COMPARATIVE GENOMICS OF THE STEM TRANSCRIPTOME FROM GRAIN AND SWEET SORGHUM

by

MARTÍN CALVIÑO TORTEROLO

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Plant Biology & Pathology

Written under the direction of

Joachim Messing

And approved by

New Brunswick, New Jersey

OCTOBER 2014

ABSTRACT OF THE DISSERTATION

Comparative Genomics Of The Stem Transcriptome

From Grain And Sweet Sorghum

By MARTÍN CALVIÑO TORTEROLO

Dissertation Director:

Joachim Messing

The current dissertation relates to comparative genomics of grain and sweet sorghum, in particular, to their stem's transcriptome at the time of flowering, when soluble sugars accumulate more abundantly in the sweet sorghum cultivar Rio than in the grain sorghum cultivar BTx623. The accumulation of soluble sugars in the stem of sorghum is a valuable agronomic trait because their fermentation into ethanol is currently being used as source of biofuel. High soluble sugar content in stems is a trait also present in the closely related grass sugarcane. Thus, it is reasonable to assume that sweet sorghum and sugarcane may use the same gene products that leads to high soluble sugar content is stems.

My dissertation consists of five chapters, the results of which are five publications as first author. In Chapter 1 I summarized the current status of sweet sorghum genomics and highlighted future research directions. My scientific contribution to the field was also mentioned. In Chapters 2 and 3 I described the

first characterization of the stem's transcriptome from grain and sweet sorghum cultivars using sugarcane Affymetrix arrays, and the use of this transcriptome data to develop molecular markers based on the differences in hybridization intensity from grain and sweet sorghum RNAs to the arrays. In Chapter 4, I described the first characterization of the small RNA component of the stem from grain sorghum BTx623 and sweet sorghum Rio cultivars, and from F2 plants derived from their cross that segregated for sugar content and flowering time. I was able to identify the microRNA family miR169, whose expression co-segregated with sugar content in stems. I also discovered nine new microRNAs in the sorghum genome. In Chapter 5 I described the genomic comparison of *MIR169* gene clusters among five different grasses and identified five new *MIR169* gene copies in the sorghum genome.

ACKNOWLEDGEMENT

My biggest gratitude is to my family members (Jusleine, Infinity, Mauricio, Walter, Elsa and Graciela) for all their support through the years as well as for their patience and unconditional love.

Special thank to the Fulbright Commission in Uruguay and The International Institute of Education (IIE) for their sponsorship of my graduate studies at Rutgers University.

I would like to thank Dr. Joachim Messing for giving me the opportunity to do research in his lab, and for having introduced me into the field of sorghum genomics. Also, I would like to thank the current and past members of the Messing lab for nurturing many intellectual discussions on crop genomics during my past years at graduate school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

OVERVIEW CHAPTER: Sorghum Comparative Genomics and Characterization of the Stem Transcriptome

My thesis concerned the study of the regulation of stem sugar accumulation using cultivars of sorghum as a model system. I have divided this thesis into five chapters that all have been published. The first chapter gives an account of sorghum as a biofuel crop that was published in *Current Opinion in Biotechnology* as an invited review article (Calviño and Messing, 2012). The second chapter lays out the concept of my thesis and the discovery of genes that could play a role in the accumulation of stem sugars. Our laboratory has been involved in the sequence analysis of the sorghum genome as the first example of the *Panicoideae*, the subfamily of the grasses that includes maize and sugarcane. Sorghum was chosen as a reference for the two because it has a smaller genome with 730 MB versus the 2,300 MB of the allotetraploid maize and the 10,000 MB polyploid/aneuploid sugarcane genome. Furthermore, sorghum exists as cultivars with low and high content of soluble sugars in stems, permitting a genetic approach to study stem sugar accumulation. Therefore, we collected sorghum cultivars that differed in the stem sugar level at the time of flowering, when sugar content peaks. To correlate gene expression with stem sugar content from a selected grain sorghum cultivar (BTx623) as reference for low sugar in stems, with a sweet sorghum cultivar (Rio) as reference for high sugar in stems, we took advantage of a commercial Affymetrix oligonucleotide array designed from 8,224 sugarcane transcripts to determine the differential expression of sorghum transcripts from stem tissue. A total of 103 genes

were down-regulated in Rio relative to BTx623, whereas 51 genes were up-regulated in Rio relative to BTx623, respectively. This work represented an examination of the stem transcriptome from two sorghum cultivars as a first step into the elucidation of the gene space possibly involved in sugar accumulation. This work also demonstrated the efficiency of a comparative hybridization approach between two closely related species such as sugarcane and sorghum that shared a common ancestor 8 to 9 million years ago (mya), with both plant species also sharing the trait of sugar accumulation in stems. Indeed, from the 8,224 transcripts represented in the sugarcane array, 70% of them gave a positive signal when hybridized with sorghum RNA samples, indicating that in both species the same gene space may be responsible for the shared trait of sugar content in stems. This work was published in the journal *RICE* in 2008 (Calviño et al., 2008).

In a follow up study, we further extended the concept of cross-species comparative genomic hybridization and used the differences in hybridization intensities of RNA samples from grain and sweet sorghum on the sugarcane array as means to identify nucleotide polymorphism in the transcribed regions of genes, known as Single Feature Polymorphisms (SFPs). By using the publicly available software GeSNP, we screened for SFPs within the 154 differentially expressed genes from grain sorghum and sweet sorghum stem samples mentioned earlier, and found that 58 genes had SFPs. With further analysis based on sequencing of gene fragments containing SFPs we could identify 19 genes where SFPs represented a true SNP between grain sorghum BTx623 and sweet sorghum Rio. Based on the SNPs identified we develop molecular makers based on Single Nucleotide Amplified

Polymorphism (SNAP) methodology. Overall, this work exemplified how a RNA hybridization study using a sugarcane array can be used to detect genome-wide polymorphisms at the DNA level between two sorghum cultivars. This approach could be a desirable methodology for orphan crops, where RNA samples from a plant species that lacks genome sequence information is hybridized to an oligonucleotide array containing probes from a plant species with its genome sequenced. This work was published in the journal *RICE* in 2009 (Calviño et al., 2009).

In a further step towards a more comprehensive characterization of the stem transcriptome in relation to sugar content from grain and sweet sorghum cultivars, we decided to combine SOLiD next-generation sequencing with a bulk segregant analysis (BSA) genetic approach. This involved the deep sequencing of stem-derived small RNAs libraries from BTx623 and Rio, and from a pool of selected F2 plants derived from their cross that segregated for soluble sugar content in stems and flowering time. This work also provided with experimental validation for the previously annotated microRNA genes in the sorghum genome, for which there was no experimental evidence at the time of our study. Indeed, we could detect the expression of 25 microRNA families from the 27 families that were previously annotated in the sorghum genome. Furthermore, we also discovered nine new microRNA genes that were expressed at very low levels in sorghum stem tissue. This was possible because we were able to sequence more than 38 million small RNA reads, from which 23 million reads matched perfectly to the BTx623 reference genome sequence, with a non-redundant set of reads equivalent to more than 2.5

million reads. The pooling of stem-derived small RNAs from F2 plants with contrasting levels of soluble sugars and flowering time allowed us to correlate high expression of microRNA 169 inherited from the BTx623 parent with low soluble sugar content in stems. Our study represented the first characterization of the small RNA component of the sorghum stem transcriptome in relation to soluble sugar content in a segregating population. This lead us to pinpoint miR169 as a candidate miRNA family involved in our trait of interest. This work was published in the journal *BMC Genomics* in 2011 (Calviño et al., 2011).

At the time of our deep sequencing analysis, 19 MIR169 gene copies were annotated in the sorghum genome, with seven gene copies arranged in three clusters. Because previous genetic analysis of soluble sugar content in stems of a recombinant inbred line population derived from BTx623 crossed with Rio uncovered a stem-sugar QTL on chromosome 7, with its closest molecular marker situated in the vicinity of a MIR169 gene cluster, we decided to analyze the process of tandem duplication as evolutionary force that gave rise to this cluster in the sorghum genome by aligning contiguous chromosomal segments from diploid *Brachypodium*, rice, foxtail millet, and the two homoeologous regions of allotetraploid maize. We found that synteny of chromosomal segments containing MIR169 gene copies was conserved between monocotyledonous species such as *Brachypodium* and sorghum but surprisingly also across the monocot barrier in dicotyledonous species such as grapevine, soybean, and cassava. Such an extensive synteny-based analysis allowed us to discover two additional MIR169 gene copies on chromosome 7 that formed an antisense miRNA gene pair, with both copies

being expressed and targeting different set of genes. We also extended the evolutionary analysis to the other two MIR169 gene clusters present in the sorghum genome and as a consequence we were able to discover three additional MIR169 gene copies that had escaped standard genome annotation. This work was published in the journal *Genome Biology & Evolution* in 2013 (Calviño and Messing, 2013).

Based on our study, it becomes clear that the characterization of each individual MIR169 gene copy in relation to soluble sugar content in sorghum stems becomes an imperative task. Our work has provided the initial framework toward future efforts into the elucidation of this complex trait and its genetic improvement in sweet sorghum as an even better feedstock for ethanol production as source of renewable fuel.

References

Calviño M, Messing J (2012) Sweet sorghum as a model system for bioenergy crops. Current Opinion in Biotechnology **23:** 1-7.

Calviño M, Bruggmann R, Messing J (2008) Screen of Genes Linked to High-Sugar Content in Stems by Comparative Genomics. Rice **1:** 166-176.

Calviño M, Miclaus M, Bruggmann R, Messing J (2009) Molecular Markers for Sweet Sorghum Based on Microarray Expression Data. Rice **2:** 129-142.

Calviño M, Bruggmann R, Messing J (2011) Characterization of the small RNA component of the transcriptome from grain and sweet sorghum stems. BMC Genomics **12:** 356-367.

Calviño M, Messing J (2013) Discovery of MicroRNA 169 gene copies in genomes of flowering plants through positional information. Genome Biology & Evolution **5 (2):** 402-417.

Chapter 1 Sweet Sorghum As Model System For Bioenergy Crops: A General Introduction

1.1. Abstract

Bioenergy is the reduction of carbon via photosynthesis. Currently, this energy is harvested as liquid fuel through fermentation. A major concern, however, is input cost, in particular use of excess water and nitrogen. Furthermore, the shortage of arable land creates competition between uses for food and fuel, resulting in increased living expenses. This introduction aims to summarize recent knowledge in genetics, genomics, and gene expression or a rising model species for bioenergy applications, sorghum. Its diploid genome has been sequenced, it has favorable low-input cost traits, and genetic crosses between different cultivars can be used to study allelic variations of genes involved in stem sugar content and incremental biomass.

1.2. Introduction

The production of biofuels (largely ethanol) in the world grew by 13.8% in 2010, and accounted for 0.5% of global primary energy consumption (BP statistical review of world energy, June 2011; URL: http://www.bp.com). Today, biofuels represent 3% of the global road transport fuel supply and are expected to account for as much as 9% by 2050 (Alternative energies for transport, Shell, June 2011; URL: http://www.shell.com). Furthermore, the International Energy Agency

estimates that biofuels will provide with 27% of the world's transport fuel by 2050 (Technology Roadmap: Biofuels for Transport, IEA 2011; URL: http://www.iea.org). Currently, Brazil and the United States are the world leaders in ethanol production. Whereas in Brazil ethanol is fermented from sucrose that accumulates in the stems of sugarcane, in the US it is produced from maize, which accumulates about 85% starch in its seeds. Although the price of oil could play a significant role in influencing the expansion of biofuels (World oil price and biofuels, World Bank Report, June 2011; URL: http://www.worldbank.org), their production costs will also depend on input costs. Thus, reductions in costs are closely tied to the prices of feedstock commodities. Indeed, for conventional biofuels today (first-generation biofuels), feedstocks account for 45-70% of total production costs (Technology Roadmap: Biofuels for Transport, IEA 2011; URL: http://www.iea.org). This is especially important for sugarcane-based and corn-based ethanol, where both crops are cultivated under high inputs conditions requiring significant amounts of water and fertilizers. Sorghum (*Sorghum bicolor* (L.) Moench) has been identified as a crop with low input costs (Belum et al., 2008). Furthermore, in contrast to maize, several sorghum genotypes (known as sweet sorghums) accumulate large amount of fermentable sugars in stems and produce high biomass. The low input costs are based on its highly drought tolerance and C4 photosynthesis. It requires minimal fertilizer for its cultivation and can therefore be planted on marginal or non-arable lands. Despite all the agronomic advantages of sorghum as a bioenergy crop, little scientific effort has been directed in the past toward the genetic and molecular elucidation of sorghum traits relevant to biofuel production compared to corn and

sugarcane research. Nonetheless, the recent sequencing of the sorghum genome (Paterson et al., 2009) not only has accelerated the pace of scientific discoveries but also has established sorghum as a model system to study the more complex genomes of other bioenergy crops such as maize, sugarcane, *Miscanthus* and switchgrass, all belonging to the same subfamily of the grasses, the Panicoideae (Okada et al., 2010; Swaminathan et al., 2010; Wang et al., 2010). The two later ones have the disadvantage that ethanol has to be produced from cellulose, raising the production costs compared to sugarcane and sweet sorghum. However, because of its polyploidy sugarcane is less a suitable model system for other species to increase stem sugar content by translational genetics. Furthermore, grain sorghum and sweet sorghum can be crossed and its desirable traits can be traced through genetic mapping of quantitative trait loci (QTLs). Here, I summarized recent advances, together with a review of my scientific contributions, in the understanding of sorghum traits relevant to bioenergy applications such as sugar content in stems, plant height, and flowering time (Fig. 1.1).

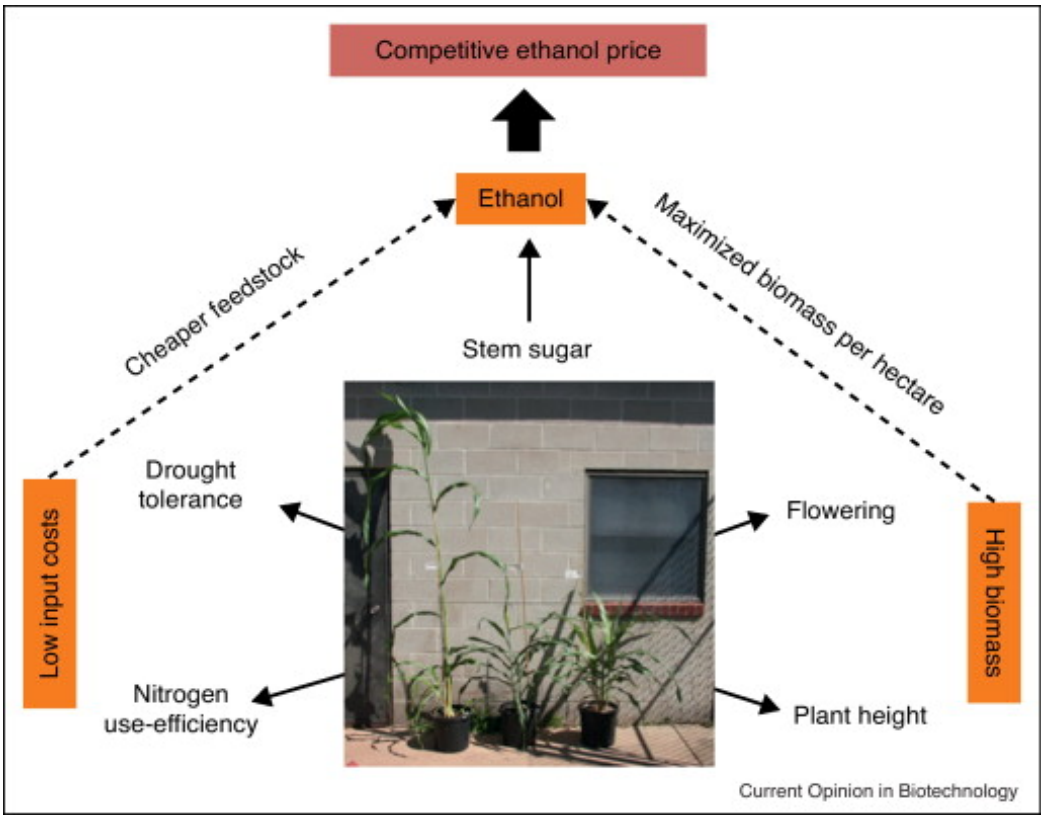


Figure 1.1. Illustration of desirable traits of bioenergy crops. The importance of input cost and biomass yield is emphasized. Samples of sorghum plants are shown in the center. On the left side is the "Rio" cultivar, known for its stem sugar. To the right is "BTx406", known also as grain sorghum. In the middle is the cultivar "R9188" from an introgression of dwarf and early flowering genes from BTx406 into Rio.

1.3. Cultivars with high stem sugar

There is natural variation for sugar content in stems within cultivated sorghum. Cultivars known as grain sorghums accumulate little sugar in contrast to sweet sorghums that accumulate sugars (mainly sucrose) up to 19% of total stem fresh weight (Wang and Liu, 2009). Despite the notorious difference in sugar

content between grain and sweet sorghums, genetic diversity studies did not support a different grouping of sweet sorghum varieties relative to grain sorghum ones, rather they clustered together (Kimberley et al., 2007). Thus, the question of how and when did the high sugar content trait appeared within cultivated sorghum remains open. The first efforts to understand sugar accumulation in sweet sorghum stems investigated the activities of sugar metabolizing enzymes during stem development (Lingle, 1987; Gudrun et al., 1996; Tarpley et al., 1996). However, recent transgenic attempts to increase sugar content in stems of sugarcane by altering the expression of carbohydrate metabolizing enzymes has proven to be less successful (Arruda, 2012), suggesting that the regulatory network responsible for high sugar content in stems is more complex that previously expected. Genetic analysis demonstrated that sugar content in stems is under polygenic inheritance and several regions in the sorghum genome affecting sugar content in stems have been identified (Murray et al., 2008; Murray et al., 2009; Shiringani et al., 2010; Yan-an et al., 2011). Until now, the genes underlying these QTLs remain to be cloned. An additional approach into the elucidation of genes responsible for high sugar content has been my characterization of the stem transcriptome from grain and sweet sorghum (Calviño et al., 2008; Calviño et al., 2009; Calviño et al., 2011). When I compared the gene expression of stem-derived RNAs from grain and sweet sorghum at the time of flowering, transcripts related to cell wall processes were down-regulated in sweet sorghum relative to grain sorghum (Calviño et al., 2008). A similar result was described for wheat where the expression of cell wall related genes was lower in cultivars with high water-soluble carbohydrate (WSC) content

compared to those with low WSC (Xue et al., 2008). These findings suggest that carbon partitioning in the stem could be a mechanism that contribute to genotypic variation in sugar content. Indeed, it was described that the cellulose and hemicellulose content of sweet sorghum stem was reduced after flowering (Zhao et al., 2009), when sugar accumulation in stems usually peak. In addition, stem cellulose and hemicellulose content negatively correlated (correlation coefficients of -0.56 and -0.36, respectively) with Brix degree in a recombinant inbred line (RIL) population derived from the cross of grain sorghum "BTx623" and sweet sorghum "Rio" (Murray et al., 2008). Furthermore, QTLs for stem cellulose and hemicellulose co-localized with a major QTL for Brix degree on chromosome 3 (Murray et al., 2008).

Given the relevance of fermentation of soluble sugars from sweet sorghum stems into ethanol, it is important to consider how high sugar content in stems is affected by other plant traits so that an integrated breeding and/or genetically engineering approach can be performed to improve sorghum as an energy crop. One of the most noted trade-offs in sweet sorghum is the partitioning of photo-assimilates into the stem (stored as soluble sugars) or the grain (stored as starch) as an explanation of the lower grain yield of sweet sorghum compared to that in grain sorghum. However, recent research appears to contradict previous notions. For instance, Murray and colleagues have suggested that both traits, high grain yield and sugar content, could be bred into a single sorghum cultivar (Murray et al., 2008). Consistent with this suggestion, Kumar and colleagues described the absence of

trade-offs in terms of Brix, sugar yield, and sucrose content in stem of several sorghum genotypes at different stages of grain maturity (Kumar et al., 2011).

1.4. Low nitrogen input

Sorghum is a crop with good nitrogen use efficiency (Gardner et al., 1994), and lack of response to nitrogen (N) application is a common phenomenon in sweet sorghum. Indeed, N application did not have an impact on sweet sorghum growth nor yield partitioning among plant organs (Barbanti et al., 2006). In addition, sugar yield did not change with N application (Wortmann et al., 2010). This implies that sweet sorghum can be cultivated with little or no N application without having a negative impact on sugar yield. In fact, sweet sorghum was found to require 36% less nitrogen than corn to attain similar ethanol yield (Geng et al., 1989). Nitrogen fertilization accounts for a big portion of energy consumed by arable crops, attaining 50% of total energy inputs (Barbanti et al., 2006). In addition, the application of nitrogen in excess usually results in nitrate leaching to deep soil layers and the release of either NH3 or N20 to the atmosphere (Barbanti et al., 2006); all these chemicals are considered an environmental hazard. Therefore, managing nitrogen application in sorghum agriculture is of importance in order to minimize environmental pollution.

1.5. Gene networks regulating stem sugar content and drought tolerance

Sweet sorghum was able to provide sufficient juice from stem with total sugar and ethanol yields from fields that were irrigated with 50-75% of the water typically applied to sorghum (Vasilakoglou et al., 2011). Despite the known resilience of sorghum to drought stress, very little is known about the gene network involved in drought tolerance. Dugas and colleagues performed a transcriptome study to investigate the gene network responding to (PEG) polyethyleneglycol-induced osmotic stress and (ABA) abscisic acid treatment in sorghum. They found that sugar-repressive motifs were enriched in promoters of genes whose expression in the shoot was reduced in response to ABA treatment and PEG-induced osmotic stress (Dugas et al., 2011). Recently, Srivasta and colleagues identified in promoters of sorghum DREB (Drought-Response-Element-Binding) genes sequences that could bind factors recognizing sugar signaling motifs (Srivasta et al., 2010). Taken together, these data suggest a crosstalk connection between drought and carbohydrate metabolism. Consistent with this, drought stress in sorghum caused an increased in the amount of leaf soluble sugars whereas starch content was drastically reduced (Kakani et al., 2011). In sorghum, two distinct drought responses have been identified, namely pre-flowering and post-flowering drought response and are controlled by different genetic loci (Harris et al., 2007). Stay-green is an important component of the post-flowering drought response because stay-green sorghum plants maintain the capacity to perform photosynthesis for a longer period than 'senescent' sorghum genotypes under severe drought conditions (Thomas and Howarth, 2000; Harris et al., 2007). Interestingly, stay green genotypes accumulated more stem sugars that senescent genotypes under drought

stress (Duncan et al., 1981). Therefore, it might be conceivable to increase sugar content in stems by manipulating genes involved in drought responses. Indeed, I have recently shown that high expression of microRNA 169 (miR169), a drought-responsive microRNA, negatively correlated with high sugar content in sorghum (see below) (Calviño et al., 2011).

1.6. Incremental biomass in sweet sorghum through plant height

The demand for biomass as source of renewable energy is leading to a paradigm shift in plant architecture in both dual-purpose crops as well as in dedicated energy crops (Salas Fernandez et al., 2009). Plant height is a relevant trait of plant architecture that is highly correlated with biomass yield. Indeed, sweet sorghum cultivars are over three meters tall and are able to produce biomass in the order of 58.3-80.5 tons of fresh stems per hectare in semi-arid zones (Wang and Liu, 2009). In sorghum, four major dwarfing genes *Dw1-Dw4* have been described (Salas Fernandez et al., 2009) and until now only one of them (*Dw3*) has been cloned (Multani et al., 2003). *Dw3* was found to encode a P-glycoprotein that regulates polar auxin transport and is orthologous to the maize *br2* gene (Multani et al., 2003). Most of the grain sorghum lines commercially used carry three of the four dwarfing genes. The combination of up to three dwarfing genes can reduce plant height from more than 3 m tall to 60 cm and this reduction in height is mainly caused by shortening of internode length without changing leaf area. The *Dw2* locus is genetically linked to the photoperiodic flowering *Ma1* locus on chromosome 6 and explained 55% of variation in plant height in the inter-specific cross *Sorghum*

bicolor x Sorghum propinquum (Lin et al., 1995), whereas *Dw1* and *Dw4* have not been conclusively mapped to any sorghum chromosome. Recently, and additional locus controlling plant height have been identified on chromosome 9 and named *Sb-HT9.1* (Brown et al., 2008). Both *Dw3* and *Sb-HT9.1* are described as the most important loci controlling plant height differences in crosses derived from tall and dwarf sorghum (Brown et al., 2008). Although the introduction of dwarfing genes into sorghum allowed for mechanical harvest and conferred lodging resistance, the yield increase was not comparable to those achieved in rice and wheat during the green revolution (George-Jaeggli et al., 2011). Jaeggli and colleagues analyzed the effect of dw3 on shoot biomass and grain yield and found that it was associated with reduced yield (reduced grain size but no grain number) and reduced biomass (3-29%) mainly from reduction in stem biomass (12-41%) (George-Jaeggli et al., 2011). The hormones brassinosteroid (BR) and gibberellin (GA) are known to regulate plant height (Salas Fernandez et al., 2009). However, despite the cloning and characterization of dwarf mutants defective in BR and GA biosynthesis and signaling in rice, maize and wheat, no mutant have been identified in sorghum yet. Still, for its application as bioenergy crop, there is enough genetic evidence that sorghum biomass could be optimized through plant height.

1.7. Sweet sorghum cultivation range widened through flowering time

Sorghum is a short-day tropical grass and thus is photoperiod sensitive, flowering later or not flowering at all in long days compared to short days. Sorghum genotypes that are day-neutral and thus flower early either in short days as well as

in long days allowed sorghum to be used as a grain crop in temperate regions of the world. There is great natural variation in photoperiod sensitivity between sorghum genotypes (Craufurd et al., 1999). Genetic analysis of photoperiod sensitivity in sorghum has identified four major flowering-time (maturity) loci, which were designated as *Ma1*, *Ma2*, *Ma3* (*PhyB*) and *Ma4* (Pao and Morgan, 1986; Childs et al., 1997). Recently, two additional flowering-time loci, *Ma5* and *Ma6* were described in forage and high-biomass sorghum hybrids, with a reported effect to increase photoperiod sensitivity and lengthen the duration of vegetative growth (Rooney and Aydin, 1999). Dominant alleles at each of these maturity loci confer delayed flowering under long days (Pao and Morgan, 1986; Childs et al., 1997; Rooney and Aydin, 1999). The maturity locus *Ma1* has the largest impact on flowering time and was found that alleles at this locus were important in sorghum domestication and dispersal from its center of origin to other parts of Africa and into Asia (Quinby, 1967). The selection of recessive alleles at *Ma1* allowed plant breeders in the US to develop early flowering sorghum cultivars to ensure sufficient time for grain maturation and to avoid frost damage enabling grain production in temperate regions (Klein et al., 2008).

Recently, the cloning of *Ma1* was described (Murphy et al., 2011), providing the first insight into the mechanism of photoperiodic control of flowering time in sorghum. The *Ma1* gene encodes a PSEUDO RESPONSE REGULATOR 37 (SbPRR37) closely related to *Arabidopsis* PRR7, barley Ppd-H1 and wheat PpD-D1a proteins (Murphy et al., 2011). In long days, SbPRR37 inhibits floral transition through the activation of the expression of the floral inhibitor CONSTANS and suppresses the

expression of the floral activators EARLY HEADING DATE 1 (Ehd1), FLOWERING LOCUS T (FT) and *Zea mays* CENTRORADIALIS 8 (ZCN8). The expression of *SbPRR37* was found to be light-dependent and under the control of the circadian clock, with peaks in RNA abundance in the morning and evening under long days. Under short days however, the evening peak in *SbPRR37* expression does not occur due to darkness, promoting flowering under this photoperiod (Murphy et al., 2011). Clearly, geographic variation in day length would require development of sweet sorghum cultivars that are optimized for different latitudes.

1.8. MicroRNA-mediated regulation of bioenergy traits

Given the complexity of networks that connects different pathways that need to be coordinately regulated from biomass to stem sugar, one could argue that part of the regulation is due to the role of small RNAs in the control of gene expression. In this respect, I have characterized the small RNA component of the transcriptome from grain and sweet sorghum stems, and from F2 plants derived from their cross that segregated for sugar content and flowering time (Calviño et al., 2011). I found that variation in miR169 expression correlated with sugar content in stems whereas variation in miR172 and miR395 expression correlated with flowering time. Although miR169 has already a known role in drought stress responses (Zhao et al., 2007; Li et al., 2008; Zhao et al., 2009; Leyva-González et al., 2012), I suggested an additional role of miR169 in sugar accumulation, at least in sorghum (Calviño et al., 2011). This notion was later corroborated when alterations in sucrose/starch balance were found in *Arabidopsis* plants over-expressing miR169n/m variants

(Leyva-González et al., 2012). Furthermore, only three MIR169 gene copies in rice (MIR169g, MIR169n and MIR169o, respectively), from the 17 gene copies present in the genome, were induced under drought and high salinity stress (Zhao et al., 2007; Zhao et al., 2009). This suggests that the rest of the rice MIR169 gene copies may have different functions other than in drought tolerance.

Another recent example of a known microRNA that unexpectedly has a potential role in carbohydrate metabolism is miR156. In maize, the *Corngrass1* (*Cg1*) mutant phenotype is caused by the overexpression of miR156 (Chuck et al., 2007). Maize *Cg1* mutant fixes plant development at the juvenile phase and displays increased biomass due to continuous initiation of tillers and leaves, in addition to delayed flowering (Chuck et al., 2007). Very recently, the maize *Cg1* gene was constitutively expressed in switchgrass (*Panicum virgatum*) (Chuck et al., 2011), and *Cg1* transgenic switchgrass plants presented 250% more starch than normal plants that resulted in higher glucose release from saccharification assays.

I found that the most expressed microRNA in grain and sweet sorghum stems at the time of flowering was miR172 (Calviño et al., 2011), and its abundance correlated well with flowering time in F2 plants. Thus allelic variation at *MIR172* gene loci may explain in part the flowering time difference between grain and sweet sorghums. Although the role of miR172 in controlling flowering is well known (Poethig, 2009; Zhu and Helliwell, 2011), recent data suggests that this miRNA might also be involved in ambient temperature responsive flowering in *Arabidopsis* (Lee et al., 2010). Indeed, *Arabidopsis* plants over-expressing miR172 showed a temperature-independent early flowering (Lee et al., 2010). In addition, in the short

vegetative phase (*svp*) mutant, an important regulator within the thermosensory pathway, the expression of miR172 was altered (Lee et al., 2010). Taken together, these results suggested a link between the thermosensory pathway and miR172. Consistent with this, I identified a *FRIGIDA-LIKE 2* (*FRL2*) gene as confirmed target of miR172 in sorghum (Calviño et al., 2011). The *FRIGIDA*-related genes are a major determinant of natural variation in the winter-annual habit between *Arabidopsis* accessions (Michaels et al., 2004; Schläppi, 2006). On the basis of this, it is tempting to speculate on a possible role of miR172-FRL2 pathway in the adaptation of sorghum to temperate climates.

Finally, I found a genotypic difference in the abundance of miR395 relative to miR395* in sorghum stems (Calviño et al., 2011), with miR395* species expressed as abundantly as miR395 in sweet sorghum but not in grain sorghum. The mechanism behind the switch in miR395 strand abundance from grain to sweet sorghum is unknown. Indeed, evidence supporting the functional role of miRNA*s in plants have recently been described (Devers et al., 2011; Meng et al., 2011). Meng and colleagues applied deep sequencing to investigate the degradome component of the *Arabidopsis* and rice transcriptome derived from the cleavage activity of miRNA*s on their predicted targets, and found significant cleavage signals that were located in the middle of the target binding sites for several miRNA*s-target interactions (Meng et al., 2011). Interestingly, a portion of the cleavage signals that were identified did not locate within the middle region of the miRNA*-target recognition sequence. Instead, the cleavage signals were found at the ends. Also, although most miRNA targets were transcription factors, this was not the case for

miRNA*s targets that were involved in a diverse range of biological functions (Meng et al., 2011). Taking this together, sweet sorghum could be a good model system to elucidate the potential function of miR395*.

1.9. Conclusions

In search for alternatives crops in the generation of bioenergy, a set of traits are coming into focus that differ from those in the production of fiber and food, illustrating the need to embark in new research directions that investigate the underlying molecular mechanisms of these traits. The creation of genetic resources in sorghum such as chemically induced mutant populations (Xin et al., 2008), the assembly and characterization of panels of sorghum germplasm suitable for association mapping (Casa et al., 2008; Murray et al., 2009; Wang et al., 2009), and the creation of protocols for sweet sorghum transformation (Raghuwanshi and Birch, 2010) provide researchers with tools to tackle bioenergy traits in sorghum. Given the intricate balance between different metabolic pathways in respect to these traits, it becomes imperative to turn to suitable genetic model for species under consideration. Sorghum emerges as such model and a number of leads based on my work are becoming apparent for future analysis (Fig. 1.2).

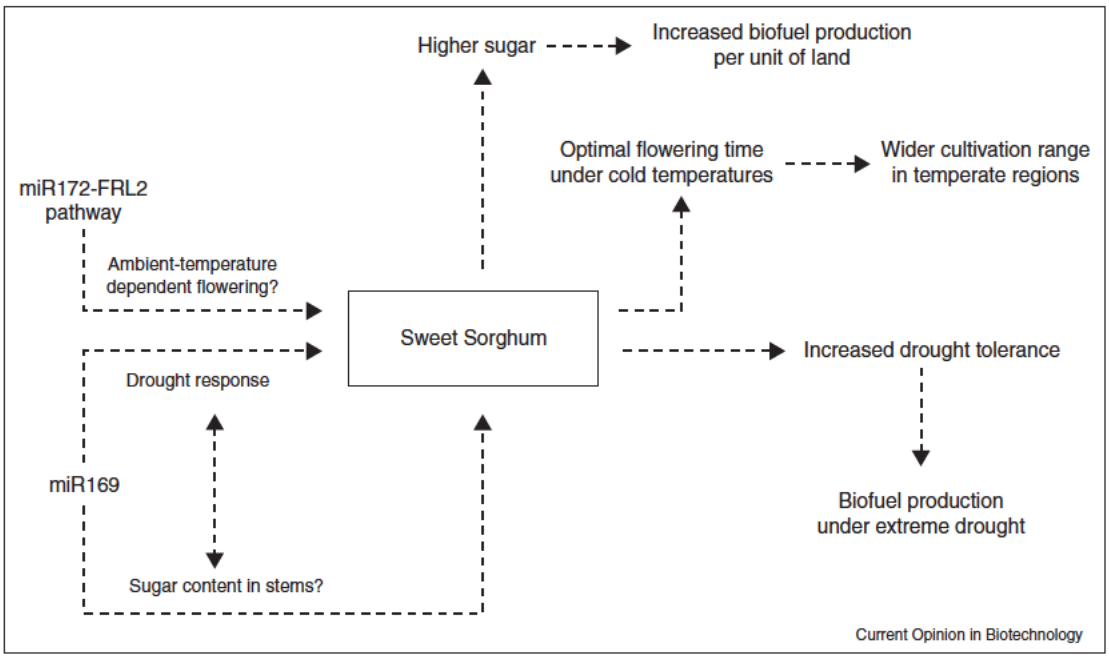


Figure 1.2. Illustration of diverse pathway connections. The potential role of microRNAs as regulatory elements in increased stem sugar accumulation is shown.

1.10. References

Arruda P (2012) Genetically modified sugarcane for bioenergy generation. Current Opinion in Biotechnology **23:** 315-322.

Barbanti L, Grandi S, Vecchi A, Venturi G (2006) Sweet and fibre sorghum (Sorghum bicolor (L.) Moench), energy crops in the frame of environmental protection from excessive nitrogen loads. European Journal of Agronomy **25:** 30-39.

Belum VSR, Ramesh S, Kumar AA, Wani SP, Ortiz R, Ceballos H, Sreedevi TK (2008) Bio-Fuel Crops Research for Energy Security and Rural Development in Developing Countries. BioEnergy Research **1:** 248-258.

Brown P, Rooney W, Franks C, Kresovich S (2008) Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. Genetics **180:** 629-637.

Calviño M, Bruggmann R, Messing J (2008) Screen of Genes Linked to High-Sugar Content in Stems by Comparative Genomics. Rice **1:** 166-176.

Calviño M, Bruggmann R, Messing J (2011) Characterization of the small RNA component of the transcriptome from grain and sweet sorghum stems. BMC Genomics **12:** 356-367.

Calviño M, Miclaus M, Bruggmann R, Messing J (2009) Molecular Markers for Sweet Sorghum Based on Microarray Expression Data. Rice **2:** 129-142.

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S (2008) Community resources and strategies for association mapping in sorghum. Crop Science **48:** 30-40.

Childs K, Miller F, Cordonnier-Pratt M, Pratt L, Morgan P, Mullet J (1997) The sorghum photoperiod sensitivity gene, Ma3, encodes a phytochrome B. Plant Physiology **113:** 611-619.

Chuck G, Cigan A, Saeteurn K, Hake S (2007) The heterochronic maize mutant Corngrass1 results from overexpression of a tandem microRNA. Nature Genetics **39:** 544-549.

Chuck G, Tobias C, Sun L, Kraemer F, Li C, Dibble D, Arora R, Bragg J, Vogel J, Singh S, Simmons B, Pauly M, Hake S (2011) Overexpression of the maize Corngrass1 microRNA prevents flowering, improves digestibility, and increases starch content of switchgrass. Proceedings of the National Academy of Sciences **108:** 17550-17555.

Craufurd PQ, Mahalakshmi V, Bidinger FR, Mukuru SZ, Chantereau J, Omanga PA, Qi A, Roberts EH, Ellis RH, Summerfield RJ, Hammer GL (1999) Adaptation of sorghum: characterisation of genotypic flowering responses to temperature and photoperiod. Theoretical and Applied Genetics **99:** 900-911.

Devers E, Branscheid A, May P, Krajinski F (2011) Stars and symbiosis: microRNA- and microRNA*-mediated transcript cleavage involved in arbuscular mycorrhizal symbiosis. Plant Physiology **156:** 1990-2010.

Dugas D, Monaco M, Olsen A, Klein R, Kumari S, Ware D, Klein P (2011) Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. BMC Genomics **12:** 514.

Duncan R, Bockholt A, Miller F (1981) Descriptive comparison of senescent and nonsenescent sorghum genotypes. Agronomy Journal **73:** 849-853.

Gardner J, Maranville J, Paparozzi E (1994) Nitrogen use efficiency among diverse sorghum cultivars. Crop Science **34:** 728-733.

Geng S, Hills FJ, Johnson SS, Sah RN (1989) Potential Yields and On-Farm Ethanol Production Cost of Corn, Sweet Sorghum, Fodderbeet, and Sugarbeet. Journal of Agronomy and Crop Science **162:** 21-29.

George-Jaeggli B, Jordan D, van Oosterom E, Hammer G (2011) Decrease in sorghum grain yield due to the dw3 dwarfing gene is caused by reduction in shoot biomass. Field Crops Research **124:** 231-239.

Gudrun H-T, Karin H, Peter N, Johannes W (1996) Sucrose accumulation in sweet sorghum stem internodes in relation to growth. Physiologia Plantarum **97:** 277-284.

Harris K, Subudhi P, Borrell A, Jordan D, Rosenow D, Nguyen H, Klein P, Klein R, Mullet J (2007) Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. Journal of Experimental Botany **58:** 327-338.

Kakani V, Vu J, Allen L, Boote K (2011) Leaf photosynthesis and carbohydrates of CO2-enriched maize and grain sorghum exposed to a short period of soil

water deficit during vegetative development. Journal of Plant Physiology **168:** 2169-2176.

Ritter K, McIntyre C, Godwin I, Jordan D, Chapman S (2007) An assessment of the genetic relationship between sweet and grain sorghums, within Sorghum bicolor ssp. bicolor (L.) Moench, using AFLP markers. Euphytica **157:** 161-176.

Klein R, Mullet J, Jordan D, Miller F, Rooney W, Menz M, Franks C, Klein P (2008) The effect of tropical sorghum conversion and inbred development on genome diversity as revealed by high-resolution genotyping. Plant Genome **1:** 12-26.

Kumar CG, Afroze F, Rao PS, Belum VSR, Abhishek R, Rao RN, Sara K, Kumar AA, Ahmed K (2011) Characterization of Improved Sweet Sorghum Genotypes for Biochemical Parameters, Sugar Yield and Its Attributes at Different Phenological Stages. Sugar Technology **12:** 322-328.

Lee H, Yoo S, Lee J, Kim W, Yoo S, Fitzgerald H, Carrington J, Ahn J (2010) Genetic framework for flowering-time regulation by ambient temperature-responsive miRNAs in Arabidopsis. Nucleic Acids Research **38:** 3081-3093.

Leyva-González M, Ibarra-Laclette E, Cruz-Ramirez A, Herrera-Estrella L (2012) Functional and Transcriptome Analysis Reveals an Acclimatization Strategy for Abiotic Stress Tolerance Mediated by Arabidopsis NF-YA Family Members. PloS One **7 (10):** e48138.

Li W-X, Oono Y, Zhu J, He X-J, Wu J-M, Iida K, Lu X-Y, Cui X, Jin H, Zhu J-K (2008) The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. Plant Cell **20:** 2238-2251.

Lin Y, Schertz K, Paterson A (1995) Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. Genetics **141:** 391-411.

Lingle SE (1987) Sucrose metabolism in the primary culm of sweet sorghum during development. Crop Science **27:** 1214-1219.

Meng Y, Shao C, Gou L, Jin Y, Chen M (2011) Construction of microRNA- and microRNA*-mediated regulatory networks in plants. RNA Biology **8:** 1124-1148.

Michaels S, Bezerra I, Amasino R (2004) FRIGIDA-related genes are required for the winter-annual habit in Arabidopsis. Proceedings of the National Academy of Sciences **101:** 3281-3285.

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

Murphy R, Klein R, Morishige D, Brady J, Rooney W, Miller F, Dugas D, Klein P, Mullet J (2011) Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. Proceedings of the National Academy of Sciences **108:** 16469-16474.

Murray S, Rooney W, Hamblim M, Mitchell S, Kresovich S (2009) Sweet sorghum genetic diversity and association mapping for brix and height. Plant Genome **2:** 48-62.

Murray S, Rooney W, Mitchell S, Sharma A, Klein P, Mullet J, Kresovich S (2008) Genetic improvement of sorghum as a biofuel feedstock: II. QTL for stem and leaf structural carbohydrates. Crop Science **48:** 2180-2193.

Murray S, Sharma A, Rooney W, Klein P, Mullet J, Mitchell S, Kresovich S (2008) Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates. Crop Science **48:** 2165-2179.

Okada M, Lanzatella C, Saha M, Bouton J, Wu R, Tobias C (2010) Complete switchgrass genetic maps reveal subgenome collinearity, preferential pairing and multilocus interactions. Genetics **185:** 745-760.

Pao C, Morgan P (1986) Genetic Regulation of Development in Sorghum bicolar: I. Role of the Maturity Genes. Plant Physiology **82:** 575-580.

Paterson A, Bowers J, Bruggmann Rm, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti A, Chapman J, Feltus F, Gowik U, Grigoriev I, Lyons E, Maher C, Martis M, Narechania A, Otillar R, Penning B, Salamov A, Wang Y, Zhang L, Carpita N, Freeling M, Gingle A, Hash C, Keller B, Klein P, Kresovich S, McCann M, Ming R, Peterson D, Mehboob ur R, Ware D, Westhoff P, Mayer K, Messing J, Rokhsar D (2009) The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

Poethig R (2009) Small RNAs and developmental timing in plants. Current Opinion in Genetics & Development **19:** 374-378.

Quinby J (1967) The maturity genes of sorghum. Advances in Agronomy **19:** 267-305.

Raghuwanshi A, Birch R (2010) Genetic transformation of sweet sorghum. Plant Cell Reports **29:** 997-1005.

Rooney WL, Aydin S (1999) Genetic control of a photoperiod-sensitive response in Sorghum bicolor (L.) Moench. Crop Science **39:** 397-400.

Salas Fernandez M, Becraft P, Yin Y, Lubberstedt T (2009) From dwarves to giants? Plant height manipulation for biomass yield. Trends in Plant Science **14:** 454-461.

Schläppi M (2006) FRIGIDA LIKE 2 is a functional allele in Landsberg erecta and compensates for a nonsense allele of FRIGIDA LIKE 1. Plant Physiology **142:** 1728-1738.

Shiringani A, Frisch M, Friedt W (2010) Genetic mapping of QTLs for sugar-related traits in a RIL population of Sorghum bicolor L. Moench. Theoretical and Applied Genetics **121:** 323-336.

Srivasta A, Mehta S, Lindlof A, Bhargava S (2010) Over-represented promoter motifs in abiotic stress-induced DREB genes of rice and sorghum and their probable role in regulation of gene expression. Plant Signaling & Behavior **5:** 775-784.

Swaminathan K, Alabady M, Varala K, De Paoli E, Ho I, Rokhsar D, Arumuganathan A, Ming R, Green P, Meyers B, Moose S, Hudson M (2010) Genomic and small RNA sequencing of Miscanthus x giganteus shows the utility of sorghum as a reference genome sequence for Andropogoneae grasses. Genome Biology **11:** R12.

Tarpley L, Vietor DM, Miller FR (1996) Metabolism of sucrose during storage in intact sorghum stalk. International Journal of Plant Sciences **157:** 159-163.

Thomas H, Howarth C (2000) Five ways to stay green. Journal of Experimental Botany **51:** 329-337.

Vasilakoglou I, Dhima K, Karagiannidis N, Gatsis T (2011) Sweet sorghum productivity for biofuels under increased soil salinity and reduced irrigation. Field Crops Research **120:** 38-46.

Wang F, Liu C-Z (2009) Development of an Economic Refining Strategy of Sweet Sorghum in the Inner Mongolia Region of China. Energy & Fuels **23:** 4137-4142.

Wang J, Roe B, Macmil S, Yu Q, Murray J, Tang H, Chen C, Najar F, Wiley G, Bowers J, Van Sluys M-A, Rokhsar D, Hudson M, Moose S, Paterson A, Ming R (2010) Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. BMC Genomics **11:** 261.

Wang M, Zhu C, Barkley N, Chen Z, Erpelding J, Murray S, Tuinstra M, Tesso T, Pederson G, Yu J (2009) Genetic diversity and population structure analysis of accessions in the US historic sweet sorghum collection. Theoretical and Applied Genetics **120:** 13-23.

Wortmann CS, Liska A, Ferguson RB, Lyon DJ, Klein R, Dweikat I (2010) Dryland performance of sweet sorghum and grain crops for biofuel in Nebraska. Agronomy Journal **102:** 319-326.

Xin Z, Wang M, Barkley N, Burow G, Franks C, Pederson G, Burke J (2008) Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population. BMC Plant Biology **8:** 103.

Xue G-P, McIntyre C, Jenkins C, Glassop D, van Herwaarden A, Shorter R (2008) Molecular dissection of variation in carbohydrate metabolism related to water-soluble carbohydrate accumulation in stems of wheat. Plant Physiology **146:** 441-454.

Yan-an G, Hai-lian W, Ling Q, Hua-wen Z, Yan-bing Y, Feng-ju G, Ru-yu L, Hong-gang W (2011) QTL mapping of bio-energy related traits in Sorghum. Euphytica **182:** 431-440.

Zhao B, Ge L, Liang R, Li W, Ruan K, Lin H, Jin Y (2009) Members of miR-169 family are induced by high salinity and transiently inhibit the NF-YA transcription factor. BMC Molecular Biology **10:** 29.

Zhao B, Liang R, Ge L, Li W, Xiao H, Lin H, Ruan K, Jin Y (2007) Identification of drought-induced microRNAs in rice. Biochemical and Biophysical Research Communications **354:** 585-590.

Zhao YL, Dolat A, Steinberger Y, Wang X, Osman A, Xie GH (2009) Biomass yield and changes in chemical composition of sweet sorghum cultivars grown for biofuel. Field Crops Research **111:** 55-64.

Zhu Q-H, Helliwell C (2011) Regulation of flowering time and floral patterning by miR172. Journal of Experimental Botany **62:** 487-495.

Chapter 2 Screen of Genes Linked to High-Sugar Content in Stems by Comparative Genomics

2.1. Abstract

One of the great advantages of the fully sequenced rice genome is to serve as a reference for other cereal genomes in particular for identifying genes linked to unique traits. A trait of great interest is reduced lignocellulose in the stem of related species in favor of fermentable sugars as a source of biofuels. While sugarcane is one of the most efficient biofuel crops, little is known about the underlying gene repertoire involved in it. Here, we took advantage of the natural variation of sweet and grain sorghum to uncover genes that are conserved in rice, sorghum, and sugarcane but differently expressed in sweet versus grain sorghum by using a microarray platform and the syntenous alignment of rice and sorghum genomic regions containing these genes. Indeed, enzymes involved in carbohydrate accumulation and cell wall metabolism could be identified.

2.2. Introduction

Comparison of genetic maps and sequences of several grass species have shown that there is global conservation of gene content and order (Gale and Devos, 1998). Therefore, grasses have been considered as a “single genetic system” (Bennetzen and Freeling, 1993). The practical aspect of such a concept is of great

importance for agronomical purposes because a useful trait in one species could be transferred to another. A relevant example could be carbohydrate partitioning and allocation. In cereals such as wheat, sorghum, and rice, the process of grain filling demands carbon from photosynthesis assimilation as well as the remobilization of pre-stored carbohydrates in the stem before and after anthesis (Yang and Zhang, 2006). It has been estimated that about 30% of the total yield in rice depends on the carbohydrate content accumulated in the stem before heading (Ishimaru et al., 2007). For these reasons, characterization of genes involved in carbohydrate metabolism and accumulation can lead to the development of improved cereal crops.

In recent years, there has been an increasing demand on biomass for the production of ethanol as a renewable resource for fuel. The biggest producers of ethanol in the world are Brazil and the United States (US) (Ragauskas et al., 2006). In Brazil, it is derived from sugarcane, while in the US, ethanol is derived from the grain of corn. Because of the use of the entire plant as a source of fermentable sugars, carbohydrate accumulation and partitioning has been extensively studied in sugarcane, probably more that in any other species (Ming et al., 2001). However, genes involved in these processes cannot easily be identified because of the complex genome of sugarcane, with several cultivars differing greatly in their ploidy levels from 2n=100 to 2n=130 chromosomes (D'Hont et al., 1996; Grivet and Arruda, 2002). Even if one could make further improvements to sugarcane, it has the disadvantage of being a crop restricted to tropical growing areas.

On the other hand, the use of corn grain for ethanol production poses a major conflict because of its dual use as food and fuel. Therefore, it has been proposed to use grain solely for food and only the stover as a source for ethanol. A major impediment to this approach is that, in contrast to sugarcane, corn stover consists mainly of lignocellulose, which is more costly to process than fermentable sugars (Chapple and Carpita, 1998). Therefore, it would be attractive to identify corn varieties with reduced lignocellulose. Interestingly, there is extensive natural intra-species variation for sugar content in sorghum, with cultivars that accumulate much less sugar in stems (referred to as grain sorghums) in contrast to those that accumulate large amounts of sugars in their stems (referred as sweet sorghums) (Ali et al., 2007). Such intra-species variation can serve as a platform to identify genes linked to increased sugar content and reduced lignocellulose (Borevitz and Chory, 2004). Moreover, if these genes are conserved by ancestry in related species, one could envision the introduction of such a trait by the import of specific regulatory regions. Conservation of gene order between closely related species permits the alignment of orthologous chromosomal segments. Non-collinear genes would constitute paralogous copies (Messing and Bennetzen, 2008). To facilitate such alignments, the use of rice with one of the smallest cereal genomes that has been sequenced (International Rice Genome Sequencing, 2005) increasingly becomes the anchor genomes for other grasses (Messing and Llaca, 1998). In this sense, we can use rice as a reference genome for biofuel crops such as sugarcane and sorghum.

While rice offers an excellent reference as a compact genome from an evolutionary point of view, it is less suitable as a reference for a phenotype related to sugar accumulation in stems. Moreover, rice is a bambusoid C3 cereal plant, and sorghum and sugarcane are panicoid C4 cereal plants, which branched out 50 million years ago (mya) (Kellogg, 2001). Sorghum and sugarcane belong to the Saccharinae clade and diverged from each other only 8-9 mya (Guimaraes et al., 1997; Jannoo et al., 2007). Therefore, sugarcane can serve as a trait reference for sorghum varieties that differ in sugar content in their stems. Consequently, we took advantage of a GeneChip that was created to study gene expression in sugarcane and its role in the accumulation of sugar in the stem during development (Casu et al., 2007) for the comparison of grain and sweet sorghum genes. One would expect that sweet sorghum and sugarcane use similar gene products for enhanced sugar accumulation in their stems. Indeed, we identified genes involved in sugar accumulation and cell wall metabolism, and also demonstrated their ancestry through the alignment of orthologous regions of the rice and sorghum genomes. Therefore, the same genes could also be used to improve other biofuel crops such as switchgrass and *Miscanthus*.

2.3. Results

2.3.1. Sugar accumulation in the stem of grain and sweet sorghum cultivars

Previous reports have indicated that in sorghum stems, sugars start to accumulate at flowering stage (Lingle, 1987; Guimaraes et al., 1997). We compared

the accumulation of sugars in the stem between six sweet sorghum lines (Dale, Della, M81-E, Rio, Top76-6 and Simon), and one line representing grain sorghum (BTx623). As an estimation of the total amount of sugars present in the juice of sorghum stems, we measured the Brix degree of each internode along the main stem at the time of flowering. We found great variation in flowering time as well as in Brix degree between the sweet sorghum lines when compared to grain sorghum BTx623 (Fig. 2.1a and b). In general, the Brix degree was lower in the mature and immature internodes of the stem, in contrast to maturing internodes. These findings are in agreement with previous studies (Lingle, 1987; Guimaraes et al., 1997). Consistent with the inability of grain sorghum to accumulate significant levels of sugars in the stem, the Brix degree in BTx623 was low and remained fairly constant for all the internodes along the stem. Among the sweet sorghum cultivars, Rio had the highest Brix degree and Simon the lowest. Furthermore, the difference in flowering time between BTx623 and Rio was smaller that in the rest of sweet sorghum lines with high Brix degrees. For this reason, we decided to perform a comparative analysis of transcripts in the stem of the Rio and BTX623 sorghum lines.

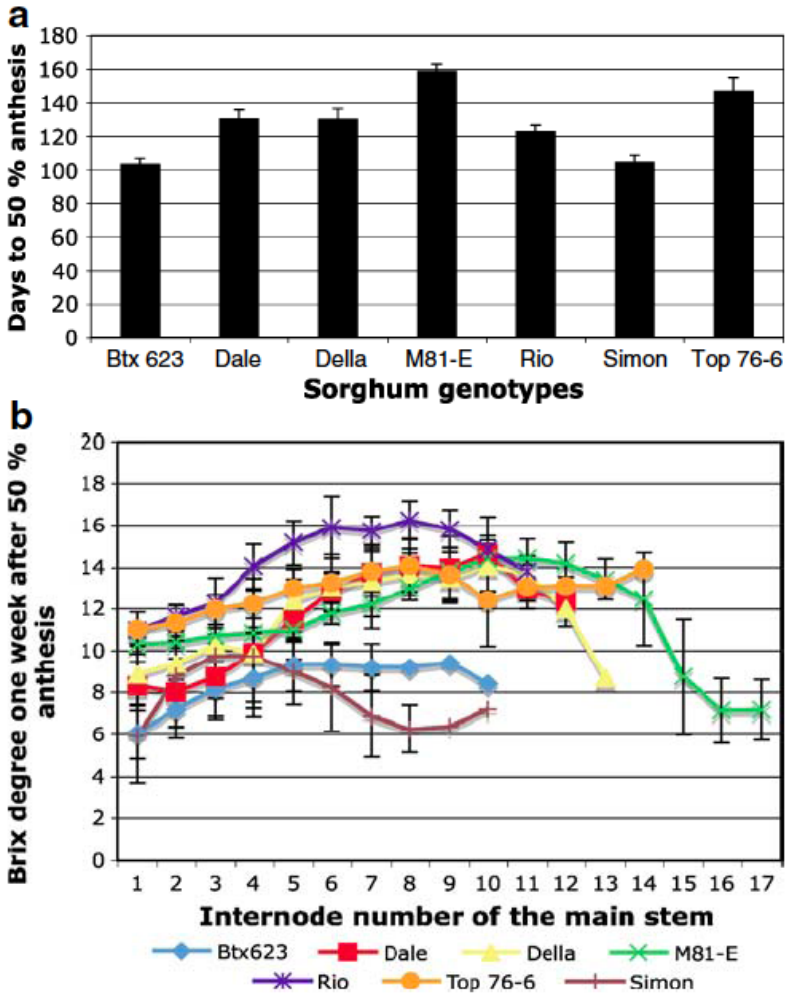


Figure 2.1. Variation in flowering time and Brix degree. (a) Comparison of flowering time between grain sorghum BTx623 and six sweet sorghum cultivars. Time to flowering was measured as days required to reach 50% anthesis. **(b)** Comparison of Brix degree along the main stem between grain sorghum BTx623 and six sweet sorghum genotypes. The Brix degree was measured for each internode, and the average of a triplicate experiment was plotted.

2.3.2. Microarray analysis of transcripts from sorghum stem tissues

In order to identify genes expressed in the stem with a potential role in sugar accumulation and reduced lignocellulose (Borevitz and Chory, 2004), we compared transcript profiles between grain (BTx623) and sweet sorghum (Rio). Such a genome-wide analysis became possible because of the recently designed GeneChip of sugarcane (Casu et al., 2007). This array was specifically developed with sequences that were obtained from several cDNA libraries representing distinct tissue types including stem from 15 different sugarcane varieties. The use of this GeneChip permitted us to directly compare gene expression data of two different sorghum cultivars with the previously generated data from sugarcane. Three independent plants for each BTx623 and Rio were grown until anthesis and RNA was extracted from the same maturing internode for all six plants. These RNAs were used to prepare biotylinated cRNAs for hybridization, each sample separatedly hybridized to one array.

The sugarcane array comprised 8,224 probe sets, of which more than 70% (5,900) gave a positive signal with sorghum RNA samples. When a twofold cut-off value was applied as criterion to distinguish differentially expressed transcripts between grain and sweet sorghum, a total of 195 transcripts were identified, with 132 transcripts being down-regulated and 63 transcripts up-regulated in Rio, respectively (Table 2.1 and Table 2.2). Because some probe sets indentify the same gene, the number of genes that was down-regulated was 103 and up-regulated 51, respectively. Based on the annotation of the sorghum genes, we were able to infer the possible function for most of the differentially expressed transcripts.

Among the transcripts that were up-regulated in Rio, a saposin-like type B gene displayed the highest differential expression. Saposins are involved in the degradation of sphingolipids (Munford et al., 1995). Other transcripts encoding stress-related proteins such as heat shock protein 70 (HSP70) and HSP90 were up-regulated, consistent with an osmotic stress imposed by high concentration of sugars (Table 2.1 and Table 2.2) (Buchanan et al., 2005). Our results showed that in Rio, down-regulated genes outnumbered those that were up-regulated by a factor of 2. The most reduced transcript encodes a fasciclin protein domain. This domain has been shown to be involved in cell adhesion (Table 2.3) (Kawamoto et al., 1998; Faik et al., 2006).

