MARKER DEVELOPMENT AND GENETIC DIVERSITY IN *SACCHARUM* AND *MISCANTHUS*

BY

BRANDON T. JAMES

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Plant Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Adviser:

Professor Ray Ming

ABSTRACT

The demand for alternative fuel sources for oil has been increasing in recent years. From 2000 to 2010, worldwide ethanol fuel production has increased from 4.49 billion gallons to 19.5 billion gallons. A large source of this ethanol comes from sugarcane produced in Brazil. Brazil is the second largest producer of ethanol fuel, accounting for 37.7% of the total ethanol fuel used globally. While sugarcane is an important crop for tropic and subtropic regions, an alternative feedstock is needed for cooler climates. *Miscanthus* is considered a promising source for cellulosic ethanol production in Europe and more recently the United Sates. Sugarcane and *Miscanthus* belong to the Andropogoneae, a tribe of C_4 grasses that are known for their high efficiency of photosynthesis and biomass production. Understanding of the genomes of these two crops is important for finding ways to improve them for agriculture. Both sugarcane and *Miscanthus* have large complex genomes that are the result of polyploidization, that hinders genomic their research. For genetic mapping and genome assembly, a large number of DNA markers are needed. Due to the relatedness of these two species, markers can be designed that work in them both. Chapter one deals with an SSR marker analysis in sugarcane, and Chapter 2 deals with the comparison of *Miscanthus* and sugarcane and an assessment of the variation within and between the two species.

To Mom and Dad

ACKNOWLEDGEMENTS

I would like to thank many people for offering me support and guidance throughout this project. I would like to thank my advisor, Ray Ming, for helping me write and make sense of my data. Thanks to the Energy Biosciences Institute and the International Consortium for Sugarcane Biotechnology for providing funding for my projects. Also, thank you to Kankshita Swaminathan for answering so many of my questions and helping me solve problems, to Ashley Spence for editing my manuscripts and helping me refine my scientific writing skills, and to my other lab mates who helped me complete this work. And finally, thanks to my family and friends for supporting me and encouraging me not to quit during difficult times.

TABLE OF CONTENTS

CHAPTER 1: DEVELOPMENT AND APPLICATION OF MICROSATELLITE MARKERS IN POLYPLOID SUGARCANE 1
1.1 ABSTRACT 1
1.2 INTRODUCTION 1
1.3 MATERIALS AND METHODS 3
1.4 RESULTS 5
1.5 DISCUSSION 7
1.6 TABLES 10
1.7 FIGURES 13

CHAPTER 2: SNP DETECTION AND ALLELIC VARIATION IN TWO CLOSELY RELATED GENERA, *MISCANTHUS* AND *SACCHARUM* 20
2.1 ABSTRACT 20
2.2 INTRODUCTION 20
2.3 MATERIALS AND METHODS 22
2.4 RESULTS 24
2.5 DISCUSSION 26
2.6 TABLES 29
2.7 FIGURES 35

REFERENCES 36

CHAPTER I
DEVELOPMENT AND APPLICATION OF MICROSATELLITE MARKERS IN POLYPLOID SUGARCANE

1.1 ABSTRACT

Sugarcane has become an increasingly important first generation biofuel crop in tropical and subtropical regions. It has a large, complex, polyploid genome that hindered the progress of genomic research and marker-assisted selection. Genetic mapping and ultimately genome assembly require a large number of DNA markers. Simple sequence repeats (SSRs) are widely used in genetic mapping because of their abundance and high rates of polymorphism. The objectives of this study were to develop SSR markers for construction of a saturated genetic map and to characterize the frequency and distribution of SSRs in a polyploidy genome. SSR markers were mined from expressed sequence tag (EST), reduced representation library genomic sequences, and bacterial artificial chromosome (BAC) sequences. These SSRs were surveyed using an agarose gel system. A total of 5,675 SSR markers were surveyed in a segregating population. The overall successful amplification and polymorphic rates were 87.9% and 16.4%, respectively. The trinucleotide repeat motifs were most abundant with tri- and hexanucleotide motifs being the most abundant for the ESTs. BAC and genomic SSRs were mostly AT rich while the ESTs were relatively GC rich due to codon bias. This large collection of SSR markers is a valuable resource for sugarcane genomic research and crop improvement.

1.2 INTRODUCTION

Sugarcane (*Saccharum* spp. L.) is a tropical perennial grass that belongs to the genus *Saccharum* in the family Poaceae. It is widely cultivated in tropical and subtropical regions primarily for sugar production and recently for biofuel production. Sugarcane has been recognized as one of the most efficient crops in

converting solar energy into chemical energy. Sugarcane has favorable energy input/output ratio at 1:3 in Hawaii and 1:8 in Brazil (Heichel 1974; Lam et al. 2009).

