genetics.uga.edu/plantmethylomes)

Through my work on the *d2* gene in pearl millet, we were able to develop resources for the pearl millet research and plant breeding community. The markers developed for the easy screening the *d2* trait can be screened on agarose gels. This is especially crucial for plant breeding programs in developing countries with limited lab resources. We have also developed a fosmid library for the pearl millet inbred, ICMP 451, a fertility restorer line, used in pearl millet hybrid production. This resource can be used for isolation of genes for other important traits associated with ICMP 451. In our efforts to confirm the identity of the *d2* gene in a heterologous system, I developed

abcb1abcb19 double mutant by crossing single mutants in *Arabidopsis*. The double mutants are not available in the Arabidopsis Biological Resource Centre (ABRC) and we could establish the identity of double mutant provided by collaborator. This particular *abcb1abcb19* mutant has a T-DNA insertion in the 3rd exon of *abcb1* which is different from published *abcb1abcb19* mutants available which house T-DNA insertions in 8th exon. This mutant will be submitted to ABRC to help the research community focused on auxin transport. Our work on studying the intron splicing efficiency in grasses generated the first data set, to our knowledge, describing the nuclear mRNA in rice and sorghum. After the publication of our chapter, this information will also be housed in a publicly accessible database.

REFERENCES

Almeida R, Allshire RC. 2005. RNA silencing and genome regulation. *Trends in Cell Biology* **15**(5): 251-258.

Clotault J, Thuillet A-C, Buiron M, Mita SD, Couderc M, Haussmann BIG, Mariac C, Vigouroux Y. 2011. Evolutionary history of pearl millet (Pennisetum glaucum [L.] R. Br.) and selection on flowering genes since its domestication. *Molecular biology and evolution*.

Coulombe-Huntington J, Majewski J. 2007. Intron loss and gain in Drosophila. *Molecular biology and evolution* **24**(12): 2842-2850.

Goron TL, Raizada MN. 2015. Genetic diversity and genomic resources available for the small millet crops to accelerate a New Green Revolution. *Frontiers in Plant Science* **6**: 157.

Gulia SK, Wilson JP, Carter J, Singh BP. 2007. Progress in grain pearl millet research and market development. In *Issues in New Crops and New Uses B2 - Issues in New Crops and New Uses*. ASHS Press, Alexandria, VA.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research* **19**(8): 1419-1428.

Johnson JC, Lowrey RS, Monson WG, Burton GW. 1968. Influence of Dwarf Characteristic on Composition and Feeding Value of near-Isogenic Pearl Millets. *Journal of Dairy Science* **51**(9): 1423-&.

Khodor YL, Menet JS, Tolan M, Rosbash M. 2012. Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. *RNA* **18**(12): 2174-2186.

Khodor YL, Rodriguez J, Abruzzi KC, Tang C-HA, Marr MT, Rosbash M. 2011. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & Development* **25**(23): 2502-2512.

Lee S-I, Kim N-S. 2014. Transposable Elements and Genome Size Variations in Plants. *Genomics & Informatics* **12**(3): 87-97.

Maloisel L, Rossignol JL. 1998. Suppression of crossing-over by DNA methylation in Ascobolus. *Genes Dev* **12**(9): 1381-1389.

Mirouze M, Lieberman-Lazarovich M, Aversano R, Bucher E, Nicolet J, Reinders J, Paszkowski J. 2012. Loss of DNA methylation affects the recombination landscape in Arabidopsis. *Proceedings of the National Academy of Sciences* **109**(15): 5880-5885.

Multani DS, Briggs SP, Chamberlin MA, Blakeslee JJ, Murphy AS, Johal GS. 2003. Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. *Science* **302**(5642): 81-84.

Rai KN, Gupta SK, Bhattacharjee R, KULKARNI VN, Singh AK, Rao AS. 2009. Morphological characteristics of ICRISAT-bred pearl millet hybrid seed parents. *Journal of SAT Agricultural Research* **7**(1): 7pp.

Roy SW, Gilbert W. 2005. The pattern of intron loss. *Proceedings of the National Academy of Sciences of the United States of America* **102**(3): 713-718.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nature genetics* **20**(1): 43-45.

Wang H, Devos KM, Bennetzen JL. 2014. Recurrent Loss of Specific Introns during Angiosperm Evolution. *PLoS Genet* **10**(12): e1004843.

Xing A, Gao Y, Ye L, Zhang W, Cai L, Ching A, Llaca V, Johnson B, Liu L, Yang X et al. 2015. A rare SNP mutation in Brachytic2 moderately reduces plant height and increases yield potential in maize. *Journal of Experimental Botany*.

Ye L, Liu L, Xing A, Kang D. 2013. Characterization of a dwarf mutant allele of Arabidopsis MDR-like ABC transporter AtPGP1 gene. *Biochemical and Biophysical Research Communications* **441**(4): 782-786.

Zhu Q, Smith SM, Ayele M, Yang L, Jogi A, Chaluvadi SR, Bennetzen JL. 2012. High-throughput discovery of mutations in tef semi-dwarfing genes by next-generation sequencing analysis. *Genetics* **192**(3): 819-829.