SUGARCANE PROBE SET ID	EXPRESSION	*S. bicolor* GENE ID	OsRAP2 GENE ID	*S. bicolor* FUNCTION	PFAM	DESCRIPTION	GO-TERM
UP-REGULATED							
Starch and sucrose metabolism							
SOF.4315.1.S1_AT	2.3	Sb03g003190.1	Os01g0190400	Hexokinase-8	PF03727 PF00349	Hexokinase Hexokinase	GO:0006096 GO:0006096
SOF.90.1.S1_AT	1.2	Sb03g040060.1	Os01g0851700	Phosphorylase	PF00343	Carbohydrate phosphorylase	GO:0005975
Sugar binding							
SOF.1513.1.A1_AT	2	Sb10g022730.1	Os06g0165200	Putative uncharacterized protein	PF01453 PF00024	D-mannose binding lectin PAN domain	GO:0005529 N/A
Cell wall catabolism							
SOF.3731.1.A1_AT	1.2	Sb01g049890.1	Os03g0110600	LysM domain containing protein	PF01476	LysM domain	GO:0016998
Transcription factor							
SOFAFFX.287.1.S1_AT	2.2	Sb10g007380.1	Os06g0217300	M21 protein	PF01486 PF00319	K-box region SRF-type transcription factor (DNA-binding and dimerisation domain)	GO:0005634 GO:0005634
SOF.2682.1.S1_AT	2	Sb01g013710.1	Os12g0612700	Class III HD-Zip protein 4	PF00046 PF01852	Homeobox domain START domain	GO:0005634 N/A

SOFAFFX.142.1.S1_AT	1.6	Sb04g005620.1	Os06g0646600	KNOX family class 2 homeodomain protein	PF03791 PF03789	KNOX2 domain ELK domain	GO:0005634 GO:0005634
SOF.3290.1.S1_AT	1.1	Sb08g016240.1	Os12g0507300	similar to Os12g0507300 protein	PF03106	WRKY DNA -binding domain	GO:0005634
Zinc-ion binding							
SOFAFFX.1438.1.A1_S_AT	2	Sb09g006050.1	Os01g0192000	Putative uncharacterized protein	PF00642	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	GO:0008270
SOF.603.1.A1_A_AT	1.6	Sb07g025220.1	Os08g0545200	Sorbitol dehydrogenase	PF08240 PF00107	N/A Zinc-binding dehydrogenase	N/A N/A
SOF.4452.1.A1_AT	1.3	Sb04g021610.1	Os02g0530300	Zinc finger A20 and AN1 domain-containing stress-associated protein 5	PF01754 PF01428	A20-like zinc finger AN1-like Zinc finger	GO:0008270 GO:0008270
SOF.1992.2.S1_AT	1.2	Sb02g039390.1	Os07g0618600	similar to Os07g0618600 protein	PF01363 PF00023	FYVE zinc finger Ankyrin repeat	GO:0008270 N/A
Oxidoreductase activity							
SOF.1594.1.S1_AT	2	Sb03g033250.1	Os01g0723400	NADP dependent malic enzyme	PF00390 PF03949	Malic enzyme, N-terminal domain Malic enzyme, NAD binding domain	GO:0016616 GO:0051287
SOF.398.1.A1_AT	1.4	Sb02g043370.1	Os07g0685800	Carbonyl reductase-like protein	PF00106	short chain dehydrogenase	
Carboxy-lyase activity							
SOF.3466.1.A1_AT	1.6	Sb07g022670.2	Os08g0465800	GAD1	PF00282	Pyridoxal-dependent decarboxylase conserved domain	GO:0019752
Translation initiation							
SOF.3301.1.S1_AT	1.4	Sb03g047210.1	Os01g0970400	Eukaryotic translation initiation factor 4E-1	PF01652	Eukaryotic initiation factor 4E	GO:0005737
Protein binding							
SOF.2770.2.S1_X_AT	1.4	Sb03g041770.1	Os01g0881900	Putative uncharacterized protein	PF00646 PF00560	F-box domain Leucine Rich Repeat	N/A
Protein catabolism							
SOFAFFX.1586.1.S1_AT	1.3	Sb03g025180.1	Os01g0550100	Ubiquitin carboxyl-terminal hydrolase	PF00240 PF00443	Ubiquitin family Ubiquitin carboxyl-terminal hydrolase	GO:0006464 GO:0006511
SOF.1683.1.S1_AT	1.2	Sb01g043060.1	Os03g0212700	Mitochondrial processing peptidase beta subunit	PF00675 PF05193	Insulinase (Peptidase family M16) Peptidase M16 inactive domain	GO:0006508 GO:0006508
Electron transport							
SOFAFFX.1192.1.S1_AT	1.3	Sb03g027710.1	Os01g0612200	Cytochrome c oxidase subunit Vb	PF01215	Cytochrome c oxidase subunit Vb	GO:0005740
SOF.2692.1.S1_AT	1	Sb08g002250.1	Os12g0119000	Cytochrome P450 51	PF00067	Cytochrome P450	GO:0006118

Membrane associated protein							
SOF.4998.1.S1_AT	1.3	Sb10g002420.1	Os06g0136000	Hypersensitive-induced reaction protein 4	PF01145	SPFH domain / Band 7 family	N/A
Alternative splicing							
SOF.3633.1.S1_AT	1.3	Sb01g046550.3	Os03g0158500	YT521-B-like family protein, expressed	PF04146	YT521-B-like family	N/A
Chaperonin activity							
SOF.3437.1.S1_AT	1.3	Sb09g022580.1	Os01g0840100	Heat shock cognate 70 kDa protein	PF00012	Hsp70 protein	N/A
Kinase activity							
SOFAFFX.494.1.S1_S_AT	1.2	Sb10g001310.1	Os06g0116100	Putative GAMYB-binding protein	PF00069 PF07714	Protein kinase domain Protein tyrosine kinase	GO:0006468 GO:0006468
Transferase activity							
SOF.1326.1.S1_A_AT	1.2	Sb02g000780.1	Os07g0108300	Alanine aminotransferase	PF00155	Aminotransferase class I and II	GO:0009058
Proton transport							
SOF.3139.1.S1_AT	1.1	Sb10g026440.1	Os02g0175400	Vacuolar ATP synthase catalytic subunit A	PF02874 PF00006	ATP synthase alpha/beta family, beta-barrel domain ATP synthase alpha/beta family, nucleotide-binding domain	GO:0016469 GO:0016469
SOFAFFX.1600.2.A1_AT	1	Sb09g027790.1	Os01g0685800	ATP synthase subunit beta, mitochondrial precursor	PF02874 PF00006	ATP synthase alpha/beta family, beta-barrel domain ATP synthase alpha/beta family, nucleotide-binding domain	GO:0016469 GO:0016469
Arginine biosynthesis							
SOFAFFX.1412.1.A1_S_AT	1	Sb08g008320.1	Os12g0235800	Argininosuccinate synthase	PF00764	Arginosuccinate synthase	GO:0006526
Metabolic process							
SOF.4917.1.S1_AT	1	Sb03g004390.1	Os05g0171000	Phospholipase D alpha 1	PF00168 PF00614	C2 domain Phospholipase D Active site motif	N/A GO:0008152
DNA methylation							
SOF.3784.1.A1_AT	1	Sb02g004680.1	Os07g0182900	Cytosine-specific methyltransferase	PF01426	BAH domain	GO:0003677
Response to stress							
SOF.2151.1.S1_AT	1	Sb09g004470.1	Os05g0157200	Putative uncharacterized protein P0676G05.12	PF00582	Universal stress protein family	GO:0006950
Vitamin C Synthesis							
SOFAFFX.630.1.S1_AT	1.1	Sb05g022890.1	Os11g0591100	GDP-mannose 3,5-epimerase 1	Pfam:N/A	Func:N/A	GO:N/A
Unknown function							
SOF.1282.2.S1_A_AT	1.4	Sb02g023980.1	Os09g0386600	Putative	Pfam:N/A	Func:N/A	GO:N/A

				uncharacterized protein			
SOF.2601.1.S1_AT	1.3	Sb08g016302.1	Os12g0508200	Expressed protein	Pfam:N/A	Func:N/A	GO:N/A
SOF.3798.1.S1_AT	1.2	Sb02g025720.1	Os09g0439200	Putative uncharacterized protein	PF06200	ZIM motif	N/A
SOF.366.1.S1_S_AT|SOF.366.2.S1_S_AT	1.1|1.3	Sb01g002220.1	Os03g0835150	Expressed protein	Pfam:N/A	Func:N/A	GO:N/A
SOF.2346.1.S1_AT	1.1	Sb03g028860.1	Os01g0633200	X1	PF03469 PF03468	XH domain XS domain	N/A N/A
SOF.32.1.S1_AT	1	Sb01g045110.1	Os03g0182400	SacIy domain containing protein, expressed	PF02383	SacI homology domain	N/A
DOWN-REGULATED							
Sucrose metabolism							
SOF.4165.1.S1_S_AT	-1.3	Sb01g033060.1	Os03g0401300	Sucrose synthase 2	PF00862 PF00534	Sucrose synthase Glycosyl transferases group 1	GO:0005985 GO:0009058
SOF.3644.2.S1_A_AT	-1.7	Sb07g001320.1	Os08g0113100	Fructokinase-2	PF00294	pfkB family carbohydrate kinase	N/A
Cell wall related							
SOF.1587.3.A1_A_AT	-1	Sb01g002050.1	Os03g0837100	Cellulose synthase-7	PF03552	Cellulose synthase	GO:0016020
SOF.5033.1.S1_AT	-1.1	Sb09g005280.1	Os05g0176100	Cellulose synthase-1	PF03552	Cellulose synthase	GO:0016020
SOF.4824.2.S1_A_AT SOFAFFX.1961.1.S1_S_AT	-1|-1.2	Sb02g006290.1	Os03g0808100	Cellulose synthase-9	PF03552	Cellulose synthase	GO:0016020
SOF.2699.2.S1_A_AT	-4.7	Sb02g025020.1	Os09g0422500	Cellulose synthase catalytic subunit 12	PF03552	Cellulose synthase	GO:0016020
SOF.3244.1.S1_A_AT	-1.8	Sb01g018400.1	Os10g0493600	Alpha-galactosidase precursor	PF02065	Melibiase	GO:0005975
SOF.4934.1.S1_AT	-2.4	Sb03g041450.1	Os05g0428100	Beta-galactosidase 3 precursor	PF02140 PF02837	Galactose binding lectin domain Glycosyl hydrolases family 2, sugar binding domain	GO:0005529 GO:0005975
SOF.3629.1.S1_AT	-2.9	Sb07g021680.1	Os09g0419200	Cinnamoyl CoA reductase	PF05368 PF01073	NmrA-like family 3-beta hydroxysteroid dehydrogenase/isomerase family	GO:0006808 GO:0006694
SOFAFFX.292.1.S1_AT|SOF.5198.2.S1_A_AT	-1.4|-2.9	Sb10g004540.1	Os06g0165800	Caffeoyl-CoA O-methyltransferase 2	PF01596	O-methyltransferase	GO:0008171
SOF.1122.2.S1_A_AT	-5.3	Sb07g028530.1	Os09g0481400	Caffeoyl-CoA O-methyltransferase	PF01596	O-methyltransferase	GO:0008171

SOF.1021.1.A1_AT	-3.5	Sb03g039520.1	Os01g0842400	Putative laccase	PF00394 F07731	Multicopper oxidase Multicopper oxidase	GO:0016491 GO:0016491
SOF.4734.1.S1_AT	-3.7	Sb04g005210.1	Os02g0177600	4-coumarate coenzyme A ligase	PF00501	AMP-binding enzyme	GO:0008152
Cell adhesion							
SOFAFFX.1406.1.S1_AT|SOF.4464.1.A1_AT	-1.9|-1.6	Sb01g005770.1	Os03g0788600	Expressed protein	PF02469	Fasciclin domain	GO:0007155
SOF.3590.1.S1_AT	-6.5	Sb09g028490.1	Os05g0563600	Fasciclin-like protein FLA15	PF02469	Fasciclin domain	GO:0007155
Carbohydrate metabolic process							
SOF.4949.1.S1_AT	-1.3	Sb03g045390.1	Os01g0939600	similar to Os01g0939600 protein	PF01210 PF07479	NAD-dependent glycerol-3-phosphate dehydrogenase N-terminus NAD-dependent glycerol-3-phosphate dehydrogenase C-terminus	GO:0005737 GO:0005975
Water transport							
SOF.863.1.S1_S_AT	-1	Sb10g008090.1	Os06g0228200	Aquaporin NIP2-3	PF00230	Major intrinsic protein	GO:0016020
Protein binding							
SOF.5088.1.S1_AT	-1	Sb04g027910.2	Os02g0748300	Kelch repeat-containing F-box-like	PF07646 PF00646	Kelch motif F-box domain	N/A N/A
SOF.4911.1.S1_AT	-1.5	Sb01g045010.1	Os03g0183800	Leucine-rich repeat transmembrane protein kinase 2	PF00560	Leucine Rich Repeat	GO:0005515
Mitochondrial envelop/electron transport							
SOF.4557.1.S1_AT	-1	Sb03g037870.1	Os01g0814900	Cytochrome b5 reductase	PF00970 PF00175	Oxidoreductase FAD-binding domain Oxidoreductase NAD-binding domain	GO:0006118 GO:0006118
DNA binding/ transcription factor							
SOF.3143.2.S1_A_AT	-1	Sb03g043690.1	Os01g0915600	Putative uncharacterized protein	PF00010	Helix-loop-helix DNA-binding domain	GO:0005634
SOF.2024.1.S1_AT	-1.4	Sb07g020090.1	Os08g0408500	DRE binding factor 1	PF00847	AP2 domain	GO:0005634
SOFAFFX.1576.1.S1_AT	-3.2	Sb03g030750.1	Os01g0672100	No apical meristem (NAM) protein-like	PF02365	No apical meristem (NAM) protein	GO:0045449
Kinase activity							
SOF.1818.1.S1_AT	-1	Sb02g037070.1	Os07g0572800	Mitogen activated protein kinase kinase	PF00069	Protein kinase domain	GO:0006468
Transferase activity							
SOF.1190.1.S1_AT	-1	Sb07g005930.1	Os08g0205900	Putative uncharacterized protein	PF00202	Aminotransferase class-III	GO:0030170

SOF.701.1.S1_AT	-1.3	Sb03g003390.1	Os01g0185300	Putative acyl transferase 3	PF02458	Transferase family	N/A
SOF.521.2.S1_AT	-1.1	Sb10g002230.1	Os06g0133900	3-phosphoshikimate 1-carboxyvinyltransferase	PF00275	EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase)	GO:0016765
SOFAFFX.409.1.S1_AT	-3.8	Sb06g021640.1	Os04g0500700	similar to OSJNBa0029H02.19 protein	PF02458	Transferase family	N/A
Nucleoside Transport							
SOF.3699.1.A1_AT	-1.4	Sb07g005850.1	Os08g0205200	Equilibrative nucleoside transporter 1	PF01733	Nucleoside transporter	GO:0016020
Cation transport							
SOF.1478.1.A1_AT	-1.4	Sb02g005440.1	Os07g0191200	Cation-transporting ATPase	PF00690	Cation transporter/ATPase, N-terminus	GO:0016020
Transporter activity							
SOF.2138.1.S1_AT	-1.9	Sb04g028300.1	Os02g0741800	Root uracil permease 1	PF00860	Permease family	GO:0016020
Zinc-ion binding							
SOF.808.1.S1_AT	-1.1	Sb09g000820.1	Os05g0106000	Putative uncharacterized protein	PF00096	Zinc finger, C2H2 type	
Metabolic process							
SOF.4186.2.S1_AT	-1.1	Sb06g015180.1	Os04g0404800	similar to H0502B11.5 protein	PF00501	AMP-binding enzyme	GO:0008152
Cysteine protease inhibitor activity							
SOF.117.1.S1_AT	-1.1	Sb09g024230.1	Os05g0494200	Cystatin	PF00031	Cystatin domain	GO:0004869
Hydrolase activity							
SOF.4601.1.S1_AT	-1.2	Sb01g041550.1	Os03g0238600	Purple acid phosphatase 1, putative, expressed	PF00149	Calcineurin-like phosphoesterase	GO:0016787
Kreb's cycle/ transferase activity							
SOF.2225.1.S1_AT	-2.2	Sb04g006440.1	Os02g0194100	Citrate synthase	PF00285	Citrate synthase	GO:0046912
Electron transport							
SOF.1998.1.A1_AT	-1.3	Sb02g036870.1	Os07g0570550	Chromosome chr5 scaffold_2, whole genome shotgun sequence	PF02298	Plastocyanin-like domain	GO:0006118
Protein translation							
SOF.4846.1.S1_AT SOF.4846.2.S1_A_AT	-1.5 -1.2	Sb04g007760.1	Os02g0220600	Elongation factor 1-gamma 1	PF00647 PF00043	Elongation factor 1 gamma, conserved domain Glutathione S-transferase, C-terminal domain	GO:0005853 N/A
SOF.3827.1.S1_S_AT	-1.7	Sb07g002560.1	Os08g0130500	60S acidic ribosomal protein P0	PF00428 PF00466	60s Acidic ribosomal protein Ribosomal protein L10	GO:0005840 GO:0005622
SOF.177.2.A1_AT	-1.8	Sb03g007840.1	Os01g0120800	Eukaryotic translation initiation factor 3 subunit 10	PF01399	PCI domain	N/A
SOFAFFX.1035.1.S1_S_AT	-2	Sb09g023950.1	Os01g0812800	60S ribosomal protein L30	PF01248	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family	N/A

SOF.1902.1.S1_S_AT	-2.6	Sb05g001680.1	Os12g0124200	40S ribosomal protein S16	PF00380	Ribosomal protein S9/S16	GO:0005840
Trypsin-alpha amylase inhibitor							
SOF.3279.1.S1_AT	-1.4	Sb08g002660.1	Os12g0115300	Non-specific lipid-transfer protein	PF00234	Protease inhibitor/seed storage/LTP family	N/A
Methionine metabolism							
SOF.3126.1.S1_AT	-1.4	Sb01g003700.1	Os03g0815200	Methylenetetrahydrofolate reductase 1	PF02219 PF00122	Methylenetetrahydrofolate reductase E1-E2 ATPase	GO:0016020
Calcium ion binding							
SOFAFFX.1248.1.S1_AT	-1.6	Sb01g048570.1	Os03g0128700	Calcium-dependent protein kinase isoform 11	PF00036 PF00036	EF hand EF hand	GO:0005509 GO:0005509
Cytoskeleton							
SOF.4093.2.S1_AT	-1.7	Sb01g009560.2	Os03g0726100	Tubulin alpha-2/alpha-4 chain	PF00091 PF03953	Tubulin/FtsZ family, GTPase domain Tubulin/FtsZ family, C-terminal domain	N/A GO:0043234
SOF.151.1.S1_AT	-1.7	Sb04g037170.1	Os02g0816500	Tubulin folding cofactor A	PF02970	Tubulin binding cofactor A	GO:0005874
SOF.110.1.A1_AT	-2	Sb06g029500.1	Os04g0629700	similar to OSJNBa0089N06.17 protein	PF00225	Kinesin motor domain	GO:0005875
Regulation of nitrogen utilization							
SOF.3747.1.S1_A_AT	-2.2	Sb03g008760.1	Os01g0106400	Isoflavone reductase homolog IRL	PF05368 PF01073	NmrA-like family 3-beta hydroxysteroid dehydrogenase/isomerase family	GO:0006808 GO:0006694
DNA binding							
SOF.4234.1.S1_A_AT	-2.4	Sb10g002040.1	Os06g0130900	Histone H3.3	PF00125	Core histone H2A/H2B/H3/H4	GO:0003677
SOF.5269.1.S1_AT	-1.7	Sb02g025440.1	Os09g0433600	Histone H4	Pfam:N/A	Func:N/A	GO:N/A
Aromatic aminoacid biosynthesis							
SOF.2944.1.S1_AT	-2.8	Sb01g033590.1	Os07g0622200	Phospho-2-dehydro-3-deoxyheptonate aldolase 1, chloroplast precursor	PF01474	Class-II DAHP synthetase family	GO:0009073
Fatty acid biosynthesis							
SOF.2629.3.S1_A_AT	-3	Sb03g012420.1	Os01g0300200	ATP citrate lyase, putative	PF00549	CoA-ligase	GO:0008152
Protein ADP-ribosylation							

SOF.4942.3.S1_A_AT|SOF.4942.2.S1_AT|SOF.4942.1.S1_AT	-2.4|-3.5|-3.5	Sb03g013840.1	Os01g0351100	Poly [ADP-ribose] polymerase 2 (EC 2.4.2.30) (PARP-2)	PF00644 PF02877	Poly(ADP-ribose) polymerase catalytic domain Poly(ADP-ribose) polymerase, regulatory domain	GO:0005634 GO:0005634
Signal transduction							
SOF.285.1.S1_AT	-3.7	Sb08g018765.1	Os12g0570000	Protein spotted leaf 11	PF00514	Armadillo/beta-catenin-like repeat	N/A
Unknown function							
SOF.4866.1.S1_AT	-1.1	Sb08g020760.1	Os12g0604800	Tetratricopeptide repeat protein, putative, expressed	PF00515	Tetratricopeptide repeat	N/A
SOF.3234.1.S1_AT	-1.1	Sb01g011740.1	Os03g0685500	similar to Putative uncharacterized protein OSJNBb0072E24.9	PF06747	CHCH domain	N/A
SOF.3225.2.S1_A_AT	-1.1	Sb02g026990.1	Os09g0465500	similar to Os02g0781700 protein	Pfam:N/A	Func:N/A	GO:N/A
SOFAFFX.794.1.S1_S_AT	-1.2	Sb09g029170.1	Os01g0652600	Putative uncharacterized protein	PF01450	Acetohydroxy acid isomeroreductase, catalytic domain	GO:0009082
SOF.849.1.A1_AT	-1.2	Sb09g023620.1	Os01g0818600	Unkown protein	PF00560	Leucine Rich Repeat	GO:0005515
SOF.5337.2.S1_AT	-1.2	Sb01g006220.1	Os07g0142000	Putative uncharacterized protein	PF02453	Reticulon	GO:0005783
SOF.4768.1.A1_AT	-1.2	Sb01g012470.1	Os03g0666700	Expressed protein	PF05967	Eukaryotic protein of unknown function (DUF887)	N/A
SOF.2335.1.S1_AT	-1.3	Sb03g026700.1	Os01g0593200	Putative uncharacterized protein	PF04570	Protein of unknown function (DUF581)	N/A
SOF.1965.1.S1_AT	-1.3	Sb09g022110.1	Os05g0451300	Putative uncharacterized protein	Pfam:N/A	Func:N/A	GO:N/A
SOF.3739.1.S1_S_AT	-1.4	Sb06g026710.1	Os04g0586200	similar to H0307D04.13 protein	PF04570	Protein of unknown function (DUF581)	N/A
SOF.1054.1.S1_AT	-1.4	Sb03g042480.1	Os01g0894700	Putative uncharacterized protein	Pfam:N/A	Func:N/A	GO:N/A
SOF.466.1.S1_AT	-1.5	Sb07g001710.1	Os08g0117900	Putative glycine-rich protein	Pfam:N/A	Func:N/A	GO:N/A
SOFAFFX.868.1.S1_S_AT	-1.7	Sb02g009980.1	Os07g0418200	Putative uncharacterized protein	Pfam:N/A	Func:N/A	GO:N/A
SOF.2471.1.S1_AT	-1.4	Sb02g006420.1	Os07g0211900	Putative bZIP protein	PF04783 PF04782	Protein of unknown function (DUF630) Protein of unknown function (DUF632)	N/A
SOF.2465.1.S1_AT	-1.4	Sb02g032470.1	Os09g0556700	similar to Os09g0556700 protein	PF00856	SET domain	GO:0005634
SOF.4919.1.S1_AT	-1.5	Sb02g022510.1	Os09g0344800	Membrane protein-like	PF01925	Domain of unknown function DUF81	GO:0016021
SOF.4946.2.S1_A_AT	-1.6	Sb03g010350.1	Os01g0265100	Putative uncharacterized protein	PF00025 PF08477	ADP-ribosylation factor family N/A	GO:0005525 GO:0005622
SOF.807.1.S1_AT	-1.7	Sb02g002940.1	Os07g0148800	weakly similar to Chromosome chr10	PF00560	Leucine Rich Repeat	GO:0005515

				scaffold_43			
SOF.4652.1.S1_AT	-1.7	Sb03g037360.2	Os05g0494500	Phosphate/phosphoenolpyruvate translocator protein-like	Pfam:N/A	Func:N/A	GO:N/A
SOF.3249.1.S1_AT	-1.8	Sb02g043510.1	Os03g0319300	Putative uncharacterized protein	PF03959 PF00036	Domain of unknown function (DUF341) EF hand	N/A GO:0005509
SOF.3649.1.A1_AT	-2	Sb01g007870.1	Os03g0751600	Expressed protein	Pfam:N/A	Func:N/A	GO:N/A
SOF.3476.1.S1_AT	-2.1	Sb03g009900.1	Os01g0257100	Putative uncharacterized protein	PF05498	Rapid ALkalinization Factor (RALF)	N/A
SOF.3418.2.S1_AT|SOF.3418.3.S1_A_AT	-2.2|-2	Sb01g009520.1	Os03g0726500	Expressed protein	Pfam:N/A	Func:N/A	GO:N/A
SOFAFFX.1105.1.S1_AT	-2.4	Sb06g022030.1	Os04g0508000	similar to OSJNBb0002J11.24 protein	PF03005	Arabidopsis proteins of unknown function	N/A
SOF.3624.1.S1_AT	-3	Sb03g005150.1	Os01g0249200	Putative uncharacterized protein	PF00190 PF07883	Cupin Cupin domain	GO:0045735 N/A
SOFAFFX.1040.1.S1_AT	-3.2	Sb03g010380.1	Os01g0265800	Putative uncharacterized protein	PF00076 PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain) RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	GO:0003676 GO:0003676
SOF.848.1.A1_AT	-3.5	Sb01g016110.1	Os03g0571900	similar to Os03g0571900 protein	PF01554 PF01554	MatE MatE	GO:0016020 GO:0016020
SOF.5314.1.A1_AT	-3.6	Sb04g025760.1	Os02g0611800	Putative uncharacterized protein	PF02458	Transferase family	N/A
SOF.2354.1.S1_A_AT	-3.9	Sb03g025160.1	Os01g0550300	Putative uncharacterized protein	Pfam:N/A	Func:N/A	GO:N/A

Table 2.1. List of differentially expressed genes between grain and sweet sorghum that have an orthologous copy in a syntenic position in rice.

The function for each gene is based on its Pfam domain and Gene Ontology (GO) term. The annotation of rice genes is based on RAP-DB (http://rapdb.dna.affrc.go.jp/). The expression is shown as Log2 mean ratio, with a

positive or negative fold change indicating increased or decreased expression in sweet sorghum Rio with respect to grain sorghum BTx623. Genes previously reported by Casu et al. (2007) are shown in red.

SUGARCANE PROBE SET ID	EXPRESSION	*S. bicolor* GENE ID	*S. bicolor* FUNCTION	PFAM	DESCRIPTION	GO-TERM
UP-REGULATED						
Cell wall related						
Sof.383.1.S1_at	1	Sb10g006290	Similar to Os11g0622800	PF00107 PF08240	Zinc-binding dehydrogenase Alcohol dehydrogenase; GroES-like domain	
Chaperonin activity						
Sof.1066.2.A1_x_at	1	Sb07g028270.1	Similar to Heat shock protein 82	PF00183 PF02518	Hsp90 protein Histidine-kinase; DNA gyrase B; HSP90-like ATPase	GO:0006457 GO:0005524
Transcription factor						
Sof.4567.2.S1_a_at/ Sof.4567.1.S1_at	1.4/ 1.3	Sb01g044810	Similar to putative MADS-domain transcription factor	PF00319	SRF-type transcription factor (DNA-binding and dimerisation domain)	GO:0005634
				PF01486	K-box region	GO:0005634
Proteolisis						
SofAffx.102.1.S1_at	1	Sb01g033620	Similar to Os03g0388900	PF00656	Caspase domain	GO:0006508
Nucleic acid binding						
Sof.3151.2.S1_a_at	1	Sb04g025670	Similar to putative uncharacterized protein	PF00076	RNA recognition motif. (a.k.a. RPM, RBD or RNP domain)	GO:0003676
Unknown function						
Sof.405.2.S1_a_at	5.7	Sb09g013990	Similar to putative uncharacterized protein		Similar to Saposin type B protein	
Sof.4787.1.A1_at	1	Sb01g026550.1	Similar to Os10g0135600	PF00561	Alpha/beta hydrolase fold	
SofAffx.403.1.S1_at	1.3	Sb10g002980	Similar to putative uncharacterized protein		Unknown	
Sof.22.1.S1_at	3.2	Sb01g041540	Similar to Purple acid phosphataase 1, putative, expressed		Unknown	
Sof.4906.1.S1_at	1	Sb01g023540	Similar to expressed protein		Unknown	
DOWN-REGULATED						
Cell wall related						

Sof.1987.1.S1_at	-1.5	Sb04g011550	Putative cinnamyl alcohol dehydrogenase	PF01073 PF01370 PF07993	3-beta hydroxysteroid dehydrogenase/isomerase family NAD dependent epimerase/dehydratase family Male sterility protein	G0:0006694 GO:0044237
Sof.1519.2.S1_at	-1.4	Sb02g006330	Putative Dolichyl-diphosphooligosaccharide-protein	PF03345	Dolichyl-diphosphooligosaccharide-protein glycosyltransferase 48kD subunit	GO:0005789
Sof.3569.2.S1_at	-1.1	Sb06g015880	Xyloglucan endo-transglycosylase/hydrolase precursor	PF00722 PF06955	Glycosyl hydrolases family 16 Xyloglucan endo-transglycosylase (XET) C-terminus	GO:0005975 GO:0048046
Sof.4258.2.S1_a_at	-1.5	Sb05g027350	Putative Xylanase inhibitor		Unknown	
Sof.4229.2.S1_a_at	-1.1	Sb02g029640	Similar to Glycoside hydrolase family 1 protein	PF00232	Glycosyl hydrolase family 1	GO:0005975
Sof.478.2.S1_at	-1.5	Sb02g004660	Similar to Putative Xylanase inhibitor protein precursor	PF00704	Glycosyl hydrolases family 18	GO:0005975
Sof.3100.1.S1_at	-2	Sb04g026520	Similar to Phenylalanine and histidine ammonia-lyase	PF00221	Phenylalanine and histidine ammonia-lyase	GO:0009058
Sof.3641.1.A1_at	-1.5	Sb02g037840	Similar to plasma membrane bound peroxidase	PF00141	Peroxidase	GO:0006979
Acyl CoA binding						
SofAffx.816.1.S1_at	-2.6	Sb07g004260	Similar to Acyl CoA binding protein	PF00887	Acyl CoA binding protein	GO:0000062
Cystein protease inhibitor activity						
SofAffx.772.1.S1_s_at	-3.2	Sb03g037370	Similar to Cystatin	PF00031	Cystatin domain	GO:0004869
Translation/Ribosome						
Sof.3035.1.S1_at	-1.3	Sb08g015010	Similar to Ribosomal protein S6 RPS6-1	PF01092	Ribosomal protein S6e	GO:0005840
Electron transport						
Sof.5340.1.S1_at	-1.7	Sb01g047640	Similar to Cytochrome P450 family protein, expressed	PF00067	Cytochrome P450	GO:0006118
Proteolisis						
Sof.15.2.S1_a_at	-1.4	Sb08g020950	Weakly similar to serine carboxypeptidase	PF00450	Serine carboxypeptidase	GO:0006508
Unknown function						
SofAffx.778.1.S1_s_at/ Sof.258.1.S1_at	-1.2	Sb09g006610	Putative uncharacterized protein	PF00069 PF07714	Protein kinase domain Protein tyrosine kinase	GO:0006468 GO:0006468
Sof.3156.2.S1_a_at	-1.5	Sb09g0200860	Unknown protein	PF03083	MtN3/saliva family	
Sof.3284.1.S1_at	-2.7	Sb10g000510	Putative uncharacterized protein	PF00234	Protease inhibitor/seed storage/LTP family	
Sof.4668.1.S1_at	-1.6	Sb07g006900	Similar to putative		Unknown	

			uncharacterized protein			
Sof.498.1.A1_at	-2.9	Sb02g003020	Similar to express protein	PF07967	C3HC zinc finger-like	GO:0005634

Table 2.2. List of differentially expressed genes between grain and sweet sorghum with no orthologous copy in a syntenic position in rice.

The function for each gene is based on its Pfam domain and Gene Ontology (GO). The expression is shown as Log2 mean ratio, with a positive or negative fold change indicating increased or decreased expression in sweet sorghum Rio with respect to grain sorghum BTx623. Gene copies previously reported by Casu et al. (2007) are shown in red.

GENE[a]	RICE	SORGHUM	EXPRESSION[b]
Starch and sucrose metabolism			
Hexokinase 8	Os01g0190400	Sb03g003190.1[c]	2.3
Hexokinase 8	*Os05g0187100*	*Sb09g005840.1*	
Carbohydrate phosphorylase	Os01g0851700	Sb03g040060.1[c]	1.2
Sucrose synthase 2	Os03g0401300	Sb01g033060.1[c]	-1.3
Sucrose synthase 2	*Os07g0616800*		
Fructokinase 2	Os08g 0113100	Sb07g001320.1[c]	-1.7
Sorbitol dehydrogenase 2	Os08g0545200	Sb07g025220.1[c]	1.6
Sugar binding			
D-mannose binding lectin	Os06g0165200	Sb10g022730.1[c]	2
CO_2 assimilation			
NADP-dependent malic enzyme	Os01g0723400	Sb03g033250.1[c]	2
Cell-wall-related			
LysM domain protein	Os03g0110600	Sb01g049890.1[c]	1.2
Cellulose synthase 7	Os03g0837100	Sb01g002050.1[c]	-1
Cellulose synthase 1	Os05g0176100	Sb09g005280.1[c]	-1.1
Cellulose synthase 9	Os07g0208500	Sb02g006290.1[c]	-1.1
Cellulose synthase 9	*Os03g0808100*	*Sb01g004210.1*	
Cellulose synthase catalytic subunit 12	Os09g0422500	Sb02g025020.1[c]	-4.7
Alpha-galactoside precursor	Os10g0493600	Sb01g018400.1[c]	-1.8
Beta-galactoside 3 precursor	Os01g0875500	Sb03g041450.1[c]	-2.4
Beta-galactoside 3 precursor	*Os05g0428100*	*Sb03g041450.1*	
Cinnamoyl CoA reductase	Os08g0441500	Sb07g021680.1[c]	-2.9
Cinnamoyl CoA reductase	*Os09g0419200*	*Sb10g005700.1*	
Laccase	Os01g0842400	Sb03g039520.1[c]	-3.5
4-Coumarate coenzyme A ligase	Os02g0177600	Sb04g005210.1[c]	-3.7
4-Coumarate coenzyme A ligase	*Os06g0656500*	*Sb10g026130.1*	

Fasciclin domain	Os03g0788600	Sb01g005770.1[c]	-1.75
Fasciclin domain	*Os07g0160600*	*Sb02g003410.1*	
Fasciclin-like protein FLA15	Os05g0563600	Sb09g028490.1[c]	-6.5
Caffeoyl-CoA O-methyltransferase 2	Os06g0165800	Sb10g004540.1[c]	-2.15
Caffeoyl-CoA O-methyltransferase	Os08g0498100	Sb07g028530.1[c]	-5.3
Caffeoyl-CoA O-methyltransferase	*Os09g0481400*	*Sb02g027930.1*	

Table 2.3. List of "Trait Specific" Genes that are syntenic with rice.

In boldface: sorghum genes that correspond to sugarcane probe set IDs previously reported by Casu et al. (2007). [a] Paralogs in italics. [b] Mean Log2 Ratio of sweet sorghum versus grain sorghum. [c] Sorghum gene to which a sugarcane probe set was mapped.

2.3.3. Genes with altered expression in carbohydrate metabolism in sweet sorghum

Based on Gene Ontology (GO) terms (http://www.geneontology.org), the sucrose and starch metabolic pathway from the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg), and the Carbohydrate-Active enzymes (CAZy) database (http://www.cazy.org), we found that almost 16% of the transcripts that were differentially expressed between BTx623 and Rio corresponded to transcripts affecting carbohydrate metabolism (Tables 2.3 and 2.4). Based on the link between hypothetical function and the sweet sorghum trait, we selected 37 candidate genes, of which differential expression could be indicative of increased sugar content in sorghum stems. Contrary to our expectations, the expression of sucrose phosphate synthase, a key enzyme in carbohydrate

metabolism, was not detected as differentially expressed in stems of grain and sweet sorghum.

Transcripts that were up-regulated included hexokinase 8 and carbohydrate phosphorylase (starch and sucrose metabolism), nicotinamide adenine dinucleotide phosphate (NADP) malic enzyme (C4 photosynthesis), a D-mannose binding lectin (sugar binding), and a lysine motif (LysM) domain protein possibly involved in cell wall degradation. Transcripts that were down-regulated included sucrose synthase 2 and fructokinase 2 (starch and sucrose metabolism), alpha galactosidase and beta-galactosidase (hydrolysis of glycosidic bonds), and cellulose synthase 1, 7, and 9 together with cellulose synthase catalytic subunit 12 (cell wall metabolism). In addition, several other transcripts with a cell-wall-related role that were down-regulated included cinnamoyl CoA reductase, cinnamyl alcohol dehydrogenase, 4-coumarate coenzyme A ligase, caffeoyl-CoA O-methyltransferase, xyloglucan endo-transglycosylase/hydrolase, peroxidase and phenylalanine, and histidine ammonia-lyase.

GENE	SORGHUM	EXPRESSION[a]
Cell-wall-related		
Alcohol dehydrogenase	Sb10g006290.1	1
Cinnamyl alcohol dehydrogenase	Sb04g011550.1	-1.5
Dolichyl-diphospho-oligosaccharide	Sb02g006330.1	-1.4
Xyloglucan endo-transglycosylase/hydrolase	Sb06g015880.1	-1.1
Putative Xylanase inhibitor	Sb05g027350.1	-1.5
Putative Xylanase inhibitor	Sb02g004660.1	-1.5
Glycoside hydrolase family 1	Sb02g029640.1	-1.1
Phenylalanine and histidine ammonia-lyase	Sb04g026520.1	-2
Peroxidase	Sb02g037840.1	-1.5
Similar to Saposin type B protein	Sb09g013990.1	5.7

Table 2.4. List of "Trait-Specific" genes that are not syntenic with rice.

In boldface: sorghum genes that correspond to sugarcane probe set IDs previously reported by Casu et al. (2007). [a] Mean Log2 Ratio of sweet versus grain sorghum.

2.3.4 Validation of microarray data by quantitative reverse transcription polymerase chain reaction

To validate the data obtained by microarray analysis, we chose 14 of the 37 candidate genes and compared their expression levels in both Rio and BTx623 by performing quantitative reverse transcription polymerase chain reaction (qRT-PCR) (Figure 2.2a). In Rio, the expression of saposin, carbohydrate phosphorylase, hexokinase 8, and NADP malic enzyme was up-regulated compared with their expression in BTx623. In contrast, the expression of fasciclin-like protein FLA15, cellulose synthase 1 and 7, fructokinase 2, 4-coumarate coenzyme A ligase, sucrose synthase 2, laccase, cinnamoyl CoA reductase, beta-galactosidase 3 precursor, and alpha galactosidase precursor were down-regulated in Rio. Although the levels of gene expression between the microarray and the qRT-PCR method differed to some extent, there was no difference in the classification of up- and down-regulated genes. The correspondence of microarray with qRT-PCR data illustrated that the microarray platform can be used as an effective method for screening large amounts of genes for a particular trait across closely related species. Before one would embark on any further experimentation, a much smaller candidate gene set can then be tested by more labor-intensive methods for the analysis of gene expression between cultivars of the same species. In order to see if the expression difference between BTx623 and Rio for the transcripts encoding a saposin-type B protein and a

fasciclin-like protein FLA15 also extended to other sweet sorghum lines, we extracted RNA from maturing stems of BTx623, Dale and Della at flowering and measured the expression of both genes by qRT-PCR. We found that the sapsin-type B gene was also highly expressed in Dale and Della when compared to grain sorghum BTx623, and the opposite was true for the expression of fasciclin-like protein FLA15, highly expressed in BTx623 compared to Dale and Della (Fig. 2.2b).

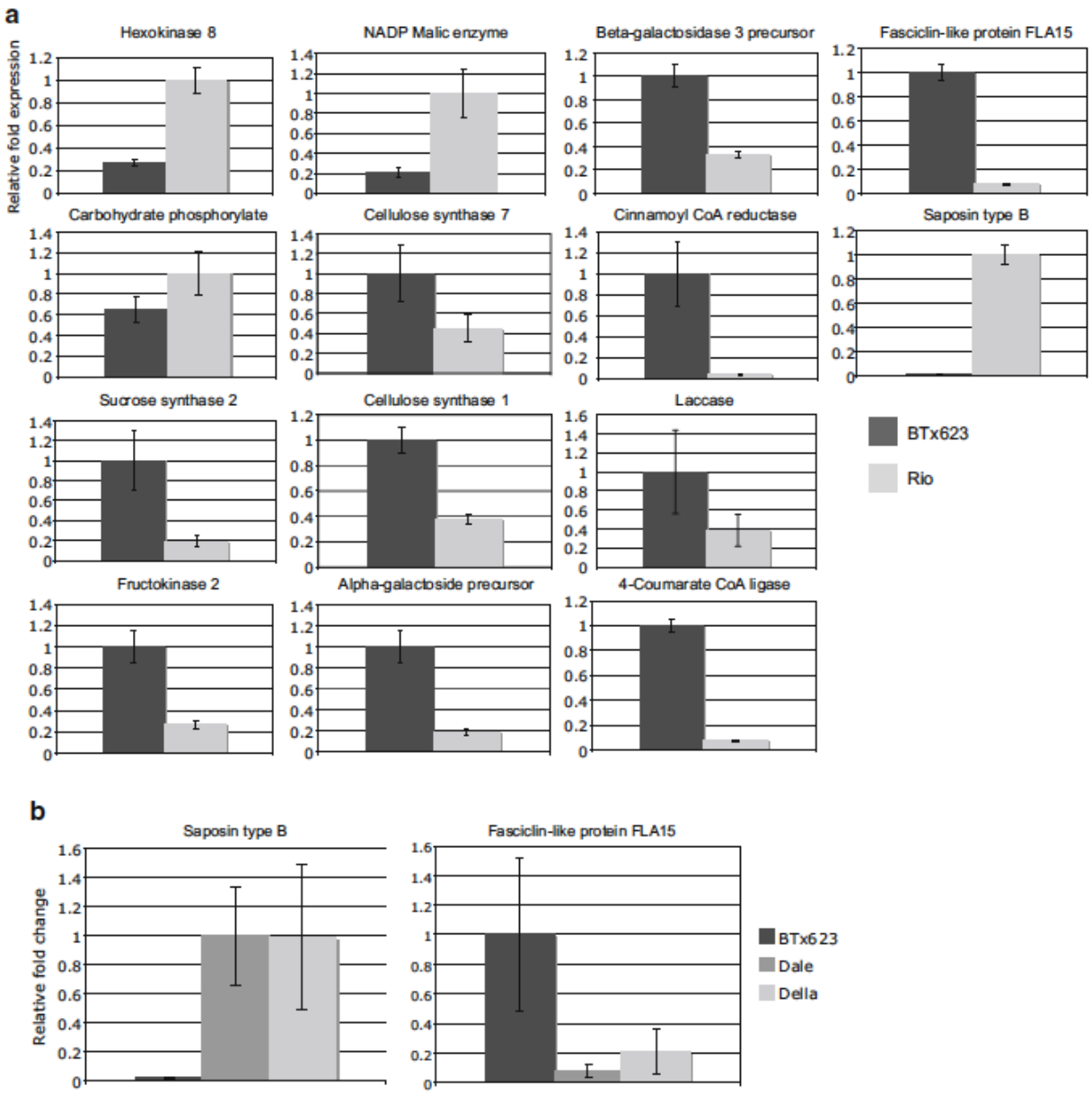


Figure 2.2. Validation of microarray data by qRT-PCR. (a) The expression of 14 genes selected from Table 2.3 was analyzed through qRT-PCR. The results of three independent experiments for both BTx623 and Rio are shown. The quantification of the mRNA abundance for each gene is presented as relative fold change expression. **(b)** qRT-PCR comparing the expression of saposin type B and fasciclin-like protein FLA15 in BTx623 and two sweet sorghum lines Della and Dale.

2.3.5. Genomic location of differentially expressed genes

In order to see if genes that were differentially expressed between grain and sweet sorghum clustered together in a particular region of the sorghum genome, We generated a "transcriptome map" (Fig. 2.3). We mapped the sequences of all up- and down-regulated sugarcane probes to the recently sequenced sorghum genome (BTx623; http://www.phytozome.net/cgi-bin/gbrowse/sorghum) using GenomeThreader (Borevitz et al., 2003). From a total of 195 probe sets, 176 could be mapped to the sorghum genome based on their alignment with a sorghum gene ("Materials and Methods"). In addition, six probe sets could be mapped to the genome but did not correlate with the current sorghum gene annotation, and for another 13 probe sets, we were not able to map them to the sorghum genome. Genes that were differentially expressed between grain and sweet sorghum did not appear to cluster in any particular region of the genome but rather reflect random distribution (Fig. 2.3).

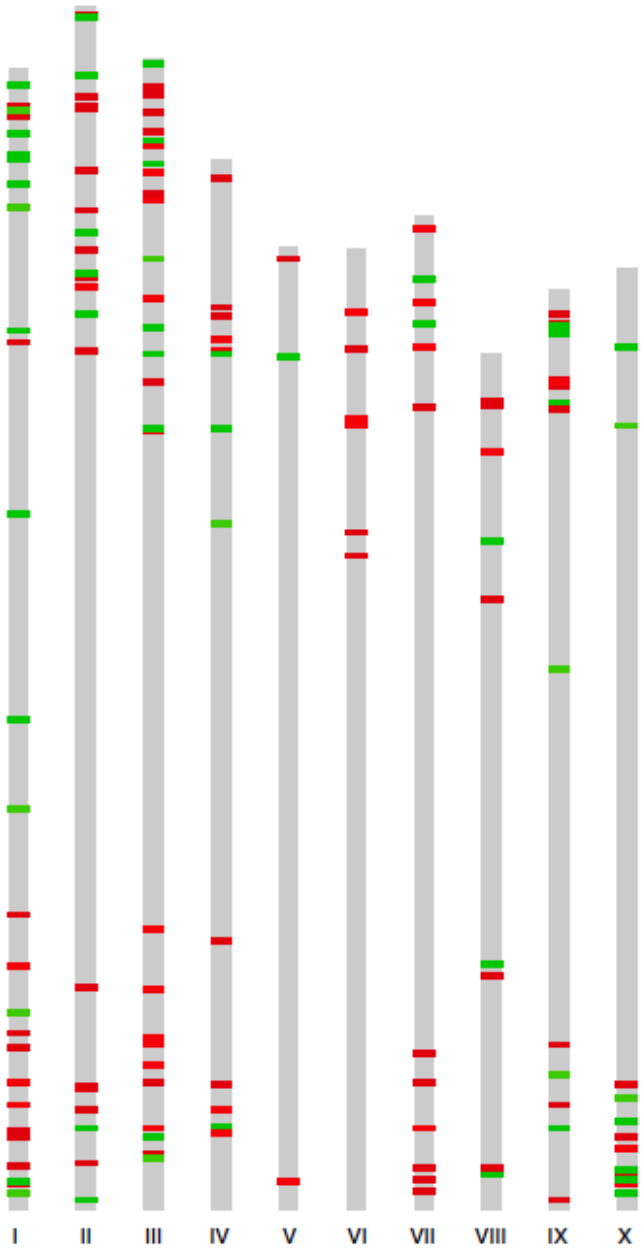


Figure 2.3. Localization of differentially expressed genes on the physical map of sorghum. Each sugarcane probe set representing a differentially expressed gene between BTx623 and Rio with a fold change of two or higher was mapped to the sorghum genome and plotted on the physical map. Up-regulated genes are in *green* and down-regulated genes are in *red*.

2.3.6. Trait-specific syntenic gene pairs between rice and sorghum

It can be considered that important gene functions have been conserved by ancestry and that divergence is mainly due to changes in regulatory control regions of genes. To determine the ancestry of genes, however, requires the alignment of syntenic regions. Because we know the positions of the sorghum genes in their respective chromosomes, we can align them with the rice genome as a reference (International Rice Genome Sequencing, 2005) and determine whether the aligned regions are collinear between rice and sorghum. Indeed, we found that from a total of 154 differentially expressed sorghum genes, 123 have an orthologous copy in syntenic positions in rice (Table 2.1). This collection included 28 candidate genes for the sweet sorghum trait (Table 2.3). Interestingly, one of these genes, sucrose synthase 2, was duplicated in rice but not in grain sorghum BTx623.

2.4. Discussion

2.4.1. Translational genomics

The non-renewable nature of fossil oil imposes an increasing pressure to develop alternative energies in order to support and secure social and economic growth in the near future (Ragauskas et al., 2006). Currently, there is a worldwide interest to develop biofuel crops, the best example being sugarcane, used in Brazil since the 1970s. Besides sugarcane, other grasses such as *Brachypodium distachyon*, *Miscanthus*, maize, rice, sweet sorghum, and switchgrass are considered as crops for biofuel research and production. Recently, the entire gene cluster of ten sorghum kafirin genes contained within a chromosomal segment of 45 kbp was intact and

stably inserted into the maize genome. Expression analysis showed that kafirins accumulated in maize endosperm in a developmental and tissue-specific manner (Song et al., 2004). Such transfer of genomic DNA between species that cannot be crossed could then be used in advance breeding techniques to introduce desirables traits from one species to another. Here, we integrated the traits of sugar accumulation in sorghum stems with genomic and expression data from three species: sugarcane, sorghum and rice. We used the an Affymetrix sugarcane genome array (Casu et al., 2007) as a tool for the identificiation of genes differentially expressed in maturing stems of grain and sweet sorghum. The intra-species variation for sugar content in sorghum is more pronounced that between sugarcane varieties, making sorghum a more suitable model to study this trait. On the other hand, because we can map sorghum genes to their chromosomal positions, we could use rice as a reference genome to identify genes by their ancestry.

2.4.2. Cross-referencing tissue-specific transcripts

Sorghum and sugarcane belong to the *Saccharinae* clade and diverged from each other only 8 to 9 million years ago (mya) (Jannoo et al., 2007), while rice is a more distant relative and separated from this clade 50 mya (Kellogg, 2001). Because sorghum and sugarcane belong to the same clade, we reasoned that, by hybridizing RNA from grain and sweet sorghum onto the sugarcane GeneChip, we could correlate changes in transcript levels with traits from sweet sorghum such as sugar content in stems. Given the tissue specificity and the rather small gene set of the sugarcane GeneChip, the positive hybridization of stem-derived RNAs from sorghum

to 5,900 sugarcane probes of a GeneChip comprising 8,224 probe sets in total was a good indication of such cross-referencing. By applying a twofold cut-off value as a parameter to filter out differentially expressed transcripts, a total of 195 probe sets were identified, of which 63 corresponded to transcripts that were up-regulated (51 genes) and 132 (103 genes) corresponded to transcripts that were down-regulated in the sweet sorghum Rio line, respectively. Each differentially expressed sorghum transcript was classified based on the Pfam domains of their encoded proteins and their GO term ("Materials and Methods").

Based on the sucrose and starch metabolic pathway from the KEGG (http://www.genome.jp/kegg/) and the CAZy database (http://www.cazy.org/), we found that almost 16% of the transcripts involved in sucrose and starch metabolism and in cell-wall-related processes were differentially expressed between BTx623 and Rio. This was particularly interesting because a previous study with cDNAs from immature and maturing stems of sugarcane identified only 2.4% of the transcripts related to carbohydrate metabolism (Casu et al., 2003). Furthermore, because sorghum stems are fully elongated at the anthesis stage, tissue samples from maturing internodes were also more suitable in profiling changes in gene expression associated with carbohydrate metabolism. The implication is that screening of differentially expressed genes can greatly be enhanced by genetic variability and selection of tissue.

2.4.3. Function of genes with elevated expression in sweet sorghum

The highest elevated transcript identified in my study encoded a saposin-like

type B domain. Increased expression was also validated and tested in other sweet sorghum lines by qRT-PCR. We found higher expression in Dale and Della compared to that in BTx623 (Fig. 2.2b). Saposins are water-soluble proteins that interact with the lysosomal membrane and are involved in the catabolism of glycosphingolipids in animals (Munford et al., 1995; Stokeley et al., 2007). Although it was unexpected that such a function could be related to a role in sugar accumulation, it underscores the value of a microarray-based screen to detect possibly new network effects. For instance, we could hypothesize that the removal of sugars from glycosphingolipids in the membrane alters its structure in such a way that it constitutes an early step in carbohydrate partitioning. Additional transcripts that were increased in sweet sorghum stems included: hexokinase 8, sorbitol dehydrogenase, and carbohydrate phosphorylase (starch phosphorylase). Hexokinase has a role not only in glycolysis but also as a glucose sensor that controls gene expression (Jang et al., 1997). Sorbitol dehydrogenase is an enzyme involved in carbohydrate metabolism that converts the sugar alcohol form of glucose (sorbitol) into fructose (Zhou et al., 2006). Increased transcript levels of carbohydrate phosphorylase suggested that enhanced starch degradation in Rio may contribute to sugar accumulation. Another increased transcript encoded a NADP-malic enzyme suggesting that carbon fixation is enhanced in the stems of sweet versus grain sorghum. Indeed, the activity of enzymes involved in photosynthesis and the expression of their transcripts are modulated by sink strength. In sugarcane, the accumulation of sucrose in the maturing and mature internodes of the stem contributed greatly to sink strength (McCormick et al., 2006). Kinetic models have been proposed to explain sucrose

accumulation in sugarcane (Rohwer and Botha, 2001; Uys et al., 2007). These models supported the notion that sucrose accumulates in the vacuole against a concentration gradient. Indeed, we found that a transcript encoding a vacuolar adenosine triphosphate (ATP) synthase catalytic subunit A had an increased expression in sweet sorghum, consistent with the role of this ATP synthase in the generation of an electrochemical gradient across the vacuolar membrane to propel the transport of sucrose.

The only cell-wall-related transcript that was up-regulated in sweet sorghum encoded a lysine motif containing protein. The LysM domain is widespread in bacterial proteins that degrade cell walls but is also present in eukaryotes. They are assumed to have a general role in peptidoglycan binding (Bateman and Bycroft, 2000).

2.4.4. Mobilization of sugars in the stems of sweet sorghum

Interestingly, genes with reduced transcript levels outpaced those with increased levels by a 2:1 margin. Down-regulated transcripts involved in the starch and sucrose metabolic pathway found in my study included alpha-galactosidase, beta-galactosidase, sucrose synthase 2, and fructokinase 2. Alpha and beta-galactosidase enzymes are O-glycosyl hydrolases that hydrolyse the glycosidic bond between two or more carbohydrates or between a carbohydrate and a non-carbohydrate moiety (Henrissat et al., 1996). Sucrose synthase is involved in the reversible conversion of sucrose to uridine diphosphate (UDP)-glucose and -fructose (Koch, 2004). UDP-glucose can then be used as a substrate for starch and

cell wall synthesis. Fructose instead is converted into fructose-6-phosphate by fructokinase and further metabolized through glycolysis (Pego and Smeekens, 2000). Our findings were in agreement with past reports showing that the onset of sucrose accumulation in Rio was accompanied by a decrease in sucrose synthase activity in stem tissue (Lingle, 1987). Similarly, Tarpley et al. (Tarpley et al., 1996) proposed that a decline in the levels of sucrose synthase, may be necessary for sucrose accumulation at stem maturity in sorghum. Consistent with my findings, Xue et al. (Xue et al., 2008) reported the down-regulation in the expression of both sucrose synthase and fructokinase genes in the stems of wheat genotypes with high water-soluble carbohydrates.

2.4.5. Reduced expression of cell-wall-related genes in sweet sorghum stems

Several transcripts involved in cell-wall-related processes were identified as down-regulated in sweet sorghum. These included cellulose synthase 1, 7, and 9 as well as cellulose synthase catalytic subunit 12, in cellulose synthesis. In the case of lignin biosynthesis, we found transcripts such as phenylalanine and histidine ammonia-lyase, cinnamoyl CoA reductase, 4-coumarate coenzyme A ligase, and caffeoyl-CoA O-methyltransferases. Interestingly, the expression of two transcripts encoding for xylanase inhibitors were also down-regulated in sweet sorghum. Xylanase inhibitor proteins belong to the group of protein inhibitors of cell wall degrading enzymes. Xylan is the major hemicellulose polymer in cereals and is degraded by plant endoxylanases (Juge et al., 2006). This suggested that, in sweet

sorghum, the degradation of hemicellulose is could be promoted by suppressing the expression of xylanases inhibitors. In other cases, a decrease in the expression of cellulose synthase genes in wheat genotypes with high water-soluble carbohydrate content was also observed (Xue et al., 2008). In addition, Casu et al. (Casu et al., 2007) characterized the expression of several cellulose synthase and cellulose synthase-like genes in sugarcane stem and found that their expression was highly variable depending on internode maturity (Casu et al., 2007).

In addition to cellulose synthesis, the geometric deposition of cellulose fibrils generally perpendicular to the axis of cell elongation is a critical step in cell wall formation. There is evidence that the orientation of cellulose deposition is somehow assisted by microtubules (Somerville et al., 2004). An example of this is the *fiber fragile* mutant *fra1* encoding a kinesin-like protein. In this mutant, cellulose deposition displayed an abnormal orientation (Burk and Ye, 2002). Consistent with these observations, the expression of two transcripts encoding tubulin alpha-2/ alpha-4 chain and tubulin folding cofactor A, in conjunction with a transcript encoding a protein with kinesin motor domain, were all down-regulated in sweet sorghum. Less clear, but also related to cell wall formation, is fasciclin. Interestingly, the most strongly down-regulated transcript in sweet sorghum encoded a protein with a fasciclin domain. Fasciclin domains are found in animal arabinogalactan proteins that have a role in cell adhesion and communication (Kawamoto et al., 1998). These proteins are structural components that mediate the interaction between the plasma membrane and the cell wall. However, their specific role in plants is still unknown (Faik et al., 2006). A loss-of-function mutant in the

Arabidopsis gene fasciclin-like arabinogalactan 4 (*AtFLA4*) displayed thinner cell walls and increased sensitivity to salinity (Yang et al., 2007).

Other transcripts that were also down-regulated encoded a peroxidase and a laccase. It has been shown that peroxidases have an important role in cell wall modification (Passardi et al., 2004). By controlling the abundance of H_2O_2 in the cell wall, a necessary step for the cross linking of phenolic compounds, peroxidases act to inhibit cell elongation and, in conjunction with laccases, are assumed to be involved in monolingol unit oxidation, a reaction necessary for lignin assembly. Furthermore, it is known that peroxidase activity can be controlled by ascorbate. Indeed, the expression of a transcript encoding a protein similar to guanosine diphosphate (GDP)-mannose3, 5-epimerase was increased in sweet sorghum. This protein catalyzes the reversible conversion of GDP-mannose either into GDP-L-galactose or a novel intermediate, GDP-gulose, a step necessary for the biosynthesis of vitamin C in plants (Wolucka and Van Montagu, 2003). In addition, GDP-mannose is used to incorporate mannose residues into cell wall polymers (Lukowitz et al., 2001). For these reasons, it is considered that GDP-mannose 3,5 epimerase could modulate the carbon flux into the vitamin C pathway as well as the demand for GDP-mannose into the cell wall biosynthesis (Wolucka and Van Montagu, 2003). Indeed, it is known that the stem of high sucrose-accumulating genotypes of sugarcane are high in moisture content and low in fiber, whereas the stem of low sucrose-accumulating genotypes are low in moisture content, thin, and fibrous (Borevitz and Chory, 2004).

2.4.6. Differential expression of genes related to osmotic stress

Consistent with the idea that high concentration of sugars imposes osmotic stress to the cell, we found increased transcripts encoding heat shock proteins HSP70 and HSP90. Additionally, a transcript encoding a poly adenosine diphosphate (ADP)-ribose polymerase 2 (*PARP 2*) was significantly down-regulated in sweet sorghum. This was in agreement with a report in which *Arabidopsis* and *Brassica napus* transgenic plants with reduced levels of *PARP 2* displayed resistance to various abiotic stresses (Vanderauwera et al., 2007). Poly ADP ribosylation involves the tagging of proteins with long-branched poly ADP-ribose polymers and is mediated by PARP enzymes (Schreiber et al., 2006). Poly ADP-ribosylation has important roles in the cellular response to genotoxic stress, influence DNA synthesis and repair, and is also involved in chromatin structure and transcription.

2.4.7. Mapping genes linked to stem-sugar content and cell wall metabolism in sorghum and rice

Although sugarcane has not been sequenced yet, we could use the sequenced genome of sorghum to construct a “transcriptome map” with the genes found in my study. Assuming that gene order has been largely conserved between these closely related species, the “transcriptome map” of sorghum served as a valuable reference for sugarcane. We could not find any particular clustering of these genes but did observe that most of the genes are located towards the telomeres and only a few of them near the centromeres. We also could not find any of these genes in the telomeric region on the long arm of chromosome six. Comparing this map with the

rice genome demonstrated that, out of 154 differentially expressed genes, 123 were in syntenic positions. With respect to the subset of genes involved in the accumulation of fermentable sugars and reduced lignocellulose, 21 genes were also found in syntenic regions, whereas nine genes appeared to be paralogous copies (Tables 2.3 and 2.4).

2.5. Outlook

Given the synteny of these genes between rice and sorghum, We can assume that they are allelic between different sorghum cultivars. Therefore, future genetic mapping experiments should provide a direct link of allelic variation and the sweet sorghum trait, that is sugar content in stems. Most likely, such allelic variations extend to the control regions of these genes because of their differential expression. Transgenic experiments could then be used to verify such functional aspects for biofuel properties. Moreover, gain of function experiments could be used to import desirable traits such as accumulation of fermentable sugars from sweet sorghum into maize. The generation of “sweet sorghum-like transgenic corn” will alleviate in part the increasing pressure of growing corn either for food or for biofuel since it would then be possible to use the grain for food and at the same time to extract fermentable sugars from the stem to use in ethanol production.

2.6. Materials and Methods

2.6.1. Plant materials and growth conditions

Seeds from both grain and sweet sorghum (*Sorghum bicolor* (L.) Moench)

were sown in pro-mix soil (Premiere Horticulture Inc., USA) and grown in greenhouse with a day length of 15 h light: 9 h dark at constant temperature of 23°C. The genotype representing grain sorghum in my study was BTx623, whereas the genotypes representing sweet sorghum were Dale, Della, M81-E, Rio, Simon, and Top76-6. The seeds from sweet sorghum were kindly provided by Dr. William L. Rooney of Texas A&M, College Station, TX, USA.

2.6.2. Measurement of "Brix degree" from sorghum stem's juice

The juice from internodes of the main stem in both grain and sweet sorghum was harvested at the time of anthesis. A section of approximately 6 cm long was dissected from the middle of each internode, and 300 μl of juice was extracted by pressing each internode with a garlic squeezer. The concentration of total soluble sugars (measured in Brix degrees) in the juice was measured with a pocket refractometer (Atago Inc., Japan).

2.6.3. Isolation of total RNA from stem tissue

Both grain sorghum BTx623 and sweet sorghum Rio were grown until anthesis and total RNA from internode #8 for each genotype (internodes were numbered from the base towards the apex of the stem) was extracted using the RNeasy Plant Mini Kit (QIAGEN Inc., USA).

2.6.4. GeneChip sugarcane genome array hybridization

Sorghum RNA from internode #8 was hybridized to the Affymetrix GeneChip Sugarcane Genome Array (Affymetrix Inc., USA). Probe set information can be found

at NetAffx Analysis Center's web page (http://www.affymetrix.com/analysis/index.affx). The One-Cycle Eukaryotic Target Labeling Assay protocol was used. The labeling, hybridization, and data collection were done at the Transcription Profiling Facility, Cancer Institute of New Jersey, Department of Pediatrics, Robert Wood Johnson Medical School.

2.6.5. Data analysis

Probe sets that were absent in all chips were eliminated. About 5,900 out of the original 8,300 probe sets passed this test. Next, a *t* test was applied to BTx623 and Rio groups (three replicates for each) with an alpha value of 0.001, and the Benjamini–Hochberg multiple-testing correction was applied. From the probe sets that passed the criteria, only those with a fold change of at least two were considered.

2.6.6. Validation of microarray data through qRT-PCR

cDNA synthesis and PCR amplification was performed in the same tube from 50 ng of total RNA using the iScript One-Step RT-PCR Kit with SYBR Green (BIORAD Laboratories, Inc.). The reaction condition used was as specified in the kit, with an annealing temperature of 55°C and 45 cycles for the data collection step. The qRT-PCR reaction was done using the MyiQ Real-Time PCR Detection System (BIO-RAD Laboratories, Inc.). Total RNA was accurately measured for each sample with the NanoDrop 1000 spectrophotometer (Thermo Scientific, Inc.). A relative quantification normalized against unit mass (50 ng of total RNA) was used to

analyze the expression data with the equation: Ratio (test/calibrator)= $2^{\Delta CT}$, as suggested in Real-Time Applications Guide from Bio-Rad. The primers for each gene were designed based on the region of homology (usually in the last exon or 3′ untranslated region) between the sugarcane probe set sequence and the sorghum gene sequence and are listed in Table 2.4. The sequence for each sugarcane probe set is freely available at the Affymetrix website: http://www.affymetrix.com/analysis/index.affx. In addition, the genomic location of each sugarcane probe set in sorghum (BTx623) identified in our work has been up-loaded to the Waksman Institute's Sorghum Genome Browser and is available at http://genlisea-rs1.waksman.rutgers.edu/cgi-bin/gbrowse/sbic/.

S. bicolor **GENE ID / SUGARCANE PROBE SET ID**	**FORWARD**	**REVERSE**
Sb09g013990.1/Sof.405.2.S1_a_at	5'TGCTGGATCACAAATCCTCA3'	5'ATAGCGCCTGGACTCCTTTT3'
Sb09g028490.1/Sof.3590.1.S1_at	5'CAGTTCAGCGAGTTCAAGCA3'	5'TCACGCAGTAGAGCACCATC3'
Sb03g040060.1/Sof.90.1.S1_at	5'GCCAAGGAGATATGGGACAT3'	5'AGCACCGTGGGTCATTATTC3'
Sb09g005280.1/Sof.5033.1.S1_at	5'TTGTCTGGTCCATCCTCCTC3'	5'TTTCCCATCTAGCCTCCTCA3'
Sb07g001320.1/Sof.3644.2.S1_a_at	5'CCTGAAGCAAAACAACGTCA3'	5'GGGTTCCGGTAGAACATGAA3'
Sb04g005210.1/Sof.4734.1.S1_at	5'ACCGAAGGCTCTGAAGTCAC3'	5'GGGGATGGATTCAGTGAAGA3'
Sb01g033060.1/Sof.4165.1.S1_s_at	5'CTTTTCCCTGGGTTTCCTTC3'	5'TCCCTCTCAACCGACTCAAC3'
Sb01g002050.1/Sof.1587.3.A1_a_at	5'TGACTGCAATATTGGGCAAA3'	5'AACTTTCTGTTCGGCTCACC3'
Sb03g003190.1/Sof.4315.1.S1_at	5'GCCATGGGTGCTTACCATAG3'	5'CCAAGCCTCGTTTTGGTTAT3'
Sb03g039520.1/Sof.1021.1.A1_at	5'CGATCTTCCCAAATGCTGAT3'	5'GTCCAGGTCAGCTAGGAACG3'
Sb07g021680.1/Sof.3629.1.S1_at	5'GCGTGAGCTAGAGGGAGATG3'	5'CAGCCAGCGAACAAACACTA3'
Sb03g033250.1/Sof.1594.1.S1_at	5'TGCATGTACAGCCCCATTTA3'	5'GCAGAACAGGACGTGAAACA3'
Sb03g041450.1/Sof.4934.1.S1_at	5'AGGCCTGTCTGAACACCAAT3'	5'CATGGGCACAGTTGTAGTGG3'
Sb01g018400.1/Sof.3244.1.S1_a_at	5'CACTCATCATTCTCGGCTCA3'	5'CACACTATGGACTCCGCTCA3'

Table 2.5. Primer sequences used in qRT-PCR reactions. Primers were designed based on the sequence from sorghum genes with homology to sugarcane Probe Set IDs.

2.6.7. Physical location of differentially expressed transcripts in the sorghum genome

The sugarcane probe sets that were up- and down-regulated in Sorghum, respectively, were mapped to the genome by using GenomeThreader (Gremme et al., 2012). Spliced alignments were only considered if 75% (score >0.75) or more bases could be aligned between the genomic sequence and a probe set. If a probe could be mapped to the genome and if it also overlapped with a sorghum gene, we assigned the annotation of the sorghum gene to the probe.

2.7. References

Bateman A, Bycroft M (2000) The structure of a LysM domain from E. coli membrane-bound lytic murein transglycosylase D (MltD). Journal of Molecular Biology **299:** 1113-1119.

Bennetzen J, Freeling M (1993) Grasses as a single genetic system: genome composition, collinearity and compatibility. Trends in Genetics **9:** 259-261.

Buchanan C, Lim S, Salzman R, Kagiampakis I, Morishige D, Weers B, Klein R, Pratt L, Cordonnier-Pratt M-Ml, Klein P, Mullet J (2005) Sorghum bicolor's transcriptome response to dehydration, high salinity and ABA. Plant Molecular Biology **58:** 699-720.

Bull T, Glasziou K (1963) The evolutionary significance of sugar accumulation in Saccharum. Australian Journal of Biological Sciences **16:** 737-742.

Burk D, Ye Z-H (2002) Alteration of oriented deposition of cellulose microfibrils by mutation of a katanin-like microtubule-severing protein. Plant Cell **14:** 2145-2160.

Casu R, Grof C, Rae A, McIntyre C, Dimmock C, Manners J (2003) Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis. Plant Molecular Biology **52:** 371-386.

Casu R, Jarmey J, Bonnett G, Manners J (2007) Identification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by Affymetrix GeneChip Sugarcane Genome Array expression profiling. Functional & Integrative Genomics **7:** 153-167.

Chapple C, Carpita N (1998) Plant cell walls as targets for biotechnology. Current opinion in plant biology **1:** 179-185.

D'Hont A, Grivet L, Feldmann P, Rao S, Berding N, Glaszmann J (1996) Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. Molecular & General Genetics **250:** 405-413.

Faik A, Abouzouhair J, Sarhan F (2006) Putative fasciclin-like arabinogalactan-proteins (FLA) in wheat (Triticum aestivum) and rice (Oryza sativa): identification and bioinformatic analyses. Molecular Genetics and Genomics **276:** 478-494.

Gale M, Devos K (1998) Comparative genetics in the grasses. Proceedings of the National Academy of Sciences of the United States of America **95:** 1971-1974.

Gremme G, Brendel V, Sparks M, Kurtz S (2005) Engineering a software tool for gene structure prediction in higher organisms. Information and Software Technology **47:** 965-978.

Grivet L, Arruda P (2002) Sugarcane genomics: depicting the complex genome of an important tropical crop. Current Opinion in Plant Biology **5:** 122-127.

Guimaraes C, Sills G, Sobral B (1997) Comparative mapping of Andropogoneae: Saccharum L. (sugarcane) and its relation to sorghum and maize. Proceedings of the National Academy of Sciences **94:** 14261-14266.

Henrissat B, Callebaut I, Fabrega S, Lehn P, Mornon J, Davies G (1996) Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. Proceedings of the National Academy of Sciences **93:** 5674.

Hoffman-Thoma G, Hinkel K, Nicolay P, Willenbrink J (1996) Sucrose accumulation in sweet sorghum stem internodes in relation to growth. Physiologia Plantarum **97:** 277-284.

International Rice Genome Sequencing P (2005) The map-based sequence of the rice genome. Nature **436:** 793-800.

Ishimaru K, Hirotsu N, Madoka Y, Kashiwagi T (2007) Quantitative trait loci for sucrose, starch, and hexose accumulation before heading in rice. Plant Physiology and Biochemistry **45:** 799-804.

Jang J, Leon P, Zhou L, Sheen J (1997) Hexokinase as a sugar sensor in higher plants. Plant Cell **9:** 5-19.

Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann J, Arruda P, D'Hont Al (2007) Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. Plant Journal **50:** 574-585.

Juge N, Nohr J, Le Gal-Coeffet M-Fo, Kramhoft B, Furniss C, Planchot Vr, Archer D, Williamson G, Svensson B (2006) The activity of barley alpha-amylase on starch granules is enhanced by fusion of a starch binding domain from Aspergillus niger glucoamylase. Biochimica et Biophysica Acta **1764:** 275-284.

Kawamoto T, Noshiro M, Shen M, Nakamasu K, Hashimoto K, Kawashima-Ohya Y, Gotoh O, Kato Y (1998) Structural and phylogenetic analyses of RGD-CAP/beta ig-h3, a fasciclin-like adhesion protein expressed in chick chondrocytes. Biochimica et Biophysica Acta **1395:** 288-292

Kellogg E (2001) Evolutionary history of the grasses. Plant Physiology **125:** 1198-1205.

Koch K (2004) Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. Current Opinion in Plant Biology **7:** 235-246.

Lingle SE (1987) Sucrose metabolism in the primary culm of sweet sorghum during development. Crop Science **27:** 1214-1219.

Lukowitz W, Nickle T, Meinke D, Last R, Conklin P, Somerville C (2001) Arabidopsis cyt1 mutants are deficient in a mannose-1-phosphate guanylyltransferase and point to a requirement of N-linked glycosylation for cellulose biosynthesis. Proceedings of the National Academy of Sciences **98:** 2262-2267.

McCormick A, Cramer M, Watt D (2006) Sink strength regulates photosynthesis in sugarcane. New Phytologist **171:** 759-770.

Messing J, Bennetzen J (2008) Grass genome structure and evolution. Genome Dynamics **4:** 41-56.

Messing J, Llaca V (1998) Importance of anchor genomes for any plant genome project. Proceedings of the National Academy of Sciences **95:** 2017-2020.

Ming R, Liu S, Moore P, Irvine J, Paterson A (2001) QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane. Genome Research **11:** 2075-2084.

Munford R, Sheppard P, O'Hara P (1995) Saposin-like proteins (SAPLIP) carry out diverse functions on a common backbone structure. Journal of Lipid Research **36:** 1653-1663.

Passardi F, Penel C, Dunand C (2004) Performing the paradoxical: how plant peroxidases modify the cell wall. Trends in Plant Science **9:** 534-540.

Pego J, Smeekens S (2000) Plant fructokinases: a sweet family get-together. Trends in Plant Science **5:** 531-536.

Ragauskas A, Williams C, Davison B, Britovsek G, Cairney J, Eckert C, Frederick W, Hallett J, Leak D, Liotta C, Mielenz J, Murphy R, Templer R, Tschaplinski T (2006) The path forward for biofuels and biomaterials. Science **311:** 484-489.

Rohwer J, Botha F (2001) Analysis of sucrose accumulation in the sugar cane culm on the basis of in vitro kinetic data. The Biochemical journal **358:** 437-445.

Schreiber Vr, Dantzer Fo, Ame J-C, de Murcia G (2006) Poly(ADP-ribose): novel functions for an old molecule. Nature Reviews in Molecular Cell Biology **7:** 517-528.

Somerville C, Bauer S, Brininstool G, Facette M, Hamann T, Milne J, Osborne E, Paredez A, Persson S, Raab T, Vorwerk S, Youngs H (2004) Toward a systems approach to understanding plant cell walls. Science **306:** 2206-2211.

Song R, Segal G, Messing J (2004) Expression of the sorghum 10-member kafirin gene cluster in maize endosperm. Nucleic Acids Research **32:** e189.

Stokeley D, Bemporad D, Gavaghan D, Sansom M (2007) Conformational dynamics of a lipid-interacting protein: MD simulations of saposin B. Biochemistry **46:** 13573-13580.

Tarpley L, Vietor DM, Miller FR (1996) Metabolism of sucrose during storage in intact sorghum stalk. International Journal of Plant Sciences**:** 159-163.

Uys L, Botha F, Hofmeyr J-HS, Rohwer J (2007) Kinetic model of sucrose accumulation in maturing sugarcane culm tissue. Phytochemistry **68:** 2375-2392.

Vanderauwera S, De Block M, Van de Steene N, van de Cotte B, Metzlaff M, Van Breusegem F (2007) Silencing of poly(ADP-ribose) polymerase in plants alters abiotic stress signal transduction. Proceedings of the National Academy of Sciences **104:** 15150-15155.

Wolucka B, Van Montagu M (2003) GDP-mannose 3',5'-epimerase forms GDP-L-gulose, a putative intermediate for the de novo biosynthesis of vitamin C in plants. Journal of Biological Chemistry **278:** 47483-47490.

Xue G-P, McIntyre C, Jenkins C, Glassop D, van Herwaarden A, Shorter R (2008) Molecular dissection of variation in carbohydrate metabolism related to water-soluble carbohydrate accumulation in stems of wheat. Plant Physiology **146:** 441-454.

Yang J, Sardar H, McGovern K, Zhang Y, Showalter A (2007) A lysine-rich arabinogalactan protein in Arabidopsis is essential for plant growth and development, including cell division and expansion. Plant Journal **49:** 629-640.

Yang J, Zhang J (2006) Grain filling of cereals under soil drying. New Phytologist **169:** 223-236.

Zhou R, Cheng L, Dandekar A (2006) Down-regulation of sorbitol dehydrogenase and up-regulation of sucrose synthase in shoot tips of the transgenic apple trees with decreased sorbitol synthesis. Journal of Experimental Botany **57:** 3647-3657.

Chapter 3 Molecular Markers For Sweet Sorghum Based On Microarray Expression Data

3.1. Abstract

Using an Affymetrix sugarcane genechip, we previously identified 154 genes differentially expressed between grain and sweet sorghum. Although many of these genes have functions related to sugar and cell wall metabolism, dissection of the trait requires genetic analysis. Therefore, it would be advantageous to use microarray data for generation of genetic markers, shown in other species as single feature polymorphisms (SFPs). As a test case, we used the GeSNP software to screen for SFPs between grain and sweet sorghum. Based on this screen, out of 58 candidate genes, 30 had single-nucleotide polymorphisms (SNPs) from which 19 had validated SFPs. The degree of nucleotide polymorphism found between grain and sweet sorghum was in the order of one SNP per 248 base pairs, with chromosome 8 being highly polymorphic. Indeed, molecular markers could be developed for a third of the candidate genes, giving us a high rate of return by this method.

3.2. Introduction

The development of molecular markers is essential for marker-assisted selection in plant breeding as well as to understand crop domestication and plant evolution (Varshney et al., 2005). Single-nucleotide polymorphisms (SNPs) have become the marker of choice because of their abundance and uniform distribution

throughout the genome (Varshney et al., 2005; Zhu and Salmeron, 2007; Gupta et al., 2008). Around 90% of the genetic variation in any organism is attributed to SNPs (Varshney et al., 2005; Zhu and Salmeron, 2007). They are discovered from genomic or expressed sequence tag sequences available in databases or through sequencing of candidate genes, PCR products, or even whole genomes (Varshney et al., 2005; Zhu and Salmeron, 2007).

Recent studies have described the use of transcript abundance data from RNA hybridizations to Affymetrix microarrays to discover genetic polymorphisms that can be utilized as markers for genotyping in mapping populations (Hazen and Kay, 2003; Varshney et al., 2005; Zhu and Salmeron, 2007; Gupta et al., 2008; Shiu and Borevitz, 2008). In an Affymetrix chip, each gene is represented by 11 different 25-bp oligonucleotides that cover features of the transcribed region of that gene (exons and 3′untranslated regions). Each of these features is described as a perfect match (PM) and mismatch (MM) oligonucleotide. The PM exactly matches the sequence of a standard genotype, whereas the MM differs from the PM by a single base substitution at the central, 13th position (Hazen and Kay, 2003; Borevitz and Chory, 2004; Zhu and Salmeron, 2007). A new aspect of this approach is to discover sequence polymorphisms in cultivars or variants of species, where one of them has been sequenced but where no sequence information is yet available from the other ones. Here, the hybridization data from microarrays not only measure differential gene expression but also can yield information on sequence variation between two inbred lines. If two genotypes differ only in the amount of mRNA in a particular tissue, this should result in a relatively constant difference in hybridization

throughout the 11 features. On the other hand, if the two genotypes contain a genetic polymorphism within a gene that coincides with one of the particular features, this will produce differential hybridization for that single feature. Such differences have been described as single-feature polymorphisms (SFPs) (Borevitz et al., 2003; Hazen and Kay, 2003; Borevitz and Chory, 2004; Zhu and Salmeron, 2007). Thus, expression microarrays hybridized with RNA are able to provide me not only with phenotypic (variation in gene expression) but also with genotypic (marker) data (Zhu and Salmeron, 2007). If two genotypes differ in the expression level of a particular gene, we can consider it as an expression level polymorphism or (ELP). Both ELPs and SFPs are dominant markers and can be mapped as alleles in segregating populations (genetical genomics), and ELPs can be considered as traits to determine expression quantitative trait loci (eQTLs) (Jansen and Nap, 2001; Coram et al., 2008).

In *Arabidopsis*, SFPs have been used for several purposes such as mapping clock mutations through bulked segregant analysis (Hazen et al., 2005), the identification of genes for flowering QTLs (Werner et al., 2005), high-density haplotyping of recombinant inbred lines (RILs) (West et al., 2006), and natural variation in genome-wide DNA polymorphism (Borevitz et al., 2007). In plant species of agronomic importance, SFPs have been utilized to identify genome-wide molecular markers in barley and rice (Rostoks et al., 2005; Kumar et al., 2007; Potokina et al., 2008) as well as markers linked to Yr5 stripe rust resistance in wheat (Coram et al., 2008). However, an impediment to SFP discovery in crop plants based on DNA hybridization to Affymetrix expression arrays could be the size of

gene families (Borevitz et al., 2003; Varshney et al., 2005; Zhu and Salmeron, 2007). Because the coding regions of many gene clusters that arose by tandem gene amplification are quite conserved, hybridization-based approaches would not be sufficient to distinguish between allelic and paralogous copies (Xu and Messing, 2008). Therefore, one would have to limit this analysis to low-copy genes. On the other hand, this approach does not aim at identifying candidate genes directly but rather linked genetic markers.

An area where gene discovery has become of general interest is the utilization of biomass for the production of alternative fuels. Because desirable traits for biofuel crops are very complex and involve many genes from different pathways, it becomes necessary to take genetic approaches to identify key genes so that molecular breeding can be employed to make performance improvements. The most successful biofuel crop today is sugarcane. However, it cannot be grown in moderate climate. Maize, which is a major biofuel crop in the USA, has a much lower yield of bioethanol per acreage than sugarcane, requires high input costs, and is a major food and feed source. A crop that bridges between the two is the close relative, sorghum. Sorghum tolerates harsher environmental conditions than sugarcane and maize, has a higher disease resistance than maize, and has a high stem sugar variant, sweet sorghum, which has potential yields of bioethanol like sugarcane. Moreover, sweet sorghum can be crossed with grain sorghum so that genetic analysis could uncover key regulatory factors that would increase sugar and decrease lignocellulose in the biomass. Therefore, sorghum could be used to identify both SFPs and ELPs linked to high sugar content.

We have recently reported the hybridization of RNAs derived from the stems of grain and sweet sorghum onto the sugarcane Affymetrix genechip (Calviño et al., 2008). A previous study demonstrated that cross-species hybridization did not affect the reproducibility of the microarray experiment (Cáceres et al., 2003). Moreover, an Affymetrix soybean genome array has been used to identify SFPs in the closely related species cowpea (Das et al., 2008).

Here, we have asked the question whether I could use the sugarcane chip analysis to extend the cross-species concept in SFP discovery in the grasses. We report the identification of SFPs in 58 sorghum genes by using the recently developed software GeSNP (Greenhall et al., 2007). These genes were described in my previous study to be differentially expressed between grain and sweet sorghum (Calviño et al., 2008). The utility of GeSNP has been successfully tested for SFP discovery in mice, humans, and chimpanzees (Greenhall et al., 2007), but there was no report on plants at the time of my study. In order to experimentally validate the SFPs identified in sorghum, we sequenced fragments from 58 genes and found SNPs in 30 of them, out of which 19 genes had a validated SFP. Furthermore, we developed molecular markers based on the SNPs found. The high experimental validation rate of SNPs of 50% of the candidate genes shows the potential of this method for the development of molecular markers and, in principle, the applicability to any trait of interest.

3.3. Results

3.3.1. SFP discovery and validation from differentially expressed genes in sorghum.

Previously, we reported the use of an Affymetrix genechip from sugarcane to identify differentially expressed genes in the stem of grain and sweet sorghum (Calviño et al. 2008). Such a cross-species hybridization (CSH) approach allowed us to identify 154 genes harboring expression level polymorphisms between grain and sweet sorghum. In order to discover single-feature polymorphisms within these genes as well, we uploaded the sugarcane Affymetrix CEL files previously obtained into the GeSNP software. Indeed, we found that, from 154 genes, 57 harbored a SFP with a *t* value ≥ 7 (Fig. 3.1 and Table 3.1). Based on existing data (Greenhall et al., 2007), we adopted a *t* value of 7 or higher as a threshold. Chromosomes 1, 2, and 3 had the highest number of genes displaying both ELPs and SFPs, whereas chromosomes 5 and 6 had the lowest number of ELPs and SFPs, respectively (Fig. 3.1).

In order to validate the SFPs discovered and calculate the SFP discovery rate (SDR) of the GeSNP software, we cloned and sequenced the fragments from 57 genes harboring both ELPs and SFPs in addition to one gene harboring only SFPs (see below) from sweet sorghum Rio and aligned the sequences against the BTx623 reference genome. The software predicted a total of 125 SFPs (on average ~2 per gene), and we could experimentally validate 32 of them (Table 3.1). We calculated the SDR as 25.6% (SDR = [Validated SFPs / Total SFPs] x 100). As expected, the SDR was dependent on the *t* value, with the lowest SDR (less than 10%) at *t* values

between 7 and 10 and the highest SDR (80%) with *t* values from 22 to 25, respectively (Fig. 3.2a).

Besides SFPs identified in genes that are differentially expressed, the GeSNP software also detected SFPs in genes that did not show differential expression under my experimental conditions. Considering the high success rate of SNPs discovered in genes having both SFPs and ELPs, we extended our screen to genes that have predicted SFPs with *t* values of 22 to 25 but no ELP. This analysis allowed us to identify 35 sugarcane probe pairs that matched the sorghum genome sequence and have a high probability of representing SNPs in genes that have no ELPs between BTx623 and Rio but were expressed in the stem (see Table 3.2). For example, one of the sugarcane probe pairs (Sof.3814.1.S1_at) matched a sorghum gene coding for fructose bisphospate aldolase. Since the protein product of this gene has a role in the sucrose and starch metabolic pathway (our trait of interest), we cloned and sequenced the fragment containing the SFPs. As it is shown in (Fig. 3.3), we found six SNPs, two of which were recognized by three sugarcane probe pairs. This result indicated that our approach was able to efficiently detect SNPs. From the 58 genes that were sequenced, 19 genes (~33%) had a validated SFP, and 11 genes (19%) harbored SNPs outside the probe pairs at different location than the one predicted by GeSNP. Therefore, the total SNP detection rate was ~52%. A list of genes with validated SFPs as well as the nature of the nucleotide change/s is provided in Table 3.3.

Most of the validated SFPs had probe pairs with *t* values from 15 to 18 and greater than 25 (Fig. 3.2b). Since the SFP validation depends on the SNP position

along the probe pair (Rostoks et al., 2005), we analyzed the SNP position from the edge of the sugarcane probe pair for those genes with validated SFPs (Fig. 3.4). We found that, from a total of 22 probe pairs (probes that recognized the same SNP were not counted), 19 of them recognized a SNP between the 6th and the 13th positions.

With regard to genes involved in our traits of interest, that is, sugar accumulation and cell wall metabolism, we validated SFPs for five of them (Figs. 3.5 and 3.5). The SFPs in the cellulose synthase 1 and dolichyl-diphospho-oligosaccharide genes was based on a SNP, whereas the SFP in the LysM gene was due to a 13-bp indel (Fig. 3.5a and 3.5b). This indel allowed us to develop an allele-specific PCR marker (Fig. 3.5d). In the case of the 4-coumarate coenzyme A ligase gene, the SFP was based on a mis-spliced intron in Rio (Fig. 3.5c).

To calculate the number of SNPs per total sequence length, I assumed a similar genome size of the Rio line to that of BTx623, the reference genome. Based on 87 SNPs in 21,612 bp of sequence from both parental lines, we concluded that there was an average of one SNP every 248 base pairs of sequence between BTx623 and Rio. Taking in consideration that the genome size is in the order of 730 Mbp (Paterson et al., 2009), we suggested that 2,938,800 SNPs could exist between grain sorghum BTx623 and sweet sorghum Rio and that at least 0.4% of the genome could be polymorphic between the two lines. We also looked at the SNP density per sorghum chromosome in order to see if there was any difference among them. Surprisingly, we found that the level of polymorphism was higher for chromosomes 8 and 9 and lower for chromosome 3 compared to the average SNP density per Kb of

sequence (4 SNPs/Kbp) (Fig. 3.6a). However, if I consider the frequency of probe pairs with *t* values between 22 and 25 for each sorghum chromosome as it is shown in (Fig. 3.6b), chromosome 3 had the highest number of probes. On the other hand, chromosome 8 had the second highest number of probes with *t* values between 22 and 25 together with a high SNP density (Fig. 3.6a and 3.6b). This might suggest an unusual level of polymorphism for this chromosome between BTx623 and Rio. However, we did not have sufficient data (genes sequenced) to test whether the SNP density differences among the chromosomes are statistically significant.

Sorghum genes harboring validated SFPs allowed us to investigate if such nucleotide substitutions were conserved or not within grain sorghum BTx623, sweet sorghum Rio, and sugarcane. Indeed, we found that from 22 SNPs discovered through 29 validated SFPs (one sugarcane probe pair can recognize more than one SNP), 15 of them were conserved between BTx623 and sugarcane, whereas only eight SNPs were conserved between Rio and sugarcane (Table 3.3).

GENE ID	#SFPs[a]	#VALIDATED SFPs	#SNPs	SEQUENCE LENGTH (bp)
Chromosome 1				
Sb01g005770	1	0	0	378
Sb01g049890	1	1	2	401
Sb01g002050	1	0	0	429
Sb01g033060	1	0	0	429
Sb01g013710	3	0	2	214
Sb01g043060	2	0	4	418
Sb01g046550	2	0	0	318
Sb01g003700	1	0	0	455
Sb01g011740	1	0	0	233
Sb01g006220	1	0	0	292
Sb01g009520	2	0	0	404
Sb01g016110	5	0	0	397
Sb01g044810	6	0	5	502
Chromosome 2				
Sb02g006330	2	1	2	191
Sb02g000780	1	1	2	273
Sb02g005440	1	0	0	464
Sb02g036870	2	0	0	225
Sb02g022510	1	0	0	552
Sb02g006420	4	2	5	731
Sb02g009980	3	2	2	363

Sb02g032470	2	0	1	438
Chromosome 3				
Sb03g039090	6	4	2	405
Sb03g037370	1	1	2	311
Sb03g009900	2	0	0	517
Sb03g037360	2	0	0	400
Sb03g013840	4	0	0	139
Sb03g012420	3	2	1	144
Sb03g007840	1	0	2	355
Sb03g037870	6	0	0	333
Sb03g045390	1	0	0	558
Sb03g027710	1	0	1	341
Sb03g003190	2	0	0	454
Chromosome 4				
Sb04g028300	1	0	0	494
Sb04g027910	2	0	0	485
Sb04g021610	1	0	0	209
Sb04g037170	1	1	2	346
Sb04g019020	8	3	6	235
Sb04g005210	1	1	1	236
Chromosome 5				
Sb05g001680	2	1	3	153
Chromosome 6				
Sb06g015180	2	0	3	314
Sb06g026710	1	0	0	277
Sb06g029500	2	0	0	486
Chromosome 7				
Sb07g001320	7	0	0	473
Sb07g005930	1	1	2	436
Chromosome 8				
Sb08g008320	1	1	7	447
Sb08g016302	1	0	3	268
Sb08g020760	1	0	3	488
Sb08g015010	4	0	0	484
Sb08g002250	6	5	4	316
Sb08g002660	1	0	0	345
Chromosome 9				
Sb09g000820	1	1	2	394
Sb09g023620	1	0	0	434
Sb09g006050	2	2	3	268
Sb09g005280	2	1	1	527
Sb09g029170	1	0	10	406
Chromosome 10				
Sb10g002230	1	0	2	398
Sb10g007380	1	1	2	374
Sb10g004540	1	0	0	255
TOTAL	125	32	87	21,612

Table 3.1. Sorghum genes with SFPs predicted by the GeSNP software. [a] SFPs with *t* values >= 7.

SUGARCANE PROBE SET	PROBE PAIR #	*S. bicolor* GENE ID	POSITION	FUNCTION
***t* value = 22**				
SOF.4093.2.S1_AT	6	NGH	Ch1_8,313,833..8,313,816	----------
SOF.4567.1.S1_AT	8	Sb01g044810	Ch1_67,980,922..67,980,946	MADS-box transcription factor
SOF.5184.2.S1_A_AT	6	Sb03g001160	`Ch3_991,187..991,163	Similar to Os02g0294700 protein
SOFAFFX.1284.1.S1_S_AT	3	Sb03g008870	Ch3_9,656,668..9,656,644	Unknown
SOF.5348.1.S1_AT	11	Sb03g003510	Ch3_3,731,533..3,731,509	Ubiquitin-conjugating enzyme E2
SOF.2770.1.S1_AT	4	Sb03g041770	Ch3_69,253,777..69,253,759	Unknown
SOF.3851.1.S1_AT	10	Sb05g004130	Ch5_4,878,250..4,878,268	60S ribosomal protein L3
SOF.2692.1.S1_AT	5	Sb08g002250	Ch8_2,360,780..2,360,756	Cytochrome P450
SOF.4985.2.S1_A_AT	10	Sb08g018480	Ch8_48,581,627..48,581,646	ATP-citrate synthase
SOFAFFX.1129.1.S1_AT	2	Sb08g021850	Ch8_53,598,165..53,598,144	Serine/threonine protein phosphatase
SOFAFFX.1129.1.S1_AT	9	Sb08g021850	Ch8_53,598,029..53,598,005	Serine/threonine protein phosphatase
SOF.4246.1.S1_A_AT	11	Sb09g005270	Ch9_6,772,194..6,772,216	Unknown
***t* value = 23**				
SOF.2535.1.A1_AT	6	Sb02g011130	Ch2_18,051,363..18,051,363	Similar to putative RES protein
SOF.1282.2.S1_A_AT	11	NGH	Ch2_57,946,767..57,946,743	----------
SOF.1664.2.S1_A_AT	1	Sb03g033760	Ch3_62,018,464..62,018,488	Putative BURP domain-containing protein
SOFAFFX.1284.1.S1_X_AT	2	Sb03g008870	Ch3_9,656,190..9,656,166	Unknown
SOF.497.2.S1_AT	7	Sb07g027480	Ch7_62,509,159..62,509,135	3-Hydroxy-3-methylglutaryl-coA reductase
SOF.1190.1.S1_AT	8	Sb07g005930	Ch7_8,393,958..8,393,934	Unknown
SOF.2692.1.S1_AT	6	Sb08g002250	Ch8_2360760..2360736	Cytochrome P450
SOF.355.1.S1_AT	8	Sb09g005570	Ch9_7,345,144..7,345,120	Heat shock protein
***t* value = 24**				
SOF.4310.1.S1_AT	3	Sb01g028500	Ch1_49,703,504..49,703,480	Senescence-associated protein like
SOF.4030.1.A1_AT	10	Sb02g003450	Ch2_3,915,697..3,915,680	Similar to B0616E02-H0507E05.5 protein
SOF.4972.1.S1_A_AT	9	NGH	Ch3_17,046,891..17,046,867	----------
SOF.1835.1.S1_AT	3	Sb03g033140	Ch3_61,527,980..61,527,956	Putative nuclear RNA binding protein A
SOF.1003.1.S1_AT	2	Sb05g002580	Ch5_2,717,665..2,717,641	Cytochrome P450
SOF.1694.1.A1_AT	9	Sb06g033460	Ch6_61,437,575..61,437,596	Similar to H0913C04.1 protein
SOF.3020.2.A1_AT	4	Sb09g002960	Ch9_3,216,665..3,216,682	Aspartic proteinase
***t* value = 25**				
SOF.2803.1.S1_AT	11	Sb01g043050	Ch1_66,375,993..66,375,971	Unknown
SOF.1537.1.S1_AT	7	Sb03g011270	Ch3_12,484,656..12,484,632	Mg-protoporphyrin IX monomethyl ester cyclase
SOF.2992.1.A1_AT	6	Sb04g037920	Ch4_67,480,989..67,481,008	Similar to Os04g0137500
SOF.1443.1.S1_AT	7	Sb04g010990	Ch4_15,758,311..15,758,334	Unknown
SOF.3814.1.S1_AT	11	Sb04g019020	Ch4_44,439,307..44,439,289	Fructose bisphosphate aldolase
SOF.3699.1.A1_AT	4	Sb07g005850	Ch7_8,311,400..8,311,376	Equilibrative nucleoside transporter 1
SOF.2286.1.A1_AT	2	Sb09g025350	Ch9_54,815,478..54,815,502	Similar to Os05g051300
SOF.1994.1.S1_X_AT	7	Sb10g005375	Ch10_4,802,664..4,802,640	Putative uncharacterized protein

Table 3.2. Sugarcane probe pairs with *t* values of 22 to 25 that identify sorghum transcripts with SFPs but not ELPs. NGH: non-genic hit.

S. bicolor GENE ID	POSITION	SUGARCANE PROBE SET	PROBE PAIR #	*t* VALUE	BTx623-Rio-Sc SNP
Sb02g006330	Ch2_7,909,203..7,909,180	SOF.1519.2.S1_AT	8	23	C-T-C
Sb02g000780	Ch2_628,587..628,568	SOF.1326.1.S1_A_AT	5	15.2	A-G-G
Sb02g006420	Ch2_8,048,752..8,048,728	SOF.2471.1.S1_AT	5	34.1	C-A-C
	Ch2_8,048,741..8,048,717		6	19.8	SAME
Sb02g009980	Ch2_14,533,601..14,533,625	SOFAFFX.868.1.S1_S_AT	9	13.7	A-T-A/C-T-C
	Ch2_14,533,610..14,533,630		10	12.9	SAME
Sb03g037370	Ch3_65,336,537..65,336,560	SOFAFFX.772.1.S1_S_AT	7	19.1	C-G-C
Sb03g012420	Ch3_14,371,043..14,371,016	SOF.2629.3.S1_A_AT	8	38.2	C-T-C
	Ch3_14,371,036..14,371,016		9	19.4	SAME
Sb03g039090	Ch3_66,876,720..66,876,744	SOF.5269.1.S1_AT	6	8.1	T-A-T/C-A-C
	Ch3_66,876,724..66,876,748		7	12	SAME
	Ch3_66,876,727..66,876,751		8	17.1	SAME
	Ch3_66,876,730..66,876,754		9	16.1	SAME
	Ch3_66,876,734..66,876,758		10	45.8	SAME
Sb04g019020	Ch4_44,439,369..44,439,345	SOF.3814.1.S1_AT	8	21.9	C-T-T
	Ch4_44,439,366..44,439,342		9	15.3	SAME
	Ch4_44,439,307..44,439,289		11	25.5	T-G-T
Sb04g037170	Ch4_66,851,287..66,851,311	SOF.151.1.S1_AT	8	19.4	G-C-G
Sb05g001680	Ch5_1,816,812..1,816,788	SOF.1902.1.S1_S_AT	6	33.1	A-G-G
Sb07g005930	Ch7_8,393,958..8,393,934	SOF.1190.1.S1_AT	8	23.3	T-G-T
Sb08g008320	Ch8_15,917,006..15,917,030	SOFAFFX.1412.1.A1_S_AT	2	15.1	T-C-C
Sb08g002250	Ch8_2,360,967..2,360,943	SOF.2692.1.S1_AT	2	16.8	A-G-A
	Ch8_2,360,780..2,360,756		5	22.1	A-G-G
	Ch8_2,360,760..2,360,736		6	23.6	T-C-C
Sb09g006050	Ch9_8,732,113..8,732,094	SOFAFFX.1438.1.A1_S_AT	3	14.9	C-G-C
	Ch9_8,732,054..8,732,030		7	82.5	C-A-C
Sb09g000820	Ch9_624,173..624,197	SOF.808.1.S1_AT	8	29	G-C-G
Sb09g005280	Ch9_6,782,917..6,782,941	SOF.5033.1.S1_AT	9	15.1	A-G-G
Sb10g007380	Ch10_7,220,153..7,220,177	SOFAFFX.287.1.S1_AT	7	14	T-C-C

Table 3.3. Nucleotide change conservation for validated SFPs between BTx623, Rio and sugarcane. SAME means that a different probe pair recognizes the same SNP. Sc: sugarcane.

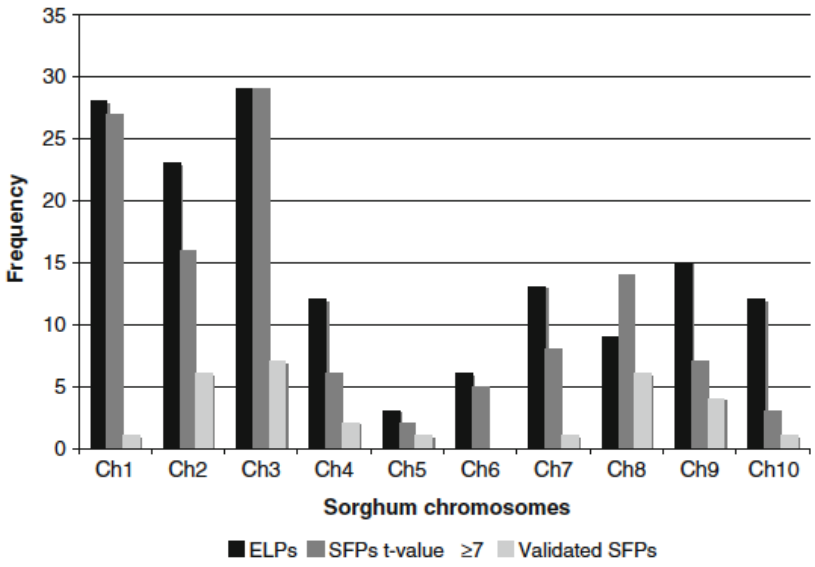


Figure 3.1. Histogram showing the proportion of ELPs and SFPs between BTx623 and Rio for each sorghum chromosome. The number of genes with ELPs that I previously reported (Calvino et al., 2008) were plotted for each chromosome along with the numbers of SFPs found in this study. Only SFPs with *t* value >=7 were taken into consideration.

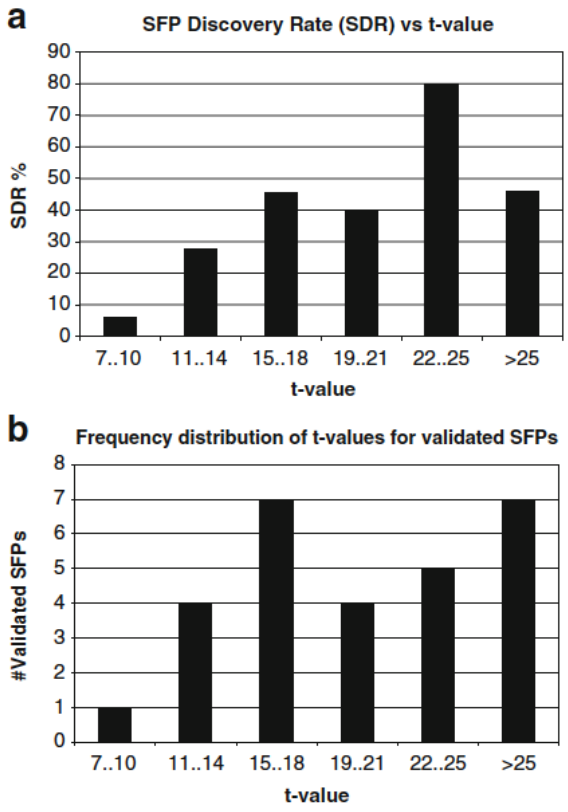


Figure 3.2. The SFP discovery rate of GeSNP is dependent on the *t* value. (a) The percentage of SFPs in sorghum genes that were validated through sequencing (and thus represented true SNPs between BTx623 and Rio) was plotted against their respective *t* values. **(b)** For the validated SFPs, I calculated the frequency distribution of their respective *t* values.

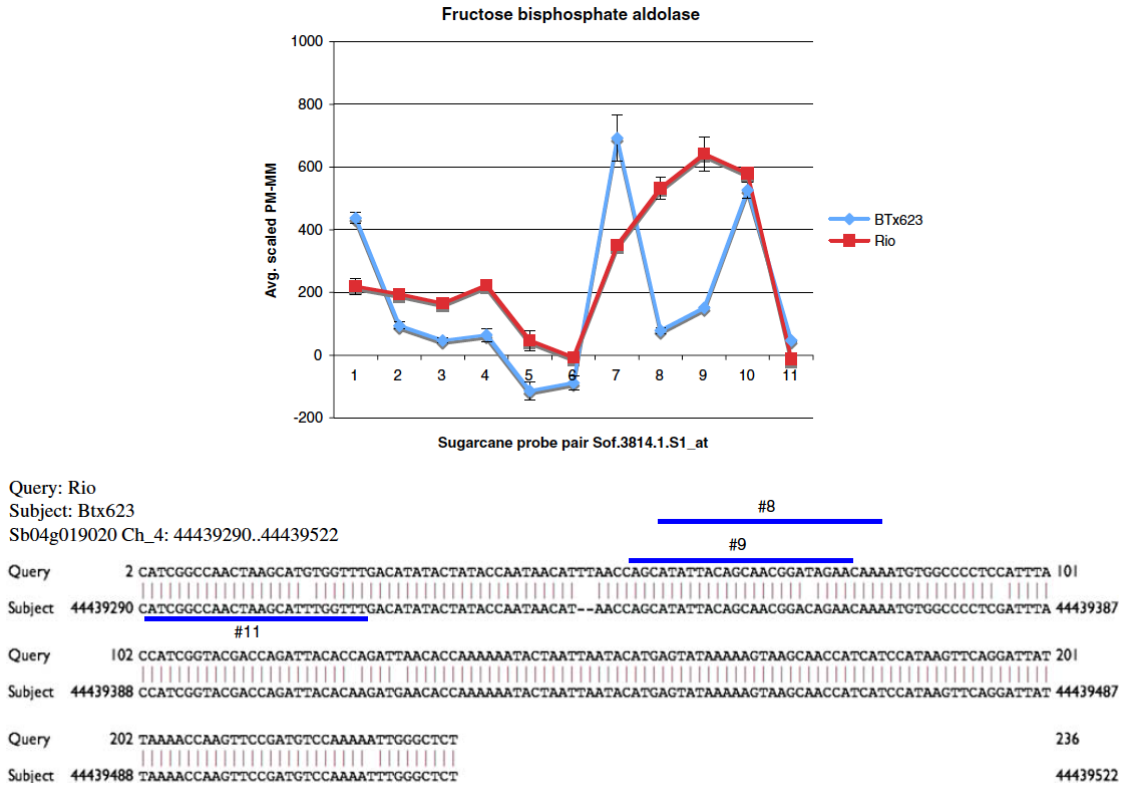


Figure 3.3. SFP validation for fructose bisphosphate aldolase. A fragment from the gene fructose bisphosphate aldolase was cloned and sequenced from both BTx623 and Rio and SNPs predicted by the sugarcane probe pairs #8, 9 and 11 were validated. The blue line bars represent the sugarcane probe pairs that are identical to either the Rio sequence (probe pairs #8 and #9) or identical to the BTx623 sequence (probe pair #11).

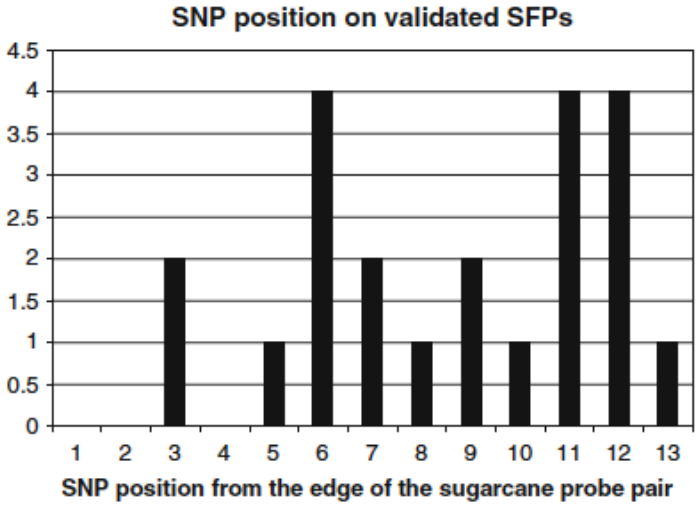
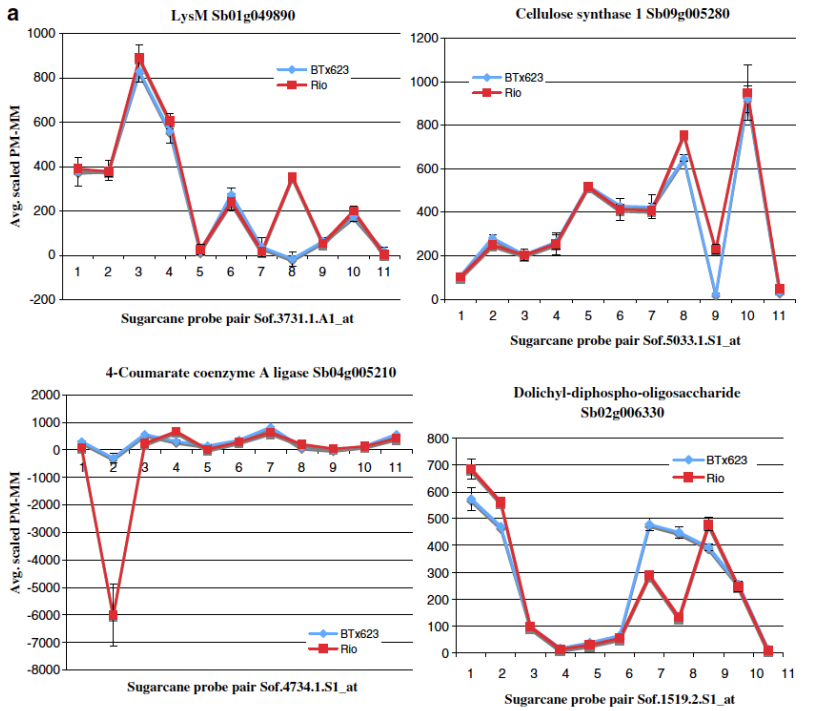


Figure.3.4. The position of the SNP along the 25mer in the sugarcane probe pair influences the SFP validation. The position of the SNP from the edge of the sugarcane probe pair was scored for each validated SFP and is shown on the *x*-axis. The number of sugarcane probe pairs is shown on the *y*-axis. Most of the SNPs located within positions 6 and 13 along the 25mer. If two or more SNPs were located on a single probe pair, their positions along the 25mer were not counted and thus not included in this graph.


a
LysM Sb01g049890
Cellulose synthase 1 Sb09g005280
4-Coumarate coenzyme A ligase Sb04g005210
Dolichyl-diphospho-oligosaccharide Sb02g006330
Avg. scaled PM-MM
BTx623
Rio
Sugarcane probe pair Sof.3731.1.A1_at
Sugarcane probe pair Sof.5033.1.S1_at
Sugarcane probe pair Sof.4734.1.S1_at
Sugarcane probe pair Sof.1519.2.S1_at

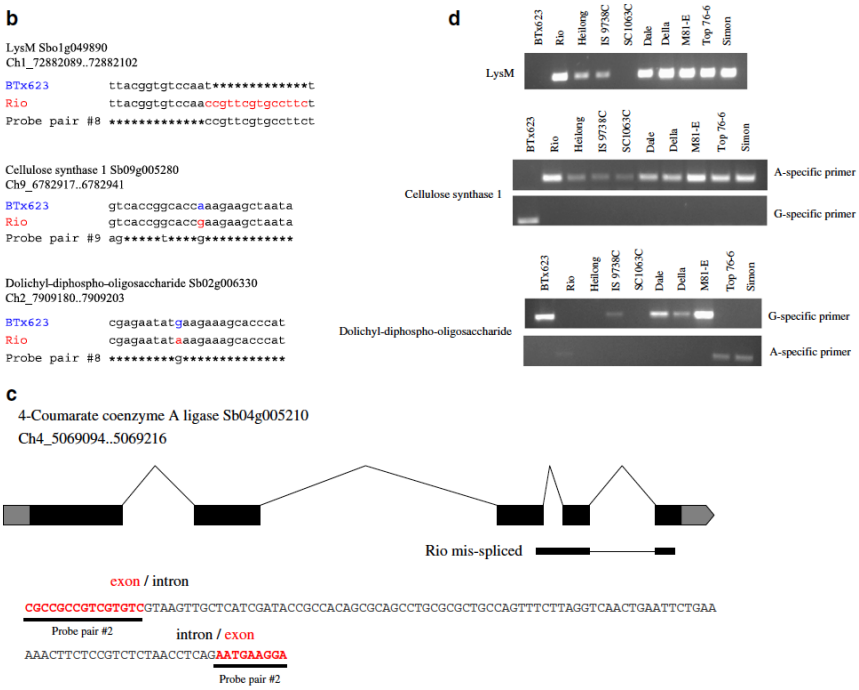

b
LysM Sbo1g049890
Ch1_72882089..72882102
BTx623 ttacggtgtccaat**********t
Rio ttacggtgtccaaccgttcgtgccttct
Probe pair #8 **********ccgttcgtgccttct
Cellulose synthase 1 Sb09g005280
Ch9_6782917..6782941
BTx623 gtcaccggcaccaaagaagctaata
Rio gtcaccggcaccgaagaagctaata
Probe pair #9 ag******t****g**********
Dolichyl-diphospho-oligosaccharide Sb02g006330
Ch2_7909180..7909203
BTx623 cgagaatatgaagaaagcacccat
Rio cgagaatataaagaaagcacccat
Probe pair #8 *********g**********
d
BTx623
Rio
Heilong
IS 9738C
SC1063C
Dale
Della
M81-E
Top 76-6
Simon
LysM
Cellulose synthase 1
A-specific primer
G-specific primer
Dolichyl-diphospho-oligosaccharide
c
4-Coumarate coenzyme A ligase Sb04g005210
Ch4_5069094..5069216
Rio mis-spliced
exon / intron
CGCCGCCGTCGTGTCGTAAGTTGCTCATCGATACCGCCACAGCGCAGCCTGCGCGCTGCCAGTTTCTTAGGTCAACTGAATTCTGAA
Probe pair #2
intron / exon
AAACTTCTCCGTCTCTAACCTCAGAATGAAGGA
Probe pair #2

Figure 3.5. GeSNP prediction of SFPs in sorghum genes related to biofuel traits. **(a)** The hybridization intensity between the perfect match (PM) and the mismatch (MM) oligonucleotides was averaged and scaled (GeSNP software output) and plotted against each sugarcane probe pair. Graphs are shown for four genes that have SFPs with *t* values of >= 7 and that I previously reported to be differentially expressed between BTx623 and Rio. **(b)** The SFP present in the LysM identified a 13-bp indel, whereas the SFPs present in cellulose synthase 1 and dolichyl-disphospho-oligosaccharide identified an A/G and G/A SNP between BTx623 and Rio, respectively. **(c)** In Rio, the third intron of the gene 4-Coumarate coenzyme A ligase was mis-spliced and detected in the sugarcane probe pair #2. **(d)** Molecular markers for the genes LysM, cellulose synthase 1, and dolichyl-diphospho-oligosaccharide were generated based on allele-specific PCR. In the case of LysM, a primer spanning the 13-bp deletion in BTx623 was used to selectively amplify the allele from Rio. In the case of cellulose synthase 1 and dolichyl-diphospho-oligosaccharide, primer pairs specific for the SNP in question were generated using the WebSNAPER software and tested empirically.

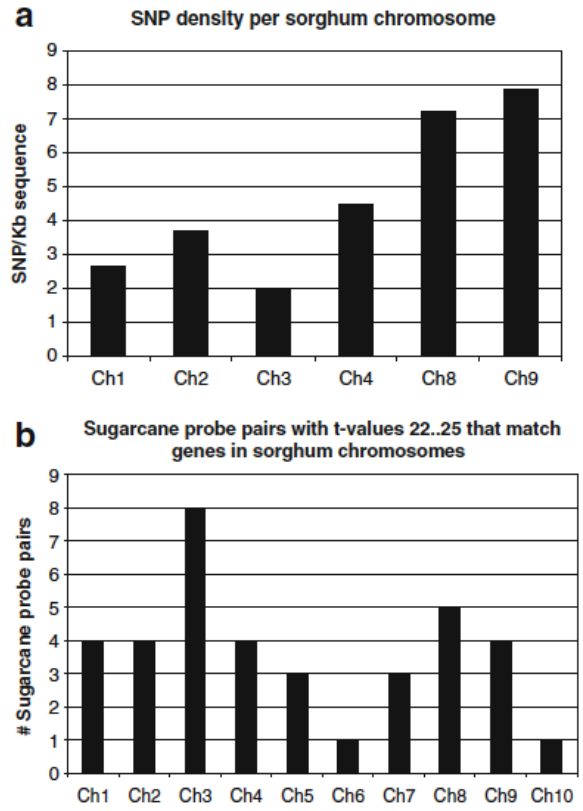


Figure 3.6. SNP density per sorghum chromosome. The number of SNPs per kb of sequence was calculated based on the number of genes sequenced belonging to a given chromosome. **(a)** Only those chromosomes with five or more genes sequenced are represented. **(b)** Frequency distribution along sorghum chromosomes of sugarcane probe pairs with *t* values between 22 and 25.

3.3.2. Development of molecular markers based on validated SFPs

The identification of SNPs between BTx623 and Rio provided a direct way to develop molecular markers that can be used in mapping populations. From 58 candidate genes, we were able to develop allele-specific PCR markers for 18 (Table 3.4). We utilized the Single Nucleotide Amplified Polymorphism (SNAP) technique

to develop markers based on SNPs (Drenkard et al., 2000), as it is shown for the gene alanine aminotransferase (Fig. 3.7). These markers were tested also in other grain and sweet sorghum lines to see whether the SNPs were conserved or not (Table 3.4). In fact, we found a marker within the gene Sb09g029170 that distinguished the grain sorghums from the sweet sorghums cultivars used in this study. The protein product encoded by this gene is a putative ketol-acid reductoisomerase enzyme that is involved in the biosynthesis of valine, leucine, and isoleucine amino acids (www.phytozome.net/cgi-bin/gbrowse/sorghum/). SNAP markers were also developed for the cellulose synthase 1 and dolichyl-diphospho-oligosaccharide genes (Fig. 3.5d).

It has been suggested that Dale and Della sweet sorghums share a common genetic background (Kimberley et al., 2007). In agreement with this, we found that from ten SNAP markers that gave a PCR product in both lines, they always represented the same allele (Table 3.4). In addition, the sweet sorghum lines Top 76-6 and Simon have been identified as attractive contrasting pairs for mapping purposes based on their difference not only in genetic distance (*D*) but also in sugar content (measured as Brix degree) (Ali et al., 2007). In my work, we identified six SNAP markers within the genes Sb01g044810, Sb03g027710, Sb04g0037170, Sb08g008320, Sb09g006050, and Sb10g002230, respectively, which were polymorphic between Top 76-6 and Simon. These markers will be useful for mapping purposes when these lines are used as parents.

S. bicolor GENE ID	ALLELE	WebSNAPER PRIMER SEQUNCES	PCR PRODUCT SIZE (bp)	ALLELE PRESENCE[a]
Sb01g043060	T	F: GTAATATACTGACGCCAAAAGAGGCGGATT R: TCAACTGCTGTTGTCGAGGACATTGG	306	BT
	A	F: TGTAATATACTGACGCCAAAAGAGGCGACTT R: TCAACTGCTGTTGTCGAGGACATTGG	307	Ri-Top
Sb01g044810	C	F: CAATCCTGCTCCCCAATCCAGACC R: GATTACGAGATCAGCGGTCTGGAAAGAAA	334	BT-Da-De-Sim
	T	F: GCAATCCTGCTCCCCAATCCAGACT R: GATTACGAGATCAGCGGTCTGGAAAGAAA	335	Ri-He-IS-SC-M81-Top
Sb02g000780	A	F: TGGAGCAATACGAGGGCTACTCCAAA R: AATCTTCAGAAACGCTCCATTTGTGCTG	118	BT
	G	F: TGGAGCAATACGAGGGCTACTCCATG R: AATCTTCAGAAACGCTCCATTTGTGCTG	118	Ri-He-IS-SC-Da-De-M81-Top-Sim
Sb02g006330	G	F: TGTGGTACAGGTACACAAGCGAGAACATG R: CCTTACAGGCATAACGAGTATGAGAGATTCATAACA	115	BT-IS-Da-De-M81
	A	F: CTTATTTGTGGTACAGGTACACAAGCGAGAATAAA R: CCTTACAGGCATAACGAGTATGAGAGATTCATAACA	121	Ri-Top-Sim
Sb03g012420	C	F: GAAGCATTCTTTCCGATACAATATGGCCTATC R: TTCGATTAAAGGATTGTTGATGAAACTAGGGG	164	BT-He-SC-M81-Top-Sim
	T	F: GAAGCATTCTTTCCGATACAATATGGCCTACT R: TTCGATTAAAGGATTGTTGATGAAACTAGGGG	164	Ri-IS-Da
Sb03g007840	C	F: CCATAAATGTCATTGTGGAGACATCCGTTC R: TGGAACGTCAAAACATTGACCGGAA	161	BT-He-IS-SC-M81-Top
	T	F: AAATGTCATTGTGGAGACATCCGGGT R: TGGAACGTCAAAACATTGACCGGAA	157	Ri-Da-Sim
Sb03g027710	T	F: GGTCATCGGTGATGGTGGAGAACCT R: GGGAATTCGATTATGTCCATCACACCC	343	BT
	G	F: AGGTCATCGGTGATGGTGGAGATCTG R: GGGAATTCGATTATGTCCATCACACCC	344	Ri-Da-Sim
Sb03g039090	C	F: CGAACCCAACAACCTGTAACAATAAGCACTAC R: GGAATTCGATTATCTCGGGGCTCATCTAC	326	BT-Da-De-Top-Sim
	A	F: GAACCCAACAACCTGTAACAATAAGCAGAAA R: GGAATTCGATTATCTCGGGGCTCATCTAC	325	Ri-M81
Sb04g0037170	G	F: CACAAGCGACTTGAAACTGCGCTG R: GGCTTGACAACTGCTTCAACCTCTGC	131	BT-IS-SC-Top
	C	F: CACAAGCGACTTGAAACTGCACCC R: GGCTTGACAACTGCTTCAACCTCTGC	131	Ri-He-Da-De-M81-Sim
Sb07g005930	T	F: CAGTTCTCCAATCCTTTCCTCTGTGGTCT R: GTGAGAAGCGTGGGATGCTCATCAG	146	BT-He-SC-Da-M81
	G	F: GTTCTCCAATCCTTTCCTCTGTGGTCG R: GTGAGAAGCGTGGGATGCTCATCAG	144	Ri-IS-Top-Sim
Sb08g020760	C	F: CAGAGGAAGCCCTTACACAGATCCGAC R: TACCCACAGGTCTGGAAAGGGCAAG	1,400	BT-M81
	T	F: CAGAGGAAGCCCTTACACAGATCCGAT R: TACCCACAGGTCTGGAAAGGGCAAG	416	Ri-He-IS-SC-Top-Sim
Sb08g008320	T	F: GCAGTGGAAGGACATCATTGCCCAT R: CTCTTCCGGGACGCGACGTTC	174	BT-He-Da-M81-Sim
	C	F: CAGTGGAAGGACATCATTGCCGTC R: CTCTTCCGGGACGCGACGTTC	173	Ri-IS-SC-Top
Sb09g005280	A	F: GCAGCACCGTCACCGGCACTA R: GAGGCTCAATCAAGATCGTCTGCCC	142	BT
	G	F: CAGCACCGTCACCGGCATCG R: GAGGCTCAATCAAGATCGTCTGCCC	141	Ri-He-IS-SC-Da-De-M81-Top-Sim

Sb09g029170	C	F: CTACTCTGAGATCATCAACGAGAGCGTGAAC R: CCTAGATCCCAGGCGAGCCGTC	124	BT-He-SC-IS
	T	F: CTACTCTGAGATCATCAACGAGAGCGTGTTT R: CCTAGATCCCAGGCGAGCCGTC	124	RI-Da-De-M81-Top-Sim
Sb09g000820	G	F: TCGAGAGCGATGCCTTCTGACATTG R: CCATATCTCCAGCCATCTTCAATGTTGTG	128	BT-Top
	A	F: CGAGAGCGATGCCTTCTGACAGCA R: CCATATCTCCAGCCATCTTCAATGTTGTG	130	Ri
Sb09g006050	C	F: ATAGAAGGCAGAATGAACGCTGGAAAGC R: GGGCAAGCAGGCCTGGAACTTC	105	BT-Top
	A	F: AGAAGGCAGAATGAACGCTGGACTGA R: GGGCAAGCAGGCCTGGAACTTC	103	Ri-He-IS-SC-Da-De-M81-Sim
Sb10g007380	T	F: GAACTACAGACATGCACAAGGATAGCAGGTT R: ATTGCATTCAGGAAGCTCGCTCGA	561	BT-Top
	C	F: GAACTACAGACATGCACAAGGATAGCAGAGC R: ATTGCATTCAGGAAGCTCGCTCGA	561	Ri-He-IS-SC-Da-De-M81
Sb10g002230	G	F: CTTCAATCCGACAACCAAGTCGCTG R: CTGGAACTGCAATGCGGCCATT	197	BT-He-IS-Top
	A	F: GCTTCAATCCGACAACCAAGTCGCTA R: CTGGAACTGCAATGCGGCCATT	197	Ri-SC-Da-De-M81-Sim

Table 3.4. Primer sequences of SNAP markers within sorghum genes. BT: BTx623; Ri: Rio; He: Heilong; IS: IS 9738C; SC: SC 1063C; Da: Dale; De: Della; M81: M81-E; Top: Top76-6; Sim: Simon. [a]Only the cultivars that gave a PCR product were scored. If a cultivar was heterogeneous for a particular allele, it was not scored.

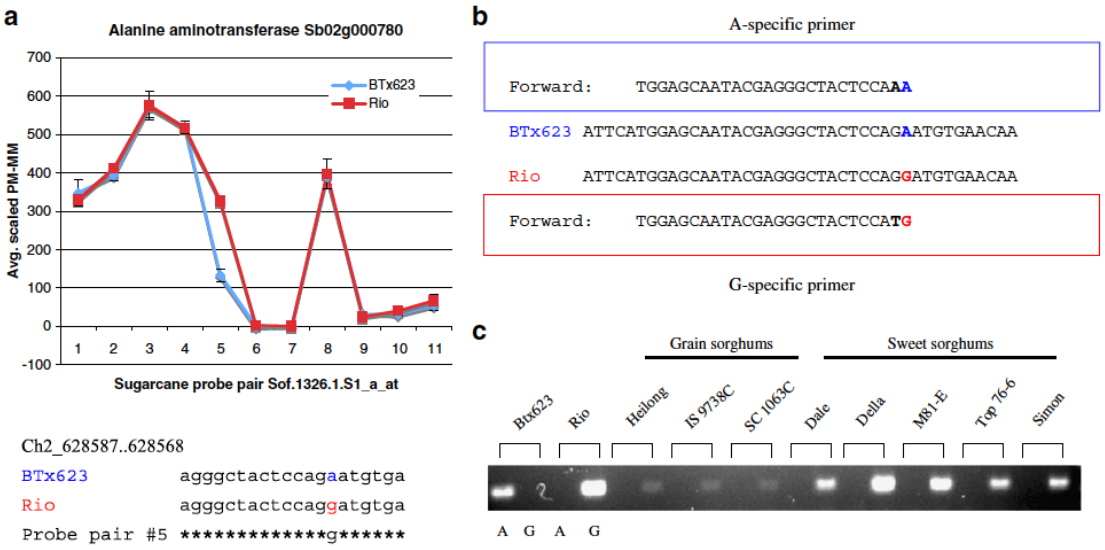


Figure 3.7. Development of a molecular marker for alanine aminotransferase based on SFP discovery and the SNAP technique. (a) The SFP detected by the probe pair #5 in the sugarcane probe set SOF.1326.1.S1_A_AT was validated through sequencing. **(b and c)** Specific primers for either A or G nucleotides were designed with WebSNAPER and tested through PCR in ten sorghum lines.

3.4. Discussion

A significant proportion of the phenotypic variation in any organism can be attributed to polymorphisms at the DNA level. Thus, these DNA polymorphisms can be used for genotyping, molecular mapping, and marker-assisted selection applications. The association of a particular trait of interest with a DNA polymorphism is essential for breeding purposes. Microarrays have been used to identify abundant DNA polymorphisms throughout the genome (Hazen and Kay, 2003; Gupta et al., 2008). In particular, ELPs and SFPs can be identified from RNA

hybridization studies. SFPs are detected by oligonucleotide arrays and represent DNA polymorphisms between genotypes within an individual oligonucleotide probe pair that is detected by the difference in hybridization affinity (Borevitz et al., 2003). In addition, SFPs present in a transcribed gene may be the underlying cause of the difference in a phenotype of interest. In most of the cases, SNPs are the cause of SFPs as have been demonstrated by sequence analysis (Borevitz et al., 2003; Rostoks et al., 2005).

Here, the goal was to identify SFPs from an Affymetrix sugarcane genechip dataset of closely related species (Calviño et al., 2008). The Affymetrix sugarcane genechip was used to survey the SFPs with the GeSNP software between two sorghum cultivars that differed in the accumulation of fermentable sugars in their stems, with the objective to develop genetic markers for mapping purposes. This was the first report to my knowledge of the use of GeSNP to identify SFPs within closely related grass species and the development of molecular markers based on validated SFPs.

We cloned and sequenced gene fragments harboring SFPs with *t* values equal or higher than 7 from 58 sweet sorghum genes comprising 125 SFPs in total. In this study, we found a SFP discovery rate of 25.6%. Still, there were several possibilities to increase the SDR. First, the number of biological replicates suggested for using the GeSNP software is 4 or more. In contrast, we had only three replicates for both grain and sweet sorghum. Second, the cross-species hybridization of sorghum RNAs to probe sets of the sugarcane array is not as sensitive as intra species hybridization. Third, false positives could be due to the cross-hybridization of paralogous gene

targets to individual probes, which may affect the specificity of the SFP calling. This problem would also arise from using next-generation sequencing for SNP detection. Nevertheless, we could show that the use of expression analysis in conjunction with GeSNP is an efficient and inexpensive way to develop new molecular markers.

The sugarcane probe pairs with *t* values between 22 and 25 had the highest SDR (80%) found in my study. One of these probe pair sets matched a sorghum gene coding for fructose bisphosphate aldolase (cytoplasmic isozyme) and the identified SFP was confirmed through DNA sequence analysis (Fig. 3.3). This gene codes for a glycolytic enzyme that catalyzes the cleavage of fructose 1,6 bisphosphate to glyceraldehyde 3-phosphate and dihydroxyacetone phosphate (Tsutsumi et al., 1994).

One third (33%) of the 58 genes that we sequenced had a validated SFP. In addition, we could detect SNPs in 19% of all sequenced genes at a different position than indicated by GeSNP. This is attributable to the fact that the probe pair set does only cover a part of the gene, which implies that any SNP that is outside this region cannot be reported by using GeSNP. We estimated the average SNP density between BTx623 and Rio to one SNP every 248 bp. This is probably an underestimation because the sugarcane probe sets were designed from genic regions and are, therefore, more conserved than other regions in the genome.

Although the sorghum chromosomes 1, 2, and 3 had the highest numbers for both ELPs and SFPs, chromosomes 8 and 9 were the most polymorphic ones, measured as the number of SNPs per Kb sequence (Figs. 3.1 and 3.6). Our data was in agreement with a previous report by Kimberley et al., (2007) in which amplified

fragment-length polymorphism markers on chromosome 8 could unambiguously distinguish grain from sweet sorghum lines (Kimberley et al., 2007). Furthermore, sugar content QTLs have been located in this chromosome with a RIL derived from a dwarf derivative of Rio as one of the parents. In addition, we found that a marker within the gene Sb09g029170 coding for a putative ketol-acid reductoisomerase could discriminate the grain sorghums from the sweet sorghum lines used in this study (Table 3.4). This enzyme is the second in the biosynthesis of branched amino acids valine, leucine, and isoleucine (Leung and Guddat, 2009). When the SNPs found through validated SFPs were compared between BTx623, Rio, and sugarcane, we found that SNPs between BTx623 and sugarcane are twice as high as between Rio and sugarcane.

Allelic genetic diversity among sweet sorghum cultivars has previously been investigated based on simple sequence repeat markers (Ali et al., 2007). This study described the correlations between allelic diversity and the degree of stem sugar. Indeed, one could envision a simpler approach, using the microarray described here by hybridizing stem-derived RNAs from these lines to the sugarcane genechip, and identify both ELPs and SFPs for subsequent mapping of sugar content QTLs. Furthermore, the SNPs identified in our study provided me with the opportunity to develop molecular markers within genes. So far, there was no report of SNP-based molecular markers in transcribed genes in sorghum at the time of my study. The SFPs generated from transcriptome studies are also useful for the development of markers in those species that lack sequence resources such as *Miscanthus* and

switchgrass, further extending the use of microarrays of one species for related ones.

3.5. Materials and methods

3.5.1. Plant Materials

The grain sorghum lines Heilong (accession number PI 563518), IS 9738C (PI 595715), and SC 1063C (PI 595741) were obtained from the National Plant Germplasm System (NPGS), USDA. The other lines used in this study were previously described (Calviño et al., 2008). Two-weeks old seedlings were harvested for the extraction of genomic DNA.

3.5.2. SFP discovery and validation from Affymetrix transcript data

The microarray analysis for differentially expressed transcripts in stems of grain and sweet sorghum with a sugarcane genechip was previously described (Calviño et al., 2008). The CEL files from the microarray work were uploaded into the publicly available GeSNP software at http://porifera.ucsd.edu/~cabney/cgi-bin/geSNP.cgi, and an excel file was obtained with all the probe sets in the array harboring an SFP together with their respective *t* values. The excel file also contained the average hybridization intensity between the PM and MM probe pairs (average scaled PM–MM) as well as their variance values that were converted to standard deviations. These values were used to generate the graphs displaying differences in hybridization intensity between BTx623 and Rio along the 11 sugarcane probe pairs for a given probe set.

From the transcripts previously described as being differentially expressed between grain sorghum BTx623 and sweet sorghum Rio, we selected those harboring SFPs with *t* values ≥ 7 for further validation through sequencing. In total, we sequenced gene fragments corresponding to 58 different genes.

Total RNA from Rio stem tissue was extracted at the time of flowering from three independent plants. RNA extraction was performed with the RNeasy Plant Mini Kit from QIAGEN. cDNA synthesis was performed for each of the three samples from 1 µg of total RNA with the SuperScript III First-Strand Synthesis kit from Invitrogen. cDNAs from Rio were pooled respectively and used for the amplification of genes with SFPs.

The reverse transcription polymerase chain reaction products were checked by agarose gel electrophoresis in order to verify that a single band amplification product from each gene was present. The PCR products were purified with the QIAquick PCR Purification kit from Qiagen and cloned into the pGEM-T easy vector from Promega. Twelve clones per gene were sequenced in order to identify any sequencing or reverse transcriptase errors. The consensus sequence for each gene was then used to find SNPs between BTx623 and Rio.

3.5.3. Development of molecular markers using WebSNAPER software

Once a SNP was identified between BTx623 and Rio for a particular gene of interest, the sequence harboring the SNP in question was uploaded into the publicly available WebSNAPER software (http://pga.mgh.harvard.edu/cgi-bin/snap3/websnaper3.cgi). The SNAP procedure has been previously described

(Drenkard et al., 2000). Several primer pairs per SNP were tested, and the ones that successfully distinguished the SNP in one line or the other were selected. The primer sequences used to distinguish SNPs are provided in Table 3.4.

Genomic DNA from 2-week-old seedlings was extracted with the PrepEase Genomic DNA Isolation kit from USB. Several concentrations of genomic DNA were tested, and 50 ng was used for testing the SNAP primer pairs through PCR. The conditions used for PCR reaction were as follows: 94°C for 2 min, then 30× [94°C 30 s, 64°C 30 s, 72°C 30 min] and a final extension at 72°C for 2 min

3.6. References

Ali ML, Rajewski JF, Baenziger PS, Gill KS, Eskridge KM, Dweikat I (2007) Assessment of genetic diversity and relationship among a collection of US sweet sorghum germplasm by SSR markers. Molecular Breeding **21:** 497-509.

Borevitz J, Chory J (2004) Genomics tools for QTL analysis and gene discovery. Current Opinion in Plant Biology **7:** 132-136.

Borevitz J, Hazen S, Michael T, Morris G, Baxter I, Hu T, Chen H, Werner J, Nordborg M, Salt D, Kay S, Chory J, Weigel D, Jones J, Ecker J (2007) Genome-wide patterns of single-feature polymorphism in Arabidopsis thaliana. Proceedings of the National Academy of Sciences **104:** 12057-12062.

Borevitz J, Liang D, Plouffe D, Chang H-S, Zhu T, Weigel D, Berry C, Winzeler E, Chory J (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Research **13:** 513-523.

Cáceres M, Lachuer J, Zapala M, Redmond J, Kudo L, Geschwind D, Lockhart D, Preuss T, Barlow C (2003) Elevated gene expression levels distinguish human from non-human primate brains. Proceedings of the National Academy of Sciences **100:** 13030-13035.

Calviño M, Bruggmann R, Messing J (2008) Screen of Genes Linked to High-Sugar Content in Stems by Comparative Genomics. Rice **1:** 166-176.

Coram T, Settles M, Wang M, Chen X (2008) Surveying expression level polymorphism and single-feature polymorphism in near-isogenic wheat lines differing for the Yr5 stripe rust resistance locus. Theoretical and Applied Genetics **117:** 401-411.

Das S, Bhat P, Sudhakar C, Ehlers J, Wanamaker S, Roberts P, Cui X, Close T (2008) Detection and validation of single feature polymorphisms in cowpea

(Vigna unguiculata L. Walp) using a soybean genome array. BMC Genomics **9:** 107.

Drenkard E, Richter B, Rozen S, Stutius L, Angell N, Mindrinos M, Cho R, Oefner P, Davis R, Ausubel F (2000) A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. Plant Physiology **124:** 1483-1492.

Greenhall J, Zapala M, Caceres M, Libiger O, Barlow C, Schork N, Lockhart D (2007) Detecting genetic variation in microarray expression data. Genome Research **17:** 1228-1235.

Gupta P, Rustgi S, Mir R (2008) Array-based high-throughput DNA markers for crop improvement. Heredity **101:** 5-18.

Hazen S, Borevitz J, Harmon F, Pruneda-Paz J, Schultz T, Yanovsky M, Liljegren S, Ecker J, Kay S (2005) Rapid array mapping of circadian clock and developmental mutations in Arabidopsis. Plant Physiology **138:** 990-997.

Hazen S, Kay S (2003) Gene arrays are not just for measuring gene expression. Trends in Plant Science **8:** 413-416.

Jansen R, Nap J (2001) Genetical genomics: the added value from segregation. Trends in Genetics **17:** 388-391.

Ritter K, McIntyre C, Godwin I, Jordan D, Chapman S (2007) An assessment of the genetic relationship between sweet and grain sorghums, within Sorghum bicolor ssp. bicolor (L.) Moench, using AFLP markers. Euphytica **157:** 161-176.

Kumar R, Qiu J, Joshi T, Valliyodan B, Xu D, Nguyen H (2007) Single feature polymorphism discovery in rice. PloS One **2:** e284.

Leung EW, Guddat L (2009) Conformational changes in a plant ketol-acid reductoisomerase upon Mg(2+) and NADPH binding as revealed by two crystal structures. Journal of Molecular Biology **389:** 167-182.

Paterson A, Bowers J, Bruggmann Rm, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti A, Chapman J, Feltus F, Gowik U, Grigoriev I, Lyons E, Maher C, Martis M, Narechania A, Otillar R, Penning B, Salamov A, Wang Y, Zhang L, Carpita N, Freeling M, Gingle A, Hash C, Keller B, Klein P, Kresovich S, McCann M, Ming R, Peterson D, Mehboob ur R, Ware D, Westhoff P, Mayer K, Messing J, Rokhsar D (2009) The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M (2008) Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant Journal **53:** 90-101.

Rostoks N, Borevitz J, Hedley P, Russell J, Mudie S, Morris J, Cardle L, Marshall D, Waugh R (2005) Single-feature polymorphism discovery in the barley transcriptome. Genome Biology **6:** R54.

Shiu SH, Borevitz J (2008) The next generation of microarray research: applications in evolutionary and ecological genomics. Heredity **100:** 141-149.

Tsutsumi K, Kagaya Y, Hidaka S, Suzuki J, Tokairin Y, Hirai T, Hu D, Ishikawa K, Ejiri S (1994) Structural analysis of the chloroplastic and cytoplasmic

aldolase-encoding genes implicated the occurrence of multiple loci in rice. Gene **141:** 215-220.

Varshney R, Graner A, Sorrells M (2005) Genomics-assisted breeding for crop improvement. Trends in Plant Science **10:** 621-630.

Werner J, Borevitz J, Warthmann N, Trainer G, Ecker J, Chory J, Weigel D (2005) Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proceedings of the National Academy of Sciences **102:** 2460-2465.

West M, van Leeuwen H, Kozik A, Kliebenstein D, Doerge R, St Clair D, Michelmore R (2006) High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. Genome Research **16:** 787-795.

Xu J-H, Messing J (2008) Organization of the prolamin gene family provides insight into the evolution of the maize genome and gene duplications in grass species. Proceedings of the National Academy of Sciences **105:** 14330-14335.

Zhu T, Salmeron J (2007) High-definition genome profiling for genetic marker discovery. Trends in Plant Science **12:** 196-202.

Chapter 4 Characterization Of The Small RNA Component Of The Transcriptome From Grain And Sweet Sorghum Stems

4.1. Abstract

Sorghum belongs to the tribe of the *Andropogoneae* that includes potential biofuel crops like switchgrass, *Miscanthus* and successful biofuel crops like corn and sugarcane. However, from a genomics point of view sorghum has compared to these other species a simpler genome because it lacks the additional rounds of whole genome duplication events. Therefore, it has become possible to generate a high-quality genome sequence. Furthermore, cultivars exists that rival sugarcane in levels of stem sugar so that a genetic approach can be used to investigate which genes are differentially expressed to achieve high levels of stem sugar. Here, we characterized the small RNA component of the transcriptome from grain and sweet sorghum stems, and from F2 plants derived from their cross that segregated for sugar content and flowering time. We found that variation in miR172 and miR395 expression correlated with flowering time whereas variation in miR169 expression correlated with sugar content in stems. Interestingly, genotypic differences in the ratio of miR395 to miR395* were identified, with miR395* species expressed as abundantly as miR395 in sweet sorghum but not in grain sorghum. Finally, we provided experimental evidence for previously annotated miRNAs detecting the expression of 25 miRNA families from the 27 known at the moment and discovered 9 new miRNAs candidates in the sorghum genome. Sequencing the small RNA component of sorghum stem tissue provided me with experimental evidence for

previously predicted microRNAs in the sorghum genome and microRNAs with a potential role in stem sugar accumulation and flowering time.

4.2. Introduction

Small RNAs (18-25 nt) regulate many developmental and physiological processes in plants through the regulation of gene expression at either the transcriptional or post-transcriptional level (Zamore and Haley, 2005; Vaucheret, 2006; Chuck et al., 2009). They can be subdivided into short-interfering RNAs (siRNAs) and microRNAs (miRNAs) (Bartel, 2004; Zamore and Haley, 2005; Vazquez, 2006).

MicroRNAs are derived from capped and polyadenylated primary (pri)-miRNA transcripts that are transcribed by RNA polymerase II and can form a hairpinloop structure by intramolecular pairing (Bartel, 2004; Lee et al., 2004). Two sequential cleavages mediated by DICER LIKE 1 (DCL1) are required to produce a mature miRNA (Bartel, 2004; Henderson et al., 2006). In the first cleavage, DCL1 cleaves near the base of the hairpin-loop stem of the pri-miRNA to produce a miRNA precursor (pre-miRNA). The second cleavage takes place near the loop of the pre miRNA to produce a miRNA/miRNA* duplex. The mature miRNA is then loaded into the RNA-induced silencing complex (RISC) and can guide the sequence-specific cleavage or translational inhibition of target mRNAs (Bartel, 2004; Henderson et al., 2006; Vaucheret, 2006; Filipowicz et al., 2008), as well as gene silencing through DNA methylation (Khraiwesh et al., 2010; Wu et al., 2010), whereas the non-incorporated miRNA* strand is usually degraded.

Through the use of next-generation sequencing, the small RNA component of the *Arabidopsis* and rice transcriptomes has been well characterized, more than in any other plant species (Lu et al., 2005; Nobuta et al., 2007). This is reflected in the miRBase database (http://www.mirbase.org, release 16: September 2010), where 213 miRNAs are described for *Arabidopsis* whereas 462 miRNAs are described for rice. Besides rice, the identification of miRNAs through deep sequencing in other grasses including maize, wheat, and *Brachypodium* have been described (Nobuta et al., 2008; Wang et al., 2009; Wei et al., 2009). The identification of rice, maize and wheat miRNAs from different tissues, developmental stages and stress-treatments (Sunkar et al., 2005; Nobuta et al., 2007; Heisel et al., 2008; Nobuta et al., 2008; Sunkar et al., 2008; Zhu et al., 2008; Wei et al., 2009; Xue et al., 2009), provides an opportunity to understand how miRNAs regulate the expression of genes influencing traits of agronomic importance. Currently, a trait of particular relevance for biofuel production is that of sugar accumulation in the stem of sorghum [Sorghum bicolor (L.) Moench] and sugarcane (*Saccharum spp.*), two closely related C4 grasses that diverged from each other about 8-9 million years ago (Jannoo et al., 2007).

In both species, sucrose is the main type of sugar and accumulates in the parenchyma tissue of the juicy stems (Glasziou and Gayler, 1972; Hoffman-Thoma et al., 1996). High sucrose content is a highly desirable trait since the accumulated sugar can be fermented to produce bioethanol as a source of renewable energy (Goldemberg, 2007). Although sugarcane has been extensively used as a source of biofuel, its use as a model system to understand the genetics of sugar accumulation

is hampered by its complex genome, with several cultivars differing greatly in their ploidy levels (Grivet and Arruda, 2002). Sorghum instead, is a diploid species and its genome has been recently sequenced (Paterson et al., 2009). In addition, the intra-species variation for sugar content is much more pronounced in sorghum than in sugarcane (Ritter et al., 2007), with sorghum cultivars known as sweet sorghums accumulating high levels of sugars relative to grain sorghums (Murray et al., 2008). This makes sorghum a more suitable system to study the genetic basis of sugar accumulation. Still, the gene repertoire involved in sugar accumulation is not well characterized in sorghum due to the low heritability of the trait and its quantitative inheritance. In addition, previous reports have suggested the existence of trade-offs between sugar content and other plant traits such as flowering time (Murray et al., 2008; Ritter et al., 2008).

We also observed that sugar accumulation (measured as Brix degree and referred herein as Brix) in the stem of grain sorghum BTx623 and sweet sorghum Rio cultivars differed at the time of flowering. Interestingly, 80% of the differentially expressed genes in stem tissue between the two cultivars had orthologous counterparts in syntenic positions in rice (Calviño et al., 2008; Calviño et al., 2009). This suggested that the ability of sorghum to accumulate soluble sugars relative to rice could not be explained by differences in their gene content but rather due to gene regulation at either the transcriptional or post-transcriptional level. To address the latter possibility, we characterized the small RNA portion of transcriptomes derived from stem tissues of grain and sweet sorghum in order to investigate the microRNAmediated regulation of genes involved in sugar

accumulation and flowering time. Using the SOLiD next generation sequencing system, we sequenced with an unprecedented depth small RNAs libraries from BTx623 and Rio, and from a pool of selected F2 plants derived from their cross that differed in sugar content and flowering time. We also reasoned that plant stems would provide us with a representative tissue to experimentally validate the previously predicted miRNAs of the sorghum genome (Paterson et al., 2009). Indeed, we not only detected the expression of 25 miRNA families from the 27 predicted families in the sorghum genome but also discovered 9 new miRNA candidates. Furthermore, we could correlate genotypic variation of miRNA expression with the sugar and flowering phenotypes. In addition, we found that the size distribution of small RNAs in sorghum stems was quite heterogeneous, characterized by RNAs with at least 25 nt in length that were mainly derived from ribosomal and transfer RNAs not annotated in the sorghum genome.

4.3. Results

4.3.1. Deep sequencing of small RNAs from grain and sweet sorghum stems

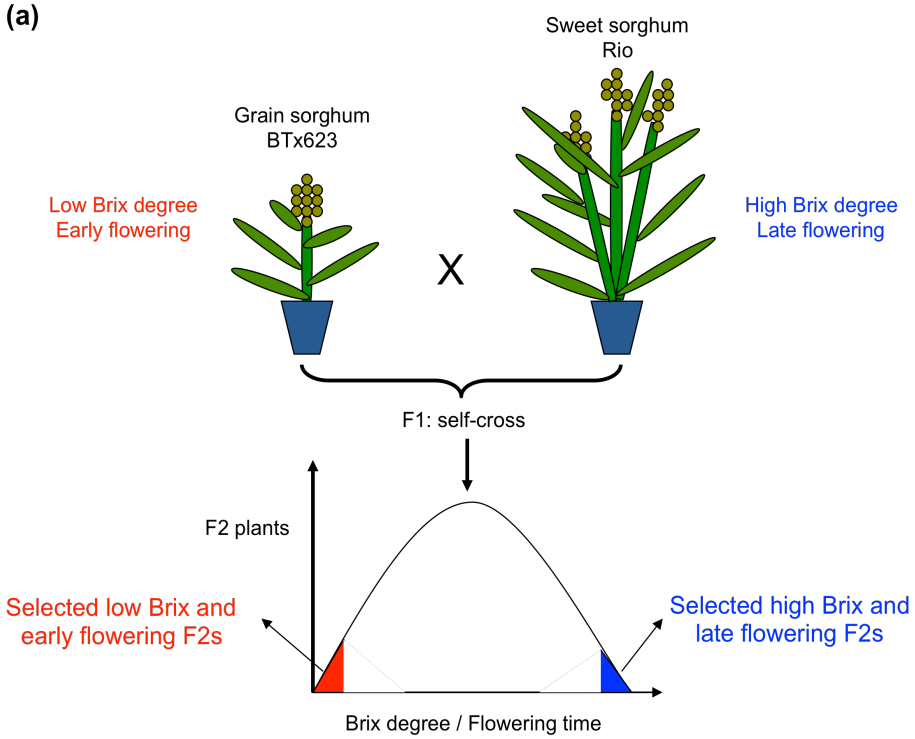
We constructed five small RNAs libraries from sorghum stem tissue at the time of flowering and sequenced them using the SOLiD platform. The libraries comprised samples from BTx623, Rio, low Brix and early flowering F2 plants (LB/EF F2s), high Brix and late flowering F2 plants (HB/LF F2s), and a "mixed library" (Mix), where small RNAs from the previous four libraries were mixed in equal proportions (Figure 4.1).

We obtained a total of 38,336,769 sequence reads, from which 23,008,945 (60%) matched perfectly to the BTx623 reference genome (Table 4.1). The reads with perfect matches that derived from repeats constituted 74 to 77% of the total reads depending on the library (Figure 4.2a). The non-redundant set of reads comprised 2,539,403 sequences, and the reads that were sequenced only once (termed here "singlets") comprised 2,167,946 sequences, corresponding only to 9% of the perfect matches (Table 4.1), suggesting that my sequencing reached a high level of saturation. If we define a cluster as two or more reads with identical sequences, the number of clusters found ranged from 20,056 in the BTx623 library to 164,623 in the HB/LF F2s library (Table 4.1).

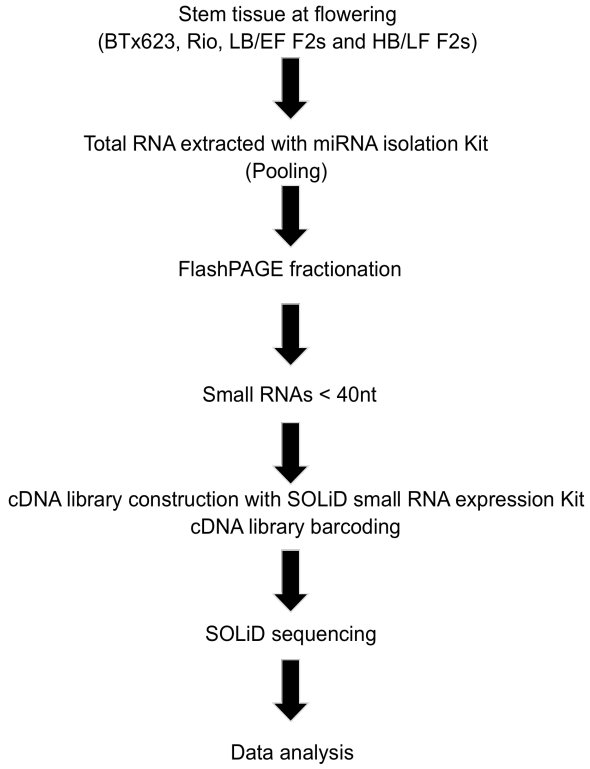
Library	# raw sequences	# perfect matches	%	# singlets	%	# clusters	Non-redundant set	%
Mix	4,023,513	2,547,108	63	276,044	11	35,083	311,127	8
BTx623	2,115,266	1,348,361	64	169,063	12	20,056	189,119	9
Rio	3,173,601	2,180,988	69	234,276	11	31,563	265,839	8
LB/EF F2s	11,974,953	7,472,940	62	653,279	9	120,132	773,411	6
HB/LF F2s	17,049,436	9,459,548	55	835,284	9	164,623	999,907	6
Total	**38,336,769**	**23,008,945**	**60**	**2,167,946**	**9**	**371,457**	**2,539,403**	**8**

Table 4.1. Deep sequencing statistics of stem-derived small RNAs

(a)


Sweet sorghum
Rio
Grain sorghum
BTx623
Low Brix degree
Early flowering
X
High Brix degree
Late flowering
F1: self-cross
F2 plants
Selected low Brix and
early flowering F2s
Selected high Brix and
late flowering F2s
Brix degree / Flowering time

(b)


Stem tissue at flowering
(BTx623, Rio, LB/EF F2s and HB/LF F2s)
Total RNA extracted with miRNA isolation Kit
(Pooling)
FlashPAGE fractionation
Small RNAs < 40nt
cDNA library construction with SOLiD small RNA expression Kit
cDNA library barcoding
SOLiD sequencing
Data analysis

(c)

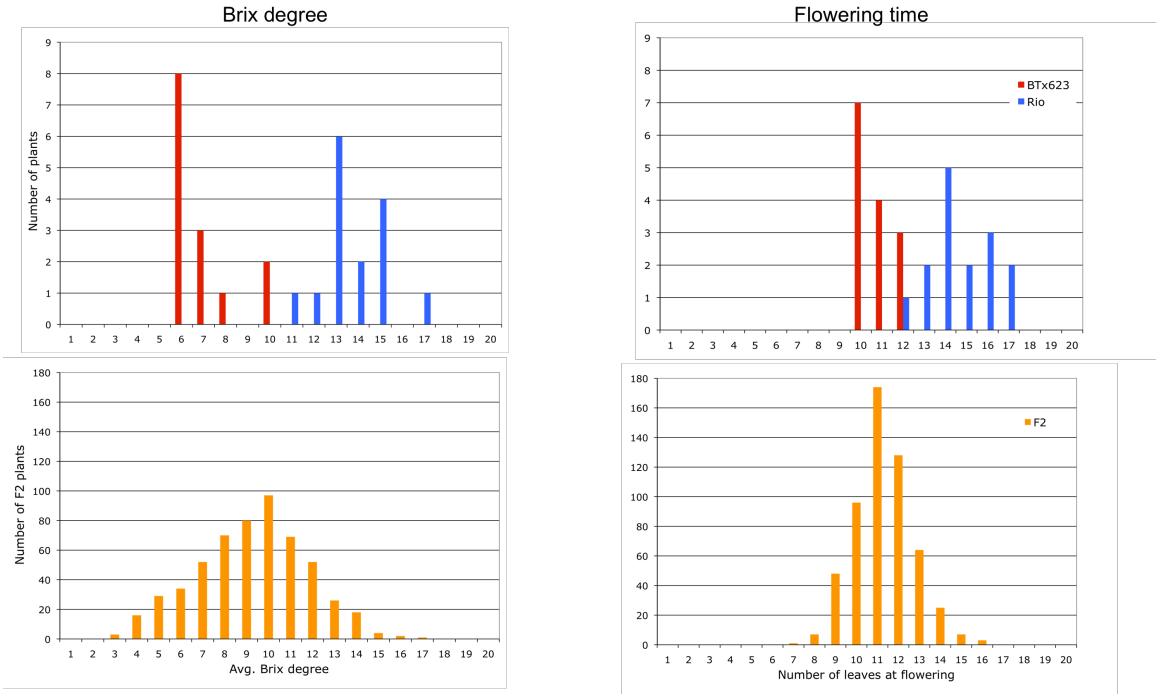
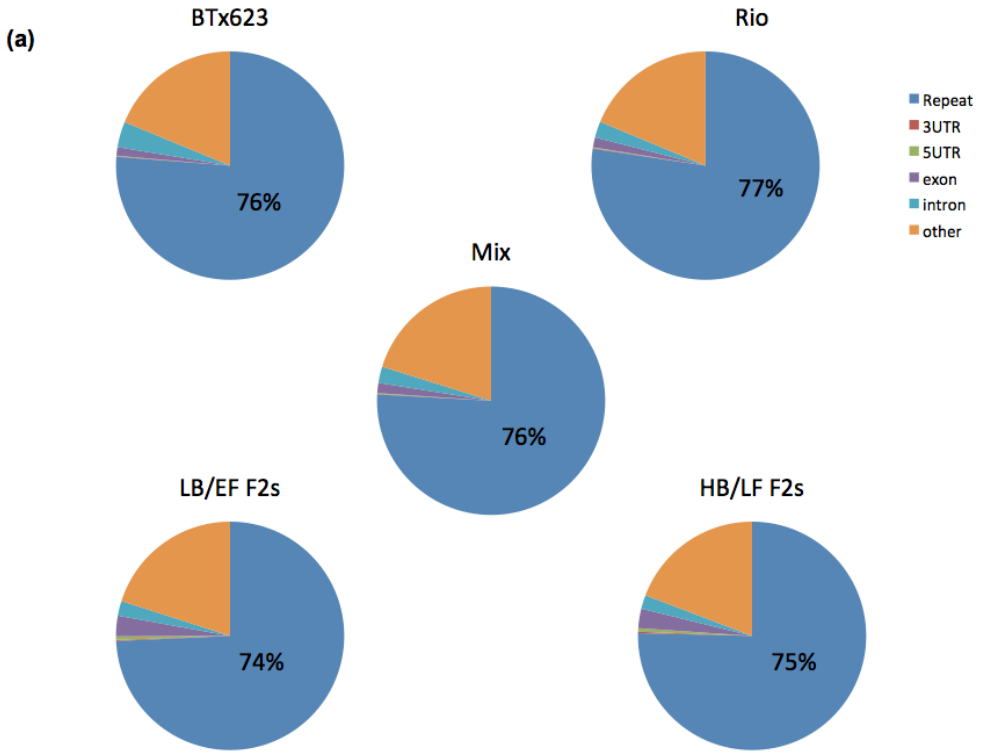
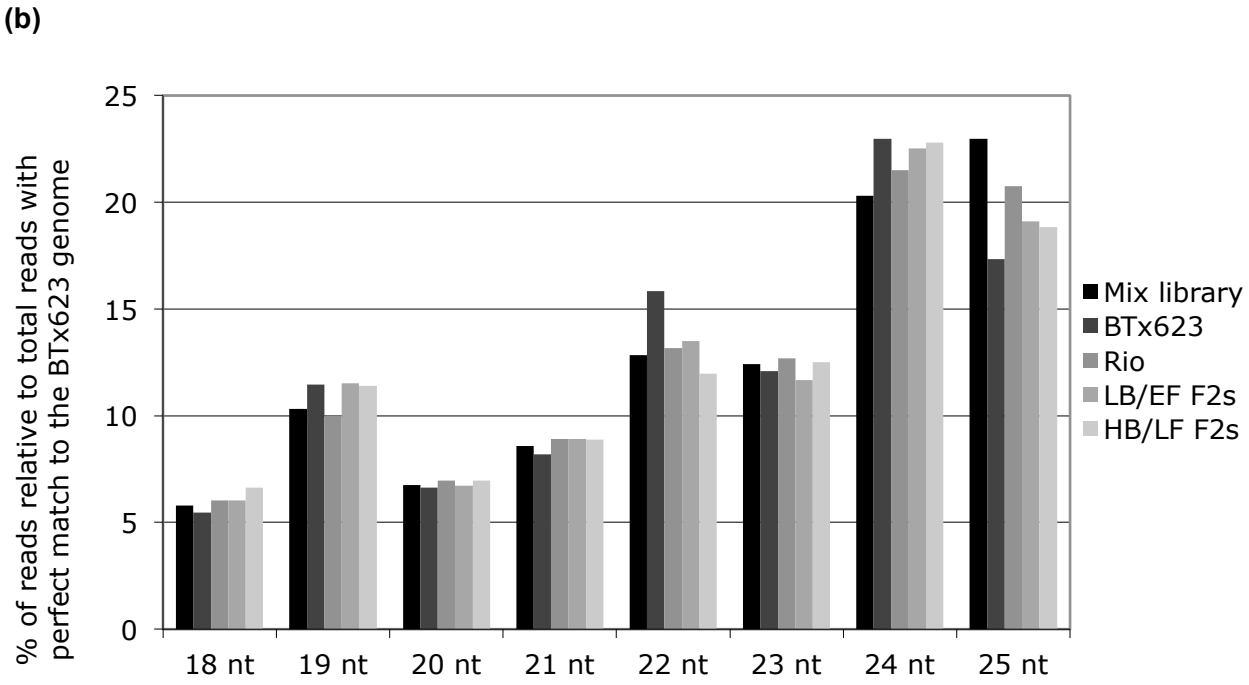


Figure 4.1. Selection of sorghum plants and construction of stem-derived small RNA libraries for deep sequencing. (a) Grain sorghum BTx623 with low Brix and early flowering phenotype, was crossed with sweet sorghum Rio with high Brix and late flowering phenotype and an F2 population was created. A total of 553 F2 plants were phenotyped for flowering time (measured as the total number of leaves at flowering) and Brix degree. Using a bulked segregant analysis (BSA) approach, I selected an equal number of F2 plants with low Brix and early flowering (LB/EF) and with high Brix and late flowering (HB/LF) phenotype, respectively. **(b)** A flow chart describing the procedure for small RNA library construction and sequencing. **(c)** Histograms displaying the Brix degree and flowering time data obtained from plants grown in the field. I selected 11 LB/EF F2s displaying Brix

degree ≤ 5 and number of leaves ≤ 9, whereas the 11 HB/LF F2s selected displayed a Brix degree ≥ 13 and number of leaves ≥ 14.


(a)
BTx623
76%
Rio
77%
Mix
76%
LB/EF F2s
74%
HB/LF F2s
75%
Repeat
3UTR
5UTR
exon
intron
other


(b)
% of reads relative to total reads with perfect match to the BTx623 genome
25
20
15
10
5
0
18 nt
19 nt
20 nt
21 nt
22 nt
23 nt
24 nt
25 nt
Mix library
BTx623
Rio
LB/EF F2s
HB/LF F2s

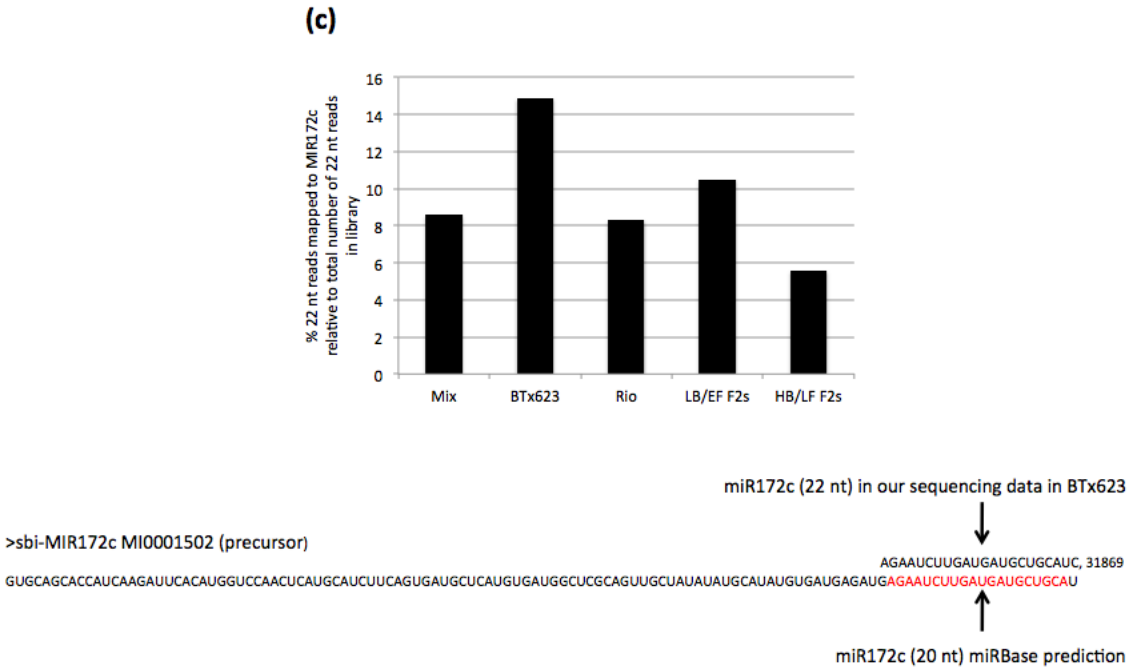


Figure 4.2. Diversity in the small RNA content of sorghum stems. **(a)** Mapping of small RNAs (18-25 nt) with perfect match to different elements of the BTx623 reference genome with the term "other" representing intergenic regions. **(b)** Frequency and size distribution of small RNAs reads. **(c)** A high proportion of 22 nt reads in each library are derived from miR172c locus. The small RNA reads derived from miR172c in sorghum stem tissue are 22 nt in length in contrast to the previously predicted length of 20 nt.

4.3.2. Diversity in the small RNA content of sorghum stems

The frequency and size distribution of small RNAs from sorghum stems revealed two interesting aspects: a peak of 25 nt small RNAs with similar abundance as the 24 nt class, and a second peak of small RNAs with 22 nt that were more abundant than the 20 and 21 nt classes, respectively (Figure 4.2b). This finding contrasted with the size distribution of small RNAs described for several monocot species (including small RNAs from sorghum inflorescence), in which the most abundant small RNAs were 21 and 24 nt in length, with maize being the exception, showing a larger 22 nt peak relative to the 21 nt peak (Nobuta et al., 2008). This led to the hypothesis that the 22 nt class of small RNAs are specific to maize (Nobuta et al., 2008). However, we have shown here that a 22 nt peak is also present in sorghum stem tissue. Furthermore, we found that a high proportion of the 22 nt reads were derived from miR172c, accounting for approximately 15% of all the 22 nt reads in the BTx623 library (Figure 4.2c). Our results differed from the predicted length of 20 nt for miR172c annotated in the miRBase database. Interestingly, MIR172c is located within the third intron of the Sb04g037375 gene.

The finding of small RNAs of 25 nt in length with such high abundance was unexpected. This prompted us to investigate whether they could be derived from ribosomal and/or transfer RNA genes that had not yet been annotated in the sorghum genome. Furthermore, since the sequencing read length of the SOLiD system at the time of our experiment was limited to a maximum of 25 nucleotides, it was possible that these RNAs were longer. In order to address this question, we analyzed several loci in the genome that accumulated more than thousand reads

(defined as 25 nt hotspots) and found indeed that they were derived from non-annotated rRNA and tRNA genes (Table 4.2).

In summary, we showed that the small RNA component from the stem transcriptome of sorghum is characterized by small RNAs of 22 nt in length that are mainly derived from miR172c, and by a size class of RNAs with at least 25 nt in length that are predominantly derived from rRNAs and tRNAs genes that had not been annotated in the sorghum genome.

Position	Length of hotspot (bp)	N⁰ of 25 nt reads	Annotation (Phytozome)	BLAST nucleotide collection (nr/nt) hit	E-value	Identity
Library: Mix						
Ch3: 72,749,847..72,749,881	35	9381	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	5E-10	100%
Ch1: 31,857,437..31,857,496	60	5652	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	2E-22	100%
Ch5: 36,051,996..36,052,067	72	4689	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	7E-29	100%
Ch10: 657,846..657,883	38	3106	Intergenic	A. thaliana At5g59055 tRNA	2E-09	97%
Ch5: 35,985,593..35,985,714	122	2882	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	4E-61	100%
Ch5: 35,931,714..35,931,863	150	2369	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	1E-77	100%
Ch3: 59,743,725..59,743,785	61	1956	Intergenic	A. thaliana At5g40545 tRNA	1E-15	93%
Ch5: 35,976,201..35,976,253	53	1691	Intergenic	Setaria italica genes for 25S rRNA, IGS and 17S rRNA	5E-18	98%
Ch8: 47,608,635..47,608,659	25	1352	Intergenic	A. thaliana At4g34975 tRNA	2E-04	100%
Library: BTx623						
Ch3: 72,749,848..72,749,881	34	3321	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	2E-09	100%
Ch5: 36,052,031..36,052,067	37	3111	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	4E-11	100%
Ch5: 35,931,716..35,931,758	43	2709	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	1E-14	100%
Ch5: 35,985,655..35,985,705	51	2287	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	2E-17	100%
Ch1: 31,863,286..31,863,315	30	1231	Intergenic	Oryza brachyantha 26S-18S rRNA intergenic spacer	3E-07	100%

Ch5: 35,997,943..35,997,972	30	1227	Intergenic	Oryza brachyantha 26S-18S rRNA intergenic spacer	3E-07	100%
Ch5: 35,976,205..35,976,252	48	1117	Intergenic	Avena sativa rDNA spacer	7E-07	100%
Library: Rio						
Ch3: 72,749,847..72,749,881	35	6727	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	5E-10	100%
Ch5: 36,052,031..36,052,067	37	6467	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	4E-11	100%
Ch5: 35,931,716..35,931,758	43	5622	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	1E-14	100%
Ch5: 35,985,655..35,985,713	59	4104	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	8E-22	100%
Ch5: 35,976,203..35,976,252	50	1583	Intergenic	Avena sativa rDNA spacer	7E-17	100%
Ch4: 50,861,835..50,861,859	25	1362	Intergenic	A. thaliana At5g46595 tRNA	2E-04	100%
Ch5: 35,981,272..35,981,333	62	1282	Intergenic	Setaria italica genes for 25S rRNA, IGS and 17S rRNA	9E-22	98%
Library: LB/EF F2s						
Ch3: 72,749,845..72,749,881	37	23470	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	4E-11	100%
Ch1: 31,857,435..31,857,497	63	14104	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	5E-24	100%
Ch5: 36,051,996..36,052,068	73	12057	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	2E-29	100%
Ch5: 35,985,593..35,985,716	124	7413	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	2E-57	100%
Ch4: 50,861,834..50,861,859	26	6443	Intergenic	A. thaliana At5g46595 tRNA	6E-05	100%
Ch5: 35,931,708..35,931,865	158	5861	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	4E-75	100%
Ch8: 47,608,634..47,608,659	26	3034	Intergenic	A. thaliana At4g34975 tRNA	6E-05	100%
Ch5: 35,937,803..35,937,851	49	3007	Intergenic	Avena sativa rDNA spacer	4E-18	100%
Ch3: 59,743,724..59,743,785	62	2116	Intergenic	A. thaliana At5g40545 tRNA	3E-17	93%
Library: HB/LF F2s						
Ch3: 72,749,845..72,749,881	37	22694	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	4E-11	100%
Ch1: 31,857,433..31,857,497	65	13314	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	4E-25	100%
Ch5: 36,051,996..36,052,068	73	11712	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	2E-29	100%
Ch4: 50,861,834..50,861,859	26	8290	Intergenic	A. thaliana At5g46595 tRNA	6E-05	100%
Ch5: 35,985,593..35,985,718	126	7099	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	1E-58	100%
Ch5: 35,931,708..35,931,863	156	5796	Intergenic	S. bicolor strain b2 internal transcribed spacer 1 5.8S rRNA	4E-75	100%
Ch8: 47,608,634..47,608,659	26	3415	Intergenic	A. thaliana At4g34975 tRNA	6E-05	100%

Ch5: 35,976,201..35,976,260	60	2976	Intergenic	Setaria italica genes for 25S rRNA, IGS and 17S rRNA	5E-20	98%
Ch3: 59,743,724..59,743,785	62	2372	Intergenic	A. thaliana At5g40545 tRNA	3E-17	93%

Table 4.2. 25 nt hotspots in the sorghum genome.

4.3.3. Genotypic variation in the expression of known miRNAs between grain and sweet sorghum correlated with sugar content and flowering time in the F2 population

The sequencing consortium of the sorghum genome identified 149 predicted miRNAs belonging to 27 miRNA families (Paterson et al., 2009), and we could detect the expression of miRNA members from 25 families based on the following criteria: a miRNA family was considered expressed only if its sequencing reads were detected in at least three libraries and with a frequency of 10 reads or more for the sum of the five libraries. A list with the reads count for each known miRNA family is provided in Table 4.3.

The most abundantly expressed miRNA family was miR172 (Figure 4.3a), comprising almost 6% of the total reads with perfect match to the BTx623 genome. The rest of the known miRNAs had abundances below 0.5% (Figure 4.3b). When the ratio of miRNA abundances between the BTx623 and Rio libraries was compared to the ratio between the LB/EF F2s and HB/LF F2s libraries, we could identify miRNA families whose expression differences between the parents were inherited in the F2 plants (Figure 4.3c). Considering a cutoff level of two-fold change in miRNA expression, we found that miR169 and miR172 were expressed higher in BTx623 relative to Rio, and higher in LB/EF F2s compared to HB/LF F2s. This means that

high expression of these miRNAs in BTx623 correlated with low Brix and early flowering in the F2 plants selected, and the opposite was true for miR395 (Figure 4.3c). Although the expression difference of miR160, miR164 and miR319 between BTx623 and Rio was inherited in the F2, and thus of interest for further analysis, it was less than two fold; so we decided to focus on miR169, miR172 and miR395 instead. The observation that high expression of miR172 correlated with early flowering was consistent with the reported role of this miRNA in the promotion of flowering (Lauter et al., 2005; Chuck et al., 2007; Mathieu et al., 2009; Wu et al., 2009; Zhu et al., 2009).

Although miR169 and miR395 have known roles in drought stress and sulphur starvation, respectively (Li et al., 2008; Kawashima et al., 2009), our data suggested a possible function for these miRNAs in sugar accumulation and flowering time. Because the pool of F2 plants used for library construction were selected based on both phenotypes, it was not possible to assign the expression inheritance pattern of both miRNAs to either sugar accumulation or flowering time alone. For this reason, additional plants from the same F2 population differing in sugar content but with similar flowering time were selected and the expression of a representative member from each miRNA family, miR169d and miR395f respectively, was quantified using the TaqMan assay. We found that high expression of miR169d in BTx623 correlated with low Brix (Figure 4.3d). This suggested that high expression levels of miR169 might lead to a reduction in stem sugar content regardless of flowering time. Surprisingly, high expression of miR395f in Rio relative to that in BTx623 did not correlate with sugar content in F2 plants (Figure 4.3e). This might

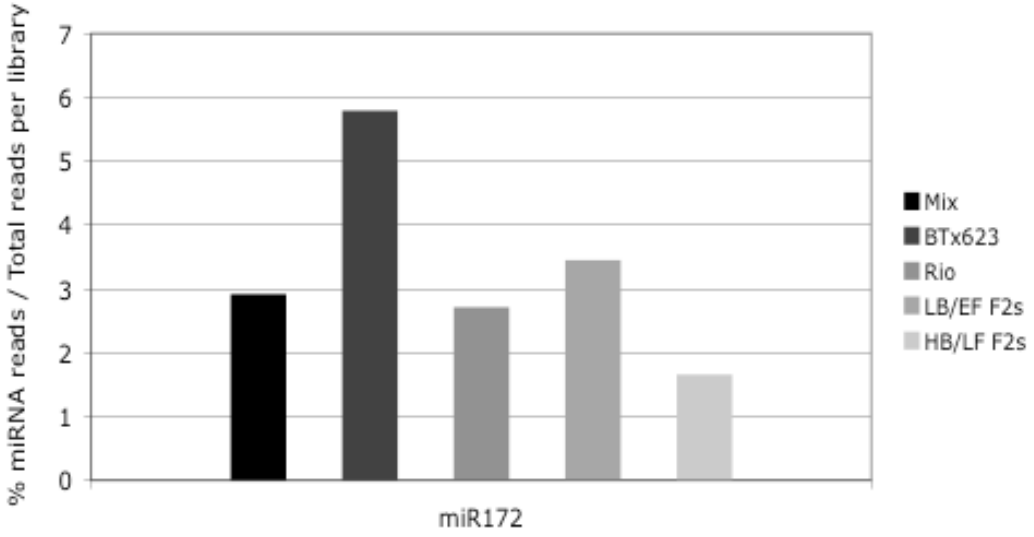
indicate that high expression of miR395 would be required for flowering regardless of sugar content in the stem. Consistent with the role of miR172 in flowering, we did not observe any difference in the expression of miR172a in F2 plants with the same flowering time but different Brix (Figure 4.3f).

In summary, high expression of miR172 in BTx623 correlated with early flowering in the F2, whereas the opposite was true for miR395, high expression of this miRNA in Rio correlated with late flowering in the F2 plants selected. Regarding sugar content in the stem, high expression of miR169 in BTx623 correlated with low Brix in the F2 plants selected.

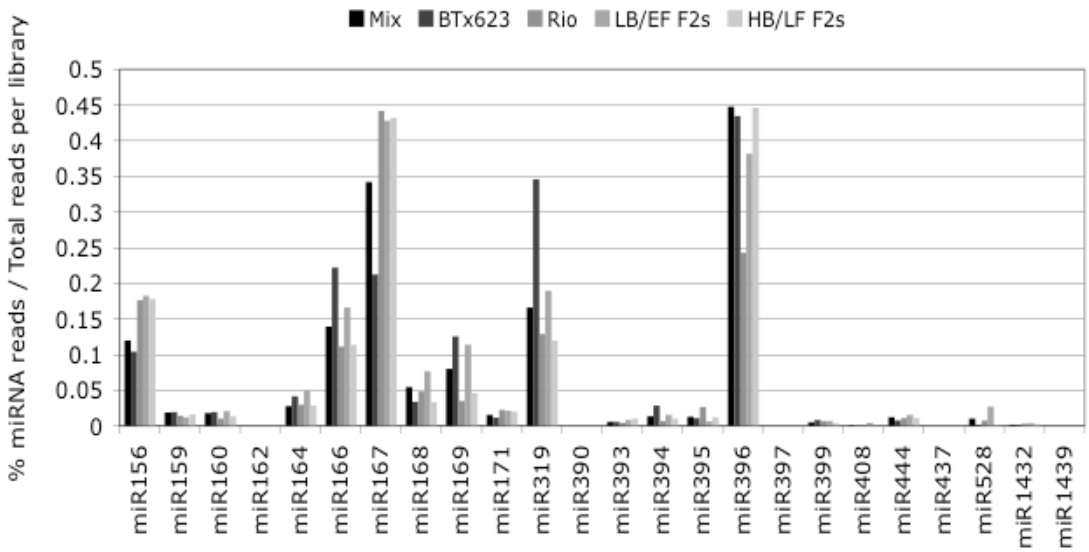
Count of mapped reads to miRNA family for each library					
miRNA family	**Mix**	**BTx623**	**Rio**	**LB/EF F2s**	**HB/LF F2s**
miR156	3058	1410	3858	13657	16807
miR159	482	267	306	916	1544
miR160	468	268	234	1563	1282
miR162	2	1	4	11	10
miR164	714	427	656	3687	2720
miR166	3559	2994	2429	12434	10781
miR167	8725	2867	9638	31997	40856
miR168	1397	459	1047	5736	3115
miR169	2044	1693	772	8503	4287
miR171	398	154	504	1590	1938
miR172	74323	78190	59332	257767	156871
miR319	4232	4665	2821	14167	11341
miR90	3	1	0	6	5
miR393	154	80	106	622	962
miR394	346	389	156	1148	1008
miR395	333	153	583	465	1181
miR395 reads	165	130	305	293	619
miR395* reads	168	23	278	172	562
miR395/miR395*	0.982142857	5.652	1.097	1.703	1.101
miR396	11415	5862	5297	28559	42214
miR397	1	0	2	8	6
miR399	129	112	156	557	393
miR408	41	5	43	364	75
miR444	313	105	238	1154	1062
miR437	1	1	0	6	5
miR528	259	26	171	2027	151
miR1432	48	26	68	280	243
miR1439	2	0	3	12	12

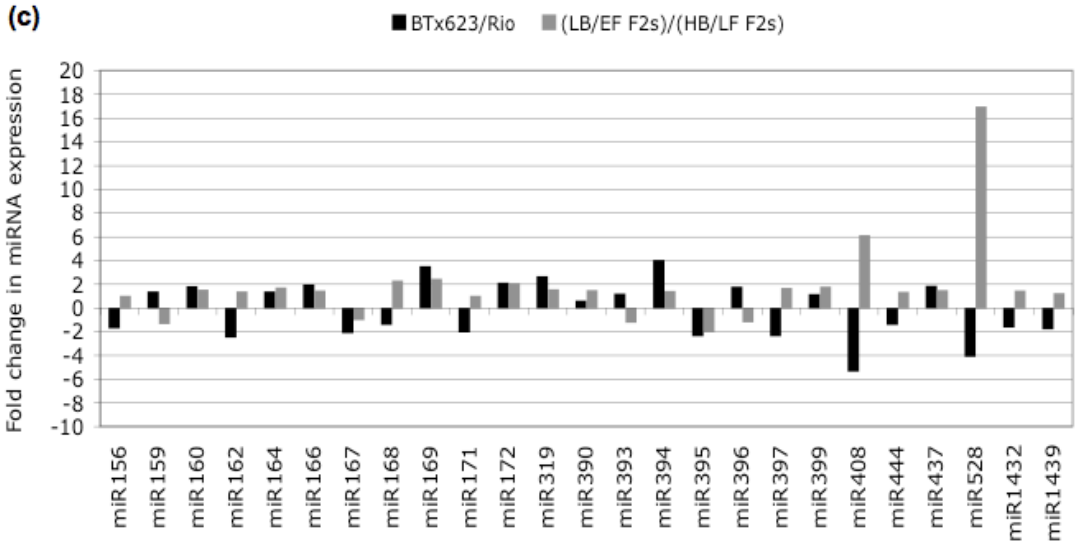
Table 4.3. Frequency counts of small RNA reads for known microRNA families.

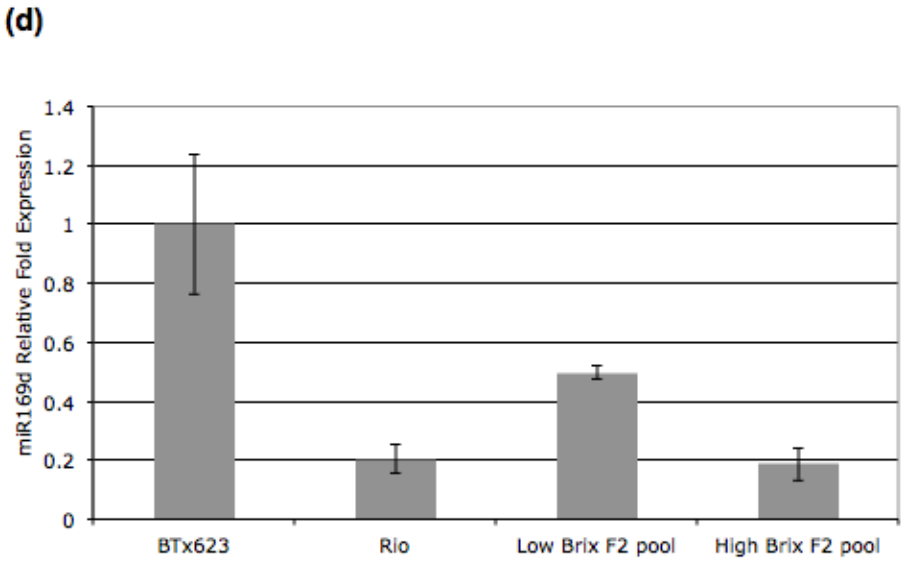
(a)


% miRNA reads / Total reads per library
7
6
5
4
3
2
1
0
miR172
Mix
BTx623
Rio
LB/EF F2s
HB/LF F2s

(b)


Mix
BTx623
Rio
LB/EF F2s
HB/LF F2s
% miRNA reads / Total reads per library
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
miR156
miR159
miR160
miR162
miR164
miR166
miR167
miR168
miR169
miR171
miR319
miR390
miR393
miR394
miR395
miR396
miR397
miR399
miR408
miR444
miR437
miR528
miR1432
miR1439


(c)
BTx623/Rio
(LB/EF F2s)/(HB/LF F2s)
Fold change in miRNA expression
20
18
16
14
12
10
8
6
4
2
0
-2
-4
-6
-8
-10
miR156
miR159
miR160
miR162
miR164
miR166
miR167
miR168
miR169
miR171
miR172
miR319
miR390
miR393
miR394
miR395
miR396
miR397
miR399
miR408
miR444
miR437
miR528
miR1432
miR1439


(d)
miR169d Relative Fold Expression
1.4
1.2
1
0.8
0.6
0.4
0.2
0
BTx623
Rio
Low Brix F2 pool
High Brix F2 pool

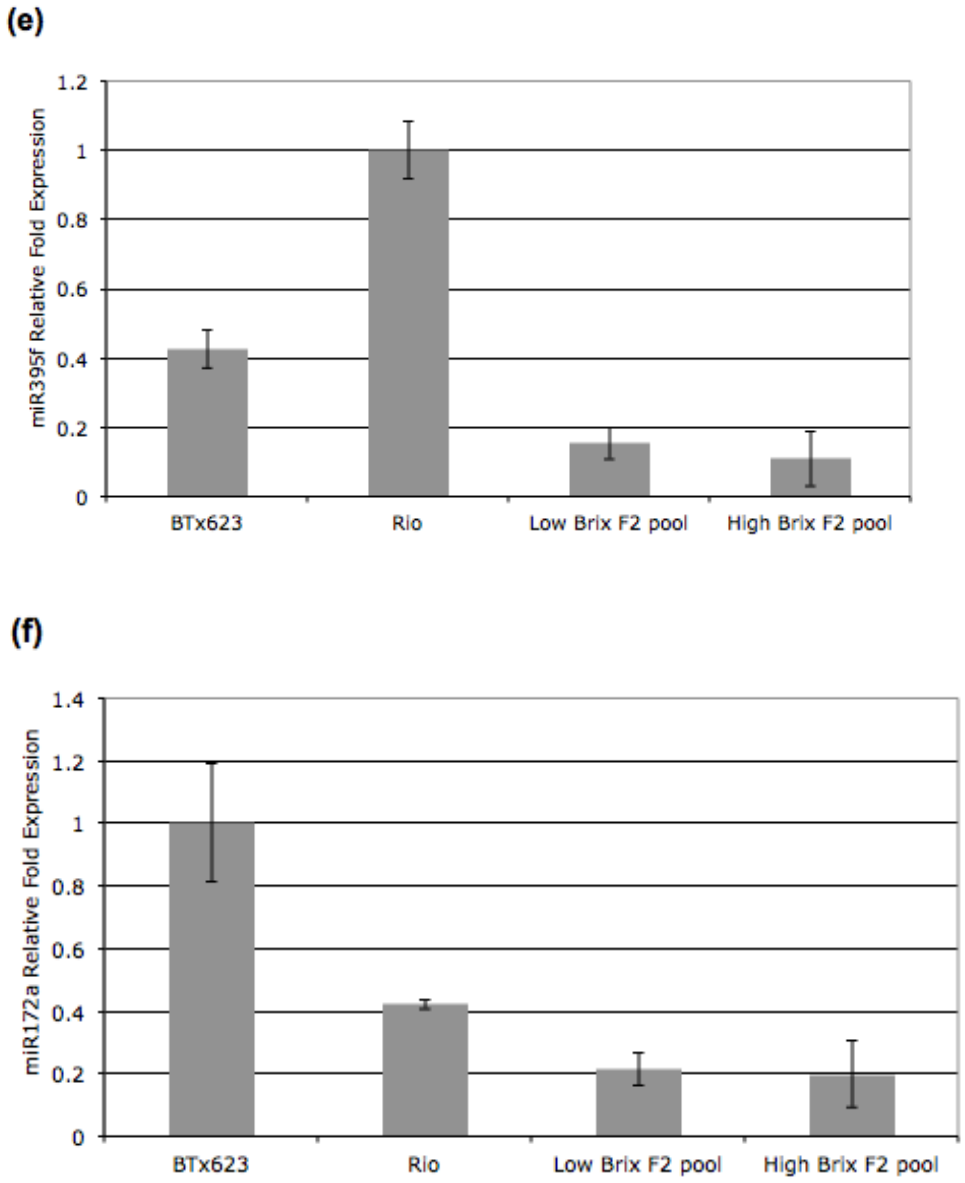


Figure 4.3. Genotypic variation in miRNA expression. **(a)** The miR172 was the most abundantly expressed miRNA in sorghum stems. **(b)** The rest of the known miRNAs were expressed at very low abundance (less that 0.5% of the total reads in the library) in stem tissue. **(c)** The miRNA abundances were used to calculate their relative fold change in expression between BTx623 and Rio, and between the LB/EF

F2s and HB/LF F2s libraries, respectively. Positive values in the y-axis of the graph denote fold changes in miRNA expression that are higher in BTx623 relative to Rio and higher in LB/EF F2s relative to HB/LF F2s libraries, respectively; the opposite is true for negative values. The expression of miR169 and miR172 was at least twice as high in BTx623 relative to that in Rio and this difference was inherited in the F2. The opposite was true for miR395 expression. **(d-f)** Quantification of miRNA expression through Taqman Assay in pools of 10 F2 plants each with similar flowering time (10-11 leaves) but different sugar content (Brix 3-5 vs Brix 13-16), respectively. **(d)** High expression of miR169d in BTx623 relative to Rio correlates with low Brix in the F2 independently of flowering time. **(e-f)** F2 plants with similar flowering time display no differences in miR395f and miR172a expression regardless Brix degree.

4.3.4. Genotypic variation in the miR395/miR395* ratio

We detected the expression of the miRNA* for all MIR395 gene copies and this was more evident in Rio compared to BTx623, and in some instances the abundance of miR395* was even higher than that of miR395 such as the case of miR395l* for instance (Figure 4.4a). Indeed, when the miR395/miR395* ratio was calculated for each library, we found that miR395 reads were approximately 6 times more abundant than miR395* reads in the BTx623 library (Table 4.3). By contrast, the abundance of miR395 relative to miR395* was in equal proportions in the Rio library. My data highlighted a genotypic difference in the ratio between miR395 and miR395*, with a switch in strand abundance from BTx623 to Rio (Figure 4.4b).

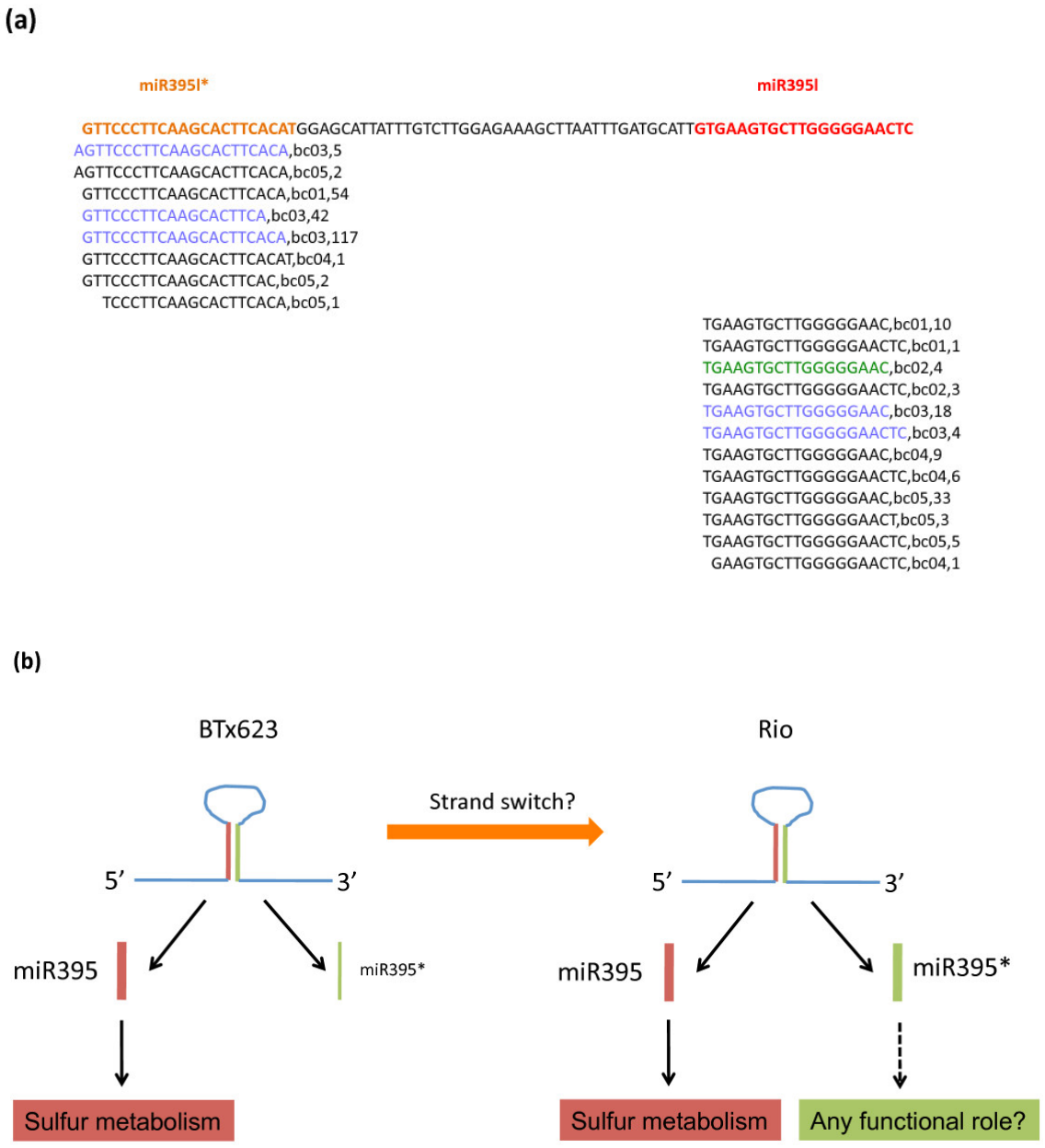


Figure 4.4. miR395* is highly abundant in Rio. **(a)** Small RNA reads derived from MIR395l are depicted. The miR395l strand sequence is shown in red whereas the miR395l* strand sequence is in orange color. In green and blue color are small RNA reads sequenced from BTx623 and Rio libraries respectively. The designation next to the small RNA reads refer to the library (bc01: Mix; bc02: BTx623; bc03: Rio;

bc04: LB/EF F2s and bc05: HB/LF F2s), followed by the number of times the small RNA read was sequenced. In the BTx623 library, only reads derived from miR395l were detected whereas in the Rio library, most of the reads where derived from miR395l* instead. **(b)** Model depicting the genotypic variation in miR395/miR395* ratio where in Rio a switch towards miR395* strand production has occurred relative to BTx623. Based on miR395* high abundance in Rio, we postulate here the hypothesis that miR395* species could have a functional role in the regulation biological processes other than the sulfur metabolism previously described for miR395.

4.3.5. The *FRL2* and *RR3* genes are novel targets of miR172

Although my data might suggest a possible function of miR169 in sugar content and miR395 in flowering time, we could not detect any predicted target related to carbohydrate metabolism and flowering time respectively (Table 4.4 and Figure 4.5). Thus, the expression of miR169 and miR395 target genes, and their correlation with Brix and flowering phenotypes remains to be elucidated. Regarding the miR172-predicted targets, we detected cleavage products for the genes *INDETERMINATE SPIKELET 1* (*IDS1*) and an AP2 transcription factor (Table 4.4; Figure 4.5; and Figure 4.6). Furthermore, when the expression of these two miR172 target genes was tested, we found that they were expressed higher in Rio compared with BTx623 as expected. However, we could not find a correlation between their expression levels with the flowering phenotype in the F2 pools of plants selected (data not shown).

A *FRIGIDA-like 2* (*FRL2*) and a *TYPE A RESPONSE REGULATOR 3* (*RR3*) were predicted as new targets of miR172 with the cleavage product of *FRL2* experimentally validated in this study (Figure 4.6). The *FRIGIDA*-related genes are a major determinant of natural variation in the winter-annual habit between *Arabidopsis* accessions (Michaels et al., 2004; Schlappi, 2006), whereas the *TYPE A RESPONSE REGULATOR 3* (*ARR3*) has a function in the circadian clock (Salome et al., 2006). Although sorghum is a crop from semi-arid regions (Paterson et al., 2009), the miR172-mediated posttranscriptional regulation of *FRL2* might have a role in the adaptation of sorghum to temperate climates. Consistent with this, a role of miR172 in the regulation of flowering time by ambient temperature in *Arabidopsis* has been recently described (Lee et al., 2010).

miRNA	TARGET GENE	GENE FUNCTION
sbi-miR169ab	Sb09g008100	Similar to Zea mays aminoacid transporter LHT1
sbi-miR169acdi	Sb08g021910	CCAAT-binding transcription factor subunit B
sbi-miR169cdi	Sb10g002400	Glycine-rich protein like
sbi-miR169cd	Sb05g026273	GRAS family transcription factor
sbi-miR169bcdefgh	Sb01g045500	CCAAT-binding transcription factor subunit B
sbi-miR169efghi	Sb01g011220	CCAAT-binding transcription factor subunit B
sbi-miR169i	Sb02g003070	TCP family transcription factor
sbi-miR172abcde	Sb01g003400	Indeterminate spikelet 1
	Sb02g007000	Indeterminate spikelet 1
	Sb06g030670	APETALA 2 transcription factor
	Sb09g002080	APETALA 2 transcription factor
sbi-miR172abcd	Sb10g025053	Glossy 15
sbi-miR172e	Sb01g044240	FRIGIDA-like protein 2
	Sb04g038320	Type A response regulator 3
sbi-miR395abcdefgh	Sb01g044100	Sulfate transporter
	Sb01g008450	ATP sulfurylase

Table 4.4. Predicted targets of miR169, miR172 and miR395. In red are miRNA-mediated cleavage of targets genes that were experimentally validated.

```
sbi-miR169cd
3' AUCCGUUCAGUAGGAACCGAU 5'
   :::::::: : : ::::::::
5' UAGGCAAGGCCUACUUGGCUA 3'
Sb10g002400 similar to Glycine-rich protein-like

sbi-miR169cd
3' AUC-CGUUCAGUAGGAACCGAU 5'
   :::-:::::::.::::::: ::
5' UAGAGCAAGUCGUCCUUGGAUA 3'
Sb05g026273 weakly similar to GRAS family transcription factor,
putative

sbi-miR169a
3' A-GCCGUUCAGUAGGAACCGAC 5'
   :-:::::: ::::::::::-:
5' UCCGGCAAAUCAUCCUUGGC-G 3'
Sb09g008100 similar to Zea mays aminoacid transporter LHT1

sbi-miR169b
3'  GGCCGUUCAGUAGGAACCGAC 5'
    ::::::: ::::::::::-:
5' UCCGGCAAAUCAUCCUUGGC-G 3'
Sb09g008100 similar to Zea mays aminoacid transporter LHT1

sbi-miR169a
3' AGCCGUUCAGUAGGAACCGAC 5'
   : ::::: ::::.::::::::
5' UAGGCAAAUCAUUCUUGGCUG 3'
Sb08g021910 similar to CCAAT-binding transcription factor subunit B
family protein, expressed

sbi-miR169b
3' GGCCGUUCAGUAGGAACCGAC 5'
   :.::::: ::::::::::::
5' CUGGCAACUCAUCCUUGGCUU 3'
Sb01g045500 similar to RAPB protein

sbi-miR169efgh
3' GUCCGUUCAGUAGGAACCGAU 5'
   ::::::: ::::.:::::::
5' CAGGCAAUUCAUUCUUGGCUU 3'
Sb01g011220 similar to CCAAT-binding transcription factor subunit B
family protein, expressed

sbi-miR169efgh
3' GUCCGUUCAGUAGGAACCG-AU 5'
   : ::::: ::::::::::-::
5' CUGGCAACUCAUCCUUGGCUUA 3'
Sb01g045500 similar to RAPB protein
```

```
sbi-miR169cd
3' AUCCGUUCAGUAGGAACCG-AU 5'
   :-::::: ::::::::::::-::
5' U-GGCAACUCAUCCUUGGCUUA 3'
Sb01g045500 similar to RAPB protein

sbi-miR169cd
3' AUCCGUUCAGUAGGAACCGA-U 5'
   ::::::: ::::.::::::::-:
5' UAGGCAAAUCAUUCUUGGCUGA 3'
Sb08g021910 similar to CCAAT-binding transcription factor subunit B
family protein, expressed

sbi-miR169i
3' A-UCCGUUCAGUAAGAACCGAU 5'
   :-:::::: ::::::::::::
5' UCAGGCAAUUCAUUCUUGGCUU 3'
Sb01g011220 similar to CCAAT-binding transcription factor subunit B
family protein, expressed

sbi-miR169i
3' AUCCGUUCAGUAAGAACCGAU 5'
    :: ::::::: :::::::::
5' GAGUCAAGUCACUCUUGGCUA 3'
Sb02g003070 similar to Os07g0152000 protein

sbi-miR172cad
3' ACGUCGUAGUAGUUCUAAGA 5'
   ::::::::::::.::::::
5' UGCAGCAUCAUCAGGAUUCU 3'
Sb01g003400 similar to Indeterminate spikelet 1

sbi-miR172b
3' ACGUCGUAGUAGUUCUAAG-G 5'
   ::::::::::::.::::::-:
5' UGCAGCAUCAUCAGGAUUCUC 3'
Sb01g003400 similar to Indeterminate spikelet 1

sbi-miR172b
3' ACGUCGUAGUAGUUCUAAGG 5'
    ::::::::::::.::::::
5' CGCAGCAUCAUCAGGAUUCC 3'
Sb10g025053 similar to Glossy15

sbi-miR172cad
3' ACGUCGUAGUAGUUCUAAGA 5'
    ::::::::::::.:::::
5' CGCAGCAUCAUCAGGAUUCC 3'
Sb10g025053 similar to Glossy15

sbi-miR172e
3' CACGU-CGUAGUAGUUCUAAGU 5'
   : :::-::::::::::::-:::
5' GCGCAGGCAUCAUCAAGA-UCA 3'
Sb01g044240 similar to FRIGIDA-like protein 2
```

```
sbi-miR172e
3' C-ACGUCGUAGUAGUUCUAAGU 5'
   :-:::::::::::: :::::
5' GCUGCAGCAUCAUCACGAUUCC 3'
Sb06g030670 similar to AP2 transcription faactor

sbi-miR172cad
3' ACGUCGUAGUAGUUCUAAGA 5'
   :::::::::::: :::::
5' UGCAGCAUCAUCACGAUUCC 3'
Sb06g030670 similar to AP2 transcription factor

sbi-miR172b
3' ACGUCGUAGUAGUUCUAAG-G 5'
   ::::::::::::.::::: :
5' UGCAGCAUCAUCAGGAUUCUC 3'
Sb02g007000 similar to Putative indeterminate spikelet 1

sbi-miR172b
3' ACGUCGUAGUAGUUCUAAGG 5'
   :::::::::::: ::::::
5' UGCAGCAUCAUCACGAUUCC 3'
Sb06g030670 similar to AP2 transcription factor

sbi-miR172e
3' C-ACGUCGUAGUAGUUCUAAGU 5'
   :-:::::-::::::::.::::::
5' GAUGCAG-AUCAUCAGGAUUCA 3'
Sb04g038320 similar to Type A response regulator 3

sbi-miR395bacde
3' CUCAAGGGGGUUUGUGAAGUG 5'
   :::::.:::::::::::::::.
5' GAGUUUCCCCAAACACUUCAU 3'
Sb01g044100 similar to Putative sulfate transporter

sbi-miR395f
3' CUCAAGGGGGUUUGUGAAGUA 5'
   :::::.:::::::::::::::::
5' GAGUUUCCCCAAACACUUCAU 3'
Sb01g044100 similar to Putative sulfate transporter

sbi-miR395bacde
3' CUCAAGGGGGUUUGUGAAGUG 5'
   :::::::.::::.::::::::.
5' GAGUUCCUCCAAGCACUUCAU 3'
Sb01g008450 similar to ATP sulfurylase

sbi-miR395f
3' CUCAAGGGGGUUUGUGAAGUA 5'
   :::::::.::::.:::::::::
5' GAGUUCCUCCAAGCACUUCAU 3'
Sb01g008450 similar to ATP sulfurylase
```

Figure 4.5. List of target genes predicted for miR169, mir172 and miR395 microRNAs. This figure displays the alignments between miR169, miR172 and miR395 microRNAs and their respective target gene sequences.

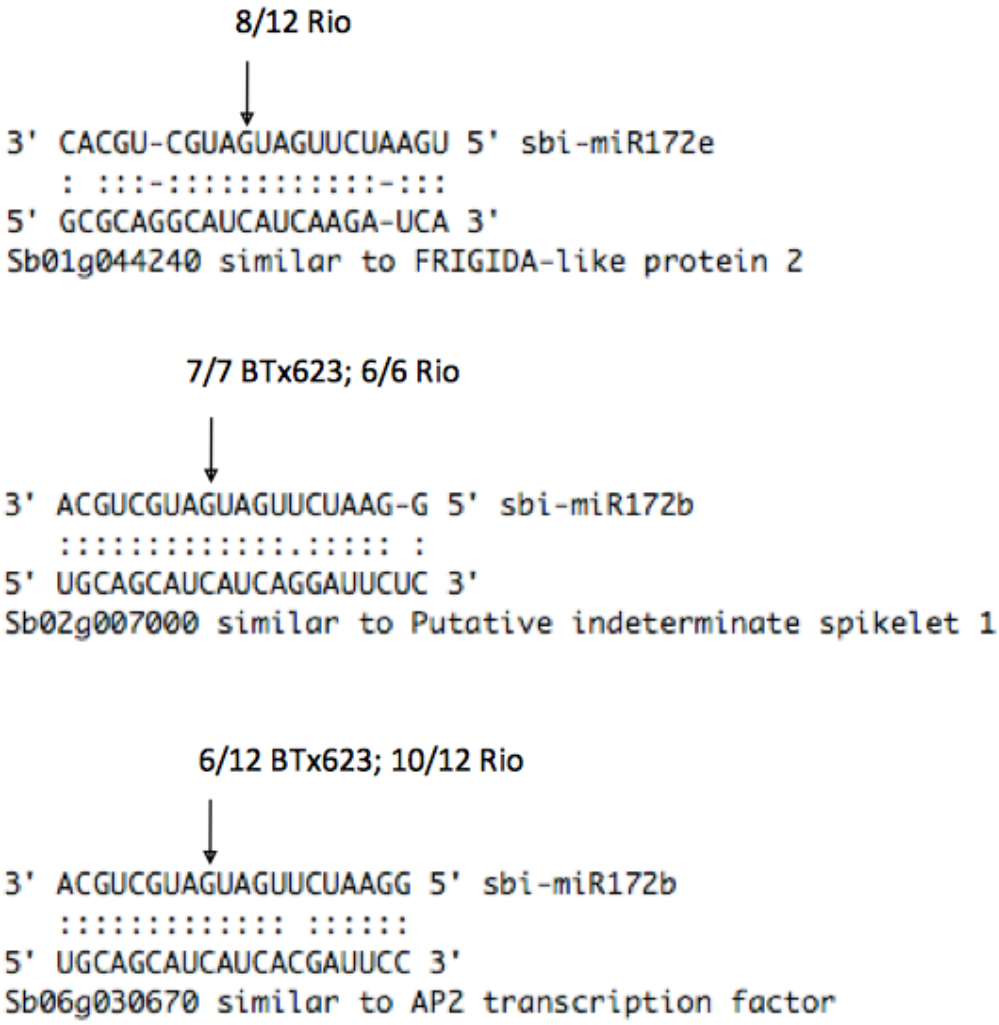


Figure 4.6. Mapping of miR172-guided cleavage sites in predicted target genes. This figure displays an alignment of miR172 with its target sequences and cleavage sites. The locations of the miRNA-cleavage sites are indicated with downward arrows and the frequency of the cleavages are indicated as the number of clones for each RACE product with respect to the total clones sequenced.

4.3.6. Identification of new miRNAs

The miRDeep pipeline (Friedlander et al., 2008) was adapted for de novo detection of miRNAs in sorghum (Figure 4.7). From an original set of 223 predicted hairpins in the sorghum genome, 9 met the miRNA annotation criteria previously established (Meyers et al., 2008), (Table 4.5; Table 4.6; and Figure 4.8). All the new miRNAs have predicted genes as targets except miR5389 (Figure 4.9). All predicted 9 miRNAs met the expression criteria used above for known miRNAs (Figure 4.10 and Table 4.6). From all miRNAs whose expression could be detected in sorghum stems, two of them were found to be within introns of protein coding genes, these included miR172c and miR437g.

From the newly identified miRNAs, miR5386, and miR5388 displayed allelic variation in expression between BTx623 and Rio that was inherited in the F2 offspring (Figure 4.10). However, the predicted target genes for miR5386 did not include any transcript involved neither in flowering nor in carbohydrate metabolism. This was a similar case with miR5388, with no predicted targets involved in flowering but with two genes involved in carbohydrate metabolism as predicted targets, encoding the beta subunit 1 and 2 of the Snf-1 related protein kinase (SnRK1) respectively (Zheng et al., 2010) (Figure 4.9).

We next attempted to experimentally validate the miRNA-mediated cleavage of predicted targets. Potential miRNA target sites were scored from 0 to 8 (see Methods), with higher scores indicating less confidence in the predictions. We tested 14 predicted targets with scores less than 4 but we could not detect the miRNA-mediated cleavage for any of them. A low rate in target validation has also

been observed for newly predicted miRNAs in tomato, with three targets validated from 65 predicted targets that were tested (Moxon et al., 2008). Recently, a similar case of very low rate in target validation was reported for predicted targets of new miRNAs identified in *Arabidopsis lyrata* (Ma et al., 2010).

microRNA gene ID	Position	Strand	miRNA size	miRNA sequence 5'-3'	miRNA* sequence 5'-3'	miRNA* size
sbi-MIR5381	Ch1: 574,388..574,497	+	19	AAGATCTGTGGCGCCGAGC	TCGGCGCTAAGATCTCTGG	19
sbi-MIR5382	Ch2: 1,930,828..1,930,937	+	18	CCAATCTAAACAGGCCCT	GACCTGTTTAGATTGGGA	18
sbi-MIR5383	Ch4: 43,242,765..43,242,874	+	24	ATGACAGAGCTCCGGCAGAGATAT	TTCTCCGCCGAGCTTATCTGTGG	23
sbi-MIR5384	Ch4: 45,785,396..45,785,505	+	18	CGCGCCGCCGTCCAGCGG	CTTGGCCGGTGCACGCGTC	19
sbi-MIR5385	Ch6: 56,307,517..56,307,626	+	22	ACCACCAACCCCACCGCTTCTC	GAAGCGGTGGTGTTGGTGGTGA	22
sbi-MIR5386	Ch7: 877,244..877,353	+	20	CGTCGCTGTCGCGCGCGCTG	GGTCAGGGCAGAGCACGCA	19
sbi-MIR5387	Ch7: 15,969,322..15,969,431	+	25	TAACACGAACCGGTGCTAAAGGATC	CCCTTTAGCACCGGTTCGTGTTACA	25
sbi-MIR5388	Ch8: 1,629,110..1,629,219	+	22	ATCTTTGCCGGGTGTCTCTGAC	CAGCAAACATTCGGCAAAGAAAA	23
sbi-MIR5389	Ch8: 4,848,342..4,848,451	+	21	GCTTGAGTTTATCAGCCGAGT	ATGGCTTATCAGCCAAGTGA	20

Table 4.5. List of new predicted microRNA genes in the sorghum genome.

Count of mapped reads to MicroRNA genes for each library					
miRNA ID	**Mix**	**BTx623**	**Rio**	**LB/EF F2s**	**HB/LF F2s**
sbi-MIR5381	10	2	2	14	14
sbi-MIR5382	8	5	9	22	37
sbi-MIR5383	3	5	0	5	5
sbi-MIR5384	2	3	3	9	12
sbi-MIR5385	41	3	42	137	182
sbi-MIR5386	7	5	4	47	8
sbi-MIR5387	4	9	12	22	31
sbi-MIR5388	0	2	1	15	9
sbi-MIR5389	6	6	2	18	19

Table 4.6. Frequency counts of small RNA reads derived from new microRNA genes in sorghum.

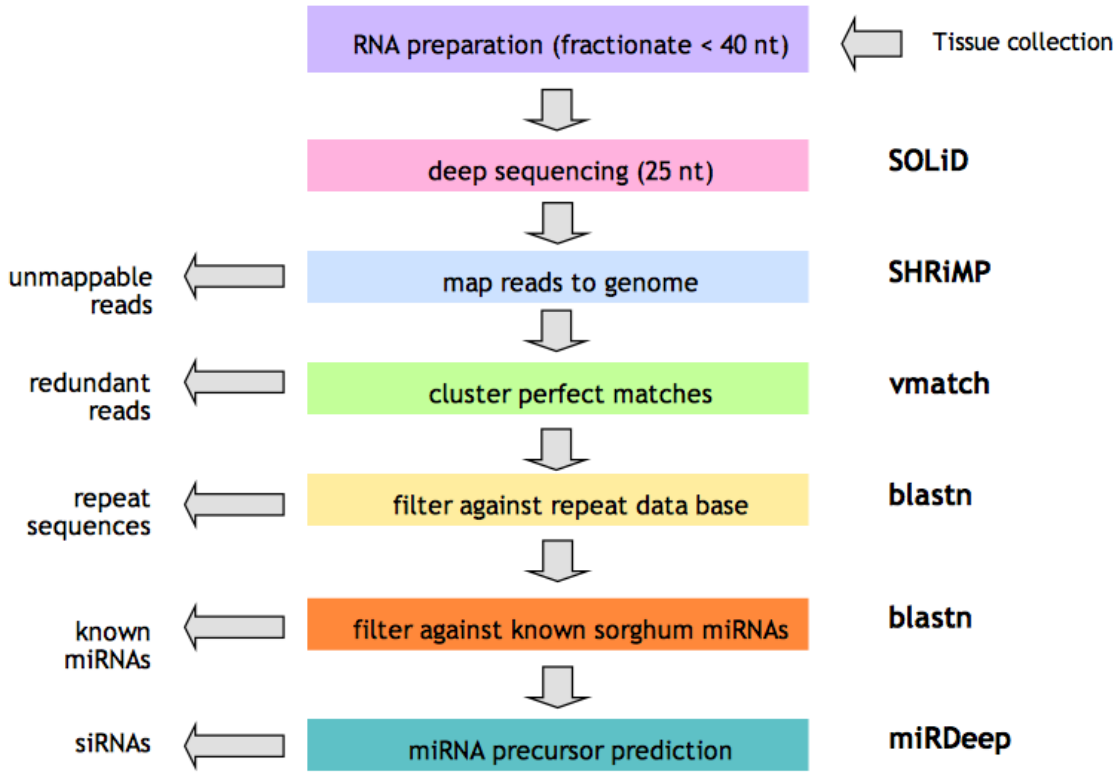
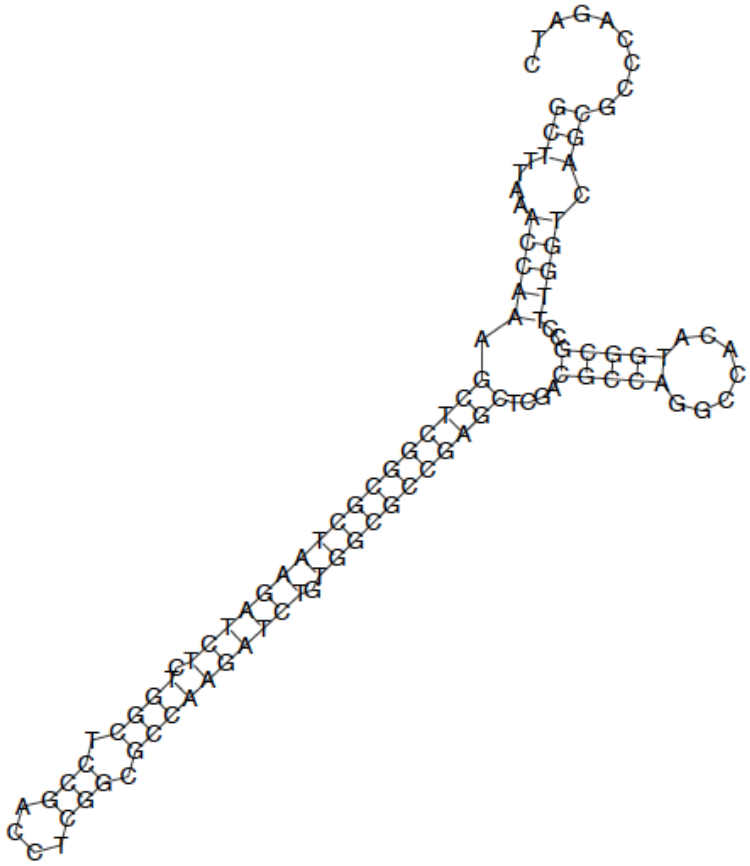


Figure 4.7. Pipeline for the *de novo* miRNA detection. This figure presents a diagram of computational steps involved in de novo miRNA detection. All reads from SOLiD sequencing were mapped in colorspace to the sorghum genome using SHRiMP. Perfect matching reads were clustered with Vmatch then filtered against the sorghum repeat sequences and compared with know sorghum miRNAs to classify them. The remaining sequences were taken for *de novo* miRNA prediction using miRDeep.

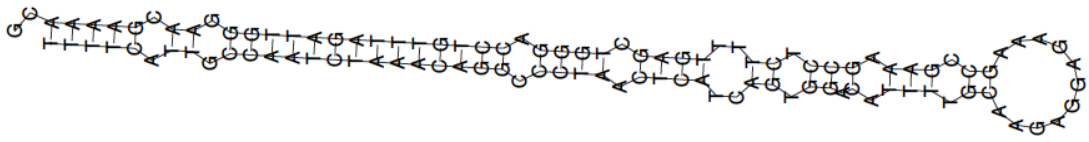
sbi-MIR5381



sbi-MIR5381

```
gctttaaaccaaagcTCGGCGCTAAGATCTCTGGctccgacctcggcgccAAGATCTGTGGCGCCGAGCtcgacgccaggccacatggcgccttggtcagcgcccagatc
     AAACCAAAGCTCGGCGCTAAGA,bc02,1
       ACCAAAGCTCGGCGCTAAGA,bc01,1
                    CTAAGATCTCTGGCTCCGACCTCG,bc04,1
                       AGATCTCTGGCTCCGACC,bc05,1
                                 TCCGACCTCGGCGCCAAGATCTG,bc04,1
                                 TCCGACCTCGGCGCCAAGATCTG,bc05,1
                                    GACCTCGGCGCCAAGATC,bc01,1
                                    GACCTCGGCGCCAAGATCT,bc05,1
                                     AGCTCGGCGCCAAGATCTGTGGC,bc01,1
                                      GCTCGGCGCCAAGATCTGT,bc04,1
                                       CTCGGCGCCAAGATCTGTG,bc05,1
                                        TCGGCGCCAAGATCTGTGGC,bc05,1
                                               CAAGATCTGTGGCGCCGA,bc01,1
                                               CAAGATCTGTGGCGCCGAG,bc05,1
                                                 AGATCTGTGGCGCCGAGCTC,bc01,1
                                                 AGATCTGTGGCGCCGAGCTCG,bc01,1
                                                 AGATCTGTGGCGCCGAGCT,bc03,1
                                                 AGATCTGTGGCGCCGAGCTC,bc03,1
                                                 AGATCTGTGGCGCCGAGCTC,bc04,1
                                                 AGATCTGTGGCGCCGAGCTCG,bc04,1
                                                 AGATCTGTGGCGCCGAGCTC,bc05,3
                                                  GATCTGTGGCGCCGAGCTC,bc01,1
                                                  GATCTGTGGCGCCGAGCTC,bc04,1
                                                  GATCTGTGGCGCCGAGCTC,bc05,1
                                                  GATCTGTGGCGCCGAGCTCG,bc05,1
                                                   ATCTGTGGCGCCGAGCTC,bc01,1
                                                   ATCTGTGGCGCCGAGCTC,bc04,1
                                                               GAGCTCGACGCCAGGCCACATG,bc01,1
                                                               GAGCTCGACGCCAGGCCACATGG,bc02,1
                                                               GAGCTCGACGCCAGGCCACA,bc04,1
                                                               GAGCTCGACGCCAGGCCACATG,bc04,2
                                                               GAGCTCGACGCCAGGCCACATGG,bc04,1
                                                               GAGCTCGACGCCAGGCCACATG,bc05,1
                                                               GAGCTCGACGCCAGGCCACATGG,bc05,1
                                                                        GCCAGGCCACATGGCGCCTTG,bc01,1
                                                                           AGGCCACATGGCGCCTTG,bc04,2
                                                                           AGGCCACATGGCGCCTTG,bc05,1
                                                                             GCCACATGGCGCCTTGGTCAGCGCC,bc04,1
```

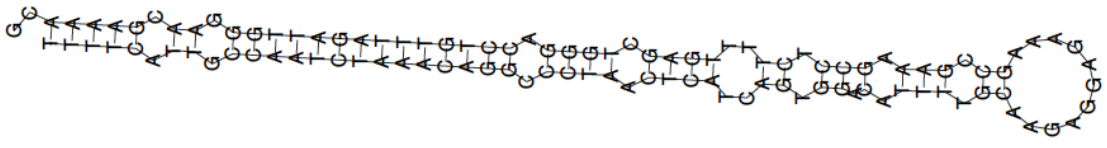
sbi-MIR5382



sbi-MIR5382

ttttcattgCCAATCTAAACAGGCCCTaactcatcagtggacattttgcaagaggagaaagccgaaagcctcttttgagctggGACCTGTTTAGATTGGGAacgaaaacg
TTCATTGCCAATCTAAACAGGC,bc05,1
TTCATTGCCAATCTAAACAGGCCC,bc05,1
TCATTGCCAATCTAAACAGGCCC,bc05,1
CATTGCCAATCTAAACAGGCCCT,bc02,1
CATTGCCAATCTAAACAGGCCC,bc04,1
CATTGCCAATCTAAACAGGCCCT,bc04,1
CATTGCCAATCTAAACAGGCCCT,bc05,1
ATTGCCAATCTAAACAGGCCCT,bc01,1
ATTGCCAATCTAAACAGGCCCT,bc02,1
ATTGCCAATCTAAACAGGCCC,bc03,1
ATTGCCAATCTAAACAGGCCC,bc04,1
ATTGCCAATCTAAACAGGCCCT,bc04,2
ATTGCCAATCTAAACAGGCCCT,bc05,1
TTGCCAATCTAAACAGGCCCT,bc01,1
TTGCCAATCTAAACAGGCCCT,bc02,2
TTGCCAATCTAAACAGGCCC,bc05,1
TTGCCAATCTAAACAGGCCCT,bc05,1
GCCAATCTAAACAGGCCCT,bc04,1
GCCAATCTAAACAGGCCCT,bc04,2
GCCAATCTAAACAGGCCC,bc05,1
GCCAATCTAAACAGGCCCT,bc05,1
CCAATCTAAACAGGCCCTAAC,bc03,2
AATCTAAACAGGCCCTAACT,bc01,1
AATCTAAACAGGCCCTAACT,bc03,1
AGGCCCTAACTCATCAGTGGACAT,bc05,1
CCCTAACTCATCAGTGGACATTT,bc05,1
CTCATCAGTGGACATTTTGCAAGA,bc05,1
ATCAGTGGACATTTTGCAAGAGGA,bc01,1
CAGTGGACATTTTGCAAGAGGAG,bc05,1
GGACATTTTGCAAGAGGAGAAAG,bc05,1
CATTTTGCAAGAGGAGAAAGCC,bc05,1
AAGCCTCTTTTGAGCTGGGACCT,bc01,2
AGCCTCTTTTGAGCTGGGACC,bc05,1
GGACCTGTTTAGATTGGGAAC,bc03,1
GGACCTGTTTAGATTGGGA,bc04,1
GGACCTGTTTAGATTGGG,bc05,1
GGACCTGTTTAGATTGGGA,bc05,1
GGACCTGTTTAGATTGGGAA,bc05,1
GGACCTGTTTAGATTGGGAAC,bc05,1
GGACCTGTTTAGATTGGGAACG,bc05,2
GGACCTGTTTAGATTGGGAACGAA,bc05,1
GACCTGTTTAGATTGGGAACGAAA,bc03,1
CCTGTTTAGATTGGGAACGA,bc01,1
CCTGTTTAGATTGGGAACGAAAA,bc01,1
CCTGTTTAGATTGGGAACGA,bc02,1
CCTGTTTAGATTGGGAACGA,bc03,1
CCTGTTTAGATTGGGAACGAA,bc03,1
CCTGTTTAGATTGGGAACGAAAA,bc03,1
CCTGTTTAGATTGGGAAC,bc04,8
CCTGTTTAGATTGGGAACG,bc04,3
CCTGTTTAGATTGGGAACGAAAA,bc04,2
CCTGTTTAGATTGGGAAC,bc05,3
CCTGTTTAGATTGGGAACG,bc05,2
CCTGTTTAGATTGGGAACGA,bc05,1
CCTGTTTAGATTGGGAACGAA,bc05,2
CCTGTTTAGATTGGGAACGAA,bc05,2
CCTGTTTAGATTGAGAACGAAAA,bc05,2
CCTGTTTAGATTGGGAACGAAAA,bc05,1
CCTGTTTAGATTGGGAACGAAAA,bc05,1

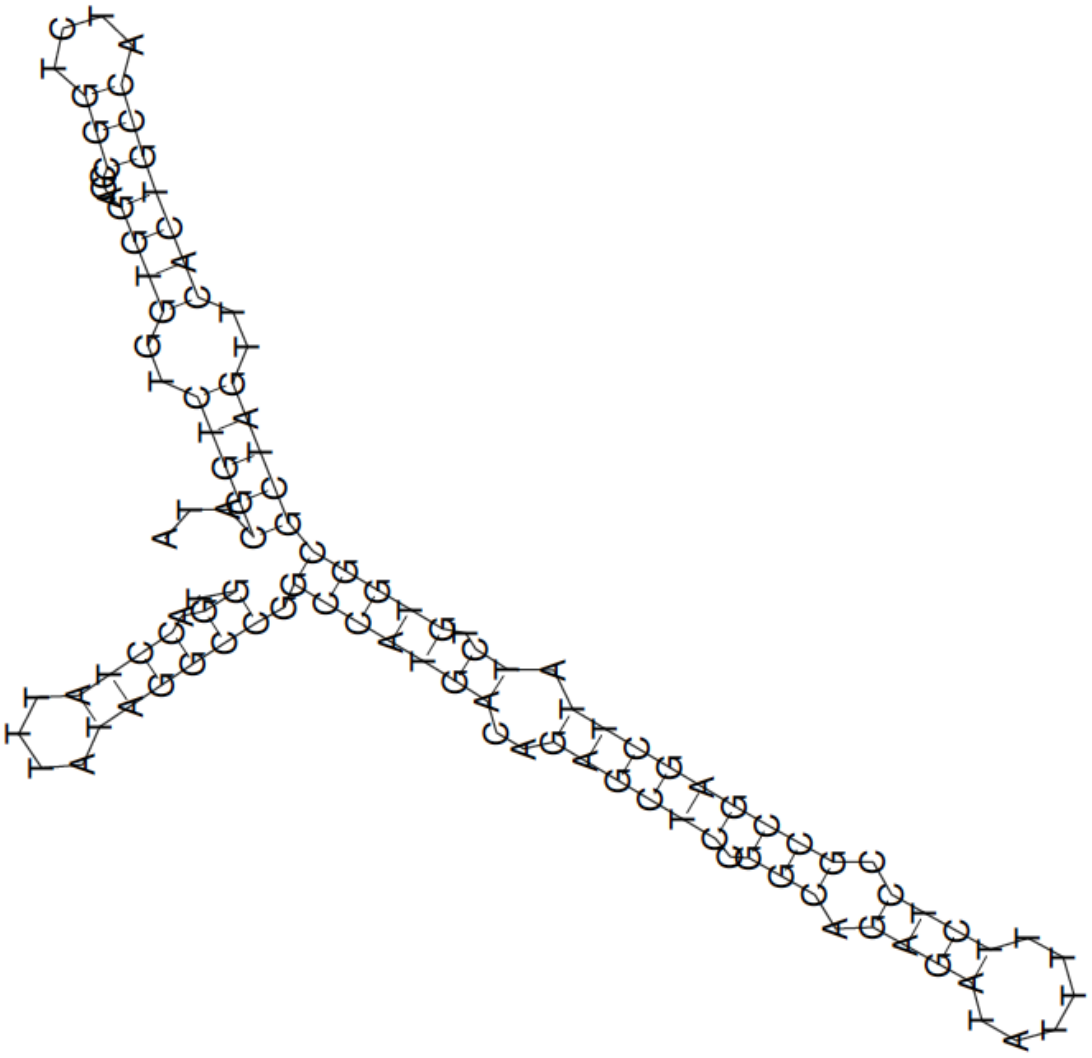
sbi-MIR5383



sbi-MIR5383

```
tggacctatttataggccggccATGACAGAGCTCCGGCAGAGATATttTTCTCCGCCGAGCTTATCTGTGGcgctagttcactgccatctggccgaggtggtctggcata
                 CGGCCATGACAGAGCTCCG,bc02,1
                  GGCCATGACAGAGCTCCGGCAGA,bc04,1
                      ATGACAGAGCTCCGGCAGAGA,bc01,1
                      ATGACAGAGCTCCGGCAGAGATAT,bc01,2
                      ATGACAGAGCTCCGGCAGAGATAT,bc02,2
                      ATGACAGAGCTCCGGCAGAGATAT,bc04,2
                      ATGACAGAGCTCCGGCAGAGATAT,bc05,5
                           AGAGCTCCGGCAGAGATA,bc04,1
                                              TCTCCGCCGAGCTTATCTGTGGCG,bc04,1
                                                TCCGCCGAGCTTATCTGTGGCGCT,bc02,1
                                                                        CCATCTGGCCGAGGTGGTCTGGC,bc02,1
```

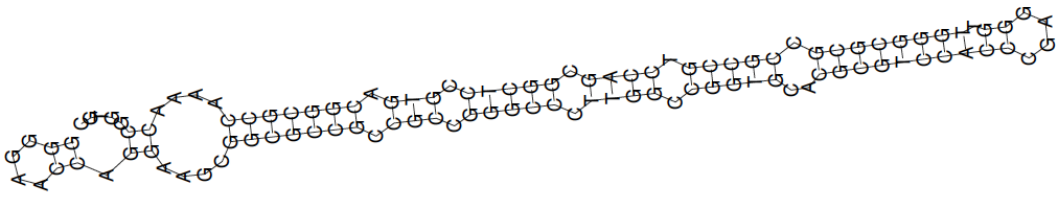
sbi-MIR5384



sbi-MIR5384


ggggaaccaggaagcggcgccgccgccgggccCTTGGCCGGTGCACGCGTCcacccgagggttgggCGCGCCGCCGTCCAGCGGctccgtgacggcgccaaaaccggtgc
CCGGGCCCTTGGCCGGTG,bc05,1
CGGGCCCTTGGCCGGTGC,bc03,1
CGGGCCCTTGGCCGGTGCGCGCG,bc05,1
TTGGCCGGTGCGCGCGTCCACCC,bc04,3
TTGGCCGGTGCGCGCGTCCACCCG,bc04,1
TTGGCCGGTGCGCGCGTCCACCCGA,bc04,1
TTGGCCGGTGCGCGCGTCCACCC,bc05,1
TGGCCGGTGCGCGCGTCCACCCG,bc05,1
GCCGGTGCGCGCGTCCACCCGAGGG,bc03,1
CGGTGCGCGCGTCCACCCGAGGGT,bc03,1
GGTGCGCGCGTCCACCCGAGGGTT,bc04,1
AGGGTTGGGCGCGCCGCC,bc05,1
GCGCGCCGCCGTCCAGCGG,bc05,1
GCGCGCCGCCGTCCAGCGGC,bc05,1
CGCGCCGCCGTCCAGCGGCT,bc01,1
CGCGCCGCCGTCCAGCGGCTCC,bc01,1
CGCGCCGCCGTCCAGCGG,bc02,2
CGCGCCGCCGTCCAGCGG,bc04,3
CGCGCCGCCGTCCAGCGG,bc05,3
CGCGCCGCCGTCCAGCGGC,bc05,1
CGCGCCGCCGTCCAGCGGCTCCGT,bc05,1
CGCCGCCGTCCAGCGGCT,bc02,1

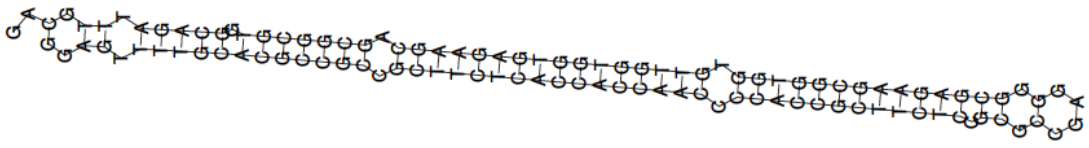
sbi-MIR5385



sbi-MIR5385

ggagttttgcacgccgccgcttctctACCACCAACCCCACCGCTTCTCcgcgccgaggggcgaGAAGCGGTGGTGTTGGTGGTGAgaagcagcggcgtggcagatttgcag
TCTCACCACCAACCCCACCG,bc03,3
TCTCACCACCAACCCCACCGC,bc03,1
TCTCACCACCAACCCCACC,bc04,14
TCTCACCACCAACCCCACCG,bc04,8
TCTCACCACCAACCCCACCGC,bc04,8
TCTCACCACCAACCCCAC,bc05,4
TCTCACCACCAACCCCACC,bc05,28
TCTCACCACCAACCCCACCG,bc05,5
TCTCACCACCAACCCCACCGC,bc05,7
CTCACCACCAACCCCACCGC,bc01,1
CTCACCACCAACCCCACCGCT,bc01,1
CTCACCACCAACCCCACCG,bc03,3
CTCACCACCAACCCCACC,bc04,1
CTCACCACCAACCCCACC,bc04,2
CTCACCACCAACCCCACCG,bc04,2
CTCACCACCAACCCCACCGC,bc04,2
CTCACCACCAACCCCACCGCT,bc04,3
CTCACCACCAACCCCACC,bc05,3
CTCACCACCAACCCCACCGCT,bc05,2
TCACCACCAACCCCACCGCTT,bc01,1
TCACCACCAACCCCACCGC,bc04,1
TCACCACCAACCCCACCGC,bc05,2
TCACCACCAACCCCACCGCT,bc05,1
CACCACCAACCCCACCGCTT,bc01,1
CACCACCAACCCCACCGCTTC,bc03,1
CACCACCAACCCCACCGC,bc04,5
CACCACCAACCCCACCGCT,bc04,1
CACCACCAACCCCACCGCTTC,bc05,9
ACCACCAACCCCACCGCT,bc02,1
ACCACCAACCCCACCGCTTC,bc04,11
ACCACCAACCCCACCGCTT,bc05,1
CCACCAACCCCACCGCTTC,bc03,1
CCACCAACCCCACCGCTTC,bc05,6
CACCAACCCCACCGCTTC,bc01,2
CACCAACCCCACCGCTTC,bc03,1
GGGCGAGAAGCGGTGGTGTTGGTGG,bc03,1
GGGCGAGAAGCGGTGGTGTTGGTGG,bc04,3
GGGCGAGAAGCGGTGGTGTTGGTGG,bc05,3
GGCGAGAAGCGGTGGTGTTGGTGGT,bc01,1
GGCGAGAAGCGGTGGTGTTGGTGG,bc05,2
GCGAGAAGCGGTGGTGTTGGTGG,bc01,1
GAGAAGCGGTGGTGTTGGTGG,bc01,3
GAGAAGCGGTGGTGTTGGTGG,bc03,3
GAGAAGCGGTGGTGTTGGTGG,bc04,8
GAGAAGCGGTGGTGTTGGTGG,bc05,16
AGAAGCGGTGGTGTTGGTGG,bc01,4
AGAAGCGGTGGTGTTGGTGG,bc03,1
AGAAGCGGTGGTGTTGGTGG,bc05,4
GAAGCGGTGGTGTTGGTGG,bc01,8
GAAGCGGTGGTGTTGGTGG,bc03,9
GAAGCGGTGGTGTTGGTGGTGA,bc03,1
GAAGCGGTGGTGTTGGTGG,bc04,11
GAAGCGGTGGTGTTGGTGGTG,bc04,1
GAAGCGGTGGTGTTGGTGG,bc05,23
GAAGCGGTGGTGTTGGTGGTG,bc05,1
AAGCGGTGGTGTTGGTGG,bc01,2
AAGCGGTGGTGTTGGTGG,bc02,1
AAGCGGTGGTGTTGGTGG,bc03,2
AAGCGGTGGTGTTGGTGG,bc04,10
AAGCGGTGGTGTTGGTGGTG,bc04,1
AAGCGGTGGTGTTGGTGG,bc05,8
AAGCGGTGGTGTTGGTGGTG,bc05,1
AAGCGGTGGTGTTGGTGGTGAG,bc05,1
AGCGGTGGTGTTGGTGGTGAG,bc01,1
AGCGGTGGTGTTGGTGGTG,bc02,1
AGCGGTGGTGTTGGTGGTG,bc04,9
AGCGGTGGTGTTGGTGGTGA,bc04,1
AGCGGTGGTGTTGGTGGTGAG,bc04,1
AGCGGTGGTGTTGGTGGTGAGA,bc04,1
AGCGGTGGTGTTGGTGGTG,bc05,7
GCGGTGGTGTTGGTGGTG,bc01,4
GCGGTGGTGTTGGTGGTGAGA,bc01,1
GCGGTGGTGTTGGTGGTG,bc03,1
GCGGTGGTGTTGGTGGTGAGA,bc03,2
GCGGTGGTGTTGGTGGTG,bc04,16
GCGGTGGTGTTGGTGGTGA,bc04,1
GCGGTGGTGTTGGTGGTGAGA,bc04,4
GCGGTGGTGTTGGTGGTG,bc05,17
GCGGTGGTGTTGGTGGTGA,bc05,1
GCGGTGGTGTTGGTGGTGAG,bc05,1
GCGGTGGTGTTGGTGGTGAGA,bc05,3
CGGTGGTGTTGGTGGTGAGA,bc03,2
CGGTGGTGTTGGTGGTGAGA,bc04,4
CGGTGGTGTTGGTGGTGAGA,bc05,5
GGTGGTGTTGGTGGTGAGA,bc01,5
GGTGGTGTTGGTGGTGAGA,bc03,1
GGTGGTGTTGGTGGTGAG,bc04,1
GGTGGTGTTGGTGGTGAG,bc04,1
GGTGGTGTTGGTGGTGAGA,bc04,1
GGTGGTGTTGGTGGTGAGA,bc05,8
GTGGTGTTGGTGGTGAGA,bc01,4
GTGGTGTTGGTGGTGAGA,bc03,7
GTGGTGTTGGTGGTGAGAA,bc03,1
GTGGTGTTGGTGGTGAGA,bc04,5
GTGGTGTTGGTGGTGAGA,bc05,10
GTGGTGTTGGTGGTGAGA,bc05,1
GGTGTTGGTGGTGAGAAGCA,bc01,1
GGTGTTGGTGGTGAGAAGCAG,bc05,1
GTTGGTGGTGAGAAGCAGCGG,bc03,2
TTGGTGGTGAGAAGCAGCGGCG,bc05,1

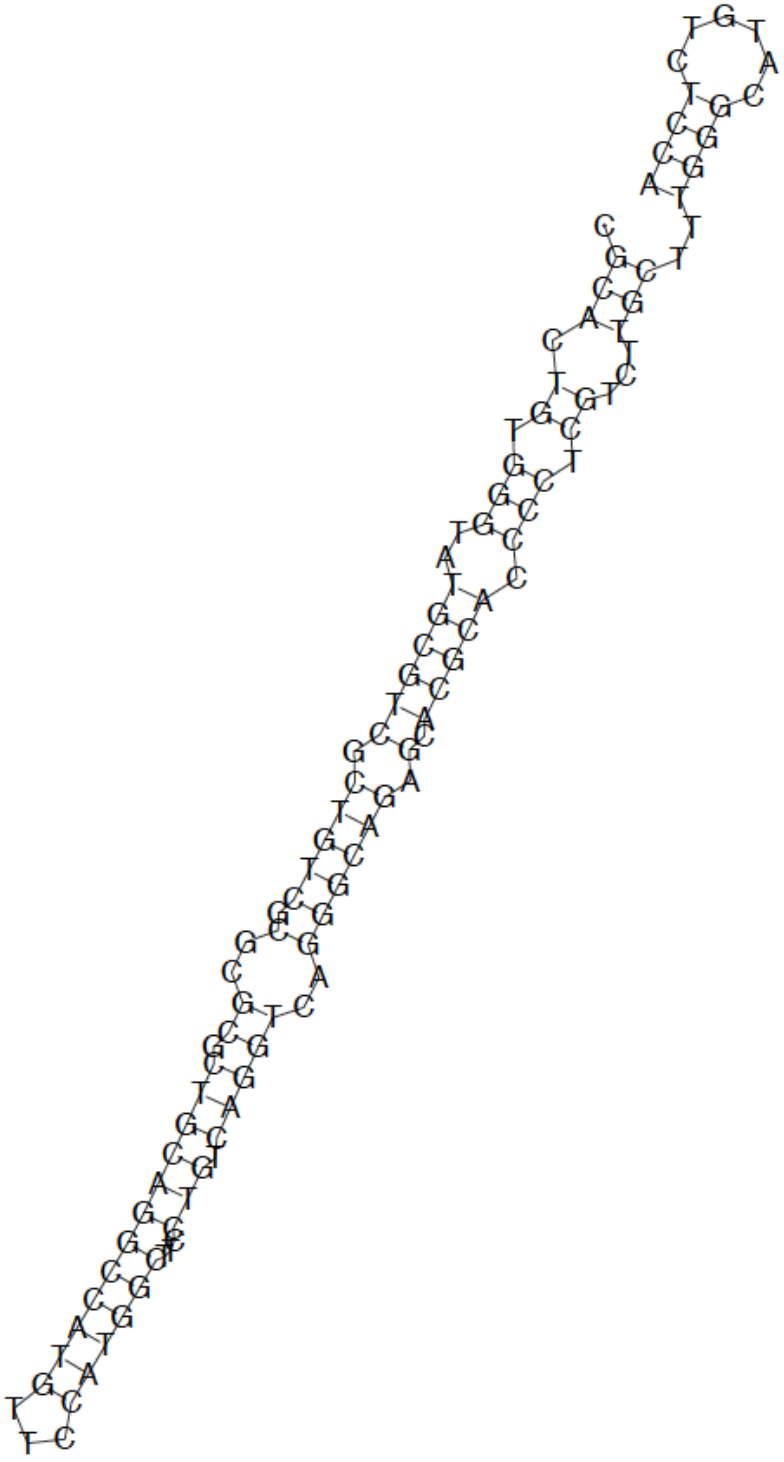
sbi-MIR5386



sbi-MIR5386

```
cgcactgtgggtatgCGTCGCTGTCGCGCGCGCTGcaggccatgttccatggcttcctgtcaGGTCAGGGCAGAGCACGCAcccctcgtcttgctttgggcatgtctcca
                                                               GTCAGGGCAGAGCACGCACCCC,bc01,1
                                                               GTCAGGGCAGAGCACGCACCCCT,bc01,1
                                                               GTCAGGGCAGAGCACGCACCCCTCG,bc01,5
                                                               GTCAGGGCAGAGCACGCACCCCTCG,bc02,5
                                                               GTCAGGGCAGAGCACGCACCCCTCG,bc03,4
                                                               GTCAGGGCAGAGCACGCAC,bc04,1
                                                               GTCAGGGCAGAGCACGCACCC,bc04,1
                                                               GTCAGGGCAGAGCACGCACCCC,bc04,4
                                                               GTCAGGGCAGAGCACGCACCCCT,bc04,9
                                                               GTCAGGGCAGAGCACGCACCCCTC,bc04,1
                                                               GTCAGGGCAGAGCACGCACCCCTCG,bc04,30
                                                               GTCAGGGCAGAGCACGCACCCC,bc05,1
                                                               GTCAGGGCAGAGCACGCACCCCT,bc05,2
                                                               GTCAGGGCAGAGCACGCACCCCTC,bc05,1
                                                               GTCAGGGCAGAGCACGCACCCCTCG,bc05,4
                                                                TCAGGGCAGAGCACGCACCCCT,bc04,1
```

sbi-MIR5387



sbi-MIR5387


tcgTAACACGAACCGGTGCTAAAGGATCttgcccaacggctactgacagctgttgttggggcagggaCCCTTTAGCACCGGTTCGTGTTACAaaccggtggtaaagggtt
CGTAACACGAACCGGTGCTAAA,bc03,1
GTAACACGAACCGGTGCT,bc02,1
GTAACACAAACCGGTGCTAAAG,bc05,1
AACACGAACCGGTGCTAA,bc03,1
ACACGAACCGGTGCTAAAGGA,bc03,1
ACACGAACCGGTGCTAAAGGA,bc05,1
CACGAACCGGTGCTAAAG,bc05,1
CACGAACCGGTGCTAAAGGATC,bc05,1
ACGAACCGGTGCTAAAGGA,bc01,1
ACGAACCGGTGCTAAAGGAT,bc04,1
CGAACCGGTGCTAAAGGA,bc01,1
GAACCGGTGCTAAAGGATC,bc05,1
AACCGGTGCTAAAGGATC,bc04,1
CTAAAGGGTCTTGCCCAACGGCT,bc02,1
CTAAAGGATCCTGCCCAACGGCT,bc04,2
AAGGATCTTGCCCAACGGCTACTG,bc04,1
AGGATCTTGCCCAACGGCTACTGAC,bc02,1
GGATCTTGCCCAACGGCTACTGA,bc02,1
GGATCTTGCCCAACGGCTACTGAC,bc02,1
ATCTTGCCCAACGGCTACTGACAGC,bc05,1
ACTGACAGCTGTTGTTGGGGCAGGG,bc03,1
GACAGCTGTTGTTGGGGCAG,bc05,1
GCTGTTGTTGGGGCAGGGACCC,bc05,1
GTTGGGGCAGGGACCCTTTAGCA,bc03,1
GTTGGGGCAGGGACCCTT,bc04,1
TGGGGCAGGGACCCTTTAG,bc03,1
GGGCAGGGATCCTTTAGCACCGG,bc04,1
GGCAGGGACCCTTTAGCACCGGTT,bc04,1
GCAGGGACCCTTTAGCACCGGTTCG,bc02,1
CAGGGACCCTTTAGCACCGGTTCGT,bc02,1
CAGGGACCCTTTAGCACCGG,bc04,1
CAGGGACCCTTTAGCACCG,bc05,1
CAGGGACCCTTTAGCACCGGTTC,bc05,2
AGGGACCCTTTAGCACCGGTTCGT,bc04,1
AGGGACCCTTTAGCACCGGTTCG,bc05,1
GGGACCCTTTAGCACCGG,bc03,1
GGGACCCTTTAGCACCGGTTCGTGT,bc03,1
GGGACCCTTTAGCACCGGTTCG,bc05,1
GGGACCCTTTAGCACCGGTTCGTGT,bc05,5
GGACCCTTTAGCACCGGTTCGTGT,bc04,1
GACCCTTTAGCACCGGTTCGTGT,bc03,1
GACCCTTTAGCACCGGTTCGTGTT,bc04,1
GACCCTTTAGCATCGGTTCGTGT,bc05,1
GACCCTTTAGCACCGGTTCATGTT,bc05,1
GACCCTTTAGCACCGGTTCGTGTTA,bc05,1
ACCCTTTAGCACCGGTTCGTGTTAC,bc05,1
CCCTTTAGCACCGGTTCGTGTTAC,bc04,1
CCTTTAGCACCGGTTCGTGT,bc04,1
CCTTTAGCACCGGTTCGTGTTAC,bc04,1
CCTTTAGCACCGGTTCGTGTT,bc05,1
CTTTAGCACCGGTTCGTGT,bc04,1
CTTTAGCACCGGTTCGTGT,bc05,1
TAGCACCGGTTCGTGTTAC,bc01,1
AGCACCGGTTCGTGTTACAA,bc04,1
AGCACCGGTTCGTGTTACCAACCG,bc04,1
AGCACCGGTTCGTGTTACCAACCG,bc05,3
GCACCGGTTCGTGTTACAAACC,bc03,1
CCGGTTCGTGTTACAAAC,bc01,1
CCGGTTCGTGTTACAAAC,bc05,1
CGGTTCGTGTTACAAACCGGTG,bc02,1
GGTTCGTGTTACAAACCGGTG,bc03,2
GGTTCGTGTTACAAACCGGTG,bc04,1
GGTTCGTGTTACAAACCGG,bc05,1
GGTTCGTGTTACAAACCGGTG,bc05,1
CGGCCCGTGTTACAAACCGGTG,bc02,1
CGTGTTACAAACCGGTGCTAAAGGG,bc04,1
GTGTTACAAACCGGTGCTAAAGGGT,bc05,1

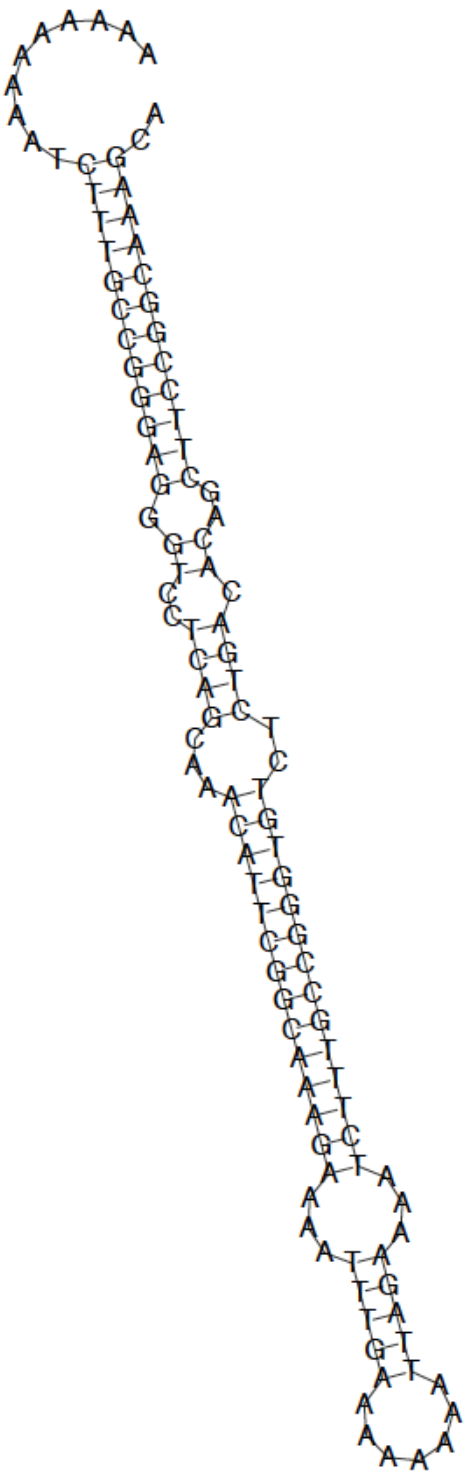
sbi-MIR5388

sbi-MIR5388

aaaaaaaaatctttgccgggagggtcctCAGCAAACATTCGGCAAAGAAAAtttgaaaaaaaattagaaaATCTTTGCCGGGTGTCTCTGACacagcttccggcaaagca
CCGGGAGGGTCCTCAGCAAACATT,bc04,1
GGGAGGGTCCTCAGCAAACATTCG,bc05,1
GGAGGGTCCTCAGCAAACA,bc04,3
GGGTCCTCAGCAAACATTCGGC,bc04,1
GGGTCCTCAGCAAACATTCGGCAAA,bc05,1
GGTCCTCAGCAAACATTCGGCAA,bc04,1
GTCCTCAGCAAACATTCGGC,bc04,1
TCCTCAGCAAACATTCGGC,bc04,1
CCTCAGCAAACATTCGGC,bc04,1
CCTCAGCAAACATTCGGCAAAGA,bc05,1
CTCAGCAAACATTCGGCAAAGAAAA,bc05,1

AGAAAATCTTTGCCGGGTG,bc04,1
AGAAAATCTTTGCCGGGTG,bc05,1
GAAAATCTTTGCCGGGTG,bc04,1
AAAATCTTTGCCGGGTGTC,bc04,1
AAATCTTTGCTGGGTGTCTCTG,bc05,1
ATCTTTGCCGGGTGTCTCTGAC,bc05,1
TTGCCGGGTGTCTCTGACACAGCT,bc02,1
TTGCTGGGTGTCTCTGACACAGCT,bc04,1
GCCGGGTGTCTCTGACACAGCTTCC,bc02,1
GCCGGGTGTCTCTGACACAGCTTCC,bc04,1
CCGGGTGTCTCTGACACAGCTTC,bc05,1
GGGTGTCTCTGACACAGCT,bc03,1
GGGTGTCTCTGACACAGCTTC,bc05,1
GTCTCTGACACAGCTCCCGGCAAA,bc04,1

sbi-MIR5389



sbi-MIR5389


tgcagccacaaacatatgctgcaagtactaatgctttgttcGCTTGAGTTTATCAGCCGAGTctgaacactactttaatttcagccATGGCTTATCAGCCAAGTGAacag
TGCTGCAAGTACTAATGCTTTGTT,bc02,1
ACTAATGCTTTGTTCGCT,bc05,1
ATGCTTTGTTCGCTTGAG,bc04,1
TTTGTTCGCTTGAGTTTA,bc01,1
TTTGTTCGCTTGAGTTTAT,bc04,1
TTTGTTCGCTTGAGTTTATC,bc04,2
TTTGTTCGCTTGAGTTTA,bc05,1
TTTGTTCGCTTGAGTTTATC,bc05,1
TTGTTCGCTTGAGTTTATC,bc01,1
TTGTTCGCTTGAGTTTATC,bc05,2
TGTTCGCTTGAGTTTATCAGCCG,bc02,3
TGTTCGCTTGAGTTTATCAGC,bc04,1
TGTTCGCTTGAGTTTATCTGCC,bc04,1
TGTTCGCTTGAGTTTATCAGCCG,bc05,2
TGTTCGCTTGAGTTTATCAGCCGA,bc05,2
GTTCGCTTGAGTTTATCAGCCGAGT,bc01,1
GTTCGCTTGAGTTTATCAGCC,bc02,1
GTTCGCTTGAGTTTATCAGCCGA,bc03,1
GTTCGCTTGAGTTTATCAG,bc04,1
GTTCGCTTGAGTTTATCA,bc05,1
GTTCGCTTGAGTTTATCAGCC,bc05,1
GTTCGCTTGAGTTTATCAGCCG,bc05,1
GTTCGCTTGAGTTTATCAGCCGA,bc05,1
GTTCGCTTGACTTTATCAGCCGAG,bc05,1
TTCGCTTGAGTTTATCAGCCGA,bc01,1
TTCGCTTGAGTTTATCAG,bc05,1
TTCGCTTGAGTTTATCAGCCG,bc05,1
TCGCTTGAGTTTATCAGCC,bc01,1
TCGCTTGAGTTTATCAGCC,bc02,1
TCGCTTGAGTTTATCAGCC,bc04,2
TCGCTTGAGTTTATCAGCCG,bc04,3
CGCTTGAGTTTATCAGCCG,bc04,1
CGCTTGAGTTTATCAGCCGA,bc05,1
GCTTGAGTTTATCAGCCG,bc05,1
CTTGAGTTTATCAGCCGA,bc03,1
TTGAGTTTATCAGCCGAGTCT,bc04,1
GAACACTACTTTAATTTCAGCC,bc01,1
GAACACTACTTTAATTTCAGCCA,bc04,1
GAACACTACTTTAATTTCAGCCAT,bc04,1
CACTACTTTAATTTCAGCCATGG,bc04,1
AGCCATGGCTTATCAGCCAA,bc05,1
GCCATGGCTTATCAGCCAAGT,bc04,1

Figure 4.8. Hairpin structures of the newly discovered miRNAs. This figure presents a collection of hairpin structures from newly discovered miRNAs. Sequences are depicted together with the frequency distribution of the small RNA reads aligned to the hairpin. The 2D hairpin structure produced by the miRDeep software is also shown.

```
Predicted targets of sbi-miR5381:


sbi-miR5381
3' CGAGCCGCGGUGUCUAGAA 5'
    :::::::::.:::::::::
5' UCUCGGCGCUGCAGAUCUU 3'
Sb06g016460.1 similar to H0525E10.16 protein

sbi-miR5381
3' CGAGCCGCGGUGUCUAGAA 5'
    ::::::::::::::::: :
5' CCUCGGCGCCACAGAUCAU 3'
Sb07g019993.1 Predicted protein

sbi-miR5381
3' CGAGCCGCGGUGUCUAGAA 5'
    ::.:::::::::: :::::
5' UCUUGGCGCCACAAAUCUU 3'
Sb04g001350.1 similar to Leaf senescence related protein-like

sbi-miR5381
3' CGAGCCGCGGUGUCUAGAA 5'
   : ::::::::.::::: :::
5' GAUCGGCGCUACAGAACUU 3'
Sb06g019590.1 similar to Putative uncharacterized protein

sbi-miR5381
3' CGAGCCGCGGUGUCUAGAA 5'
   ::::::: .:::::::::
5' GCUCGGCAUCACAGAUCUC 3'
Sb03g043060.1 similar to SMC2 protein

sbi-miR5381
3' CGA-GCCGCGGUGUCUAGAA 5'
   :: -:::::::-::::::::
5' GCCCCGGCGCC-CAGAUCUU 3'
Sb01g001300.1 similar to Uncharacterized protein At4g14147.1


sbi-miR5381
```

```
3'  CGAGCCGCGGUGUCUAGAA 5'
    :::::::::: :::-:::: 
5' AGCUCGGCGCCUCAG-UCUU 3'
Sb03g047490.1 similar to MDR-like ABC transporter

sbi-miR5381
3' CG-AGCCGCGGUGUCUAGAA 5'
   ::- :::::::-:::::::: 
5' GCGGCGGCGCC-CAGAUCUU 3'
Sb04g035440.1 similar to OSJNBa0042L16.3 protein

sbi-miR5381
3' CG-AGCCGCGGUGUCUAGAA 5'
   ::- :::::::-:::::::: 
5' GCGGCGGCGCC-CAGAUCUU 3'
Sb04g035465.1 similar to OSJNBa0042L16.3 protein

sbi-miR5381
3' CGA-GCCGCGGUGUCUAGAA 5'
   :::-:: ::::-:::::::: 
5' GCUGCGUCGCC-CAGAUCUU 3'
Sb04g037550.1 similar to LOB domain protein-like

sbi-miR5381
3' CGAGCCGCGGU-GUCUAGAA 5'
   :.::::::::-::::-::: 
5' GUUCGGCGCCAGCAGA-CUU 3'
Sb09g004010.1 similar to Putative uncharacterized protein

sbi-miR5381
3' CG-AGCCGCGGUGUCUAGAA 5'
   ::-::::::::-::: :::: 
5' GCUUCGGCGCC-CAGCUCUU 3'
Sb09g024180.1 similar to Cyclin IbZm

sbi-miR5381
3' CGAGCCGCGGUGUCUAGAA 5'
   ::::: :::::: : :::: 
5' GCUCGUCGCCACCGCUCUU 3'
Sb02g020846.1 Predicted protein

sbi-miR5381
3' CGAGCCGCGGUGU-CUAGAA 5'
   ::::::-:::::::-:: ::: 
5' GCUCGG-GCCACAUGAGCUU 3'
Sb03g012700.1 similar to Putative dihydropterin pyrophosphokinase
/dihydropteroate synthase

sbi-miR5381
3' CGAGCCGCGGUGUCU-AGAA 5'
   ::::-:::::::::::- ::: 
5' GCUC-GCGCCACAGACACUU 3'
Sb03g026530.1 similar to ATP-dependent Zn proteases-like protein



sbi-miR5381
```

```
3' CGAGCCGCGGUGUCUAGAA 5'
   :: :::::::::  :::::
5' GCGCGGCGCCACUUAUCUU 3'
Sb04g008550.1 similar to Pentatricopeptide (PPR) repeat-containing
protein-like

sbi-miR5381
3' CGAGCCGCGGUGUCU-AGAA 5'
   :::::::::: : ::-::::
5' GCUCGGCGCCCCUGACUCUU 3'
Sb05g003270.1 similar to Eukaryotic-type carbonic anhydrase family
protein, expressed

sbi-miR5381
3' CGAGCCGCGGUGUCUAGA-A 5'
   :::::::::: ::: ::-:
5' GCUCGGCGCCAAAGAGCUCU 3'
Sb10g003875.1 weakly similar to H0215A08.3 protein

sbi-miR5381
3' CGAGC-CGCGGUGUCUAGAA 5'
   :::::- :::::-:::::::
5' GCUCGUUCGCCA-AGAUCUU 3'
Sb10g024900.1 similar to Putative uncharacterized protein
OSJNBa0019I19.51

Predicted targets of sbi-miR5382:

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   ::::::::::::::::::
5' AGGGCCUGUUUAGAUUGG 3'
Sb01g031390.1 similar to Putative uncharacterized protein

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   ::::::::::::::::::
5' AGGGCCUGUUUAGAUUGG 3'
Sb02g022960.2 weakly similar to Putative uncharacterized protein
OJ1506_A04.20

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   ::::::::::::::::::
5' AGGGCCUGUUUAGAUUGG 3'
Sb03g045940.1 similar to UMP synthase

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   ::::::::::::::::::
5' AGGGCCUGUUUAGAUUGG 3'
Sb06g020380.1 similar to H0510A06.16 protein

sbi-miR5382
3' UCCC-GGACAAAUCUAACC 5'
   : ::-::::::::::::::
5' AAGGCCCUGUUUAGAUUGG 3'
Sb02g018630.1 similar to Os09g0279300 protein
```

```
sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   :::::::::::::::: :
5' AGGGCCUGUUUAGAUUUG 3'
Sb10g014850.1 similar to Putative oxidative-stress responsive

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   ::::.:::::::::::: :
5' AGGGUCUGUUUAGAUUUG 3'
Sb01g001160.1 similar to Cytochrome P450 71E1

sbi-miR5382
3'  UCCCGGACAAAUCUAACC 5'
    :::: :::::::::-:::::
5' AAGGGGCUGUUUAG-UUGG 3'
Sb01g014880.1 similar to Putative uncharacterized protein

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
    ::::.::::::::: :::
5' UGGGCUUGUUUAGAGUGG 3'
Sb02g007080.1 similar to Putative uncharacterized protein

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   : ::::::::::.:::::
5' AAGGCCUGUUUGGAUUGU 3'
Sb10g005250.2 similar to Pyrophosphate-energized vacuolar membrane
proton pump (EC 3.6.1.1) (Pyrophosphate-energized inorganic
pyrophosphatase) (H(+)-PPase)

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   : ::::::::::.:::::
5' AAGGCCUGUUUGGAUUGU 3'
Sb10g005250.1 similar to Pyrophosphate-energized vacuolar membrane
proton pump (EC 3.6.1.1) (Pyrophosphate-energized inorganic
pyrophosphatase) (H(+)-PPase)

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
    :::::: ::::: ::::
5' UGGGCCUAUUUAGUUUGG 3'
Sb05g003890.1 similar to Expressed protein

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   ::: ::::::::. :::::
5' AGGCCCUGUUUGCAUUGG 3'
Sb08g021580.1 similar to Os07g0179300 protein

sbi-miR5382
3' UCCCGG-ACAAAUCUAACC 5'
   ::::::-::::::::-::::
5' AGGGCCUUGUUUAG-UUGG 3'
Sb09g002060.1 similar to Putative transcription factor IIIA
```

```
sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   :::: :::::::.: ::::
5' AGGGGCUGUUUGGUUUGG 3'
Sb09g024730.1 similar to Putative uncharacterized protein

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   : :::.:::::::::::: :
5' AAGGCUUGUUUAGAUUUG 3'
Sb06g017670.1 similar to 12-oxo-phytodienoic acid reductase

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   ::: :::::::::::: :
5' AGGCCCUGUUUAGAUCGC 3'
Sb02g025470.1 weakly similar to Putative uncharacterized protein

sbi-miR5382
3'  UCCCGGACAAAUCUAACC 5'
    :::::::::::::-:: :
5' UAGGGCCUGUUUAG-UUCG 3'
Sb02g026560.1 similar to Remorin protein-like

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   ::::::::::::::: :
5' AGGGCCUGUUUGGAUCGU 3'
Sb01g027960.1 similar to Xyloglucan endotransglucosylase/hydrolase
protein 28precursor, putative, expressed

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   :::: :::::::. :::::
5' AGGGACUGUUUGCAUUGG 3'
Sb01g027640.1 similar to Pleckstriny domain-containing protein,
putative, expressed

sbi-miR5382
3' UCC-CGGACAAAUCUAACC 5'
   :::- ::::::::: ::::
5' AGGCACCUGUUUAGCUUGG 3'
Sb01g037790.1 similar to Kinesin-4, putative, expressed

sbi-miR5382
3' UCCCGGACAAAU-CUAACC 5'
   :: :-:::::::-::::::
5' AGUG-CUGUUUAUGAUUGG 3'
Sb02g004740.1 similar to Putative ATP-dependent DNA helicase II 70 kDa
subunit

sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   :::::.::::::.:::: :
5' AGGGCUUGUUUGGAUUAG 3'
Sb03g030740.1 similar to Auxin response factor 2
```

```
sbi-miR5382
3' UCCCGGACAAAUCUAACC 5'
   ::: ::::::::: :::
5' AGGCCCUGUUUAGUUUGC 3'
Sb10g004720.2 similar to Putative uncharacterized protein

sbi-miR5382
3' UCCCGGACA-AAUCUAACC 5'
    ::::::::-:::::-:::
5' UGGGCCUGUAUUAGA-UGG 3'
Sb03g038020.2 similar to 2,3-bisphosphoglycerate-independent
phosphoglycerate mutase

sbi-miR5382
3' UCCCG-GACAAAUC-UAACC 5'
   : :::-:::::::::-:::::
5' AAGGCUCUGUUUAGCAUUGG 3'
Sb06g000572.1 similar to Putative uncharacterized protein

Predicted targets of sbi-miR5383:

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   ::::.::::..:::::::: ::::
5' AUAUUUCUGUUGGAGCUCUCUCAU 3'
Sb03g035680.1 similar to Putative prolylcarboxypeptidase, isoform 1

sbi-miR5383
3' UAUAGAGACGGCCUCGA-GACAGUA 5'
   ::::::.:::.:::::: ::: :::
5' AUAUCUUUGCUGGAGCUGCUGACAU 3'
Sb04g002415.1 weakly similar to Os02g0128300 protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   :::::: : ::: :::::::::.::
5' AUAUCUAUCCCGUAGCUCUGUUAU 3'
Sb02g036790.1 similar to Pentatricopeptide (PPR) repeat-containing
protein-like

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   :::::: : ::: :::::::::.::
5' AUAUCUAUCCCGUAGCUCUGUUAU 3'
Sb03g047110.1 similar to Pentatricopeptide (PPR) repeat-containing
protein-like

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   : ::: :: :.:::::.:::::::
5' AGAUC-CUUCUGGAGUUCUGUCAC 3'
Sb10g006300.1 similar to Os11g0622800 protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   ::::. :: :: :::.:::::::::
5' AUAUUGCU-CCUGAGUUCUGUCAU 3'
Sb04g002830.1 similar to Mitogen-activated protein kinase 13
```

```
sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   ::::.:.:::.  ::::::::::::
5' AUAUUUUUGCU-UAGCUCUGUCAU 3'
Sb02g037730.1 similar to Os07g0585100 protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACA-GUA 5'
    :: .:::::.:::::::::: :::
5' CUACUUCUGCUGGAGCUUUGUGCAU 3'
Sb02g024680.1 similar to Os09g0415800 protein

sbi-miR5383
3' UAUAGAG-ACGGCCUCGAGACAGUA 5'
   ::.:: : : :.::: :::::::::
5' AUGUCACAUCCUGGAACUCUGUCAU 3'
Sb10g010013.1 similar to Os03g0435200 protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   :::::::: ..::: :::::::.:
5' AUAUCUCUCUUGGAUCUCUGUUAG 3'
Sb01g029740.1 similar to Putative uncharacterized protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   .  :::. :::::::: :::::::
5' GCUUCUUCGCCGGAGCACUGUCAU 3'
Sb03g007990.1 Predicted protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   ::: :: : :::: :::.::::::
5' AUAGCU-UCCCGGUGCUUUGUCAU 3'
Sb08g021890.1 similar to Expressed protein

sbi-miR5383
3' UAUAGAGACGGCC-UCGAGACAGUA 5'
   :: :::.: :.:: ::::::::::
5' AUUUCUUUUCUGGAAGCUCUGUCAA 3'
Sb01g038930.1 similar to Expressed protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
    :::. :::: :::::::: :.::
5' CUAUUGCUGCAGGAGCUCUUUUAU 3'
Sb01g018220.1 weakly similar to Putative uncharacterized protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACA-GUA 5'
    ::::::: :. :::.:::::: :::
5' UUAUCUCUCCUUGAGUUCUGUACAU 3'
Sb07g022960.1 similar to DEAD-box ATP-dependent RNA helicase 16



sbi-miR5383
```

```
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   .: :. ::::.:::::: ::::
5' GUCUUCCUGCUGGAGCUC-GUCAC 3'
Sb10g028280.1 similar to Putative uncharacterized protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   ::. : :: :::::::::.::::.:
5' AUGAC-CUCCCGGAGCUUUGUUGU 3'
Sb02g037850.1 similar to Putative uncharacterized protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   ::.:.:.: : : ::::.::::::
5' AUGUUUUU-CAGAAGCUUUGUCAU 3'
Sb08g001550.1 Predicted protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   ::.:. .: :::::::::: .::
5' AUGUUGUUCCCGGAGCUCUGAUAU 3'
Sb02g036900.1 similar to Putative uncharacterized protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   . ::: .:::. :::::.::::::
5' GGAUCCUUGCU-GAGCUUUGUCAU 3'
Sb04g037420.1 similar to Os02g0819400 protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   . ::: .:::. :::::.::::::
5' GGAUCCUUGCU-GAGCUUUGUCAU 3'
Sb04g037420.2 similar to Os02g0819400 protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   ::: :: : .::: :::.::::::
5' AUAGCU-UCUCGGUGCUUUGUCAU 3'
Sb01g008777.1 similar to GTP-binding protein-like

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   : ::: .:: .::::.:: :::::
5' AAAUCCUUGAUGGAGUUCAGUCAU 3'
Sb07g020950.1 weakly similar to Os08g0430000 protein

sbi-miR5383
3' UAU-AGAGACGGCCUCGAGACAGUA 5'
   ::. ::::  :.: ::::::::::.:
5' AUGUUCUCAUCUGCAGCUCUGUCGU 3'
Sb10g010050.1 similar to Hd1

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   :: :::::::.:::::: ::: ::
5' AU-UCUCUGUUGGAGCCCUGCCAA 3'
Sb02g005862.1 Predicted protein
```

```
sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   .:::::  ::.::::::: :: :::
5' GUAUCU-GGCUGGAGCUAUGCCAU 3'
Sb01g040750.1 similar to Beta-galactosidase 6 precursor

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   :: :::.: .: ::::: ::::::
5' AUUUCUUU-UCAGAGCUGUGUCAU 3'
Sb03g031370.1 similar to Origin recognition complex subunit 4

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   .::: :::::.:: ::.:.:::::
5' GUAUAUCUGUCGAAGUUUUGUCAG 3'
Sb06g028560.1 similar to OSJNBa0008M17.16 protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   : ::::.:::.:::::::  :.::
5' AAAUCUUUGCUGGAGCUC-AUUAU 3'
Sb01g043800.1 weakly similar to Expressed protein

sbi-miR5383
3' UAUAGAGACGGCCUCGAGACAGUA 5'
   ::.: ::: :.::: ::::::: :
5' AUGUAUCUCCUGGAACUCUGUCUU 3'
Sb06g029335.1 similar to Transcriptional activator-like

Predicted targets of sbi-miR5384:

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
   .::::::.::::::::::
5' UCGCUGGGCGGCGGCGCG 3'
Sb02g025810.1 similar to Putative uncharacterized protein

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
     ::::::::::::::::
5' GAGCUGGACGGCGGCGCG 3'
Sb06g020045.1 similar to OSIGBa0106G07.15 protein

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
   :.::::::::.::::::
5' CUGCUGGACGGUGGCGCG 3'
Sb07g003120.1 weakly similar to Putative uncharacterized protein

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
   ::::::::::::::::::
5' CCGCUGGACGGCGGCGCC 3'
Sb02g039070.1 similar to Os07g0612400 protein

sbi-miR5384
```

```
3'  GGCGACCUGCCGCCGCGC 5'
    :::-::::::::::::::
5' UCCG-UGGACGGCGGCGCG 3'
Sb03g000870.1 similar to Putative uncharacterized protein P0485D09.14

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
   :.::::::.::::::::::
5' CUGCUGGGCGGCGGCGCG 3'
Sb01g014800.1_chromosome_1_sbi similar to Putative uncharacterized
protein

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
   :.: ::::::::::::::
5' CUGAUGGACGGCGGCGCG 3'
Sb10g008610.1 similar to Arm repeat-containing protein-like

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
    .:: ::::::::::::
5' GUGCGGGACGGCGGCGCG 3'
Sb03g035500.1 similar to Probable indole-3-acetic acid-amido synthetase
GH3.2

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
    : : ::::::::::::
5' ACACAGGACGGCGGCGCG 3'
Sb07g005890.1 similar to Os08g0205800 protein

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
    :::::::::::::: :
5' ACGCUGGACGGCGGCGAG 3'
Sb07g022750.1 similar to Nitrate reductase [NADH] 1

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
   .::::::::::::::: :
5' UCGCUGGACGGCGGCGAG 3'
Sb08g015320.1 similar to H0207B04.10 protein

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
     :::::.::::::::::
5' GGGCUGGGCGGCGGCGCG 3'
Sb09g022030.1 similar to Putative uncharacterized protein B1122D01.3

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
    : : ::::::::::::
5' ACCCAGGACGGCGGCGCG 3'
Sb07g005920.1 similar to Os08g0205800 protein



sbi-miR5384
```

```
3' GGCGACCUGCCGCCGCGC 5'
   ::::::::: ::::::::
5' CCGCUGGACCGCGGCGCG 3'
Sb07g006090.1 similar to Putative uncharacterized protein

sbi-miR5384
3'  GGCGACCUGCCGCCGCGC 5'
    :::-:::::::::::::.:
5' GCCG-UGGACGGCGGCGUG 3'
Sb03g002390.1 similar to Carnitine racemase-like

sbi-miR5384
3' GGC-G-ACCUGCCGCCGCGC 5'
   :::-:-:::::::::::::::
5' CCGUCGUGGACGGCGGCGCG 3'
Sb03g004020.1 similar to GA 3beta-hydroxylase

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
    ::: ::.::::::::::
5' GCGCAGGGCGGCGGCGCG 3'
Sb03g002460.1 similar to Putative uncharacterized protein P0451C06.8

sbi-miR5384
3' GGC-GACCUGCCGCCGCGC 5'
   : :-:::::::::::::::
5' CGGCCUGGACGGCGGCGCG 3'
Sb02g026180.1 weakly similar to Putative uncharacterized protein

sbi-miR5384
3' GG-CGACCUGCCGCCGCGC 5'
   ::-:::::::::::::.:
5' CCUGCUGGACGGCGGCGUG 3'
Sb01g046350.1 similar to Putative uncharacterized protein

sbi-miR5384
3'  GGCGACCUGCCGCCGCGC 5'
    :::- ::::::::::::
5' UCCG-GGGACGGCGGCGCG 3'
Sb02g001170.1 Predicted protein

sbi-miR5384
3'  GGCGACCUGCCGCCGCGC 5'
    :::-:::::::.::::::
5' UCCG-UGGACGGUGGCGCG 3'
Sb02g022220.1 similar to Polygalacturonase isoenzyme 1 beta subunit-
like

sbi-miR5384
3'-GGCGACCUGCCGCCGCGC-5'
   . ::::::::::::.:::
5'-UGGCUGGACGGCGGUGCG-3'
Sb02g023160.1 similar to Os09g0365900 protein



sbi-miR5384
```

```
3' GGCGACCUGCCGCCGCGC 5'
    ::: :::.::::::::::
5' GCGCCGGAUGGCGGCGCG 3'
Sb02g024750.1 similar to BHLH transcription factor PTF1-like protein

sbi-miR5384
3'  GGCGACCUGCCGCCGCGC 5'
    :: :-::::::::::::::
5' ACCAC-GGACGGCGGCGCG 3'
Sb02g033600.1 similar to Expansin-A26 precursor

sbi-miR5384
3' GGCGACCUGCCGCCGCGC 5'
    :  ::::::::::::::
5' ACCAUGGACGGCGGCGCG 3'
Sb01g006830.1 similar to AP2 domain containing protein, expressed

sbi-miR5384
3'  GGCGACCUGCCGCCGCGC 5'
    ::::-::::::::::::
5' GCCGC-GGACGGCGGCGCC 3'
Sb01g008810.1 similar to Putative uncharacterized protein 110K5.12


Predicted targets of sbi-miR5385:

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :.: :: : ::::::::::::::
5' GGGCAGGGAUGGGGUUGGUGGU 3'
Sb06g016600.1 weakly similar to Os06g0484800 protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :. .:::::::::: :::::::
5' GGUGAGCGGUGGGGAUGGUGGU 3'
Sb01g038870.1 weakly similar to Putative uncharacterized protein

sbi-miR5385
3' CUCUU-CGCCACCCCAACCACCA 5'
   :::.: : ::::: :::::::::
5' GAGGACGAGGUGGUGUUGGUGGU 3'
Sb03g005056.1 similar to PREDICTED: hypothetical protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :::.:::::::: : :::::::
5' GAGGAGCGGUGGUGGUGGUGGU 3'
Sb09g002110.1 similar to Os05g0121900 protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :.:. :.::::::::: ::::::
5' GGGGUGUGGUGGGGUGGGUGGU 3'
Sb01g031230.1 similar to Pollen ankyrin, putative, expressed


sbi-miR5385
```

```
3' CUCUUCGCCACCCCAACCACCA 5'
    .:::::::::: ::::::::
5' AGGAAGCGGUGGAGUUGGUGGC 3'
Sb10g024570.1 similar to Putative uncharacterized protein

sbi-miR5385
3' CUCUUC-GCCACCCCAACCACCA 5'
   :::.::  ::::: :::::::::
5' GAGGAGAGGGUGGUGUUGGUGGU 3'
Sb02g002890.1 similar to Os07g0147800 protein

sbi-miR5385
3' CU-CUUCGCCACCCCAACCACCA 5'
   :: :. :::::::: ::::::::
5' GACGGCGCGGUGGGCUUGGUGGU 3'
Sb06g000630.1 weakly similar to B0616E02-H0507E05.9 protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :.:::  :: ::::::::::::
5' GGGAAAGGGAGGGGUUGGUGGU 3'
Sb04g009090.1 similar to Os02g0250400 protein

sbi-miR5385
3' CUCUUCGCC-ACCCCAACCACCA 5'
   :::.:: :: ::::: :::::::
5' GAGGAGGGGAUGGGGAUGGUGGU 3'
Sb03g026730.1 similar to Extensin-like

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :.::.:.::: ::::::::::
5' GGGAGGUGGU-GGGUUGGUGGC 3'
Sb03g008050.1 similar to Fructose-bisphosphate aldolase

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :. ..:.::::::: :::::::
5' GGUGGGUGGUGGGGAUGGUGGU 3'
Sb01g017795.1 similar to AGAP001055-PA

sbi-miR5385
3' CUC-UUCGCCACCCCAACCACCA 5'
   ::: : : :::: ::::::::::
5' GAGCAUGAGGUGAGGUUGGUGGU 3'
Sb07g000520.1 similar to Putative uncharacterized protein

sbi-miR5385
3' CUC-UUCGCCACCCCAACCACCA 5'
   ::: : : :::: ::::::::::
5' GAGCAUGAGGUGAGGUUGGUGGU 3'
Sb07g000530.1 similar to Putative uncharacterized protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
    .::.::  ::::::::::::
5' CGGAGGC-CUGGGGUUGGUGGG 3'
Sb03g014130.1 similar to Multidrug resistance-associated protein MRP1
```

```
sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   : :::::  :::: ::::::::
5' GUGAAGC-UUGGGAUUGGUGGU 3'
Sb09g030200.1 similar to Putative uncharacterized protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :.: .::::::::: ::::::
5' GGGCGGCGGUGGGGGUGGUGGG 3'
Sb02g006180.1 similar to Putative uncharacterized protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :::::: :: :: :::::::::
5' GAGAAGGGGAGGAGUUGGUGGU 3'
Sb03g029220.1 weakly similar to Integral membrane protein-like

sbi-miR5385
3' CUCUUCG-CCACCCCAACCACCA 5'
   ::::::: ::::: : :::::::
5' GAGAAGCAGGUGGUGGUGGUGGU 3'
Sb08g021020.1 Predicted protein

sbi-miR5385
3' CUCU-UCGCCACCCCAACCACCA 5'
   : :: .:.:::::::: :::::::
5' GUGAUGGUGGUGGGGAUGGUGGU 3'
Sb03g019110.1 Predicted protein

sbi-miR5385
3' CUCU-UCGCCACCCCAACCACCA 5'
   : :: .:.:::::::: :::::::
5' GUGAUGGUGGUGGGGAUGGUGGU 3'
Sb04g010710.1 Predicted protein

sbi-miR5385
3' CUCUUCG-CCACCCCAACCACCA 5'
   :.: .:: :::::::::: ::::
5' GGGCGGCGGGUGGGGUUGCUGGU 3'
Sb05g006550.1 Predicted protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :.::.:: :::::: ::::::
5' GGGAGGC-GUGGGGGUGGUGGA 3'
Sb10g008870.1 similar to Os02g0738900 protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
    .: .:::::::::::::: :::
5' CGGCGGCGGUGGGGUUGG-GGU 3'
Sb03g026050.1 similar to Putative uncharacterized protein B1097D05.32



sbi-miR5385
```

```
3' CUCUUCGCCACCCCAACCACCA 5'
   :: :::: :::::: ::::::
5' GAUAAGCAGUGGGGAUGGUGGC 3'
Sb09g025110.2 similar to Os05g0510300 protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :: :::: :::::: ::::::
5' GAUAAGCAGUGGGGAUGGUGGC 3'
Sb09g025110.1 similar to Os05g0510300 protein

sbi-miR5385
3' CUCUUCG-CCACCCCAACCACCA 5'
    :: ::: ::::::: :::::::
5' AAGUAGCAGGUGGGGGUGGUGGU 3'
Sb03g007020.1 similar to Putative uncharacterized protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   : :::::::: ::::::: :::
5' GCGAAGCGGUAGGGUUGG-GGU 3'
Sb02g040600.1 similar to Putative uncharacterized protein

sbi-miR5385
3' CUCUUCGCCACCCCAACCACCA 5'
   :. .:: :::::::: ::::::
5' GGCGAG-GGUGGGGUCGGUGGU 3'
Sb03g005230.1 Predicted protein

sbi-miR5385
3' CUCUUCGCC-ACCCCAACCACCA 5'
   :::: : :: :::::: ::::::
5' GAGACGGGGAUGGGGUGGGUGGU 3'
Sb01g007370.1 similar to Chromosome chr1 scaffold_5, whole genome
shotgun sequence

Predicted targets of sbi-miR5386*:

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   ::::  : ::::::: ::::::::.
5' CGAGACGGGCGUGCUGUGCCCUGAU 3'
Sb07g003540.1 similar to Putative uncharacterized protein P0702G08.11

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
    :::: ::::::::.:::: :::::
5' GGAGGAGUGCGUGUUCUG-CCUGAG 3'
Sb02g043600.1 Predicted protein

sbi-miR5386*
3' GCUCCCCACGCACG-AGACGGGACUG 5'
   : ::::::::: ::  ::::::::::
5' CCAGGGGUGCGGGCGGCUGCCCUGAC 3'
Sb02g015770.1 similar to Putative uncharacterized protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
```

```
   ::::: : ::::::::  ::.::::
5' CGAGGAGGGCGUGCUCGACCUUGAC 3'
Sb06g027150.1 similar to OSIGBa0142I02-OSIGBa0101B20.26 protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   ::::: :: :: ::::::: .::.:
5' CGAGGAGUUCGAGCUCUGCGUUGGC 3'
Sb06g033000.1 similar to OSJNBa0064G10.23 protein

sbi-miR5386*
3' GCU-CCCCACGCACGAGACGGGACUG 5'
   ::: :: :: ::::::::::. :.:
5' CGAUGGUGUCCGUGCUCUGCCUCGGC 3'
Sb01g001900.1 similar to Putative uncharacterized protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   ::. : :::: :: ::::::::::..
5' CGGUGAGUGCCUGAUCUGCCCUGGU 3'
Sb03g041690.1 similar to Leaf senescence related protein-like

sbi-miR5386*
3' GCUCCCCACG-CACGAGACGGGACUG 5'
   .:::  :::: ::::::::::.::: .
5' UGAGUAGUGCUGUGCUCUGCUCUGCU 3'
Sb03g000230.1 Predicted protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   . :: :::: :::. :::::::::.
5' UUAGUGGUG-GUGUGCUGCCCUGAU 3'
Sb03g004790.1 similar to Os05g0165400 protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   .:: : :::::::: :: ::.::::
5' UGAUGAGUGCGUGCGCUUCCUUGAC 3'
Sb01g023200.1 similar to Chromosome chr6 scaffold_3, whole genome
shotgun sequence

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   ::. :: ::::::::::::.  :::
5' CGGCGGAUGCGUGCUCUGCUGCGAC 3'
Sb09g026010.1 weakly similar to Chromosome chr6 scaffold_3, whole
genome shotgun sequence

sbi-miR5386*
3' GCU-CCCCACGCACGAGACGGGACUG 5'
   ::: ::::::  ::::::::.:. ::.
5' CGACGGGGUGGAUGCUCUGUCUGGAU 3'
Sb05g020290.1 similar to Expressed protein



sbi-miR5386*
```

```
3' GCUCCCCACGCACGAGACGGGACUG 5'
   ::.: ::::: ::::: ::: ::.:
5' CGGGCGGUGCUUGCUC-GCCGUGGC 3'
Sb08g020180.1 similar to Expressed protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   ::::: :::::::: ::::  ::.:
5' CGAGGCGUGCGUGCGCUGCAAUGGC 3'
Sb04g030400.1 similar to Putative uncharacterized protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   :: :: ::::::::. ::: :::::.:
5' CGUGGAGUGCGUGUACUG-CCUGGC 3'
Sb02g033330.1 similar to Putative uncharacterized protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   : :::::: :::::: :::.: :.:
5' CAAGGGGUACGUGCUGUGCUC-GGC 3'
Sb10g028890.1 similar to Putative uncharacterized protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
   :: ::::: ::::.:::: ::: .:
5' CGCGGGGU-CGUGUUCUGGCCUCGC 3'
Sb07g004000.1 Predicted protein

sbi-miR5386*
3' GCUCCCCACGCACGAGACGGGACUG 5'
    :::::::: : :::. ::.::::.:
5' GGAGGGGUGGGGGCUUAGCUCUGGC 3'
Sb02g009640.1 similar to Chromosome chr2 scaffold_176, whole genome
shotgun sequence

Predicted targets of sbi-miR5387:

sbi-miR5387
3' CUAGGAAAU-CGUGGCCAAGCACAAU 5'
   ::::.:::: :.:: ::::.::::::
5' GAUCUUUUAUGUACAGGUUUGUGUUA 3'
Sb07g004285.1 similar to Amino acid permease family protein, putative,
expressed

sbi-miR5387
3' CUAGGAAAUCGU-GGCCAAGCACAAU 5'
   :: :.::::. :: .:::::::::::::
5' GACCUUUUGCCAGUCGGUUCGUGUUA 3'
Sb05g004310.1 similar to Putative uncharacterized protein

sbi-miR5387
3' CUAGGAAAUCGU-GGCCAAGCACAAU 5'
   :: :.::::. :: .:::::::::::::
5' GACCUUUUGCCAGUCGGUUCGUGUUA 3'
Sb05g004390.1 similar to Putative uncharacterized protein

sbi-miR5387
```

```
3' CUAGGAAAUCGUGGCCAAGCACAAU 5'
   :::::.:   :::::::::: ::::.
5' GAUCUUCAUGCACCGGUUCCUGUUG 3'
Sb09g027440.1 similar to P0482D04.17 protein

sbi-miR5387
3' CUAGGAAAUCGUGGCCAAGCACAAU 5'
   :::::.:   :::::::::: ::::.
5' GAUCUUCAUGCACCGGUUCCUGUUG 3'
Sb03g044530.1 similar to P0482D04.17 protein

sbi-miR5387
3' CUAGGAAAUCGU-GGCCAAGCACAAU 5'
   :: :.::::. :: ..:::::::::::
5' GACCUUUUGCCAGUUGGUUCGUGUUA 3'
Sb05g004370.1 similar to Putative uncharacterized protein

sbi-miR5387
3' CUAGGAAAUCG-UGGCCAAGCACAAU 5'
    .::::: ::: :::::: ::::::.
5' CGUCCUUCAGCGACCGGUACGUGUUG 3'
Sb06g025980.1 similar to OSIGBa0147H17.8 protein

sbi-miR5387
3' CUAGGAAAUCGUGGCC-AAGCACAAU 5'
   :.::::::   :::::: :::::::::.
5' GGUCCUUCUCCACCGGCUUCGUGUUG 3'
Sb09g022500.1 similar to P0482D04.17 protein

sbi-miR5387
3' CUAGGAAAUCGUGGCCAAGCACAAU 5'
   :.::: ::: :::: ::::::: ::
5' GGUCCAUUA-CACCAGUUCGUGAUA 3'
Sb02g025030.1 similar to Streptococcal hemagglutinin-like protein

sbi-miR5387
3' CUAGGAAAUCGUGGCCAAGC-ACAAU 5'
   :::.:::: :::  :::::: ::::.
5' GAUUCUUUUGCAAGGGUUCGUUGUUG 3'
Sb05g004200.1 weakly similar to H0311C03.6 protein

sbi-miR5387
3' CUAGGAAAUCGUGGCCAAGCACAAU 5'
   :: ::: :::.::: ::: :::::.
5' GACCCUCUAGUACCUGUU-GUGUUG 3'
Sb07g028060.1 Predicted protein

sbi-miR5387
3' CUAGGAAAUCGUGGCCAAGCACAAU 5'
   :: ::: :::.::: ::: :::::.
5' GACCCUCUAGUACCUGUU-GUGUUG 3'
Sb07g028060.2 Predicted protein

sbi-miR5387
3' CUAGGAAAUCGUGGCCAAGCACAAU 5'
   :::::::::: .:: :::::: : :
5' GAUCCUUUAGGGCCCGUUCGUUUCA 3'
Sb04g007190.1 similar to Putative fatty acid elongase
```

Predicted targets of sbi-miR5388:

```
sbi-miR5388
3' CAG-UCUCUGUGGGCCGUUUCUA 5'
   ::: ::::::::::::::.:.::
5' GUCAAGAGACACCCGGCGAGGAC 3'
Sb0010s010920.1 putative protein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   :: :::: :::::::::: :::
5' GUGAGAG-CACCCGGCAACGAU 3'
Sb05g002160.1 Predicted protein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
     : :: ::::::::::::::::
5' AACUGAAACACCCGGCAAAGAU 3'
Sb03g001720.1 similar to Tetratricopeptide repeat protein-like

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
    :.. ::::::.::::::::::
5' CUUGAAGACACUCGGCAAAGAG 3'
Sb10g020140.1 weakly similar to Putative uncharacterized protein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   : .: : :.:::.::::::::::
5' GAUA-ACAUACCUGGCAAAGAU 3'
Sb08g018460.1 similar to Chromosome chr14 scaffold_27, whole genome
shotgun sequence

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   :  :::::::::: :::::::
5' GAAAGAGACACCC-GCAAAGAA 3'
Sb04g028080.1 weakly similar to Putative uncharacterized protein

sbi-miR5388
3' CAGUCUCUG-UGGGCCGUUUCUA 5'
   :::::.:.: .:::::::::.::
5' GUCAGGGGCUGCCCGGCAAGGAA 3'
Sb08g020270.1 similar to WRKY DNA binding domain containing protein,
expressed

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   :  :::::::::: :::::::
5' GAAAGAGACACCC-GCAAAGAA 3'
Sb05g003710.1 similar to Transposon protein, putative, CACTA, En/Spm
sub-class



sbi-miR5388
```

```
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   : :::: :::: :::::::::
5' GACAGA-ACACACGGCAAAGAC 3'
Sb03g043640.1 weakly similar to Chromosome chr5 scaffold_67, whole
genome shotgun sequence

sbi-miR5388
3' CAG-UCUCUGUGGGCCGUUUCUA 5'
   :::  :::..:::.:::.:::::
5' GUCUUGAGGUACCUGGCGAAGAU 3'
Sb02g029010.1 similar to Os09g0500900 protein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   : ::: :: ::::::::.:.:::
5' GGCAGCGA-ACCCGGCGAGGAU 3'
Sb01g017610.1 weakly similar to Putative uncharacterized protein
OSJNBb0015K05.12

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
    :.::::: ..:.:::::::::
5' AUUAGAGAGGUCUGGCAAAGAU 3'
Sb09g024060.1 similar to Beta subunit 2 of SnRK1

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
    :::::  ..:::::.::::::
5' UUCAGAACUGCCCGGUAAAGAU 3'
Sb06g030515.1 similar to Putative gag-pol polyprotein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   ::. :::: :.:::: ::::::
5' GUU-GAGAAAUCCGGAAAAGAU 3'
Sb02g043500.1 similar to Beta subunit 1 of SnRK1

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   :: ::::. ::::::::..:::
5' GUGAGAGG-ACCCGGCGGAGAG 3'
Sb05g025540.1 similar to Leucine Rich Repeat family protein, expressed

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   :::::. ::.:::::::::: .:
5' GUCAGGUACGCCCGGCAAACGU 3'
Sb05g003100.1 similar to Cytochrome P450 family protein, expressed

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   : :. :::::::::::::.::.
5' GGCGUAGACACCCGGCAGAGGG 3'
Sb09g017250.1 similar to Putative uncharacterized protein
OSJNBa0036C12.16


sbi-miR5388
```

```
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   ::::::.:: :::.:: .:::::
5' GUCAGGGAAACCUGG-GAAGAU 3'
Sb09g022360.1 similar to Os01g0847700 protein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   :::.  :: ::::::::.:.:::
5' GUCGUCGA-ACCCGGCGAGGAU 3'
Sb02g005470.1 weakly similar to Putative uncharacterized protein

sbi-miR5388
3' CAGUCUCUGUGG-GCCGUUUCUA 5'
   ::::::. ::::: ::::::.:::
5' GUCAGGAACACCACGGCAGAGAA 3'
Sb03g027420.1 similar to Putative uncharacterized protein
OJ1723_B06.129

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   :::.:.:. ::::: :.::::::
5' GUCGGGGG-ACCCGCCGAAGAU 3'
Sb10g009690.1 similar to Expressed protein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
     ::::::::::.::::. :::
5' AGCAGAGACACCUGGCAGUGAU 3'
Sb03g022870.1 similar to Putative uncharacterized protein B1074C08.29

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
    :.:: ::.:.:.:::::::::
5' AUUAG-GAUAUCUGGCAAAGAG 3'
Sb02g042000.1 similar to Os07g0661500 protein

sbi-miR5388
3' CA-GUCUCUGUGGGCCGUUUCUA 5'
   :: :::::: ::..:: ::::::
5' GUGCAGAGAAACUUGGAAAAGAU 3'
Sb07g026450.1 Predicted protein

sbi-miR5388
3' CA-GUCUCUGUGGGCCGUUUCUA 5'
   :: :::::: ::..:: ::::::
5' GUGCAGAGAAACUUGGAAAAGAU 3'
Sb07g025785.1 Predicted protein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   ::..   ::::.:::::::::::
5' GUUGCUCACACUUGGCAAAGAU 3'
Sb02g043380.1 similar to Os05g0239200 protein



sbi-miR5388
```

```
3' CAGUCUCUGUGGGCCGUUUCUA 5'
    :.:: ::.:: :::::.:::: 
5' UUUAG-GAUACACGGCAGAGAU 3'
Sb03g004270.1 similar to Chromosome chr11 scaffold_13, whole genome
shotgun sequence

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
     :::: ::::::::::.::.
5' AACAGA-ACACCCGGCAGAGGA 3'
Sb02g006440.1 similar to MutT-like protein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   :::.::: . :::::: ::::: 
5' GUCGGAGCUUCCCGGC-AAGAU 3'
Sb08g007720.1 similar to Os12g0231000 protein

sbi-miR5388
3' CAGUCUCUGUGGGCCGUUUCUA 5'
   : ::: :::.:: :::::.::: 
5' GGCAG-GACGCCAGGCAAGGAU 3'
Sb10g003730.1 similar to Os06g0152700 protein
```

Figure 4.9. Predicted targets for the newly discovered miRNAs in sorghum. This figure displays a list of alignments btween the newly discovered microRNAs and their predicted target sequences in sorghum.

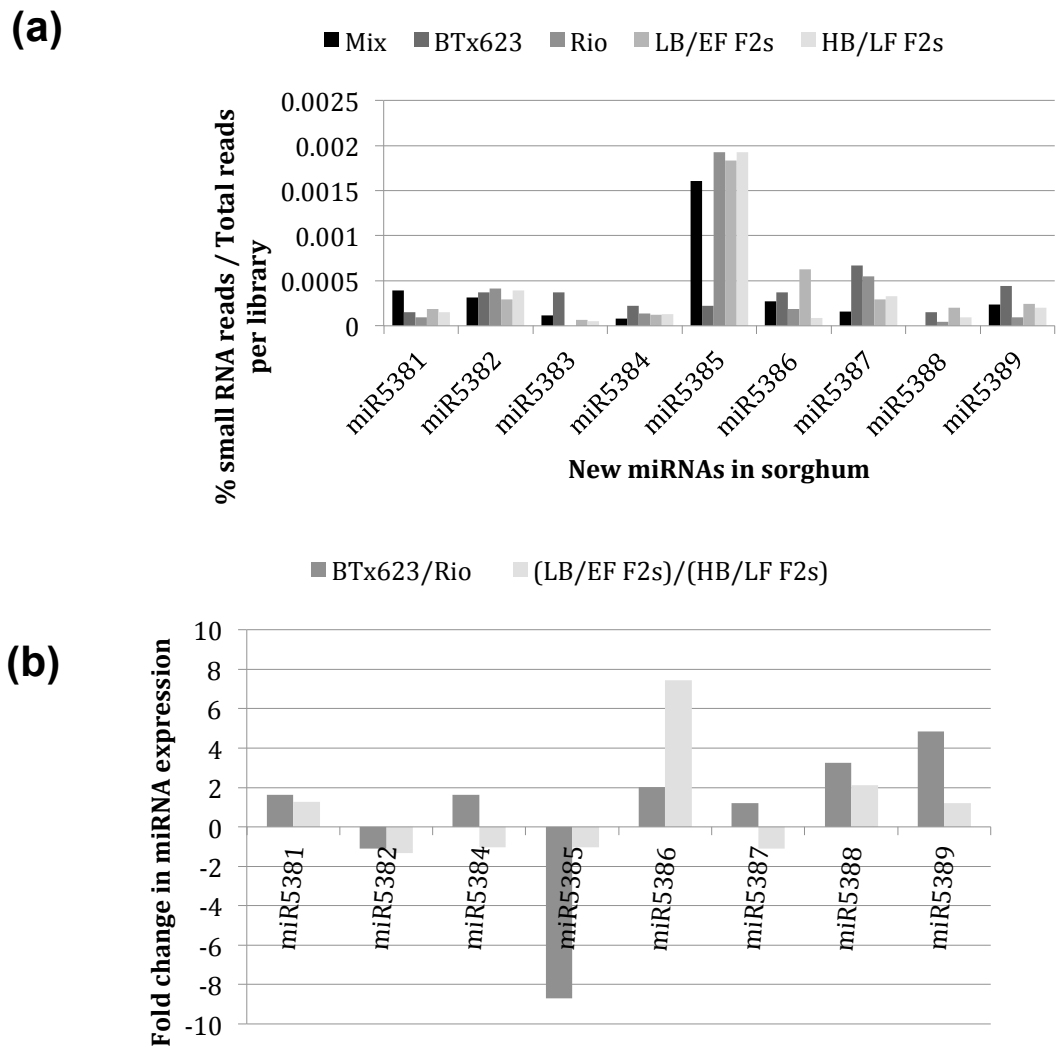


Figure 4.10. Genotypic variations in the expression of new miRNAs. **(a)** The frequency count of small RNAs for each new miRNA was used to calculate its abundance. **(b)** The miRNA abundances were used to calculate their relative fold change in expression between BTx623 and Rio, and between the LB/EF F2s and HB/LF F2s libraries, respectively. Positive values in the y-axis of the graph denote fold changes in miRNA expression that are higher in BTx623 relative to Rio and higher in LB/EF F2s relative to HB/LF F2s libraries, respectively; the opposite is

true for negative values. The miRNA miR5383 was not included in the graph because it was not detected in the Rio library (see Table 4.6).

4.4. Discussion

Here we have described the first characterization of the small RNA component of the transcriptome from sorghum stems. The choice of stems as plant material was interesting not only because it was the tissue were fermentable sugars did accumulated, but it was also the venue for the movement of small RNA duplexes (siRNAs and miRNAs) from source to sink tissues, as it was recently demonstrated (Dunoyer et al., 2010; Molnar et al., 2010). Thus, we could expect the small RNA component of the stem to be quite diverse or heterogeneous. Indeed, the unexpected finding of a high abundance peak of RNAs with 25 nt or more in length lead me to the finding of rRNA and tRNA genes that have not been annotated yet in the sorghum genome. We also shown that the abundance of the 22 nt small RNAs in sorghum stem tissue was greater than the 20 and 21 nt small RNAs respectively. My results contrasted the recently proposed notion that the 22 nt peak of small RNAs was exclusive of maize (Nobuta et al., 2008). Furthermore, we found that up to 15% of all the 22 nt small RNAs in the BTx623 library were derived from miR172c, which was previously predicted to have a length of 20 nt (Paterson et al. 2009). Recently, 22 nt miRNAs were described to trigger siRNA biogenesis from target transcripts in *Arabidopsis* (Chen et al., 2010; Cuperus et al., 2010). Thus, it would be interesting to test if miR172c can also trigger siRNA biogenesis in sorghum.

As expected, the specific genetic material, tissue sample and developmental

stage used in our study, allowed us to capture a broad spectrum of the small RNA component of the sorghum transcriptome. On the other hand, the specificity of the material also permitted us to gain new insights into how complex traits like sugar accumulation and flowering time might be regulated at the post-transcriptional level. Such regulation of gene expression could provide an opportunity to manipulate biofuel traits, where stem sugar rather than cellulose and increased biomass because of delayed flowering could be enhanced (Torney et al., 2007). By taking a genetic approach in conjunction with deep-sequencing of stem-derived small RNAs, we was able to correlate variation in miRNA expression between grain and sweet sorghum, with the sugar and flowering phenotypes of selected F2 plants derived from their cross. However, analysis of the differential accumulation of potential target genes did not exhibit a simple correlation with miRNA levels. Therefore, further studies will be required to unveil the underlying mechanisms between genotype and phenotype.

In the case of miR395, it is interesting to note that there was genotypic variation in the miR395/miR395* ratio, with the Rio genotype expressing both strands at equal proportions in contrast to a clear predominance of miR395 abundance over miR395* in BTx623 (Figure 4b). This is reminiscent of the recently proposed "arm switching" model of miRNA evolution described for nematodes species (de Wit et al., 2009), in which the mature miRNA is produced from the 5' arm of the miRNA hairpin in a particular species but in a different nematode species the 5' arm of the same MIR gene gives rise to the miRNA* instead. Interestingly, it has been shown recently that miRNA* species have physiological relevance in

Drosophila, since a significant number of them are well conserved, can be loaded into the RISC complex through their preferential association with ARGONAUTE2 (AGO2) rather that AGO1, and can also regulate the expression of target genes (Ghildiyal et al., 2010). Furthermore, the regulatory potential of miRNA* species in vertebrates has been recently demonstrated as well (Yang et al., 2011).

4.5. Conclusions

Based on the above, several interesting questions can be formulated. First, does miR395* have any regulatory potential? Second, what is the mechanism behind the genotypic difference in miR395/miR395* ratio? Third, is this ratio altered in a developmental and/or tissue dependent manner? Fourth, is this an example of a general phenomenon? If this is the case, we would envision that other miRNAs families as well will display differences in their miRNA/miRNA* ratio dependent on the genotype and/or condition. Future work will be required to provide a better understanding of miR395' s involvement in processes other than its previously described role in sulfur metabolism.

4.6. Methods

4.6.1. Plant material

The grain (BTx623) and sweet (Rio) sorghum cultivars together with F2 plants derived from their cross were grown in the field of the Waksman Institute during the summer of 2008. The juice from three internodes of the main stem was harvested at the time of flowering and the Brix degree measured as previously

described (Calviño et al., 2008; Calviño et al., 2009). The average Brix degree from three internodes per plant was used. Flowering time was measured as the number of leaves in the main stem at the time of anthesis. In total, 15 plants for each parent and 553 F2 plants were scored for Brix degree and flowering time. The F2 plants selected for sequencing had either low Brix (Brix ≤ 5)/early flowering (N0 leaves ≤ 9) or high Brix (Brix ≥ 13)/late flowering (NO leaves ≥ 14).

4.6.2. Construction of small RNA libraries

Total RNA from internode tissue was extracted at the time of flowering with the mirVana miRNA isolation kit (Ambion). RNA extraction was performed in 5 independent plants for each BTx623 and Rio, and 11 independent plants for each low Brix/early flowering and high Brix/late flowering F2 plants respectively. The total RNA (1 **μ** g per sample) was pooled and then fractionated with the flash-Page fractionator (Ambion) to isolate RNAs smaller that 40 nt in length. The isolated small RNAs were used to construct small RNA cDNA libraries with the SOLiD small RNA library construction kit (Ambion). The sequencing was carried out at the Waksman genomics laboratory on the SOLiD 3 platform, which has a read length limit of 25 nt http://solid.rutgers.edu.

4.6.3. Bioinformatics analysis

We mapped the 25 nt long reads to the sorghum genome using the SHRiMP program version 1.0.5 (Rumble et al., 2009), with default parameter settings except that the number of matches was limited to 10. SHRiMP allowed us to perform the

alignments in SOLiD' s colorspace. For the further analyses we used only alignments that matched perfectly to the genome starting from the first position in the read up to the sequencing primer. Because the SOLiD 3 platform had a read length limit of 25 nucleotides, adaptor sequences did not have to be trimmed prior mapping to the genome. As a consequence, we could estimate the length of an individual sequence read by one base with a probability of 0.25. These reads were then clustered with Vmatch http://vmatch.de/ to reduce the number of identical reads for downstream analyses. We required 100% identity among the sequences of a cluster. We have further filtered the clustered reads against the repetitive elements of sorghum and used the remaining sequences for de novo prediction of miRNA using miRDeep. We defined a 25 nt " hotspot" as those loci in the genome that displayed a high coverage of 25 nt reads, in our case thousand reads. The length of the hotspot was determined as each consecutive interrogated base that had more than thousand 25 nt reads mapped to it.

4.6.4. Quantification of miRNA expression

The TaqMan MicroRNA Assays (Applied Biosystems) was used to quantify the expression of miR172a, and the Custom TaqMan Small RNA Assays (Applied Biosystems) was used to quantify the expression of miR169d and miR395f respectively. The qRT-PCR reaction was done using the MyiQ Real-Time PCR Detection System (BIO-RAD Laboratories, Inc.). A relative quantification normalized against unit mass (10 ng total RNA) was performed as previously described (Calviño et al., 2008).

4.6.5. *De novo* discovery of sorghum miRNAs

For de novo prediction of potential miRNAs, we have used the miRDeep package (Friedlander et al., 2008). As miRDeep does not take colorspace alignment as input, we had to reshape the output to miRDeep' s blastparse format. Moreover, the SHRiMP alignment scores and the score used had to be recalculated to fit miRDeep' s blastparse format. We used the same formula and method as described (Goff et al., 2009). At this point, we also had to translate the color space two base encoding sequences into standard nucleotide base space sequences. As we considered only perfectly matching reads after the initial alignment to the genome, we could easily translate from color space to base space sequence format. The subsequent de novo calling of miRNAs was carried out as described (Friedlander et al., 2008; Goff et al., 2009). Finally, the coordinates of de novo miRNAs that were predicted on the minus strand were corrected as miRDeep refers the coordinates to the 5' end of the minus strand. Though, conventionally the coordinates refer always to the 5' end of the plus strand. The GenBank accession numbers for the new miRNAs are sbi-MIR538.sqn sbi-MIR5381 JN205291; sbi-MIR538.sqn sbi-MIR5382 JN205292; sbi-MIR538.sqn sbi-MIR5383 JN205293; sbi-MIR538.sqn sbi-MIR5384 JN205294; sbi-MIR538.sqn sbi-MIR5385 JN205295; sbi-MIR538.sqn sbi-MIR5386 JN205296; sbi-MIR538.sqn sbi-MIR5387 JN205297; sbi-MIR538.sqn sbi-MIR5388 JN205298; sbi-MIR538.sqn sbi-MIR5389 JN205299.

We have also validated all potential new miRNAs according to the annotation criteria proposed by (Meyers et al., 2008).

4.6.6. Target prediction and validation

We have used the novel miRNAs for a target prediction. Firstly, we compared the sequences to the unspliced transcripts of sorghum (Paterson et al., 2009), with BLASTN using these parameters: -F F -W 7 -e 1 -q -2 -G -1. We scored each base of the alignment according to these criteria: match as 0; GU pairs as 0.5; gaps as 2; all other pairs were scored as 1. We doubled the score within the first 13 bases of the miRNA/alignment. We considered the gene as a potential target if the total score of the alignment was equal to or less than 8.

The miRNA-mediated cleavage of mRNAs was performed through a modified procedure of the RLMRACE protocol from Invitrogen. The sequences of the reverse primers used in the modified RACE are: Sb01g044240 (5' GCCCATATGGACGGAAGATA 3'); Sb02g007000 (5' CTGGTAGCCGGAGAACAACT 3') and Sb06g030670 (5' TTTCATCAGTGCTTGCCAAT 3'). The validation of predicted targets was performed in BTx623 or Rio cultivars only. Annotation of the miRNA gene targets was based on the Phytozome database http://www.phytozome.net.

4.7. References

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

Calviño M, Bruggmann R, Messing J (2008) Screen of Genes Linked to High-Sugar Content in Stems by Comparative Genomics. Rice **1:** 166-176.

Calviño M, Miclaus M, Bruggmann R, Messing J (2009) Molecular Markers for Sweet Sorghum Based on Microarray Expression Data. Rice **2:** 129-142.

Chen H, Chen L, Patel K, Li Y, Baulcombe D, Wu S (2010) 22-Nucleotide RNAs trigger secondary siRNA biogenesis in plants. Proceedings of the National Academy of Sciences **107:** 15269-15274.

Chuck G, Candela H, Hake S (2009) Big impacts by small RNAs in plant development. Current Opinion in Plant Biology **12:** 81-86.

Chuck G, Meeley R, Irish E, Sakai H, Hake S (2007) The maize tasselseed4 microRNA controls sex determination and meristem cell fate by targeting Tasselseed6/indeterminate spikelet1. Nature Genetics **39:** 1517-1521.

Cuperus J, Carbonell A, Fahlgren N, Garcia-Ruiz H, Burke R, Takeda A, Sullivan C, Gilbert S, Montgomery T, Carrington J (2010) Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. Nature Structural Molecular Biology **17:** 997-1003.

De Wit E, Linsen S, Cuppen E, Berezikov E (2009) Repertoire and evolution of miRNA genes in four divergent nematode species. Genome Research **19:** 2064-2074.

Dunoyer P, Schott G, Himber C, Meyer D, Takeda A, Carrington J, Voinnet O (2010) Small RNA duplexes function as mobile silencing signals between plant cells. Science **328:** 912-916.

Filipowicz W, Bhattacharyya S, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Review Genetics **9:** 102-114.

Friedlander M, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nature Biotechnology **26:** 407-415.

Ghildiyal M, Xu J, Seitz H, Weng Z, Zamore P (2010) Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. RNA **16:** 43-56.

Glasziou K, Gayler R (1972) Storage of sugars in stalks of sugarcane. The Bottanical Review **38:** 471-490.

Goff L, Davila J, Swerdel M, Moore J, Cohen R, Wu H, Sun Y, Hart R (2009) Ago2 immunoprecipitation identifies predicted microRNAs in human embryonic stem cells and neural precursors. PloS One **4:** e7192.

Goldemberg J (2007) Ethanol for a sustainable energy future. Science **315:** 808-810.

Grivet L, Arruda P (2002) Sugarcane genomics: depicting the complex genome of an important tropical crop. Current Opinion in Plant Biology **5:** 122-127.

Heisel S, Zhang Y, Allen E, Guo L, Reynolds T, Yang X, Kovalic D, Roberts J (2008) Characterization of unique small RNA populations from rice grain. PloS One **3:** e2871.

Henderson I, Zhang X, Lu C, Johnson L, Meyers B, Green P, Jacobsen S (2006) Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. Nature Genetics **38:** 721-725.

Hoffman-Thoma G, Hinkel K, Nicolay P, Willenbrink J (1996) Sucrose accumulation in sweet sorghum stem internodes in relation to growth. Physiologia Plantarum **97:** 277-284.

Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann J, Arruda P, D'Hont A (2007) Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. Plant Journal **50:** 574-585.

Kawashima C, Yoshimoto N, Maruyama-Nakashita A, Tsuchiya Y, Saito K, Takahashi H, Dalmay T (2009) Sulphur starvation induces the expression of microRNA-395 and one of its target genes but in different cell types. Plant Journal **57:** 313-321.

Khraiwesh B, Arif M, Seumel G, Ossowski S, Weigel D, Reski R, Frank W (2010) Transcriptional control of gene expression by microRNAs. Cell **140:** 111-122.

Lauter N, Kampani A, Carlson S, Goebel M, Moose S (2005) microRNA172 down-regulates glossy15 to promote vegetative phase change in maize. Proceedings of the National Academy of Science **102:** 9412-9417.

Lee H, Yoo S, Lee J, Kim W, Yoo S, Fitzgerald H, Carrington J, Ahn J (2010) Genetic framework for flowering-time regulation by ambient temperature-responsive miRNAs in Arabidopsis. Nucleic Acids Research **38:** 3081-3093.

Lee Y, Kim M, Han J, Yeom K, Lee S, Baek S, Kim V (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO Journal **23:** 4051-4060.

Li W, Oono Y, Zhu J, He X, Wu J, Iida K, Lu X, Cui X, Jin H, Zhu J (2008) The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. Plant Cell **20:** 2238-2251.

Lu C, Tej S, Luo S, Haudenschild C, Meyers B, Green P (2005) Elucidation of the small RNA component of the transcriptome. Science **309:** 1567-1569.

Ma Z, Coruh C, Axtell M (2010) Arabidopsis lyrata small RNAs: transient MIRNA and small interfering RNA loci within the Arabidopsis genus. Plant Cell **22:** 1090-1103.

Mathieu J, Yant L, Murdter F, Kuttner F, Schmid M (2009) Repression of flowering by the miR172 target SMZ. PLoS Biology **7:** e1000148.

Meyers B, Axtell M, Bartel B, Bartel D, Baulcombe D, Bowman J, Cao X, Carrington J, Chen X, Green P, Griffiths-Jones S, Jacobsen S, Mallory A, Martienssen R, Poethig R, Qi Y, Vaucheret H, Voinnet O, Watanabe Y, Weigel D, Zhu J (2008) Criteria for annotation of plant MicroRNAs. Plant Cell **20:** 3186-3190.

Michaels S, Bezerra I, Amasino R (2004) FRIGIDA-related genes are required for the winter-annual habit in Arabidopsis. Proceedings of the National Academy of Sciences **101:** 3281-3285.

Molnar A, Melnyk C, Bassett A, Hardcastle T, Dunn R, Baulcombe D (2010) Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells. Science **328:** 872-875.

Moxon S, Jing R, Szittya G, Schwach F, Rusholme R, Moulton V, Dalmay T (2008) Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. Genome Research **18:** 1602-1609.

Murray S, Sharma A, Rooney W, Klein P, Mullet J, Mitchell S, Kresovich S (2008) Genetic Improvement of Sorghum as a Biofuel Feedstock: I. QTL for Stem Sugar and Grain Nonstructural Carbohydrates. Crop Science **48:** 2165-2179.

Nobuta K, Lu C, Shrivastava R, Pillay M, De Paoli E, Accerbi M, Arteaga-Vazquez M, Sidorenko L, Jeong D, Yen Y, Green P, Chandler V, Meyers B (2008) Distinct size distribution of endogeneous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. Proceedings of the National Academy of Sciences **105:** 14958-14963.

Nobuta K, Venu R, Lu C, Belo A, Vemaraju K, Kulkarni K, Wang W, Pillay M, Green P, Wang G, Meyers B (2007) An expression atlas of rice mRNAs and small RNAs. Nature Biotechnology **25:** 473-477.

Paterson A, Bowers J, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti A, Chapman J, Feltus F, Gowik U, Grigoriev I, Lyons E, Maher C, Martis M, Narechania A, Otillar R, Penning B, Salamov A, Wang Y, Zhang L, Carpita N (2009) The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

Ritter K, Jordan D, Chapman S, Godwin I, Mace E, McIntyre C (2008) Identification of QTL for sugar-related traits in a sweet x grain sorghum (Sorghum bicolor L. Moench) recombinant inbred population. Molecular Breeding **22:** 367-384.

Ritter K, McIntyre C, Godwin I, Jordan D, Chapman S (2007) An assesment of the genetic relationship between sweet and grain sorghums within Sorghum bicolor ssp. bicolor (L.) Moench using AFLP markers. Euphytica **157:** 161-176.

Rumble S, Lacroute P, Dalca A, Fiume M, Sidow A, Brudno M (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Computational Biology **5:** e1000386.

Salome P, To J, Kieber J, McClung C (2006) Arabidopsis response regulators ARR3 and ARR4 play cytokinin-independent roles in the control of circadian period. Plant Cell **18:** 55-69.

Schlappi M (2006) FRIGIDA LIKE 2 is a functional allele in Landsberg erecta and compensates for a nonsense allele of FRIGIDA LIKE 1. Plant Physiology **142:** 1728-1738.

Sunkar R, Girke T, Jain P, Zhu J (2005) Cloning and characterization of microRNAs from rice. Plant Cell **17:** 1397-1411.

Sunkar R, Zhou X, Zheng Y, Zhang W, Zhu J (2008) Identification of novel and candidate miRNAs in rice by high throughput sequencing. BMC Plant Biology **8:** 25.

Torney F, Moeller L, Scarpa A, Wang K (2007) Genetic engineering approaches to improve bioethanol production from maize. Current Opinion in Biotechnology **18:** 193-199.

Vaucheret H (2006) Post-transcriptional small RNA pathways in plants: mechanisms and regulations. Genes and Development **20:** 759-771.

Vazquez F (2006) Arabidopsis endogenous small RNAs: highways and byways. Trends in Plant Science **11:** 460-468.

Wang X, Elling A, Li X, Li N, Peng Z, He G, Sun H, Qi Y, Liu X, Deng X (2009) Genome-wide and organ-specific landscapes of epigenetic modifications and

their relationships to mRNA and small RNA transcriptomes in maize. Plant Cell **21:** 1053-1069.

Wei B, Cai T, Zhang R, Li A, Huo N, Li S, Gu Y, Vogel J, Jia J, Qi Y, Mao L (2009) Novel microRNAs uncovered by deep sequencing of small RNA transcriptomes in bread wheat (Triticum aestivum L.) and Brachypodium distachyon (L.) Beauv. Functional & Integrative Genomics **9:** 499-511.

Wu G, Park M, Conway S, Wang J, Weigel D, Poethig R (2009) The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. Cell **138:** 750-759.

Wu L, Zhou H, Zhang Q, Zhang J, Ni F, Liu C, Qi Y (2010) DNA methylation mediated by a microRNA pathway. Molecular Cell **38:** 465-475.

Xue L, Zhang J, Xue H (2009) Characterization and expression profiles of miRNAs in rice seeds. Nucleic Acids Research **37:** 916-930.

Yang J, Phillips M, Betel D, Mu P, Ventura A, Siepel A, Chen K, Lai E (2011) Widespread regulatory activity of vertebrate microRNA* species. RNA **17:** 312-326.

Zamore P, Haley B (2005) Ribo-gnome: the big world of small RNAs. Science **309:** 1519-1524.

Zheng Z, Xu X, Crosley R, Greenwalt S, Sun Y, Blakeslee B, Wang L, Ni W, Sopko M, Yao C, Yau K, Burton S, Zhuang M, McCaskill D, Gachotte D, Thompson M, Green T (2010) The protein kinase SnRK2.6 mediates the regulation of sucrose metabolism and plant growth in Arabidopsis. Plant Physiology **153:** 99-113.

Zhu Q, Spriggs A, Matthew L, Fan L, Kennedy G, Gubler F, Helliwell C (2008) A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. Genome Research **18:** 1456-1465.

Zhu Q, Upadhyaya N, Gubler F, Helliwell C (2009) Over-expression of miR172 causes loss of spikelet determinacy and floral organ abnormalities in rice (Oryza sativa). BMC Plant Biology **9:** 149.

Chapter 5 Discovery Of MicroRNA169 Gene Copies In Genomes Of Flowering Plants Through Positional Information

5.1. Abstract

Expansion and contraction of microRNA (miRNA) families can be studied in sequenced plant genomes through sequence alignments. Here, I focused on miR169 in sorghum because of its implications in drought tolerance and stem-sugar content. I was able to discover many miR169 copies that have escaped standard genome annotation methods. A new miR169 cluster was found on sorghum chromosome 1. This cluster is composed of the previously annotated *sbi-MIR169o* together with two newly found MIR169 copies, named *sbi-MIR169t* and *sbi-MIR169u*. I also found that a miR169 cluster on sorghum chr7 consisting of *sbi-MIR169l*, *sbi-MIR169m*, and *sbi-MIR169n* is contained within a chromosomal inversion of at least 500 kb that occurred in sorghum relative to *Brachypodium*, rice, foxtail millet, and maize. Surprisingly, synteny of chromosomal segments containing *MIR169* copies with linked bHLH and CONSTANS-LIKE genes extended from *Brachypodium* to dictotyledonous species such as grapevine, soybean, and cassava, indicating a strong conservation of linkages of certain flowering and/or plant height genes and microRNAs, which may explain linkage drag of drought and flowering traits and would have consequences for breeding new varieties. Furthermore, alignment of rice and sorghum orthologous regions revealed the presence of two additional miR169 gene copies (miR169r and miR169s) on sorghum chr7 that formed an antisense miRNA gene pair. Both copies were expressed and targeted different set of

genes. Synteny-based analysis of microRNAs among different plant species should lead to the discovery of new microRNAs in general and contribute to our understanding of their evolution.

5.2. Introduction

Several mechanisms have been proposed to explain the evolutionary origin of microRNA (miRNA) genes. For instance, they can be derived from miniature-inverted repeat transposable elements (MITEs) because the inverted repeat with a short internal sequence can be transcribed and form a hairpin structure that can be processed into small RNAs. Indeed, several miRNA genes derived from MITEs have been described in *Arabidopsis* and rice (Piriyapongsa and Jordan, 2008). It has also been proposed that miRNA genes can originate from spontaneous mutations in hairpin-like structures in the genome, and several miRNAs in *Arabidopsis* appeared to have originated this way (Fenselau de Felippes et al., 2008). The third and probably the most accepted explanation for the origin of microRNAs is based on the inverted duplication of genes, which when transcribed would form hairpin structures capable of generating small RNAs with perfect complementarity to the parental transcripts (Allen et al., 2004; Axtell and Bowman, 2008). Over time, the accumulation of mutations erodes the extensive homology with the parental transcripts and the accuracy of small RNA processing improves, eventually leaving a single segment (the mature miRNA) that retains complementarity (Allen et al., 2004; Axtell and Bowman, 2008). This hypothesis is supported with evidence where extended complementarity between plant miRNAs and target mRNAs is more evident in less-conserved and younger loci (Fahlgren et al., 2007).

Duplication of a newly formed miRNA eventually results in the creation of a multigene miRNA family, with evolutionary old and conserved miRNAs having more than one gene copy in the genome, whereas new and thus nonconserved (or species-specific) miRNAs being usually single copy (Allen et al., 2004; Fahlgren et al., 2007; Ma et al., 2010). Similar to protein-coding genes, duplication and subsequent divergence of miRNA gene copies can lead to loss of function (pseudogenes), keep current function (gene redundancy), gain a new function (neofunctionalization), or acquire a more specialized function (subfunctionalization) (Maher et al., 2006). Consistent with this, diversification in the sequence of duplicated miRNA gene copies was accompanied by changes in spatial and temporal expression patterns (Jiang et al., 2006; Maher et al., 2006). MicroRNA genes that undergo events of tandem duplication result in the formation of paralogous miRNA gene copies located in close proximity to each other on the same chromosome and thus forming miRNA clusters. Recently, (Sun et al., 2012) analyzed miRNAs that had amplified through tandem duplication in *Arabidopsis*, poplar (*Populus thricocarpa*), rice (*Oryza sativa*), and sorghum (*Sorghum bicolor*) genomes and found that 248 miRNAs in total belonging to 51 miRNA families arose by tandem duplication. This study showed the importance of tandem duplication events as a major force in the creation of new miRNA gene copies and into the expansion of miRNA families. Interestingly, the average miRNA copy number in tandemly duplicated regions from eudicots *A. thaliana* and *P. thricocarpa* was lower (2.8 copies/tandem) than in monocots *O. sativa* and *S. bicolor* (3.4 copies/tandem), suggesting that tandem duplications might have been more common in rice and sorghum (Sun et al., 2012). Despite this finding,

there is a lack of knowledge on the evolutionary fate of miRNA gene clusters across the grass family.

Here, I analyzed the process of tandem duplication that gave rise to MIR169 gene clusters in sorghum (*S. bicolor* [L.] Moench) and traced its evolutionary path by aligning contiguous chromosomal segments of diploid *Brachypodium*, rice, foxtail millet, and the two homoeologous regions of allotetraploid maize. I chose miR169 as an example because of its possible role in stem-sugar accumulation in sorghum besides its previously described role in drought stress response in several plant species. I discovered allelic variation in miR169 expression between grain and sweet sorghum, suggesting that miR169 could also play a role in the sugar content of sorghum stems (Calviño et al., 2011). Although high sugar content in stems is a trait shared by sorghum and sugarcane (Calviño et al., 2008; Calviño et al., 2009), this trait seems to be silent in other grasses (Calviño and Messing, 2012). This prompted me to investigate the evolution and dynamic amplification of miR169 gene copies in grass genomes. I found that synteny of chromosomal segments containing *MIR169* gene copies was conserved between monocotyledonous species such as *Brachypodium* and sorghum but surprisingly also across the monocot barrier in dicotyledonous species such as grapevine, soybean, and cassava. Furthermore, linkage of *MIR169* copies with a bHLH gene similar to *Arabidopsis bHLH137* and with a CONSTANS-LIKE gene similar to *Arabidopsis COL14* was conserved in all the grasses examined as well as in soybean and cassava (linkage between *MIR169* and *bHLH* genes) and grapevine (linkage between MIR169 and COL14 genes). We discuss the importance of this finding for breeding crops with

enhanced bioenergy traits.

5.3. Results

5.3.1. New *MIR169* gene copies in the rice, sorghum, and maize genomes

A miRNA cluster as defined in the miRBase database (release 19, August 2012) is composed of two or more miRNA gene copies that are located on the same chromosome and separated from each other by a distance of 10 kb or less. The distance set to define a miRNA cluster is arbitrary though, as evidenced by a cluster composed of 16 copies of *MIR2118* distributed over an 18-Kb segment on rice chr4 (Sun et al., 2012). The sequencing of the sorghum genome allowed the identification of 17 *MIR169* gene copies, from which five were arranged in two clusters, one located on chr2 (*sbi-MIR169f* and *sbi-MIR169g*) and the other located on chr7 (*sbi-MIR169l*, *sbi-MIR169m*, and *sbi-MIR169n*, respectively (Paterson et al., 2009) (Table 5.1 and Figure 5.1).

I first analyzed the region containing the *MIR169* cluster on sorghum chr7 because it had the highest number of gene copies. The alignment of sorghum genes flanking *MIR169* copies to the rice genome permitted me the identification of a collinear region on rice chr8 also containing a cluster of *MIR169* gene copies (Figure 5.2). Interestingly, the cluster on rice chr8 was composed of five *MIR169* gene copies, whereas the orthologous cluster on sorghum chr7 contained only three annotated *MIR169* gene copies. Further investigation based on reciprocal Blastn analysis revealed that *osa-MIR169l* and *osa-MIR169q* are orthologous to a region on sorghum chr7, where there was no previous annotation of *MIR169* genes. Indeed, by

taking the sorghum DNA segment highly similar to *osa-MIR169l* and *osa-MIR169q* and subjecting it to an RNA folding program (RNAfold: http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) to identify hairpin-like structures characteristic of microRNA precursors, I was able to discover two new *MIR169* gene copies in sorghum that I named *sbi-MIR169r* and *sbi-MIR169s*, respectively (Figure 5.2 and Figure 5.3). Independent support for the new annotation of *sbi-MIR169r* and *sbi-MIR169s* was achieved through orthologous alignment of a third species, maize, through *zma-MIR169e* and *zma-MIR169h* gene copies (Figure 5.4).

To identify additional *MIR169* gene copies in sorghum that might have arisen by tandem duplication, I took each of the annotated *MIR169* genes and performed Blastn analysis against the sorghum genome to search for new copies located in close proximity to any of the previously annotated ones. Such analysis identified two new *MIR169* copies on sorghum chromosome 1 (chr1) when *sbi-MIR169o* was used as query that I named *sbi-MIR169t* and *sbi-MIR169u*, respectively (Figure 5.3). Thus, *sbi-MIR169o* together with *sbi-MIR169t* and *sbi-MIR169u* constituted a new *MIR169* cluster of the sorghum genome (Table 5.1). The segment containing the newly identified *MIR169* cluster on sorghum chr1 was collinear with an orthologous segment of rice chr3 (Figure 5.5), although no *MIR169* gene had previously been found in this region. By performing reciprocal Blastn analysis with *sbi-MIR169o* against the rice genome, I could identify the corresponding orthologous *MIR169* copy on rice chr3 that I named *osa-MIR169r* (Figure 5.3 and Figure 5.5). Furthermore, *osa-MIR169r* is contained within a segment that is collinear with an orthologous region of chr1 of a fourth species, *Brachypodium*, corresponding to *bdi-*

MIR169k (Figure 5.5). Comparison between sorghum and maize revealed that the *MIR169* cluster on sorghum chr1 is collinear with a segment on maize chr1 that contains *zma-MIR169l* (Figure 5.6). Indeed, *sbi-MIR169u* and *zma-MIR169l* are also orthologous gene copies. Finally, when the cluster on sorghum chr2 containing *sbi-MIR169f* and *sbi-MIR169g* was analyzed, collinearity with the segment on sorghum chr7 containing the *sbi-MIR169r/s* and *sbi-MIR169l-n* cluster revealed the existence of an additional *MIR169* copy on sorghum chr2 that I named *sbi-MIR169v* (Figure 5.2 and Figure 5.3 and Table 5.1). Furthermore, the *sbi-MIR169f/g/v* cluster is syntenic with a region on maize chr7 containing *zma-MIR169k* and its homoeologous region on maize chr2 containing *zma-MIR169j* and the newly identified *zma-MIR169s* gene copy (Figure 5.3 and Figure 5.7 and Table 5.1).

In summary, by aligning sorghum chromosomal segments containing *MIR169* clusters with orthologous regions of *Brachypodium*, rice, and maize, we were able to identify five additional *MIR169* copies in sorghum and an additional copy in rice and maize, respectively.

Chromosome	Gene ID[a]	Coordinates[b]	Strand	Distance between genes flanking the cluster[c]
Brachypodium Distachyon				
Chr1	bdi-*MIR169*k	1,175,425..1,175,598	+	
Chr3	bdi-*MIR169*e bdi-*MIR169*g	43,441,526..43,441,689 43,444,486..43,444,666	+ +	Cluster 1: bdi-*MIR169*e to bdi-*MIR169*g = 2,960 bp
Oryza sativa				
Chr3	osa-*MIR169r*	35,782,397..35,782,553	+	

Chr8	osa-*MIR169i* osa-*MIR169h* osa-*MIR169m* osa-*MIR169l* osa-*MIR169q*	26,891,154..26,891,261 26,895,354..26,895,475 26,901,902..26,902,039 26,905,493..26,905,600 26,905,600..26,905,493	+ + + + -	Cluster 1: osa-*MIR169*i to osa-*MIR169*q = 14,446 bp
Chr9	*osa-MIR169j* *osa-MIR169k*	19,788,861..19,788,985 19,792,133..19,792,288	+ +	Cluster 2: osa-*MIR169*j to osa-*MIR169*k = 3,272 bp
Setaria italic				
Chr9	*sit-MIR169o*	526,081..525,981	-	
Chr2	sit-*MIR169f* sit-*MIR169g* sit-*MIR169h*	36,921,078..36,921,205 36,923,991..36,924,143 36,924,215..36,924,361	+ + +	Cluster 1: sit-*MIR169*f to sit-*MIR169*h = 3,137 bp
Chr6	sit-*MIR169i* sit-*MIR169j* sit-*MIR169k* sit-*MIR169r* sit-*MIR169s*	33,994,480..33,994,680 33,997,832..33,997,997 34,001,008..34,001,109 34,003,536..34,003,402 34,003,402..34,003,536	+ + + - +	Cluster 2: sit-*MIR169*i to sit-*MIR169*s = 8,922 bp
Sorghum bicolor				
Chr1	sbi-*MIR169o* sbi-*MIR169t* sbi-*MIR169u*	1,029,916..1,029,814 1,030,265..1,030,155 1,037,237..1,037,096	- - -	Cluster 1: sbi-*MIR169*o to sbi-*MIR169*u = 7,321 bp
Chr2	sbi-*MIR169f* sbi-*MIR169g* sbi-*MIR169v*	64,603,670..64,603,817 64,606,503..64,606,654 64,606,719..64,606,868	+ + +	Cluster 2: sbi-*MIR169*f to sbi-*MIR169*v = 3,049 bp
Chr7	sbi-*MIR169r* sbi-*MIR169s* sbi-*MIR169l* sbi-*MIR169m* sbi-*MIR169n*	61,058,625..61,058,750 61,058,750..61,058,625 61,062,736..61,062,640 61,068,118..61,068,027 61,071,181..61,071,273	+ - - - -	Cluster 3: sbi-*MIR169*r to sbi-*MIR169*n = 12,648 bp
Zea mays				
Chr1	zma-*MIR169*l	298,277,019..298,277,107	+	
Chr2	zma-*MIR169*j zma-*MIR169s*	192,700,339..192,700,489 192,700,616..192,700,748	+ +	Cluster 1: zma-*MIR169*j to zma-*MIR169*s = 277 bp
Chr4	zma-*MIR169*i zma-*MIR169*d zma-*MIR169*h zma-*MIR169e*	47,241,963..47,242,153 47,454,177..47,454,304 47,513,567..47,513,694 47,513,695..47,513,568	+ - + -	Cluster 2: zma-*MIR169*i to zma-*MIR169*e = 271,605 bp
Chr7	zma-*MIR169*k	135,706,179..135,706,311	-	
Vitis vinifera				
Chr1	vvi-*MIR169*y	22,233,573..22,233,820	+	
Chr14	vvi-*MIR169z* vvi-*MIR169*e	25,082,612..25,082,498 25,082,865..25,082,717	- -	Cluster 1: vvi-*MIR169*z to vvi-*MIR169*e = 367 bp
Chr17	vvi-*MIR169*x	355,713..355,837	-	
Glycine max				

Chr6	gma-*MIR169*w	13,783,352..13,783,225	-	
Chr8	gma-*MIR169*x gma-*MIR169*y	717,092..717226 724,205..724,340	+ +	Cluster 1: gma-*MIR169*o to gma-*MIR169*p = 7,248 bp
Manihot esculenta				
scaffold01701	mes-*MIR169*w	436,633..436,794	+	
scaffold09876	mes-*MIR169*y	536,510..536,709	-	

Table 5.1. Summary of *MIR169* gene copies described in this study. [a] In green are microRNA genes identified for the first time in this study. [b] Chromosomal positions are based on Phytozome annotation for all the species except rice that is based on RAPDB annotation. [c] Distance within the cluster was calculated from the beginning of the first microRNA gene to the beginning of the last microRNA gene in the cluster.

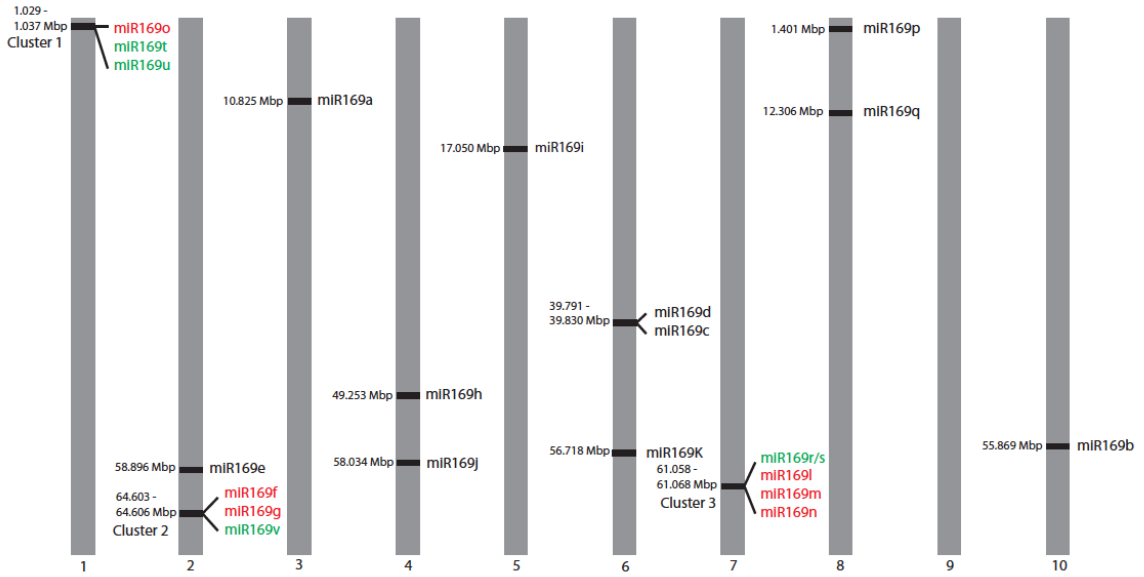


Figure 5.1. Distribution of MIR169 gene copies in the genome of *Sorghum bicolor* cultivar BTx623. A total of 22 *MIR169* gene copies are shown, with 17 copies previously annotated by the sorghum genome-sequencing consortium (shown in black and red) (Paterson et al., 2009) and with five additional *MIR169* copies described in this study for the first time (shown in green). The evolutionary trajectory of sorghum MIR169 gene copies arranged in clusters 1, 2, and 3 are described in this study.

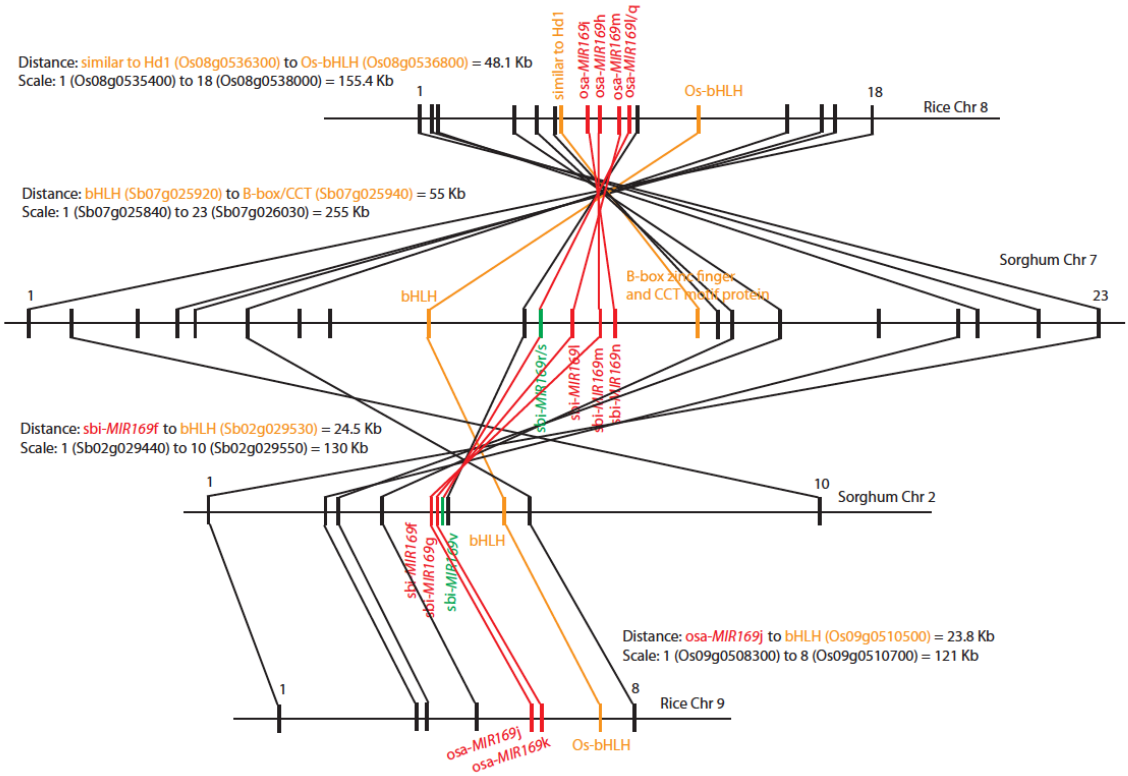
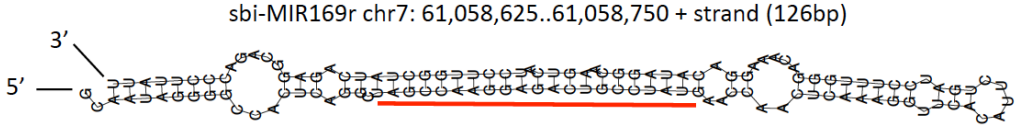


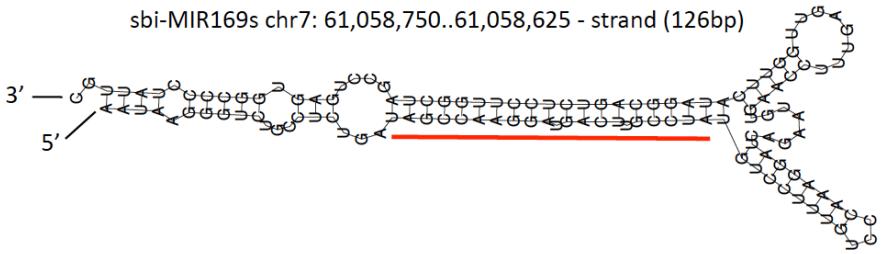
Figure 5.2. Syntenic alignment of rice and sorghum chromosomal segments containing MIR169 gene clusters. Sorghum *MIR169* gene clusters on chr2 and chr7 together with their flanking protein coding genes were aligned with rice by orthologous gene pairs. Rice and sorghum chromosomes are represented as horizontal lines, whereas genes along the chromosome are represented as rectangle bars. Known *MIR169* gene copies are shown as red bars, whereas new *MIR169* gene copies described in this study are shown as green bars. The bHLH and B-box zinc finger and CCT motif (B-box/CCT) genes are represented as yellow bars. All other protein coding genes in the chromosomal regions under study are represented as black bars. Orthologous gene pairs are indicated as lines connecting bars, with red lines indicating orthology between *MIR169* gene pairs and yellow lines indicating orthology between bHLH and B-box/CCT gene pairs, respectively. All other

orthology between rice and sorghum protein coding genes are indicated as black lines connecting black bars. The physical distance between bHLH and B-box/CCT genes and/or between bHLH or B-Box/CCT genes to the flanking *MIR169* copy is indicated. To provide a scale of the chromosomal segments highlighted in the figure, the physical distance between the first and the last gene in the segment is indicated and thus serves as a reference to observe expansion and contraction of genomic regions. An inversion event on sorghum chr7 containing the *MIR169* cluster occurred relative to the orthologous regions on sorghum chr2 and rice chr8 and chr9 respectively.

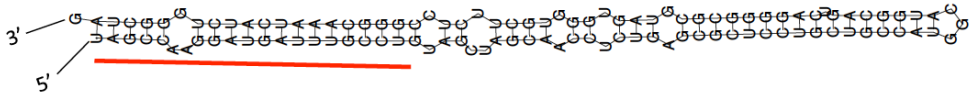
sbi-MIR169r chr7: 61,058,625..61,058,750 + strand (126bp)



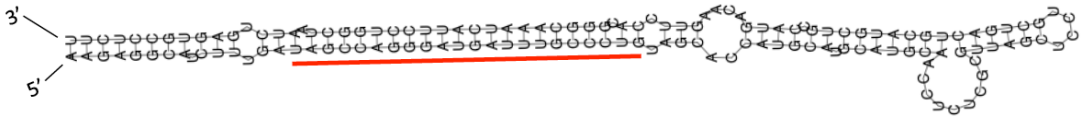
sbi-MIR169s chr7: 61,058,750..61,058,625 - strand (126bp)



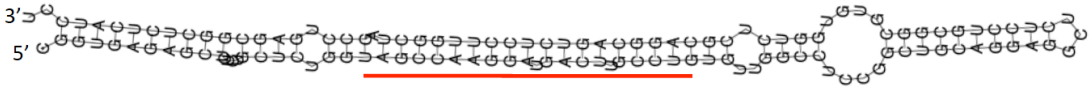
sbi-MIR169t chr1: 1,030,265..1,030,155 (111 bp) - strand



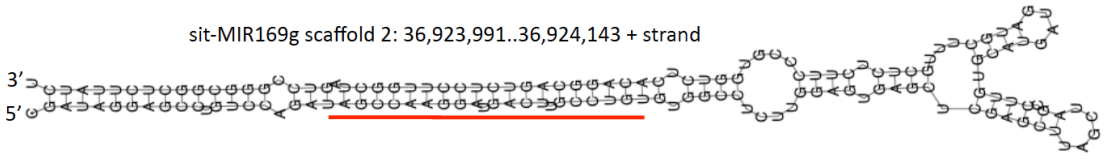
sbi-MIR169u chr1: 1,037,237..1,037,096 (142 bp) - strand



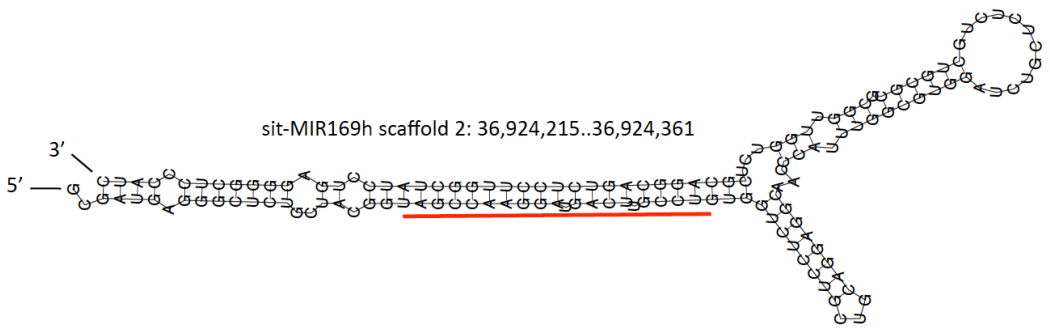
sit-MIR169f scaffold 2: 36,921,078..36,921,205 + strand



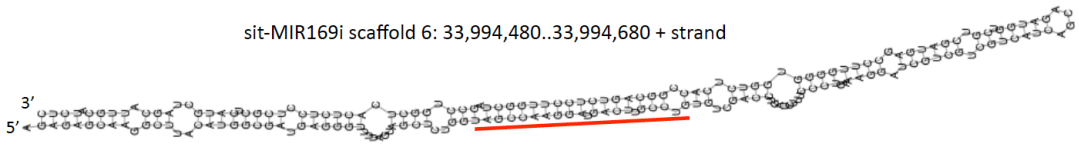
sit-MIR169g scaffold 2: 36,923,991..36,924,143 + strand



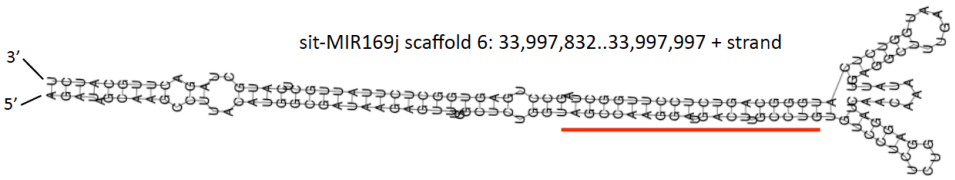
sit-MIR169h scaffold 2: 36,924,215..36,924,361



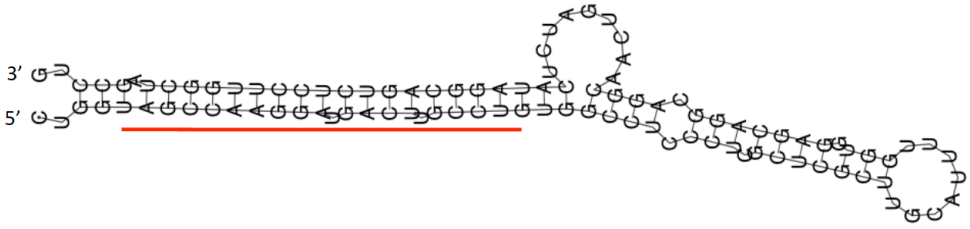
sit-MIR169i scaffold 6: 33,994,480..33,994,680 + strand



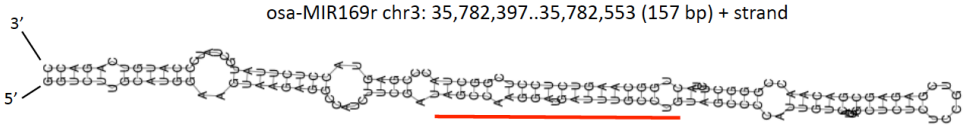
sit-MIR169j scaffold 6: 33,997,832..33,997,997 + strand



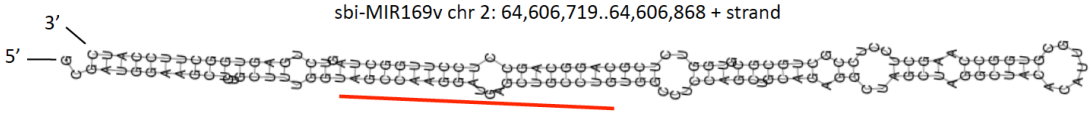
sit-MIR169k scaffold 6: 34,001,008..34,001,109 + strand



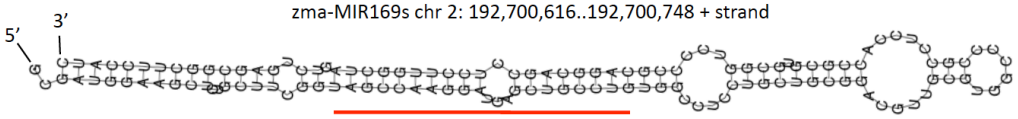
osa-MIR169r chr3: 35,782,397..35,782,553 (157 bp) + strand



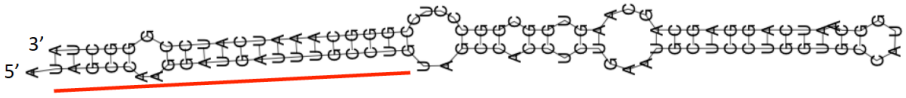
sbi-MIR169v chr 2: 64,606,719..64,606,868 + strand

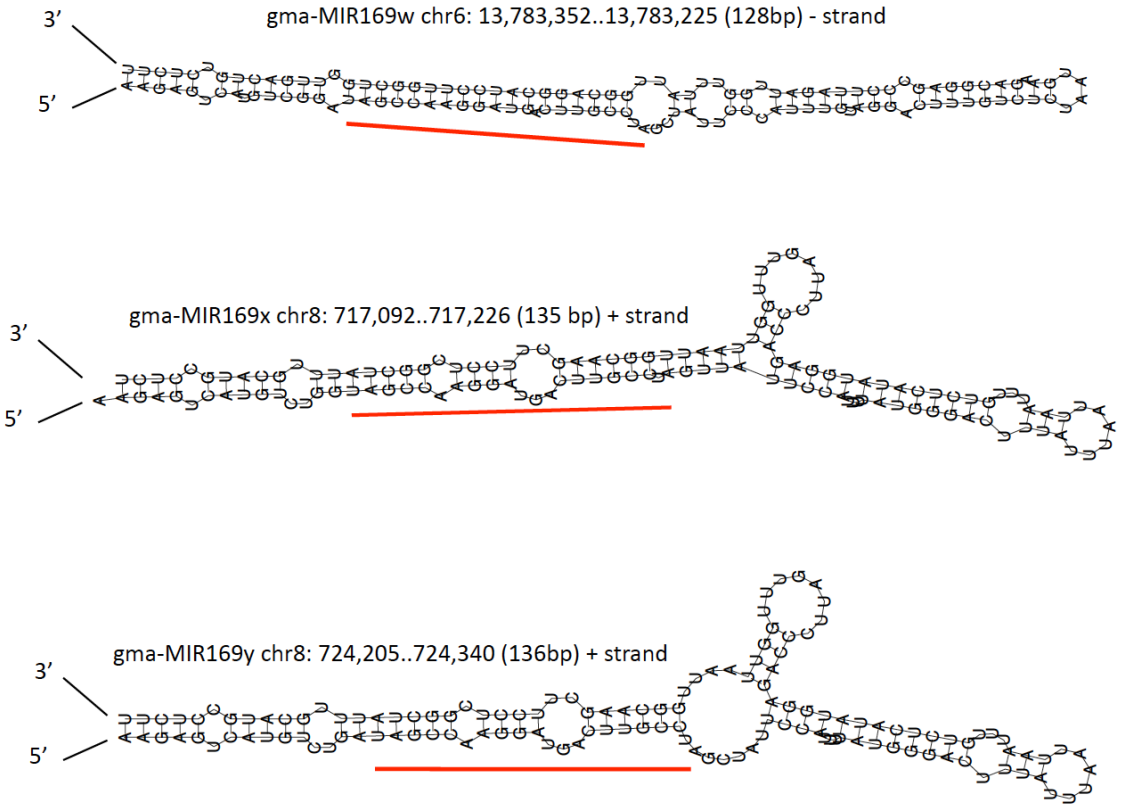


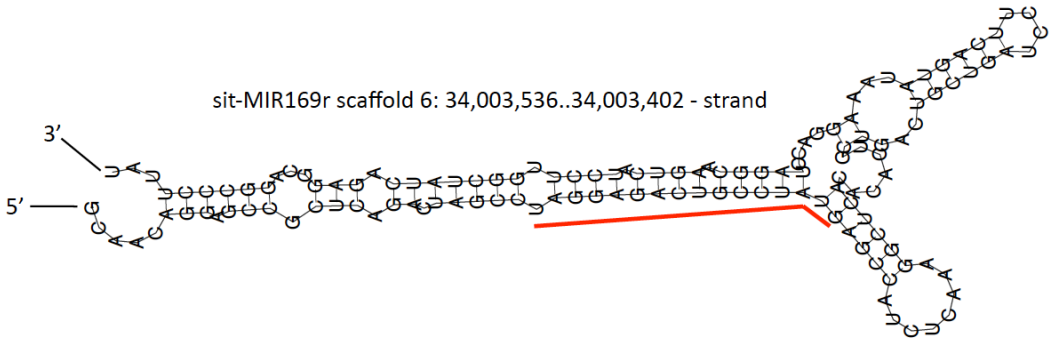
zma-MIR169s chr 2: 192,700,616..192,700,748 + strand

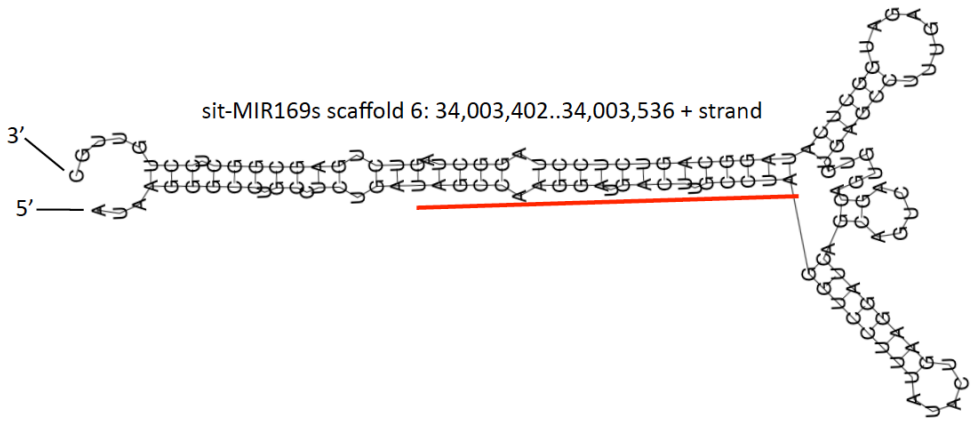


sit-MIR169o scaffold 9: 526,081..525,981 -strand




3’
5’
gma-MIR169w chr6: 13,783,352..13,783,225 (128bp) - strand
3’
5’
gma-MIR169x chr8: 717,092..717,226 (135 bp) + strand
3’
5’
gma-MIR169y chr8: 724,205..724,340 (136bp) + strand


sit-MIR169r scaffold 6: 34,003,536..34,003,402 - strand
3’
5’


sit-MIR169s scaffold 6: 34,003,402..34,003,536 + strand
3’
5’

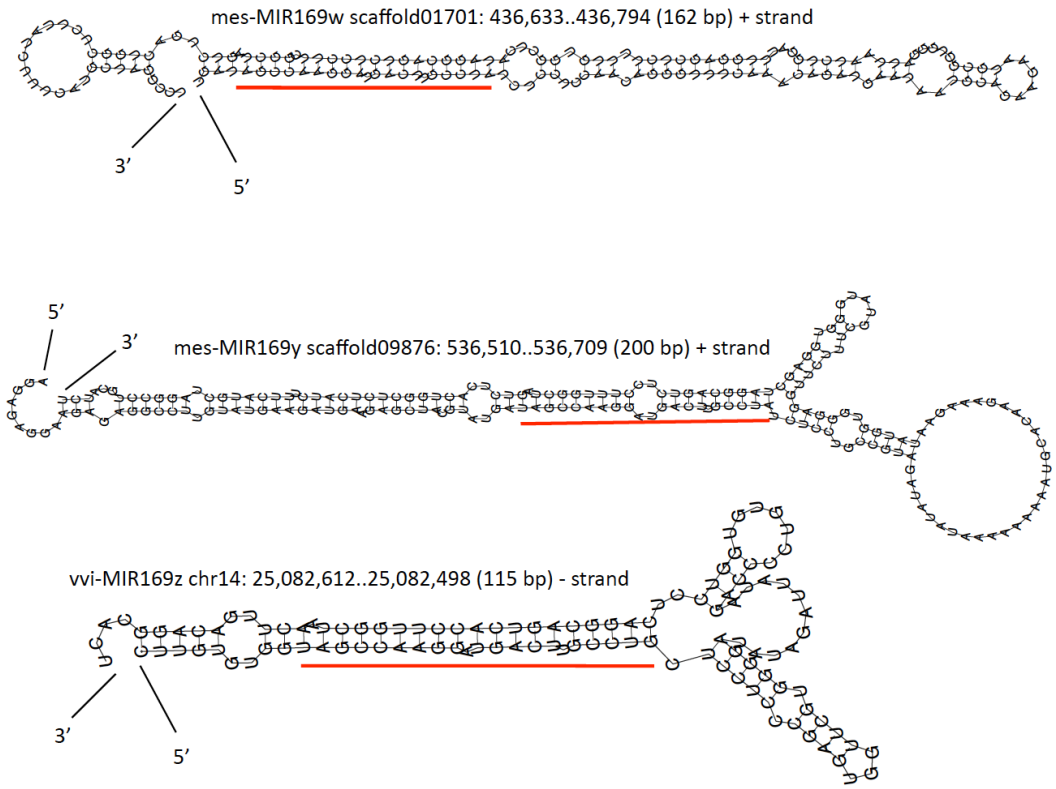


Figure 5.3. Stem-loop precursor sequences of newly predicted MIR169 copies in rice, sorghum, foxtail millet and maize. The genomic location for each MIR169 stem-loop precursor is given. The predicted mature miR169 sequence is indicated with a red bar.

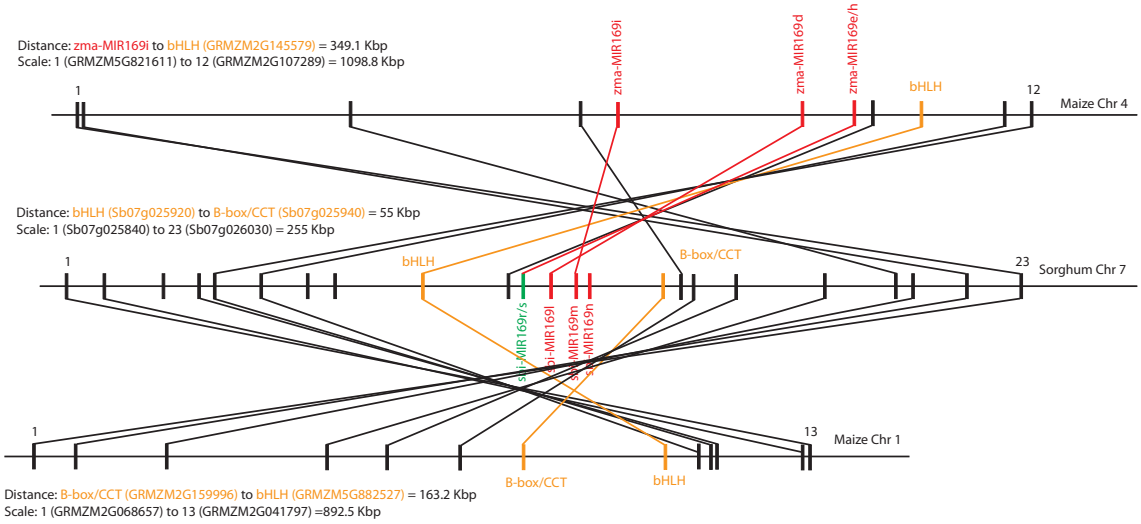


Figure 5.4. Sequence alignment of sorghum chr7 segment containing *MIR169* gene cluster to homoeologous chromosomal segments from maize. Sorghum sbi-*MIR169r/s*, sbi-*MIR169l* and sbi-*MIR169m* genes on chr7 are orthologous to maize zma-*MIR169e/h*; zma-*MIR169d* and zma-*MIR169i* respectively on chr4. Notice that the *MIR169* cluster on the homoelogous region on maize chr1 was deleted although its flanking genes remained. The orthologous copy of sorghum B-box/CCT gene flanking the *MIR169* gene cluster was lost on maize chr4 but retained on the homoelogous segment on chr1. Expansion in the maize genome relative to sorghum is clear when regions on maize chr1 and sorghum chr7 are compared. The region on sorghum chr7 is inverted relative to maize.

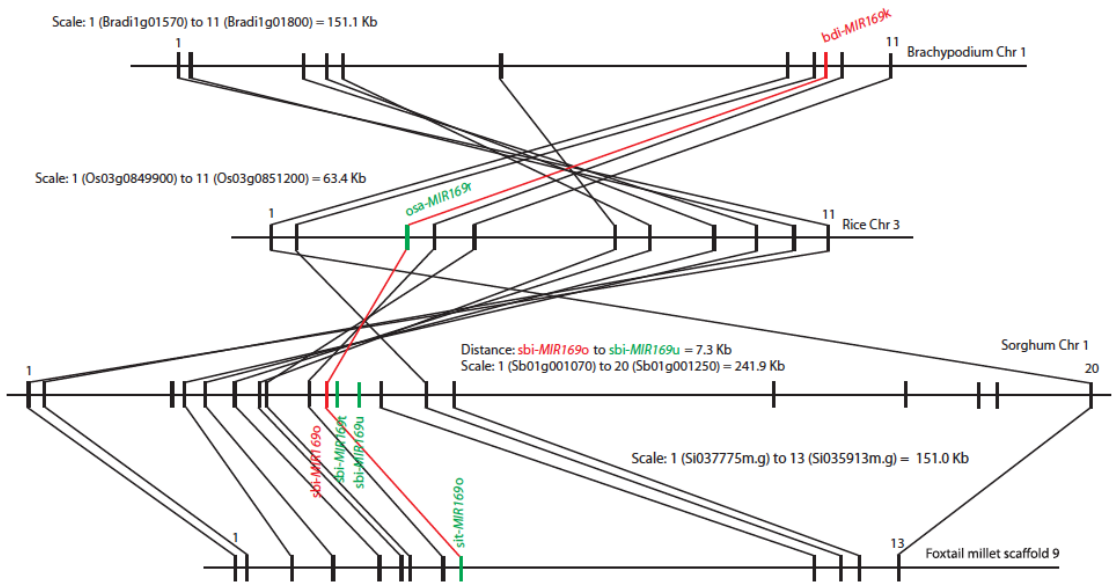


Figure 5.5. Sequence alignment of sorghum *MIR169* cluster on chr1 with orthologous regions from *Brachypodium*, rice and foxtail millet. The *sbi-MIR169o* copy in sorghum allowed the identification of the orthologous *osa-MIR169r* copy in rice and *sit-MIR169o* copy in foxtail millet, respectively. For the region containing *sbi-MIR169o/t/u* on chr1, we could not find sufficient conservation of synteny to identify an orthologous region in sorghum, thus a synteny graph is only shown with sorghum chr1. An inversion event on rice chr3 occurred relative to *Brachypodium*, foxtail millet, and sorghum.

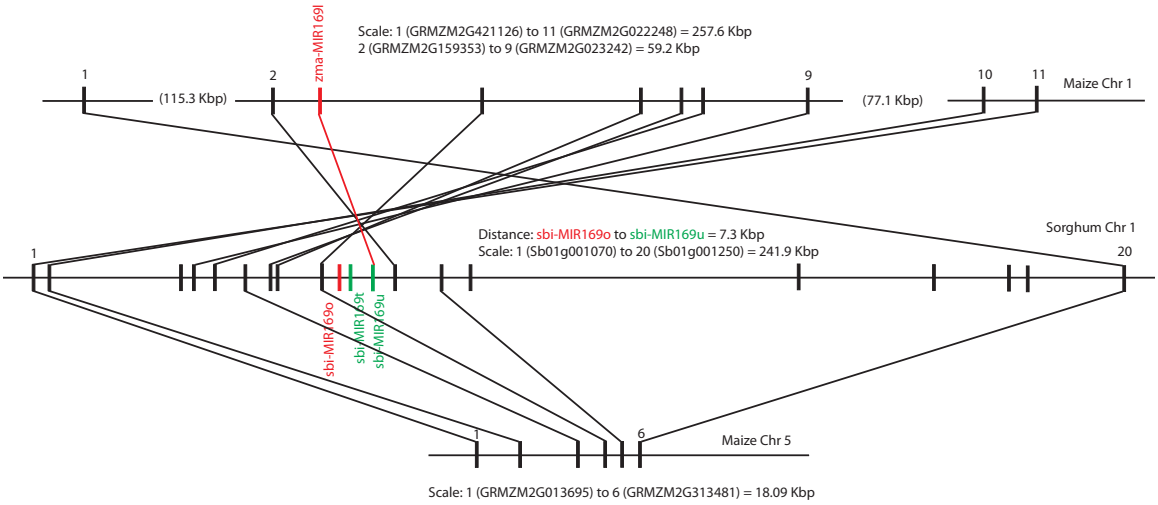


Figure 5.6. Sequence alignment of sorghum MIR169 cluster on chr1 with orthologous regions from maize. Sorghum sbi-*MIR169u* and maize zma-*MIR169l* are orthologous copies. There isn't any orthologous *MIR169* copy on maize homoeologous chr5. The region on maize chr1 is expanded (comprising a total of 257.6 Kbp) relative to the homoeologous region on chr5 (comprising 18.09 Kbp only). An inversion event occurred on maize homeologous region on chr1.

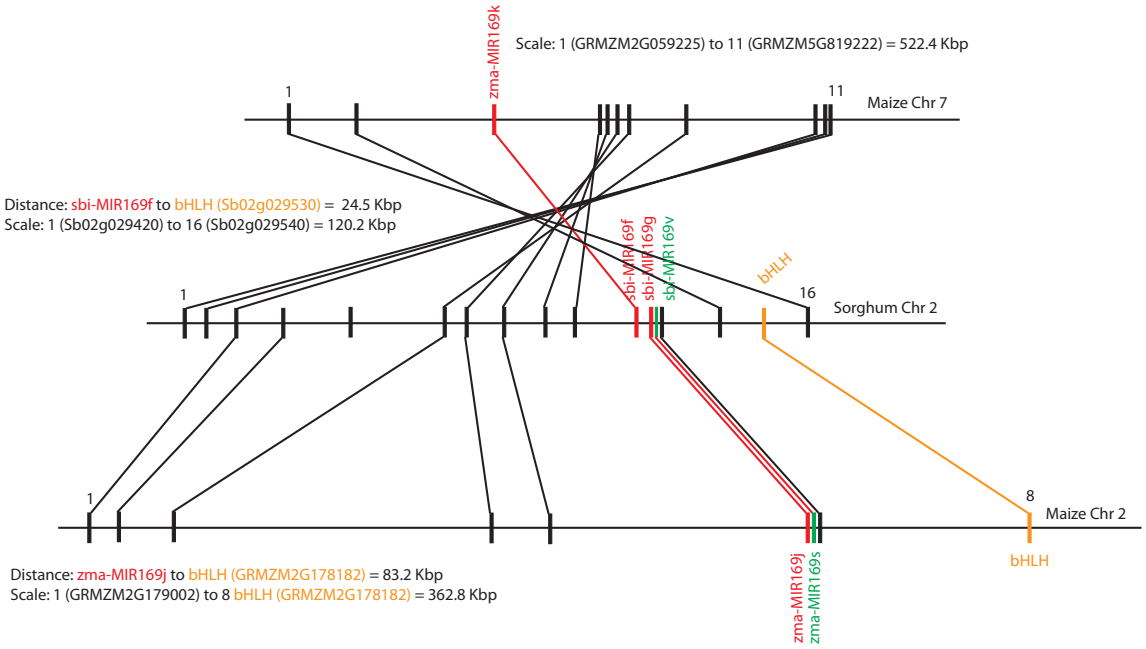


Figure 5.7. Sequence alignment of sorghum *MIR169* cluster on chr2 with orthologous regions from maize. Sorghum *MIR169* gene cluster on chr2 is collinear with a region on maize chr7 that contains zma-*MIR169k*, and with the homeologous region on maize chr2 that contains the previously annotated zma-*MIR169j* and the new copy zma-*MIR169s* that is described in this study. Although the *MIR169* gene cluster on maize chr2 is physically adjacent to the bHLH gene, similarly with the *MIR169* gene cluster on sorghum chr2, the homeologous region containing zma-*MIR169k* lacked the bHLH gene copy. An inversion event on maize chr7 occurred relative to its homeologous region on chr2 and to sorghum chr2.

5.3.2. New *MIR169* clusters in the recently sequenced foxtail millet genome

The recent release of the complete reference genome sequence for foxtail millet (*Setaria italica*) (Bennetzen et al., 2012; Zhang et al., 2012) greatly enhances comparative genomics analysis within the *Poaceae*, with genome sequences available from five species. Foxtail millet provided me with additional information to study syntenic relationships with sorghum because they split from each other approximately 26 Ma (Bennetzen et al., 2012; Zhang et al., 2012). Indeed, 19 collinear blocks were found between foxtail millet and sorghum, which comprised approximately 72% of the foxtail millet genome (Zhang et al., 2012). Consequently, I could use sorghum to identify and predict *MIR169* gene copies in the foxtail millet genome. I identified and predicted *MIR169* copies in foxtail millet, collinear with sorghum *MIR169* copies, arranged in clusters on chr1, chr2, and chr7. The sorghum *MIR169* cluster on chr1 was collinear with a segment on chr9 of foxtail millet, from which *sit-MIR169o* was identified as the ortholog of *sbi-MIR169o* (Figure 5.3 and Figure 5.5 and Table 5.1). The sorghum *MIR169* copies arranged in cluster on chr7 were collinear with a segment on chr6 from foxtail millet that harbored the newly identified orthologous *MIR169* copies *sit-MIR169i*, *sit-MIR169j*, *sit-MIR169k*, *sit-MIR169r*, and *sit-MIR169s* (Figure 5.8 and Figure 5.3 and Table 5.1). Finally, tandem sorghum *MIR169* copies on chr2 were collinear with a segment on foxtail millet chr2 that contained the three newly predicted *MIR169* copies *sit-MIR169f*, *sit-MIR169g*, and *sit-MIR169h* (Figure 5.9 and Figure 5.3 and Table 5.1).

In summary, I used sorghum as a reference genome to identify and predict

nine *MIR169* gene copies that were collinear with foxtail millet. The prediction of *MIR169* genes in the foxtail millet genome will greatly facilitate their experimental validation through the sequencing of small RNAs from different tissues and developmental stages.

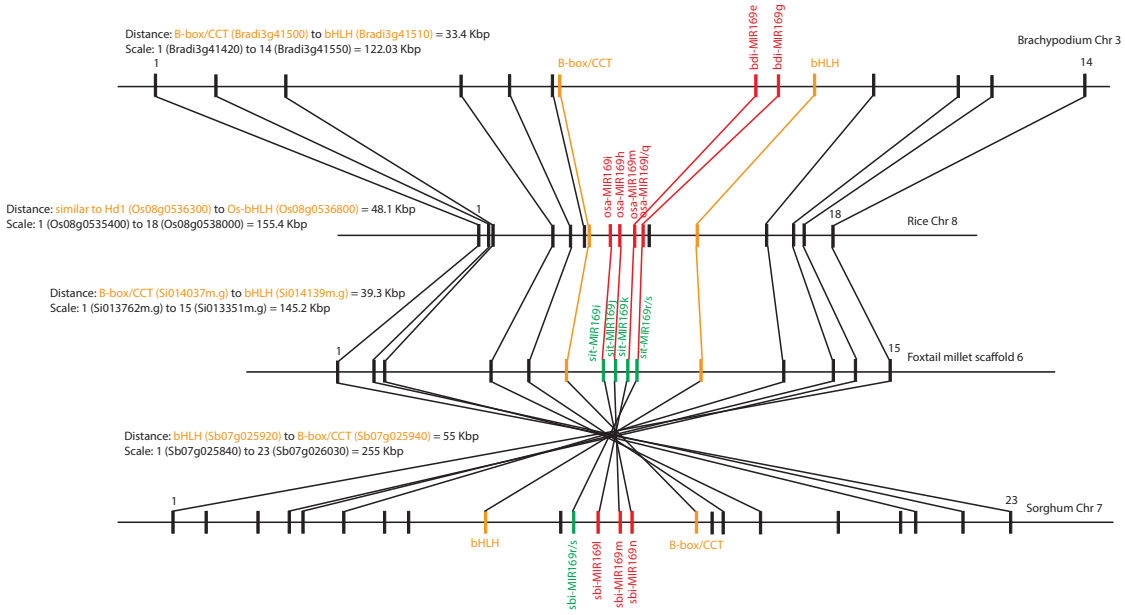


Figure 5.8. Sequence alignment of sorghum *MIR169* cluster on chr7 with orthologous regions from *Brachypodium*, rice, and foxtail millet. Rice and sorghum *MIR169* gene copies were used to identify and annotate five *MIR169* genes in foxtail millet (shown in green). The bHLH and B-box/CCT genes were physically adjacent to MIR169 gene copies in the four species examined. The region examined on sorghumchr7 expanded relative to the orthologous region from the other three grasses and was inverted only in sorghum.

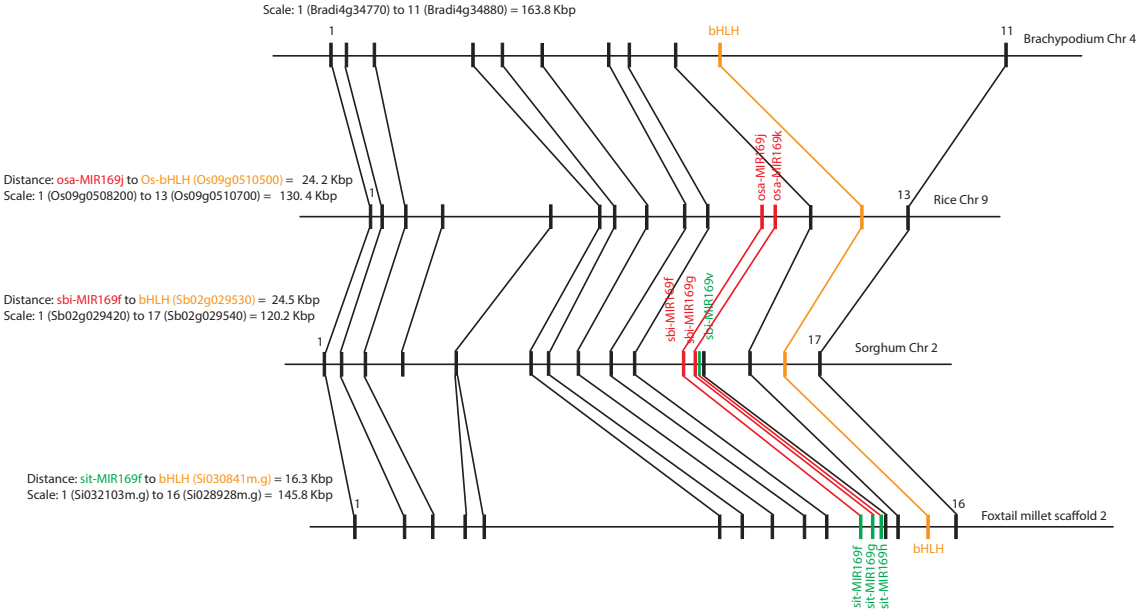


Figure 5.9. Sequence alignment of sorghum *MIR169* cluster on chr2 with orthologous regions from *Brachypodium*, rice, and foxtail millet. *MIR169* gene copies were deleted from *Brachypodium* chr4 but the flanking genes remained. The *MIR169* gene cluster in rice was composed of two copies, whereas in sorghum and foxtail millet, the cluster comprised three copies. The bHLH gene was present in all four grasses and was physically adjacent to *MIR169* gene copies in rice, sorghum, and foxtail millet. Sorghum *MIR169* gene copies were used to identify and annotate the orthologous copies on foxtail millet scaffold 2 (shown in green).

5.3.3. Gain and losses of *MIR169* gene copies during grass evolution

To determine expansion and contraction of the *MIR169* gene clusters, I aligned collinear chromosomal segments of diploid *Brachypodium*, rice, and foxtail millet and the two homoeologous regions of allotetraploid maize. Based on nucleotide substitution rates, the cluster of *MIR169* copies on sorghum chr7 was likely preserved from an ancestral grass chromosome and comprised five MIR169 gene copies, from which three of them were deleted in *Brachypodium* after the split of *Brachypodium* from the ancestor of rice, foxtail millet, and sorghum (Figure 5.8 and 5.10A and B). The number of *MIR169* genes (five copies per cluster) was unchanged in rice, sorghum, and foxtail millet, whereas in maize, four copies were retained on orthologous homoeologous region on chr4, but none on the homoeologous region on chr1 (Figure 5.4 and 5.10A). Although the *MIR169* copies were deleted from maize chr1, the flanking genes remained intact.

In the case of the *MIR169* cluster on sorghum chr2, its evolution can be explained according to two models (Figure 5.10A). In the first one, the ancestor of the grasses had two *MIR169* copies and they were conserved before the split of *Brachypodium* and rice, with *Brachypodium* losing these two *MIR169* copies, where as rice maintained them. An additional copy was gained in the common ancestor of foxtail millet, sorghum, and maize, giving rise to a cluster with three *MIR169* gene copies. Phylogenetic analysis suggested that the new copy in the ancestor of foxtail millet, sorghum, and maize was the ancestral copy that gave rise to *sit-MIR169h*, *sbi-MIR169v*, and *zma-MIR169s*, respectively (Figure 5.10C). I estimated that the time at which this copy arose in the progenitor of foxtail millet, sorghum, and maize was

approximately 41.1 Ma (see Materials and Methods for estimation of time of duplication). Alternatively, the common ancestor of the grasses could have three *MIR169* gene copies, and one copy was lost in the common ancestor of *Brachypodium* and rice, with a subsequent loss of two additional *MIR169* gene copies in *Brachypodium* relative to rice (Figure 5.10A).

Regarding the cluster of *MIR169* copies on sorghum chr1, I favor a model where the ancestor of the grasses had a single *MIR169* copy because *Brachypodium*, rice, and foxtail millet all have a single *MIR169* copy (Figure 5.10D). Thus, the additional two *MIR169* copies present in the sorghum cluster could have arisen by duplication events. Phylogenetic analysis suggested that the ancestral copy in the cluster was *sbi-MIR169o*, from which *sbi-MIR169t* subsequently duplicated 8.5 Ma (see Materials and Methods) (Figure 5.10D). Thus, *sbi-MIR169t* was acquired specifically in the sorghum lineage. Because *sbi-MIR169u* and *zma-MIR169l* are highly related but distantly related from *sbi-MIR169o* and *sbi-MIR169t* (Figure 5.10D), I postulate that the ancestral copy of *sbi-MIR169u* and *zma-MIR169l* was inserted next to the other *MIR169* gene copies in the progenitor of sorghum and maize. In the maize lineage, diploidization after allotetraploidization led to the deletion of the corresponding orthologous *MIR169* copy from the homoeologous segment on chr5, whereas the flanking genes remained conserved (Figure 5.6).

In summary, differences in *MIR169* copy number between clusters from *Brachypodium*, rice, foxtail millet, sorghum, and maize arose by duplication of ancestral *MIR169* genes that were retained or lost during grass evolution. Overall, sorghum gained eight *MIR169* copies relative to *Brachypodium*, three copies relative

to rice, two copies relative to foxtail millet, and three copies relative to maize.

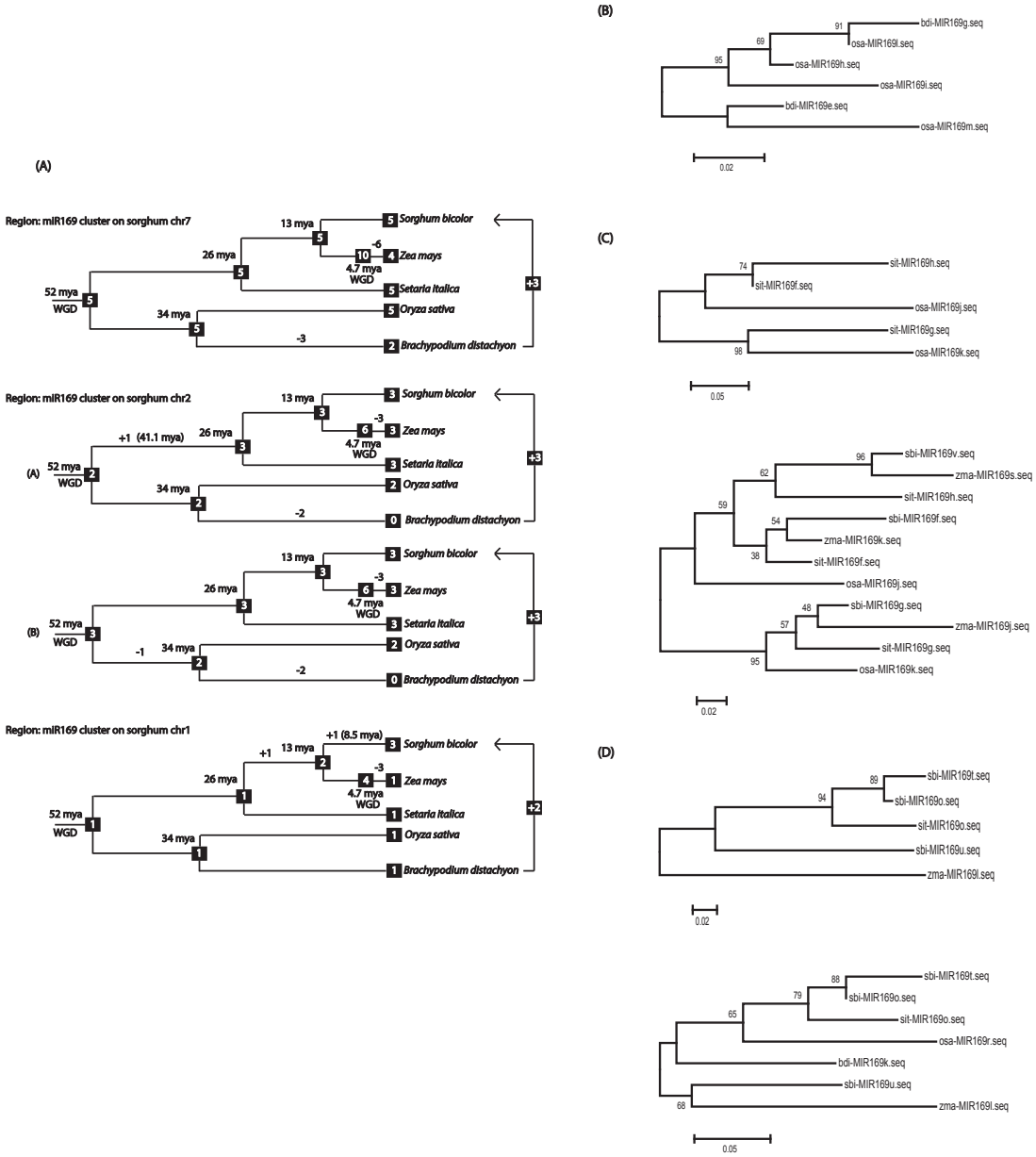


Figure 5.10. Gains and losses of *MIR169* gene copies during grass evolution. **(A)** Phylogenetic distribution of *MIR169* gene copies in ancestral and current species with gain and losses of *MIR169* copy number during grass evolution. Numbers in squares represent the number of *MIR169* gene copies for a given cluster

in each species. Numbers along each line represent gains (+) and losses (-) of *MIR169* gene copies. The estimated divergence time for each species is given at each node in the tree according to (Paterson et al., 2009; Brachypodium-Sequencing-Initiative., 2010; Bennetzen et al., 2012; Zhang et al., 2012). The gain in *MIR169* copy number of sorghum relative to *Brachypodium* is depicted. Note: WGD in maize is used as a term to represent the allotetraploidy event that took place. NJ phylogenetic trees with bootstrap support are shown depicting the relationships of *MIR169* stem-loop sequences from the grass species shown in (A). **(B)** NJ phylogenetic tree with *Brachypodium* (bdi) and rice (osa) *MIR169* stem-loop sequences orthologous to sorghum *MIR169* copies on chromosome 7. **(C)** NJ phylogenetic tree with rice (osa) and foxtail millet (sit) *MIR169* stem-loop sequences (top) and rice, foxtail millet, sorghum (sbi), and maize (zma) *MIR169* stem-loop sequences (bottom) orthologous to *MIR169* copies on sorghum chromosome 2. **(D)** NJ phylogenetic tree depicting the relationship of foxtail millet and maize *MIR169* copies orthologous to sorghum *MIR169* copies on chromosome 1 (top), and *Brachypodium*, rice, foxtail millet, and maize *MIR169* copies orthologous to sorghum *MIR169* copies on chromosome 1 (bottom).

5.3.4. Polymorphisms in chromosomal inversions containing *MIR169* clusters

Through the analysis of three chromosomal regions in sorghum containing *MIR169* clusters and their alignment with the genomes of *Brachypodium*, rice, foxtail millet, and maize, I was able to identify four chromosomal inversions in total, one in rice chr3 containing *osa-MIR169r* (Figure 5.5); a second on sorghum chr7 containing *sbi-MIR169r*, *sbi-MIR169s*, *sbi-MIR169l*, *sbi-MIR169m*, and *sbi-MIR169n* (Figure 5.2); a third on maize chr1 containing *zma-MIR169l* (Figure 5.6); and the fourth on maize chr7 containing *zma-MIR169k* (Figure 5.7), respectively. The inversion on rice chr3 was absent from the corresponding collinear regions on *Brachypodium* chr1, sorghum chr1, and foxtail millet chr9 (Figure 5.5), indicating that the inversion happened after the split of rice from the common ancestor of sorghum and foxtail millet. The region on sorghum chr1 containing *sbi-MIR169o*, *sbi-MIR169t*, and *sbi-MIR169u* that was collinear with the inverted segment on rice chr3 was also collinear with an inverted segment on the homoeologous region of maize chr1 containing *zma-MIR169l* (Figure 5.6). However, the inversion did not occur on the homoeologous region on maize chr5, indicating that the inversion occurred after the allotetraploidization event that took place in maize. The inversion on sorghum chr7 containing *sbi-MIR169r*, *sbi-MIR169s*, *sbi-MIR169l*, *sbi-MIR169m*, and *sbi-MIR169n* cluster only occurred in this species (Figure 5.2 and Figure 5.4), suggesting that it took place after the split of sorghum from the common ancestor of sorghum and maize. The *MIR169* cluster on sorghum chr2 was collinear with an inverted region on maize chr7 containing *zma-MIR169k* (Figure 5.7). The homoeologous region on

chr2 did not exhibit the inversion, suggesting that it took place after the allotetraploidization event that occurred in maize.

In summary, four inversions containing *MIR169* copies were found in total, one in rice, one in sorghum, and two in maize. These inversions were lineage specific as none of them was present in a collinear region in the genome of a second grass species, indicating that these inversions happened after the species were formed.

5.3.5. Validation of newly identified *MIR169* gene copies in sorghum and maize

To experimentally validate the new *MIR169* gene copies found in sorghum through my syntenic analysis among grasses, I mapped previously sequenced small RNAs from sorghum stems (Calviño et al., 2011) to the newly predicted *MIR169t/u/v/r/s* hairpins. Similarly, to validate the newly described *zma-MIR169s* gene copy in maize, I constructed small RNA libraries from endosperm tissue belonging to cultivars B73, Mo17, and their reciprocal crosses (Table 5.2). Maize endosperm derived small RNAs were then mapped to the new *MIR169s* hairpin annotated in this study. I could effectively map small RNA reads to the stem-loop sequences of all five predicted microRNA169 in sorghum (with respect to *sbi-MIR169r/s*, see next section). In the case of *sbi-MIR169t* and *sbi-MIR169u*, the most abundant small RNA reads were derived from the miR169* sequence (Figure 5.11), although small RNAs derived from the canonical miR169 sequence were also found but in less abundance. The experimental validation of *sbi-MIR169v* was supported with mapping of small RNAs to the corresponding predicted mature miR169v

sequence (Figure 5.11). Regarding the experimental validation of the predicted *zma-MIR169s* copy in maize, I was able to detect small RNA reads derived from miR169s although their abundance was very low (Figure 5.11).

Library	#Raw sequences	#Sequences with perfect match to B73 genome	%
B73	14,371,575	3,805,955	26.48
Mo17	16,207,393	7,688,661	47.44
B73 x Mo17	13,051,982	5,985,649	45.86
Mo17 x B73	19,924,315	6,514,306	32.7

Table 5.2. Deep sequencing statistics of maize endosperm-derived small RNAs

```
>sbi-MIR169t
TAGCCAAGGATGATTTGCCTGtagctagcaacctctgagcgctcctgctgccatggcatggcagtcaggggcgcgtagtgggtgcttctccGGGCAAATCATCTGGGCTAG
 AGCCAAGGATGATTTGC      bc01,4                                                                  GGGCAAATCATCTGGGC    bc01,2
  GCCAAGGATGATTTGCC     bc01,3                                                                  GGGCAAATCATCTGGGCT   bc01,6
 AGCCAAGGATGATTTGCCTG   bc01,4                                                                  GGGCAAATCATCTGGGCTA  bc01,8
  GCCAAGGATGATTTGCC     bc02,2                                                                  GGGCAAATCATCTGGGCTAG bc01,6
 AGCCAAGGATGATTTGC      bc04,6                                                                  GGGCAAATCATCTGGGCTA  bc02,4
 AGCCAAGGATGATTTGC      bc02,1                                                                  GGGCAAATCATCTGGGCTAG bc02,4
 AGCCAAGGATGATTTGCC     bc02,2                                                                  GGGCAAATCATCTGGGCT   bc04,10
 AGCCAAGGATGATTTGCCTG   bc02,2                                                                   GGCAAATCATCTGGGCTA  bc04,2
TAGCCAAGGATGATTTGCCT    bc02,2                                                                  GGGCAAATCATCTGGGC    bc04,16
  GCCAAGGATGATTTGCC     bc04,4                                                                  GGGCAAATCATCTGGGCTA  bc04,32
 AGCCAAGGATGATTTGCC     bc04,2                                                                  GGGCAAATCATCTGGGCTAG bc04,52
 AGCCAAGGATGATTTGCCT    bc04,4                                                                CCGGGCAAATCATCTGG      bc05,2
TAGCCAAGGATGATTTGCCT    bc04,2                                                                  GGGCAAATCATCTGGGC    bc05,1
 AGCCAAGGATGATTTGCCTG   bc04,2                                                                  GGGCAAATCATCTGGGCT   bc05,4
 AGCCAAGGATGATTTGCCTGT  bc04,2                                                                    GCAAATCATCTGGGCTAG bc05,2
TAGCCAAGGATGATTTGCCTG   bc04,2                                                                  GGGCAAATCATCTGGGCTA  bc05,18
TAGCCAAGGATGATTTGCCTGT  bc04,8                                                                  GGGCAAATCATCTGGGCTAG bc05,8
TAGCCAAGGATGATTTGCCTGTA   bc04,2
TAGCCAAGGATGATTTGCCTGTAGC bc04,4
TAGCCAAGGATGATTTGCCTGT    bc05,2
                     AGCTAGCAACCTCTGAGCG         bc01,1
                         AGCAACCTCTGAGCGCTCCTGC  bc01,1
                     AGCTAGCAACCTCTGAGCGCTCC     bc02,1
                     AGCTAGCAACCTCTGAG           bc04,1
                     AGCTAGCAACCTCTGAGC          bc04,2
                     AGCTAGCAACCTCTGAGCGCT       bc04,1
                     AGCTAGCAACCTCTGAGCGCTCC     bc04,3
                     AGCTAGCAACCTCTGAGC          bc05,1
                     AGCTAGCAACCTCTGAGCGC        bc05,1


>sbi-MIR169u
aagaggcatctttgaTAGCCAGGGATGATTTGCCCTGtagcaccatgcatgcatgcaacctctcgcgttagctcctgctgactgcatgctgccatgacaagttccacggGCAAATCATTCCTGGCTAATCtgagtgcctctt
               TAGCCAGGGATGATTTGCCC    bc01,2                                                                  CAAATCATTCCTGGCTA    bc01,1
               TAGCCAGGGATGATTTG       bc04,2                                                                 GCAAATCATTCCTGGCT     bc01,1
               TAGCCAGGGATGATTTGC      bc04,1                                                                 GCAAATCATTCCTGGCTA    bc01,4
               TAGCCAGGGATGATTTGCCC    bc04,20                                                                GCAAATCATTCCTGGCTAA   bc01,8
               TAGCCAGGGATGATTTG       bc05,1                                                                 GCAAATCATTCCTGGCTAA   bc02,1
               TAGCCAGGGATGATTTGC      bc05,2                                                                 GCAAATCATTCCTGGCTAATC bc02,1
               TAGCCAGGGATGATTTGCCC    bc05,7                                                                 GCAAATCATTCCTGGCT     bc03,1
                                                                                                              GCAAATCATTCCTGGCT     bc04,26
                                                                                                               CAAATCATTCCTGGCTA    bc04,3
                                                                                                              GCAAATCATTCCTGGCTA    bc04,27
                                                                                                               CAAATCATTCCTGGCTAA   bc04,11
                                                                                                              GCAAATCATTCCTGGCTAA   bc04,80
                                                                                                               CAAATCATTCCTGGCTAAT  bc04,2
                                                                                                              GCAAATCATTCCTGGCTAAT  bc04,4
                                                                                                                AAATCATTCCTGGCTAATCT bc04,1
                                                                                                              GCAAATCATTCCTGGCTAATC  bc04,11
                                                                                                            ACGGGCAAATCATTCCTGGCTA     bc04,1
                                                                                                                 CAAATCATTCCTGGCTAATCTG bc04,1
                                                                                                              GCAAATCATTCCTGGCT       bc05,2
                                                                                                               CAAATCATTCCTGGCTA      bc05,1
                                                                                                              GCAAATCATTCCTGGCTA      bc05,5
                                                                                                               CAAATCATTCCTGGCTAA     bc05,4
                                                                                                              GCAAATCATTCCTGGCTAA     bc05,10
                                                                                                               CAAATCATTCCTGGCTAATC   bc05,1
                                                                                                               CAAATCATTCCTGGCTAATCTG bc05,1


>sbi-MIR169v
gcgatggaagctctgctttggTAGCCAAGGATGAGCTGCCTGtggcctccagctgcagaggctagctaggctacacattgcgtggccaagctcctccgctgcgcgtggtctcgcaGGCAGCCTCCTTGGCTAGTctgagtggcttccatc
                     TAGCCAAGGATGAGCTG       bc01,3                              TGGCCAAGCTCCTCCGCT bc02,1
                     TAGCCAAGGATGAGCTG       bc02,1                               GGCCAAGCTCCTCCGCT bc04,1
                     TAGCCAAGGATGAGCTG       bc04,8                                       CCTCCGCTGCGCGTGGTC bc05,1
                     TAGCCAAGGATGAGCTGCCTG   bc04,5
                     TAGCCAAGGATGAGCTG       bc05,9
                     TAGCCAAGGATGAGCTGCC     bc05,2
                     TAGCCAAGGATGAGCTGCCTG   bc05,1
    TGGAAGCTCTGCTTTGGTAGCCAA                 bc04,1
           TCTGCTTTGGTAGCCAA                 bc05,1
                                                    CTAGCTAGGCTACACAT bc05,1
```

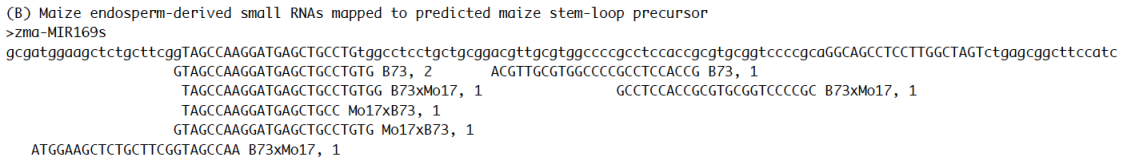


Figure 5.11. Experimental validation of predicted MIR169 stem-loop precursors in sorghum and maize. (A) Sorghum stem-derived small RNAs were mapped to sbi-*MIR169t*, sbi-*MIR169u* and sbi-*MIR169v* stem-loop sequences. Only sequences with perfect match to the BTx623 genome are shown. Predicted mature and star miR169 sequence is highlighted in capital letters on the stem-loop sequence. To the left side of each small RNA sequence a label is shown with information about the small RNA library from which it was sequenced (bc01: Mix library; bc02: BTx623 library; bc03: Rio library; bc04: low Brix and early flowering F2 library; bc05: high Brix and late flowering F2 library), together with the abundance of the small RNA read indicated by a number. **(B)** Maize endosperm-derived small RNAs were mapped to predicted stem-loop precursor zma-*MIR169s*.

5.3.6. Antisense MicroRNA169 gene pair generated small RNAs that targeted different set of genes

In rice, *osa-MIR169l* and *osa-MIR169q* were annotated as antisense microRNAs and small RNA reads derived from both strands were identified (Xue et al., 2009). In sorghum, *sbi-MIR169r*, and *sbi-MIR169s* are collinear with *osa-MIR169l/q* (Figure 5.2 and Figure 5.8) and are antisense microRNAs as well (Figure 5.3 and Figure 5.12). Despite the lack of Expressed Sequence Tag (EST) evidence for *sbi-MIR169r* and *sbi-MIR169s* annotation, my previously generated small RNA library from sorghum stem tissue (Calviño et al., 2011) supported the transcription from both strands based on small RNA reads mapped to both *sbi-MIR169r* and *sbi-MIR169s*, respectively (Figure 5.12). Similarly, EST evidence supported the transcription from opposite strands in the microRNA antisense pair *zma-MIR169e/h* (ESTs ZM_BFb0354L14.r and ZM_BFb0294A24.f, respectively). Because small RNAs derived from *zma-MIR169e/h* had not been previously reported (miRBase database: release 19, August 2012), I used the SOLiD system to sequence small RNAs from endosperm tissue derived from B73 and Mo17 cultivars and their reciprocal crosses; however, I could not detect small RNA reads derived from them, at least in endosperm tissue. Thus, antisense microRNAs from *MIR169* gene copies are being actively produced in rice and sorghum, and possibly in maize.

With respect to the *sbi-MIR169r/s* antisense gene pair, I found that the small RNA reads mapped to *sbi-MIR169r* were predominantly associated with the miR169r* sequence (Figure 5.12). The mature miRNA sequences for sbi-miR169r* and sbi-miR169s differed from each other in seven nucleotides (Figure 5.12).

Moreover, they would have different set of genes as targets based on their sequences (Figure 5.13 and Figure 5.14). Moreover, the assumption that also microRNA* have functional roles was recently described (Meng et al., 2011; Yang et al., 2011).

```
>sbi-MIR169r
gcaauaggggccacucaggcUAGCCAAGGAGACUGCCUAUGaaccaacucaaaggguucacauucugauccuuugggacaaaggacaUAGGCAAGUCAUCCUUGGCUAucagaggcagacccuuauu
                  GGCUAGCCAAGGAGACUGCC bc04, 15                                                UAGGCAAGUCAUCCUUGGCUA bc01, 2
                   GCUAGCCAAGGAGACUGCC bc02, 2                                                  AGGCAAGUCAUCCUUGGCU  bc01, 17
                   GCUAGCCAAGGAGACUGCC bc03, 1                                                  AGGCAAGUCAUCCUUGGCUA bc01, 34
                   GCUAGCCAAGGAGACUGCC bc05, 3                                                  AGGCAAGUCAUCCUUGGCUA bc02, 26
                    CUAGCCAAGGAGACUGCC bc05, 14                                                 AGGCAAGUCAUCCUUGGC   bc04, 3
                                                                                                 GGCAAGUCAUCCUUGGCUA bc03, 22
                                                                                                 GGCAAGUCAUCCUUGGCUA bc05, 103
                                                                                                 GGCAAGUCAUCCUUGGCUA bc04, 430
                                                                                                 GGCAAGUCAUCCUUGGCUA bc02, 88
                                                                                                  GCAAGUCAUCCUUGGCUA bc03, 2

>sbi-MIR169s
aauaagggucugccucugaUAGCCAAGGAUGACUUGCCUAuguccuuugucccaaaggaucagaaugugaaccuuugaguugguucauaGGCAGUCUCCUUGGCUAGCCUgaguggccccuauugc
                    UAGCCAAGGAUGACUUGCCUA bc04, 29                                              GGCAGUCUCCUUGGCUAG   bc01, 4
                     AGCCAAGGAUGACUUGCC   bc02, 1                                               GGCAGUCUCCUUGGCUAG   bc02, 2
                     AGCCAAGGAUGACUUGCCUA bc02, 2                                               GGCAGUCUCCUUGGCUAGCC bc02, 1
                      GCCAAGGAUGACUUGCCU  bc01, 1
                       CCAAGGAUGACUUGCCUA bc04, 1
```

Supplemental Figure 6B

```
sbi-miR169r*.seq TAGGCAAGTCATCCTTGGCTA
sbi-miR169s.seq  TAGCCAAGGATGACTTGCCTA
```

Figure 5.12. Antisense MIR169r/s gene pair generates small RNAs. Although sequencing of stem-derived small RNAs from grain and sweet sorghum were previously described (Calvino et al., 2011), I mapped small RNAs from my sequenced libraries to the newly annotated sbi-*MIR169r* and sbi-*MIR169s* hairpin structures. **(A)** The most abundant small RNA reads mapped to sbi-*MIR169r* corresponded to the miR169r* sequence, whereas the most abundant small RNA reads mapped to sbi-*MIR169s* corresponded to miR169s, respectively. **(B)** Nucleotide polymorphism between miR169r* and miR169s.

```
>Expectation: 1.5
miR169r*                  20 UCGGUUCCUACUGAACGGAU 1
                             ::::::: :::::::::::
Sb08g021910.1 3'UTR      716 AGCCAAGAAUGAUUUGCCUA 735
CCAAT-binding transcription factor subunit B family protein

>Expectation: 2.5
miR169r*                20 UCGGUUCCUACUGAACGGAU 1
                           ::::::: :::: ::::::.
Sb01g011220.1 3'UTR   1259 AGCCAAGAAUGAAUUGCCUG 1278
CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B

>Expectation: 3.5
miR169r*                  20 UCGGUUCCUACUGAACGGAU 1
                             ::::::::.  ::::::.::
Sb01g035610.1 3'UTR    1128 AGCCAAGGGAUACUUGUUUA 1147
ATP-dependent Clp protease proteolytic subunit

>Expectation: 3.5
miR169r*                     21 AUCGGUUCCUACUGAACGGAU 1
                                :. ::::: :::.::::::::.
Sb02g026600.1 7th exon      1648 UGCCCAAGCAUGGCUUGCCUG 1668
similar to Abscisic acid 8'-hydroxylase 3

>Expectation: 3.0
miR169r*                    19 AUCGGUUCCUACUGAACGG 1
                               :. ::::: :::.:::::::
Sb02g026600.1 7th exon   1648 UGCCCAAGCAUGGCUUGCC 1666
similar to Abscisic acid 8'-hydroxylase 3

>Expectation: 3.5
miR169r*                     21 AUCGGUUCCUACUGAACGGAU 1
                                : : .::::::::.:::::::.
Sb01g047950.1 3rd exon      980 UUGAUAAGGAUGGCUUGCCUG 1000
similar to Ankyrin repeat protein, chloroplast, putative, expressed

>Expectation: 3.5
miR169r*                  20 UCGGUUCCUACUGAACGGAU 1
                             ::::::::::  ::::::::
Sb06g001950.1 6th exon   582 AGCCAAGGAGAACUUGUCUU 601
Phosphoglycerate mutase

>Expectation: 2.0
miR169r*           18 UCGGUUCCUACUGAACGG 1
                      .::::: ::::::.::::::
Sb01g043590.1  2076  GGCCAAUGAUGAUUUGCC 2093
similar to CUE domain containing protein, expressed

>Expectation: 2.5
miR169r*                  18 UCGGUUCCUACUGAACGG 1
                             ::: :::::::::.::::
Sb01g043450.1 3rd exon  582 AGCAAAGGAUGAUUUGCA 599
```

```
Pfam: Syntaxin 6, N-terminal

>Expectation: 2.5
miR169r*                  18 UCGGUUCCUACUGAACGG 1
                             ::::.: :::::.::::.
Sb06g024340.1 4th exon   834 AGCCGAUGAUGAUUUGCU 851
similar to DNAJ heat shock N-terminal domain-containing protein-like

>Expectation: 2.5
miR169r*                 18 UCGGUUCCUACUGAACGG 1
                             ::::: ::::::::::.
Sb03g028620.1 3rd exon  949 CGCCAAAGAUGACUUGCU 966
similar to Cytochrome P450

>Expectation: 3.0
miR169r*                    18 UCGGUUCCUACUGAACGG 1
                                ::::: ::::::::::
Sb08g004540.1 2nd exon      669 UGCCAAUGAUGACUUGCA 686
similar to 4-alpha-L-fucosyltransferase

>Expectation: 3.0
miR169r*                   19 AUCGGUUCCUACUGAACGG 1
                              :::::::::: ::.::: :
Sb01g035620.1 25th exon  3062 UAGCCAAGGAAGAUUUGGC 3080
similar to AAA-type ATPase family protein, putative, expressed

>Expectation: 3.0
 miR169r*                 19 AUCGGUUCCUACUGAACGG 1
                             :. : .:::::::::::::
Sb01g032770.1 3rd exon   680 UGCCGGAGGAUGACUUGCC 698
weakly similar to OSMYB3

>Expectation: 3.0
miR169r*                 19 AUCGGUUCCUACUGAACGG 1
                            :.:::::::::::.:: .:
Sb03g001440.1 6th exon 3634 UGGCCAAGGAUGAUUUCUC 3652
similar to Ethylene insensitive 2

>Expectation: 3.5
miR169r*             19 AUCGGUUCCUACUGAACGG 1
                        :::: :.::::::::: ::
Sb01g038240.1        32 UAGCGAGGGAUGGCUUCCC 50
Mitochondrial import inner membrane translocase, subunit TIM22

>Expectation: 1.5
miR169r*                18 UCGGUUCCUACUGAACGG 1
                           ::::::: ::::.:::::
Sorghum EST: TC130929  391 AGCCAAGAAUGAUUUGCC 408 EST mapped to a segment of Sb01g013430.1
CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B

>Expectation: 3.5
miR169r*                      18 UCGGUUCCUACUGAACGG 1
                                   :::::::::::::: ::
Sb01g036110.1 17th exon      2618 GUCCAAGGAUGACUUACC 2635


similar to Insulinase containing protein, expressed
```

Figure 5.13. List of predicted targets of sbi-miR169r*. The psRNATarget program was used to predict mRNAs targeted by sbi-miR169r*. The miR169r*-target alignment is shown together with the expectation level of the prediction with 1 as high confident and 3.5 less confident. The annotation for each predicted gene is shown in conjunction with the region where the miR169r* recognition sequence is located (exon or 3'UTR).

```
>Expectation: 3.0
miR169s                20 UCCGUUCAGUAGGAACCGAU 1
                          ::::: ::::::::::::
Sb01g045500.1 3'UTR   935 UGGCAACUCAUCCUUGGCUU 954
CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B

>Expectation: 3.5
miR169s                 20 UCCGUUCAGUAGGAACCGAU 1
                           :::::. :..: ::::::::
Sb01g027540.1 3'UTR   1668 AGGCAGCUUGUACUUGGCUA 1687
similar to Serine carboxypeptidase family protein, expressed

>Expectation: 4
miR169s                    21 AUCCGUUCAGUAGGAACCGAU 1
                              ::  ::: ::::.:::::::.
Sb10g005870.1 10th exon  1508 UAAUCAAAUCAUUCUUGGCUG 1528
similar to Serine carboxypeptidase II-2 precursor (EC 3.4.16.6) (CP-MII.2)

>Expectation: 2.0
miR169s                21 AUCCGUUCAGUAGGAACCGAU 1
                          ::::::: ::::.:::::::.
Sb08g021910.1 3'UTR   716 UAGGCAAAUCAUUCUUGGCUG 736
CCAAT-binding transcription factor subunit B family protein, expressed

>Expectation: 2.5
miR169s                20 UCCGUUCAGUAGGAACCGAU 1
                          :::::: ::::.:::::::
Sb01g011220.1 3'UTR  1260 AGGCAAUUCAUUCUUGGCUU 1279
CCAAT-binding transcription factor (CBF-B/NF-YA) subunit B
```

Figure 5.14. List of predicted targets of sbi-miR169s. The psRNATarget program was used to predict mRNAs targeted by sbi-miR169s. The miR169s-target alignment is shown together with the expectation level of the prediction with 1 as high confident and 3.5 less confident. The annotation for each predicted gene is shown in conjunction with the region where the miR169s recognition sequence is located (exon or 3'UTR).

5.3.7. Linkage of *MIR169* gene copies with flowering and plant height genes

Based on the alignment of collinear regions containing *MIR169* genes located on sorghum chr2 and chr7, I noticed a tight linkage of *MIR169* copies with two genes encoding a bHLH protein, and a B-box zinc finger and CCT-motif protein that were similar to *Arabidopsis* bHLH137 and CONSTANSLIKE 14 proteins (Figures 5.2, 5.4, 5.7, 5.8, and 5.9). The *Arabidopsis bHLH137* and *COL14* genes were described to have a role in gibberellin signaling (mutations in genes involved in gibberellin signaling and/or perception affects plant height (Fernandez et al., 2009) and flowering time, respectively (Griffiths et al., 2003; Wenkel et al., 2006; Zentella et al., 2007). The physical linkage of *MIR169* gene copies to bHLH and COL genes (or any of the two) was present in all the five grasses examined. We hypothesized that the physical association of *MIR169* to either of these flowering and/or plant height genes could be of relevance because of previously reported trade-offs in sorghum between sugar content in stems and plant height and flowering time, respectively (Murray et al., 2008). For breeding purposes, the introgression of a particular gene/phenotype from a specific cultivar into another would consequently also bring in the neighboring gene, a process known as linkage drag. Furthermore, linkage drag between *MIR169* copies and the bHLH and COL genes could also be of ecological importance because a single chromosomal segment comprises genes involved in drought tolerance, sugar accumulation, and flowering. If this is the case, linkage of *MIR169* copies to either bHLH or COL genes could have been preserved even after the monocotyledonous diversification. Indeed, I was able to find collinearity

between chromosomal segments containing *MIR169* and bHLH genes from *Brachypodium*, sorghum, soybean, and cassava (Figure 5.15). Moreover, I found that the physical linkage between *MIR169* and the bHLH gene on sorghum chr7 was retained in collinear regions of soybean chr6 and cassava scaffold 01701, respectively (Figure 5.15). Similarly, the physical/genetic association of *MIR169* with the bHLH gene from sorghum chr2 was retained in the corresponding collinear regions from soybean chr8 and cassava scaffold 09876 (Figure 5.16). Interestingly, the linkage between *MIR169* and the COL gene that was present in *Brachypodium* chr3 and sorghum chr7 was broken in the corresponding collinear regions of soybean chr6 and cassava scaffold 01701 (Figure 5.15). I then compared the two *MIR169* clusters from sorghum chr2 and chr7 with the grapevine genome because grapevine and sorghum are more closely related than sorghum to soybean and cassava, respectively. My comparison revealed a two-to-three relationship between sorghum and grapevine (Figure 5.17), and this is consistent with the paleo-hexaploidy event that took place in the grapevine genome (Jaillon et al., 2007). The physical/genetic linkage of *MIR169* copies with the COL gene on sorghum chr7 was preserved in two of the three homoeologous chromosomal segments in grapevine on chr1 and chr14, whereas the third homoeologous segment on chr17 retained the close association of *MIR169* with the bHLH gene.

The finding of microsynteny conservation between monocots and dicots species in chromosomal segments containing *MIR169* gene copies together with bHLH and COL genes is remarkable because the estimated time of divergence between monocots and dicots is of approximately 130 to 240 Ma (Wolfe et al., 1989;

Jaillon et al., 2007). Such microsynteny conservation permitted the discovery of new *MIR169* gene copies in soybean (*gma MIR169w*, *gma-MIR169x* and *gma-MIR169y*), cassava (*mes-MIR169w* and *mes-MIR169y*), and grapevine (*vvi-MIR169z*).

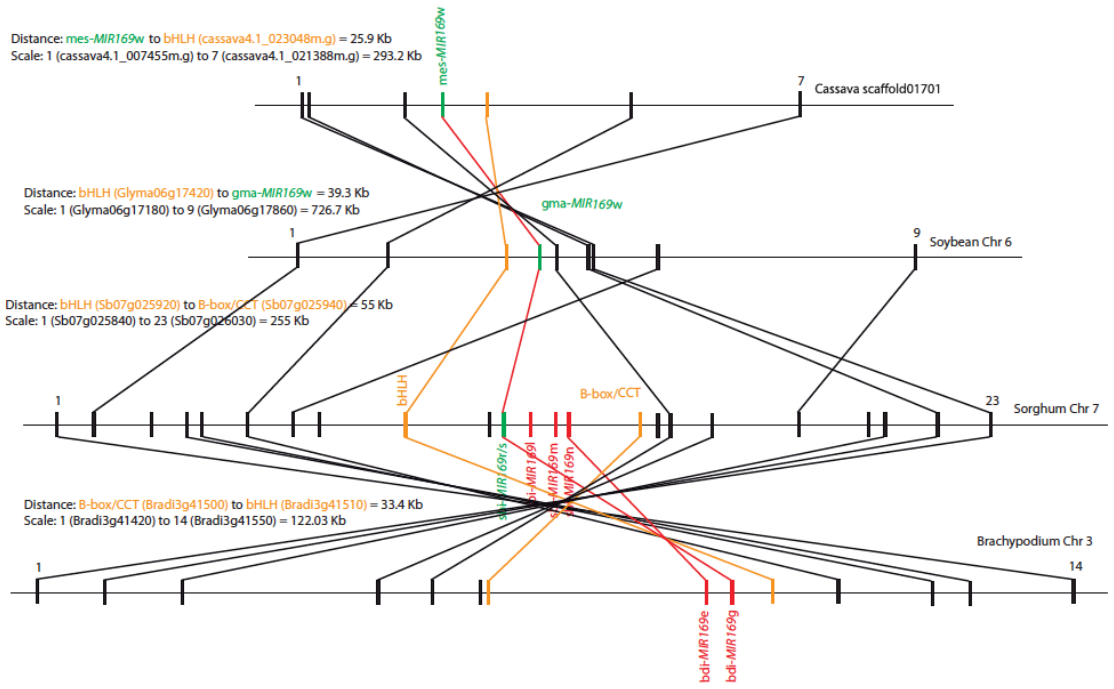


Figure 5.15. Sequence alignment of sorghum *MIR169* cluster on chr7 with orthologous regions from *Brachypodium*, soybean, and cassava. There is conservation of synteny between monocot species *Brachypodium* and sorghum and dicot species soybean and cassava when chromosomal segments containing *MIR169* gene copies and their flanking genes are aligned. Conservation of synteny allowed the identification of new *MIR169* gene copies on soybean chromosome 6 (gma-*MIR169w*) and cassava scaffold 01701 (mes-*MIR169w*), respectively. Physical association on the chromosome between *MIR169* and the flanking bHLH gene was retained in soybean and cassava as well. Notice the inversion on soybean chr6.

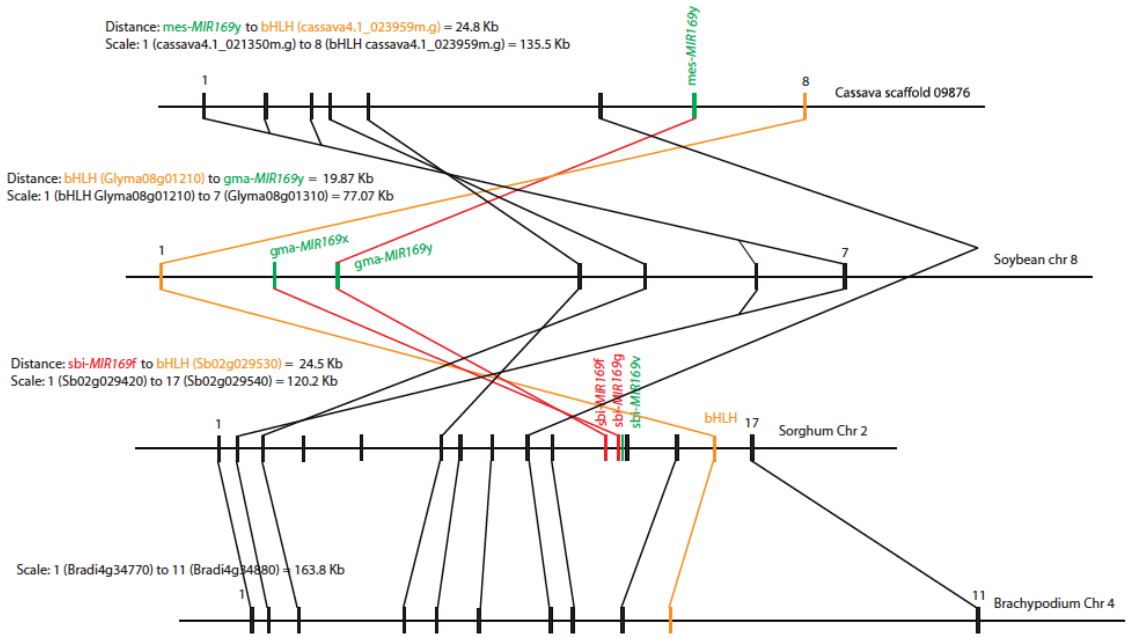


Figure 5.16. Sequence alignment of sorghum *MIR169* cluster on chr2 with orthologous regions from *Brachypodium*, soybean, and cassava. The alignment of the sorghum *MIR169* cluster on chr2 with soybean chr8 and cassava scaffold 09876 allowed the identification of two new *MIR169* gene copies in soybean (gma-*MIR169x* and gma-*MIR169y*) and one new copy in cassava (mes-*MIR169y*), respectively. The physical association of *MIR169* gene copies with the bHLH was retained in soybean and cassava. An inversion occurred on soybean chr8.

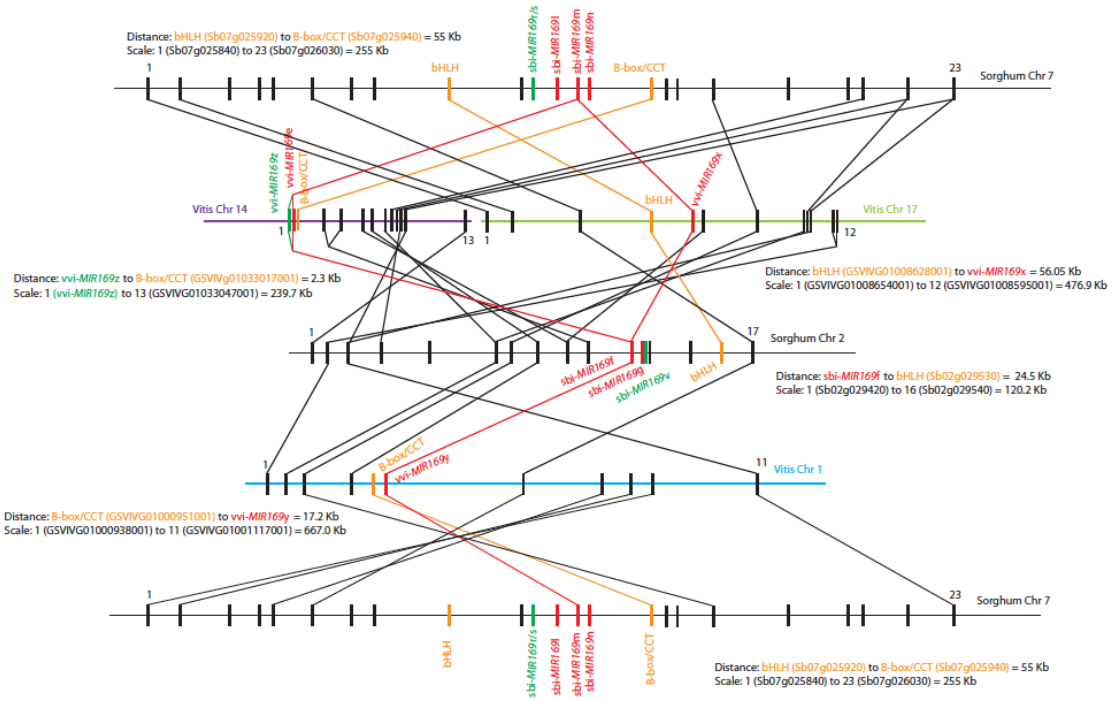


Figure 5.17. Conservation of synteny between sorghum and grapevine chromosomal segments containing *MIR169* gene copies. Sorghum segments containing *MIR169* gene clusters from chr2 and chr7 were aligned to the grapevine genome based on orthologous gene pairs. Because grapevine is a hexopaleo-polyploid, we found a 2:3 chromosomal relationship between sorghum and grapevine. Collinearity allowed the identification of a new MIR169 copy (vvi-*MIR169z*) in grapevine chr14. Different grapevine chromosomes are represented in colors, whereas sorghum chromosomes are in black. Relative to sorghum chr2, grapevine had an inversion event on chr14 and chr17. The association of *MIR169* with its flanking COL gene was maintained on grapevine chr14 and chr1, whereas the association of *MIR169* with the bHLH gene was maintained on chr1.

5.3.8. Subfunctionalization of the bHLH gene in the *MIR169* cluster of *Brachypodium*

The microsynteny in chromosomal segments containing miR169 gene copies flanked by the bHLH gene among such distantly related species such as *Brachypodium* and cassava suggested that the linkage between miR169 and bHLH resulted from selection because of the divergence from a common ancestor approximately 130–240 Ma. In support of this interpretation, the bHLH gene on *Brachypodium* chr4, where the miR169 cluster had been deleted (Figure 5.9), appeared to have undergone subfunctionalization. First, the bHLH copy on *Brachypodium* chr4 involved the loss of the basic domain, which is involved in DNA binding (Toledo-Ortiz, 2003) and thus evolved into a HLH protein (Figure 5.18). Because bHLH proteins act as homo- and/or heterodimers, where the basic domain of each bHLH protein bind to DNA, HLH proteins homo- or heterodimerize and prevent the binding of the complex to DNA and thus becomes a negative regulator (Toledo-Ortiz, 2003). Second, *Brachypodium* has a redundant intact orthologous copy on chr3, also a miR169 cluster next to it (Figure 5.8). Third, the synonymous and nonsynonymous substitution rate of the HLH orthologous gene pairs was higher than the synonymous and nonsynonymous substitution rate in the bHLH orthologous gene pairs, respectively (Figure 5.18). Fourth, when I run a test for detecting adaptive evolution (calculated as the number of replacement mutations per replacement sites [dN] divided by the number of silent mutations per silent site [dS]) in the bHLH and HLH coding sequences, I found evidence on purifying selection on the HLH gene sequence (dN/dS ratio of −4.647).

Conservation of synteny between sorghum and grapevine showed that the linkage between *MIR169* gene copies and the COL gene was maintained in both species. Both COL genes in grapevine, on chr14 and on chr1, lost the B-box and zinc finger domain, whereas the orthologous copy in sorghum retained it (Figure 5.19). Similarly, foxtail millet COL protein lost the B-box and zinc finger domain, whereas *Brachypodium*, rice, and maize retained it. The B-box and zinc finger domain are thought to mediate protein–protein interactions, whereas the CCT domain acts as a nuclear localization signal, with mutations in both domains causing flowering time phenotypes (Griffiths et al., 2003; Wenkel et al., 2006; Valverde, 2011). Although the COL gene on grapevine chr14 has been recently identified as a candidate gene for a flowering Quantitative Trait Loci (QTL) (Duchêne et al., 2012), the function of its corresponding orthologous copy on sorghum chr7 remains to be elucidated.

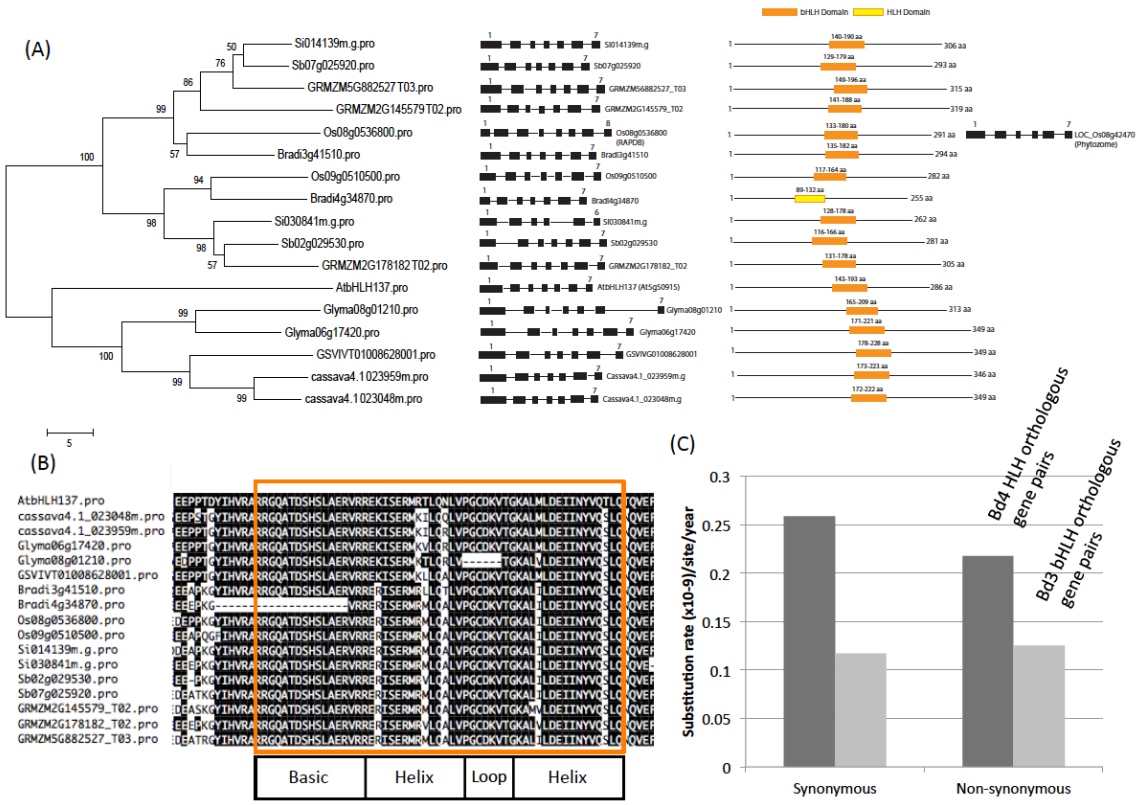


Figure 5.18. Sub-functionalization of *Brachypodium* bHLH gene copy. (A) Left: Neighbor Joining (NJ) phylogenetic tree of orthologous bHLH proteins with the *Arabidopsis* bHLH137 protein as reference. Middle: a representation of the gene structure in exons (boxes) and introns (lines) (5' and 3' UTRs not included). Right: graphic representation of the linear protein with the bHLH domain represented as an orange box and the HLH domain as a yellow box with orange border. **(B)** Protein alignment highlighting the bHLH motif with *AtbHLH137* protein as reference. The *Brachypodium* protein encoded by the gene Bradi4g34870 lost most of the basic domain, becoming a HLH protein instead. **(C)** Graph depicting the average synonymous and non-synonymous substitution rate of the bHLH Bradi3g41510 orthologous gene pairs compared to HLH Bradi4g34870 orthologous gene pairs.

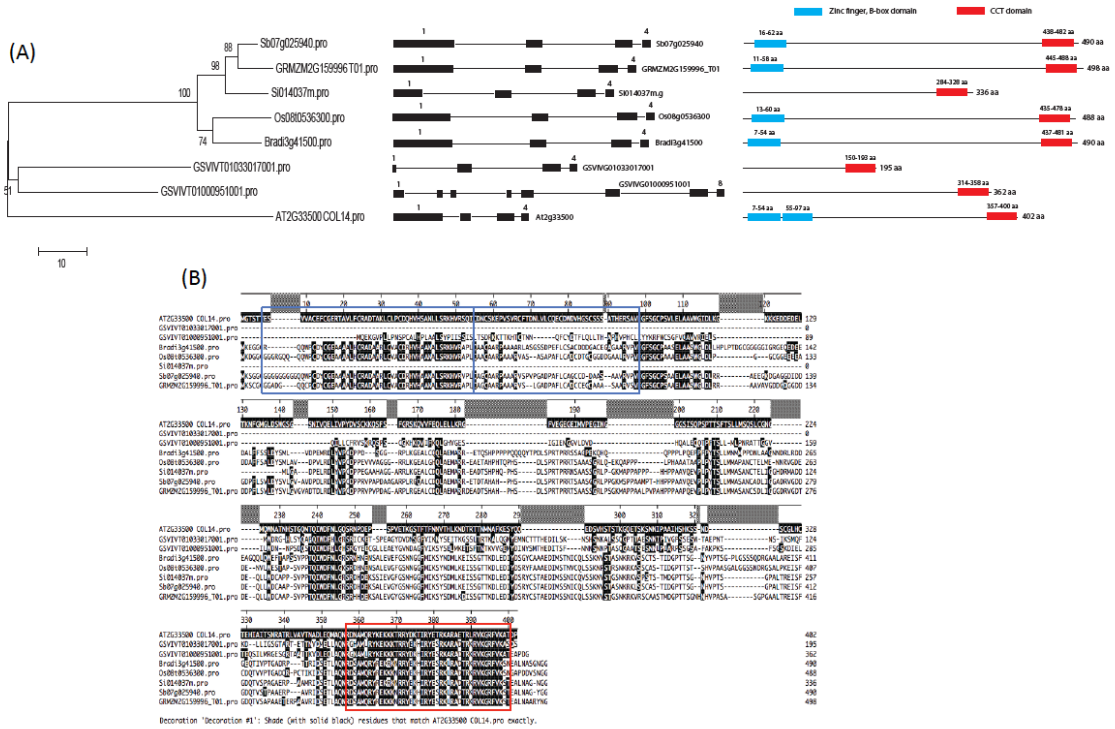


Figure 5.19. Evolution of the Zinc finger, B-box and CCT domain protein. (A) Left: Neighbor Joining (NJ) phylogenetic tree of B-box and CCT motif orthologous proteins with *Arabidopsis* COL14 protein as reference. Center: graphic representation of the B-box and CCT motif gene structure for each species with exons as boxes and introns as lines (5' and 3' UTRs not shown). Right: linear representation of the B-box and CCT motif protein for each species with the Zinc finger, B-box domain shown as a blue box where the CCT domain is shown as a red box. **(B)** Protein alignment highlighting the Zinc finger, B-box domain in blue boxes (*Arabidopsis* COL14 has two) and the CCT domain in a red box.

5.4. Discussion

I described the alignment of 25 chromosomal regions with orthologous gene pairs from eight different plant species. These regions contained a total of 48 *MIR169* gene copies, from which 22 of them were described and annotated here for the first time. The alignment of sorghum chromosomal regions containing *MIR169* clusters to their corresponding orthologous regions from *Brachypodium*, rice, foxtail millet, and maize, respectively, allowed me not only to better understand the differential amplification of *MIR169* gene copies during speciation but also to identify new *MIR169* gene copies not previously annotated in the rice, sorghum, and maize genomes. My work highlighted the usefulness of this approach in the discovery of microRNA gene copies in grass genomes and surprisingly also in dicotyledonous genomes such as those from grapevine, soybean, and cassava. In addition, collinearity among grasses was used to predict and annotate *MIR169* hairpin structures in the foxtail millet genome *de novo*, from which no current microRNA annotation was available from the miRBase database at the time (Release 19: August 2012). My work suggested that synteny-based analysis should complement (whenever possible) homology-based searches of new microRNA gene copies in plant genomes.

My analysis of *MIR169* gene copies organized in clusters in the sorghum genome revealed that sorghum acquired eight *MIR169* gene copies after *Brachypodium* split from a common ancestor, primarily due to gene losses (up to 5 MIR169 gene copies) in the *Brachypodium* lineage and new gene copies (up to 3) in the sorghum lineage (Figure 5.10). I propose that differences in MIR169 gene copy

number between sorghum and *Brachypodium* is based on selective amplification in sorghum. Because diploidization of the maize genome resulted in the deletion of duplicated gene copies after allotetraploidization approximately 4.7 Ma (Messing et al., 2004; Swigonova et al., 2004), also resulted in selective amplification in sorghum. Maize lost more than half, 9 of 16 *MIR169* gene copies, after allotetraploidization. Single gene losses in maize appeared to be caused by short deletions that were predominantly in the 5–178 bp size range, with these deletions being approximately 2.3 times more frequent in one homoeologous chromosome than in the other (Woodhouse et al., 2010). This observation is particularly relevant to maize microRNAs genes with average length distributions at the 5′-regions of their primary microRNAs (pri-miRNAs) in the order of 100–300 nt (Zhang et al., 2009). Although I detected chromosome breaks of the *MIR169* neighboring gene *COL14* on the maize homoeologous chr1–chr4 pair (Figure 5.4) and the bHLH gene on maize homeologous chr2–chr7 pair (Figure 5.7), retention of the bHLH gene copy on both homoeologous regions from chr1 and chr4 was observed (Figure 5.4). It was observed that transcription factors were preferentially retained after whole-genome duplication (WGD) (Xu and Messing 2008; Murat et al., 2010), with a recent study showing that from 2,943 sorghum–maize syntenic shared genes, 43% of them were retained as homoeologous pairs in maize, from which transcription factors were 4.3 times more frequently among retained genes than other functions (Woodhouse et al., 2010).

Alignment of sorghum regions containing *MIR169* gene copies on chr2 and chr7 with their respective collinear regions from *Brachypodium*, rice, foxtail millet,

and maize revealed the close linkage of *MIR169* gene copies with their flanking *COL14* and bHLH genes in all five grasses examined. Furthermore, collinearity of *MIR169* gene copies with either the *COL14* and/or the bHLH genes extended to dicot species such as grapevine, soybean, and cassava. Previously, it was suggested that conservation of collinearity between monocot and dicot species is rather rare because of the dynamic genomic rearrangements in genomes over 130–240 Ma (Wolfe et al., 1989; Jaillon et al., 2007). Still, conservation of synteny between rice and grapevine was also previously observed (Tang et al., 2010). Therefore, I hypothesized that preservation of collinearity in rare cases was subject to selection even after WGD events. In support of this hypothesis, the subfunctionalization and higher protein divergence rate of the HLH gene in *Brachypodium* chr4, where the *MIR169* cluster was deleted, occurred in comparison to the orthologous bHLH copy on chr3 with the *MIR169e* and *MIR169g* copies next to it. Indeed, trade-offs between sugar content and flowering time/plant height were reported in sorghum (Murray et al., 2008). When two genes controlling linked phenotypes are in close proximity on the chromosome for selection to act on both of them, the loss of one gene releases selection pressure on the other gene, allowing it to diverge. On the basis of its similarity to Arabidopsis *bHLH137*, which was postulated as putative DELLA target gene that functions in the GA response pathway (Zentella et al., 2007), I hypothesized that the grass homolog may function either in flowering and/or plant height, which future research will have to confirm. On the other hand, the importance of COL family proteins in the regulation of flowering time is well known (Griffiths et al., 2003; Wenkel et al., 2006). Collinearity between sorghum and

grapevine revealed the tight association of *COL14* with *vvi-MIR169z* and *vvi-MIR169e* on grapevine chr14, with the three genes contained within a 2.3 Kb interval. Furthermore, *COL14* has been recently considered a candidate gene for a flowering QTL in grapevine (Duchêne et al., 2012). With such a short physical distance between a flowering time gene and two *MIR169* gene copies, it is tempting to propose that grapevine breeding for late or early flowering time could have brought different *COL14* alleles together with its neighboring *MIR169* genes, a process known as linkage drag. Interestingly, although we could not find extensive collinearity between sorghum and *Arabidopsis thaliana* as to draw a synteny graph, I did find a close association on chr5 between *COL4* gene and *ath-MIR169b*, separated each other 61.7 kb (data not shown).

On the basis of these considerations, I proposed a hypothesis where the linkage of *MIR169* gene copies with the neighboring COL gene could have coevolved (Figure 5.20). This hypothesis was based on the findings presented here, together with a previous report describing that CO and COL proteins can interact through their CCT domains with proteins belonging to the NF-Y (HAP) family of transcription factors (Wenkel et al., 2006); specifically, it was described that CO together with COL15 interacted with NF-YB and NF-YC displacing NF-YA from the ternary complex. The mRNAs encoded by the NF-YA gene family are known targets of miR169 (Li et al., 2008). Thus, the association on the chromosome of a COL gene with a *MIR169* gene or gene cluster would ensure that miR169 would reduce the expression of the NF-YA mRNA and thus its protein levels, so that the COL protein can replace NF-YA in the ternary complex and drive transcription of CCAAT box

genes. Furthermore, this hypothesis could provide a genetic framework where to test the previously known drought and flowering trade-offs: When plants were exposed to drought stress during the growing season, they flowered earlier than control plants under well-watered environments (Franks et al., 2007), with the response being genetically inherited. For this reason, I decided to term my model the "Drought and Flowering Genetic Module Hypothesis."

I can envision a prominent role of linkage drag in breeding sorghum for enhanced biofuel traits such as high sugar content in stems and late flowering time for increased biomass. Under the *MIR169-bHLH* and/or *MIR169-COL* linkage drag model, any breeding scheme in sweet sorghum whose aim is to increase plant biomass through delayed flowering by crossing cultivars with different COL and/or bHLH alleles on either chr7 or chr2, respectively, should take into account the allelic variation at the neighboring *MIR169* gene copies as they may affect sugar content in stems and drought tolerance. The same can be said in breeding sorghum for grain production where the norm is to increase germplasm diversity among grain sorghums through the introduction of dwarf and early flowering genes from a donor line into exotic tall and late flowering lines with African origins (Brown et al., 2008).

On the basis of my results from comparative genomics analysis, I envision that any conservation in collinearity between closely associated genes (in this particular study between a microRNA and a protein-coding gene) controlling related phenotypes that is conserved among several plant species might be subject to linkage drag through breeding, opening a new area of research in genomics assisted breeding. In support of this notion, the early development of conserved ortholog set

markers (referred as COS markers) among different plant species (Fulton et al., 2002) highlighted the existence of a set of genes with synteny conservation because of the early radiation of dicotyledonous plants that can be used in mapping through comparative genomics. In addition, conservation in linkage between candidate genes for seed glucosinolate content and SSR markers between *Arabidopsis* and oilseed rape (*Brassica napus ssp. napus*) were used in marker-assisted selection in breeding oilseed rape for total glucosinolate content (Hasan et al., 2008).

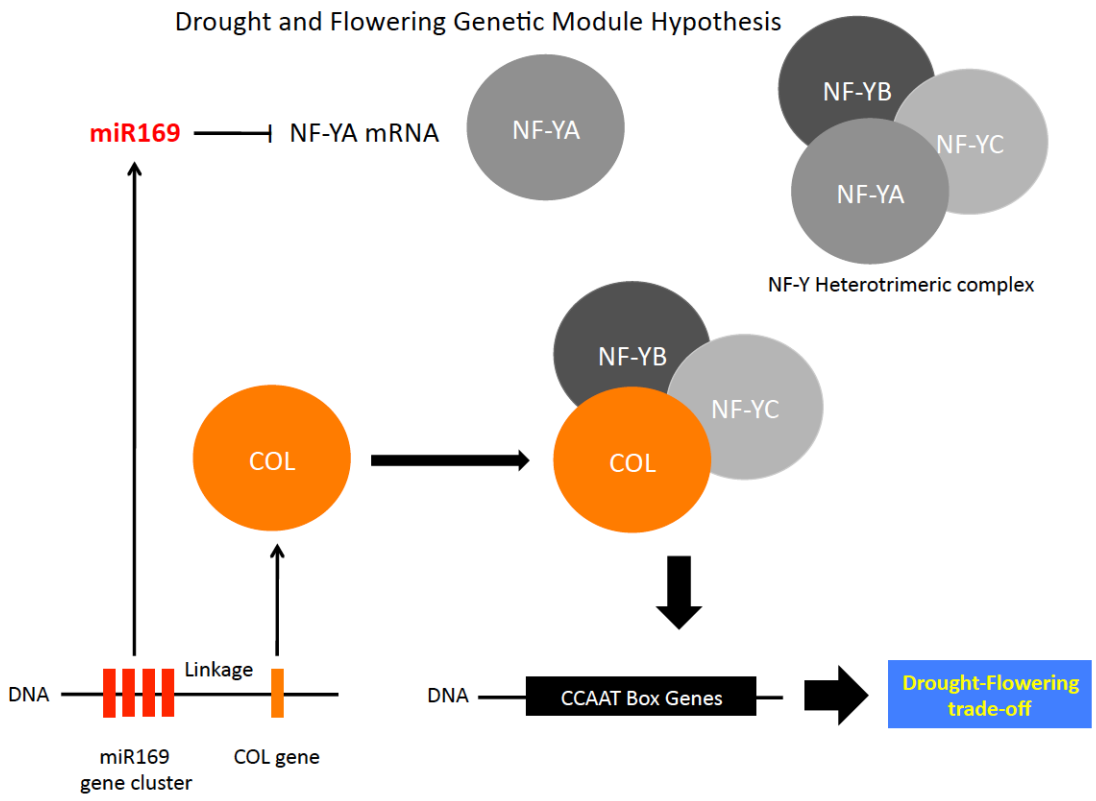


Figure 5.20. The "Drought and Flowering Genetic Module Hypothesis". Here I suggest that trade-offs between drought stress and flowering time could be explained in part by the genetic linkage of MIR169 and COL genes. In this model, a given COL gene genetically linked to a MIR169 gene will be positively selected over any other COL gene located somewhere else in the genome. This is so because COL proteins can replace the NF-YA (HAP2) subunit from the NF-YA, NF-YB (HAP3) and NF-YC (HAP5) hetero-trimeric transcription factor complex (Wenkel et al., 2006), with NF-YA mRNA targeted by miR169 (Li et al., 2008). Thus, depending on water availability, plants can adjust their flowering time according to the severity of drought during the growing season by modulating the expression of miR169 and COL genes. Under this scenario, high miR169 expression lower NF-YA mRNA levels,

consequently decreasing NF-YA protein levels, which may in turn increase the frequency of COL protein to interact with NF-YB and NF-YC subunits and thus guide the transcription complex toward the expression of CCAAT box genes involved in flowering. The current model establishes a genetic framework to explain the observation that plants flowered earlier under drought compared to well-watered environments (Franks et al., 2007).

5.5. Materials and methods

5.5.1. DNA sequences

Rice sequences were downloaded from the Rice Annotation Project Database website (http://rapdb.dna.affrc.go.jp/), whereas *Brachypodium*, foxtail millet, sorghum, maize, grapevine, soybean, and cassava sequences were downloaded from the Join Genome Institute website (www.phytozome.net). MicroRNA sequences were downloaded from the miRBase database (http://www.mirbase.org/).

5.5.2. *MIR169* gene prediction and annotation

Stem-loop precursors/hairpin structures from previously annotated *MIR169* genes were used in reciprocal Blastn analysis during the process of creating synteny graphs. Previously known *MIR169* stem-loop precursors were used as query sequences with Blastn. When the corresponding target sequences identified matched a genomic region where there was no any previous annotation of a *MIR169* gene copy, I took a 100–300 bp segment and fed it into an RNA folding program (RNAfold web server: http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) to look for

signatures of hairpin-like structures typical of microRNAs. Guidelines in microRNA gene prediction were followed as suggested by (Meyers et al., 2008).

5.5.3. Experimental validation of predicted *MIR169* genes

I took advantage of my previously sequenced small RNA libraries from sorghum stems (Calviño et al., 2011) and mapped small RNAs to the newly predicted *MIR169r/s/t/u/v* hairpin sequences. To validate the newly predicted *MIR169s* in maize, I used the SOLiD platform to sequence small RNAs derived from endosperm tissue from B73 and Mo17 inbred lines as well as endosperm tissue derived from their reciprocal crosses. Small RNA reads were then mapped to *zma-MIR169s* stem-loop precursor.

5.5.4. Prediction of miR169 targets

Target prediction was conducted in sorghum for the newly discovered miR169r* and miR169s microRNAs using the Small RNA Target Analysis Server psRNATarget (Dai and Zhao, 2011) at http://plantgrn.noble.org/psRNATarget/. In addition to the sorghum genome sequence incorporated into psRNATarget (Sorghum DCFI Gene Index SBGI Release 9) as preloaded transcripts, I also uploaded a FASTA file from phytozome (http://www.phytozome.net/dataUsagePolicy.php?org=Org_Sbicolor) with all sorghum genes coding sequences and used this data set for target prediction as well. Target prediction was conducted for the annotated 21 nt miR169 and for the most abundant small RNA reads different from 21 nt in size that matched the predicted

miR169 sequence (miR169 variants).

5.5.5. Estimation of MIR169 gene number in ancestral species

To estimate the numbers of *MIR169* genes in ancestral species of the grass family together with gains and losses of *MIR169* copies during grass evolution, I took the parsimony approach as described previously by (Nozawa et al., 2012).

5.5.6. Estimation of substitution rates in *MIR169* genes and ancient duplication time

To study the rate of nucleotide substitution in *MIR169* genes, I aligned *MIR169* stem-loop sequences using MUSCLE, available with the MEGA5 software package (Tamura et al., 2011). When I analyzed the gained *MIR169* gene copy that gave rise to *sit-MIR169h*, *sbi-MIR169v*, and *zma-MIR169s* copies (Figure 5.10A: region miR169 cluster on sorghum chr2), I first computed the average (Jukes and Cantor) distance (D_a) between *zma-MIR169s/sbi-MIR169v* and *zma-MIR169s/sit-MIR169h* gene pairs. The substitution rate (R) was subsequently calculated with the formula $R = D_a /2T$, where T was the divergence time (in this case 26 million years ago [Ma]), when the ancestor of maize and sorghum diverged from foxtail millet. I then calculated the ancient duplication time at which *sit-MIR169h* arose by using the formula $t = d_a/2R$, where t is the divergence time of two sequences and d_a is the average distance between sequences in the miR169 cluster (the average of pairwise distances between *sit-MIR169h/sit-MIR169g* and *sit-MIR169h/sit-MIR169f*, respectively). A similar rationale was applied for the calculation of the ancient

duplication time of *sbi-MIR169t* in the sorghum miR169 cluster 1 (Figure 5.10A).

5.5.7. Rate of synonymous and nonsynonymous substitutions of the bHLH orthologous gene pairs

I used gene exon sequences to estimate synonymous and nonsynonymous substitutions using the MEGA5 program (Tamura et al., 2011). The synonymous and nonsynonymous substitution rate was calculated for a given bHLH orthologous gene pair (*Brachypodium*–rice; *Brachypodium*–foxtail millet; *Brachypodium*–sorghum; and *Brachypodium*–maize), where *Brachypodium* bHLH gene *Bradi3g41510* was compared with the HLH gene *Bradi4g34870*.

5.5.8. Phylogenetic analysis

Phylogenetic analysis were performed by creating multiple alignments of nucleotide or amino acid sequences using MUSCLE and Clustal_W, respectively, and phylograms were drawn with the MEGA5 program using the neighbor joining (NJ) method (Tamura et al., 2011). Multiple alignments of microRNA 169 stem-loop sequences were improved by removing the unreliable regions from the alignment using the web-based program GUIDANCE (http://guidance.tau.ac.il), and NJ phylogenetic tress were created with 2,000 bootstrap replications, and the model/method used was the maximum composite likelihood.

5.6. References

Allen E, Xie Z, Gustafson AM, Sung G-H, Spatafora JW, Carrington JC (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. Nature Genetics **36:** 1282-1290.

Axtell MJ, Bowman JL (2008) Evolution of plant microRNAs and their targets. Trends in Plant Science **13:** 343-349.

Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J, Jenkins J, Barry K, Lindquist E, Hellsten U, Deshpande S, Wang X, Wu X, Mitros T, Triplett J, Yang X, Ye C-Y, Mauro-Herrera M, Wang L, Li P, Sharma M, Sharma R, Ronald PC, Panaud O, Kellogg EA, Brutnell TP, Doust AN, Tuskan GA, Rokhsar D, Devos KM (2012) Reference genome sequence of the model plant *Setaria*. Nature Biotechnology **30:** 555-561.

Brachypodium-Sequencing-Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature **463:** 763-768.

Brown PJ, Rooney WL, Franks C, Kresovich S (2008) Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. Genetics **180:** 629-637.

Calviño M, Bruggmann R, Messing J (2008) Screen of Genes Linked to High-Sugar Content in Stems by Comparative Genomics. Rice **1:** 166-176.

Calviño M, Bruggmann R, Messing J (2011) Characterization of the small RNA component of the transcriptome from grain and sweet sorghum stems. BMC genomics **12:** 356.

Calviño M, Messing J (2012) Sweet sorghum as a model system for bioenergy crops. Current opinion in biotechnology **23:** 1-7.

Calviño M, Miclaus M, Bruggmann R, Messing J (2009) Molecular Markers for Sweet Sorghum Based on Microarray Expression Data. Rice **2:** 129-142.

Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Research **39:** W155-W159.

Duchêne E, Butterlin G, Dumas V, Merdinoglu D (2012) Towards the adaptation of grapevine varieties to climate change: QTLs and candidate genes for developmental stages. Theoretical and Applied Genetics **124:** 623-635.

Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, Carrington JC (2007) High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. PloS One **2:** e219.

Fenselau de Felippes F, Schneeberger K, Dezulian T, Huson DH, Weigel D (2008) Evolution of Arabidopsis thaliana microRNAs from random sequences. RNA **14:** 2455-2459.

Fernandez MGS, Becraft PW, Yin Y, Luebberstedt T (2009) From dwarves to giants? Plant height manipulation for biomass yield. Trends in Plant Science **14:** 454-461.

Franks SJ, Sim S, Weis AE (2007) Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. Proceedings of the National Academy of Sciences **104:** 1278-1282.

Friedlander M, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol **26:** 407 - 415

Fulton T, Van der Hoeven R, Eannetta N, Tanksley S (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. The Plant Cell **14:** 1457-1467.

Griffiths S, Dunford RP, Coupland G, Laurie DA (2003) The Evolution of CONSTANS-Like Gene Families in Barley, Rice, and Arabidopsis. Plant Physiology **131:** 1855-1867.

Hasan M, Friedt W, Pons-Kuhnemann J, Freitag N, Link K, Snowdon R (2008) Association of gene-linked SSR markers to seed glucosinolate content in oilseed rape (Brassica napus ssp. napus). Theoretical and Applied Genetics **116:** 1035-1049.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon A-F, Weissenbach J, Quetier F, Wincker P, Public F-I (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463-467.

Jiang D, Yin C, Yu A, Zhou X, Liang W, Yuan Z, Xu Y, Yu Q, Wen T, Zhang D (2006) Duplication and expression analysis of multicopy miRNA gene family members in Arabidopsis and rice. Cell Research **16:** 507-518.

Li W-X, Oono Y, Zhu J, He X-J, Wu J-M, Iida K, Lu X-Y, Cui X, Jin H, Zhu J-K (2008) The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. The Plant Cell **20:** 2238-2251.

Ma Z, Coruh C, Axtell MJ (2010) Arabidopsis lyrata small RNAs: transient MIRNA and small interfering RNA loci within the Arabidopsis genus. The Plant Cell **22:** 1090-1103.

Maher C, Stein L, Ware D (2006) Evolution of Arabidopsis microRNA families through duplication events. Genome Research **16:** 510-519.

Meng Y, Shao C, Gou L, Jin Y, Chen M (2011) Construction of microRNA- and microRNA*-mediated regulatory networks in plants. RNA Biology **8:** 1124-1148

Messing J, Bharti AK, Karlowski WM, Gunlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, Wing RA (2004) Sequence composition and genome organization of maize. Proceedings of the National Academy of Sciences **101:** 14349-14354.

Meyers B, Axtell M, Bartel B, Bartel D, Baulcombe D, Bowman J, Cao X, Carrington J, Chen X, Green P, Griffiths-Jones S, Jacobsen S, Mallory A, Martienssen R, Poethig R, Qi Y, Vaucheret H, Voinnet O, Watanabe Y,

Weigel D, Zhu J (2008) Criteria for annotation of plant MicroRNAs. The Plant Cell **20:** 3186-3190.

Murat F, Xu J-H, Tannier E, Abrouk M, Guilhot N, Pont C, Messing J, Salse G (2010) Genome Research **20:** 1545-1557.

Murray S, Sharma A, Rooney W, Klein P, Mullet J, Mitchell S, Kresovich S (2008) Genetic Improvement of Sorghum as a Biofuel Feedstock: I. QTL for Stem Sugar and Grain Nonstructural Carbohydrates. Crop Science **48:** 2165 - 2179.

Nozawa M, Miura S, Nei M (2012) Origins and evolution of microRNA genes in plant species. Genome Biology and Evolution **4:** 230-239.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-Ur-Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

Piriyapongsa J, Jordan IK (2008) Dual coding of siRNAs and miRNAs by plant transposable elements. RNA **14:** 814-821.

Sun J, Zhou M, Mao Z, Li C (2012) Characterization and Evolution of microRNA Genes Derived from Repetitive Elements and Duplication Events in Plants. PloS One **7:** e34092.

Swigonova S, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J (2004) Close split of sorghum and maize genome progenitors. Genome Research **14:** 1916-1923.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution **28:** 2731-2739.

Tang H, Bowers JE, Wang X, Paterson AH (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proceedings of the National Academy of Sciences **107:** 472-477.

Toledo-Ortiz G (2003) The Arabidopsis Basic/Helix-Loop-Helix Transcription Factor Family. The Plant Cell **15:** 1749-1770.

Valverde F (2011) CONSTANS and the evolutionary origin of photoperiodic timing of flowering. Journal of Experimental Botany **62:** 2453-2463.

Wenkel S, Turck F, Singer K, Gissot L, Le Gourrierec J, Samach A, Coupland G (2006) CONSTANS and the CCAAT Box Binding Complex Share a Functionally Important Domain and Interact to Regulate Flowering of Arabidopsis. The Plant Cell **18:** 2971-2984.

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M (2010) Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. PLoS Biology **8:** e1000409.

Xu J-H, Messing J (2008) Organization of the prolamin gene family provides insight into the evolution of the maize genome and gene duplications in grass species. Proceedings of the National Academy of Sciences of the United States of America **105:** 14330-14335

Xue L-J, Zhang J-J, Xue H-W (2009) Characterization and expression profiles of miRNAs in rice seeds. Nucleic Acids Research **37:** 916-930.

Yang J, Phillips M, Betel D, Mu P, Ventura A, Siepel A, Chen K, Lai E (2011) Widespread regulatory activity of vertebrate microRNA* species. RNA **17:** 312-326.

Zentella R, Zhang Z-L, Park M, Thomas SG, Endo A, Murase K, Fleet CM, Jikumaru Y, Nambara E, Kamiya Y, Sun T-p (2007) Global Analysis of DELLA Direct Targets in Early Gibberellin Signaling in Arabidopsis. The Plant Cell **19:** 3037-3057.

Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W, Tao Y, Bian C, Han C, Xia Q, Peng X, Cao R, Yang X, Zhan D, Hu J, Zhang Y, Li H, Li H, Li N, Wang J, Wang C, Wang R, Guo T, Cai Y, Liu C, Xiang H, Shi Q, Huang P, Chen Q, Li Y, Wang J, Zhao Z, Wang J (2012) Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nature Biotechnology 30: 549-554.

Zhang L, Chia J-M, Kumari S, Stein JC, Liu Z, Narechania A, Maher CA, Guill K, McMullen MD, Ware D (2009) A genome-wide characterization of microRNA genes in maize. PLoS Genetics **5:** e1000716.