Modern sugarcane cultivars are hybrids derived from interspecific crosses between *S. officinarum* and *S. spontaneum* and result in a high degree of polyploidy and frequent aneuploidy, making the construction of genetic maps by conventional methods difficult. *S. officinarum* is the high sugar content species and *S. spontaneum* provides stress and disease tolerance and high fiber for biomass "energy cane". Covering the large number of chromosomes, ranging from 110 to 140, in modern sugarcane and energy cane cultivars requires a large number of markers.

Prior studies identified hundreds of Restriction Fragment Length Polymorphism (RFLP) and Amplified Fragment Length Polymorphism (AFLP) markers, which have been used to generate genetic maps. The first attempts at mapping sugarcane genomes were based on RFLP markers (da Silva et al. 1993; Al-Janabi et al. 1993; da Silva et al. 1995; Grivet et al. 1996; Ming et al. 1998). However, comparing these maps to the sorghum linkage map showed that all of the sugarcane maps are incomplete due to the low density of markers (Ming et al. 1998). AFLP markers have been reported to have high rates of polymorphism in sugarcane and have been used to construct genetic maps (Hoarau et al. 2001; Aitken et al. 2005). However, these marker systems are labor-intensive and can be difficult to be applied across laboratories.

Microsatellites or SSRs are highly polymorphic, abundant, and widely dispersed in the genomes of eukaryotes. SSRs consist of tandemly repeated units of usually one to six nucleotides in length. Flanking these tandem repeats are unique sequences that are used to design primers. Amplifying the products of these primers and then resolving them on gels can reveal polymorphisms. The mechanisms that cause length polymorphisms in SSRs are replication slippage,

unequal crossing over, and point mutations (Goldstein and Schlotterer 1999). The high frequency of these variations reflects the rapid rate at which SSRs accumulate mutations. Experts suggest this high mutability plays a crucial role in the evolution of genomes by the creation and maintenance of genetic variations (Kashi et al. 1997). SSRs have proven to be a versatile tool for discovering new molecular markers because of their high degree of allelic variation, ease of PCR based detection and reproducibility, co-dominant transmission, wide genome coverage, and transferability across species and genera (Gupta et al. 1996).

Previous studies have used expressed sequence tag (EST) derived markers in *Saccharum* (Pinto et al. 2006), and a limited number of genomic markers have been developed and evaluated (Pinto et al. 2004; Aitken et al. 2005; Parida et al 2009; da Silva et al. 2001). However, due to high levels of polyploidy in the sugarcane genome, more markers are needed for advancements in genetics, genomics, and breeding. One objective of this study, therefore, was to develop a large collection of SSR markers for future construction of a saturated genetic map using EST, genomic, and BAC sequences. Another objective was to characterize the frequency and distribution of SSRs in a polyploid genome.

1.3 MATERIALS AND METHODS

1.3.1 Sequence sources

The EST data was downloaded from TIGR Plant Transcript Assemblies database (http://plantta.jcvi.org/) providing 117 Mb of sugarcane sequences. The sugarcane tissue used to make the EST library was obtained from the varieties SP70-1143, SP80-3280, SP80-87432, PB5211 x P57150-4, CB47-89, RB855205, RB845298, and RB805028 (Vettore et al. 2003). To increase the number of SSRs available, we made reduced representation libraries using the 650 to 850 bp size fractions obtained after a double digest with HinP1I and NcoI from *S. officinarum* 'LA purple' (LAP) and *S. robustum* Molakai (Mol) 6081. HinP1I is a methylation sensitive enzyme and NcoI cuts at CATG sites. The

libraries were sequenced using 454 Titanium and assembled using GS assembler from Roche. A total of 188,672 contigs were produced for LAP and 203,282 contigs for Mol 6081 yielding 69.1 Mb and 76.8 Mb of sequences respectively. Sequences were obtained from 20 sugarcane BACs (cultivar R570) (Wang et al. 2010). Each of these BACs corresponded to one sorghum chromosome arm.

1.3.2 SSR mining

A total of 2,640 SSR markers were developed from the ESTs. The SSRs with flanking sequences that were greater than or equal to 20 bp were identified from the EST sequences using the SSR finder software (http://www.Maizemap.org). The flanking sequences were used for designing primers using Primer3 software (Rosen and Skaletsky 2000) implemented in primer design Perl-script in the SSR finder software package. Genomic reduced representation sequences generated 2,991 genomic markers from LAP and Mol 6081. For marker discovery, the LAP and Mol 6081 sequences were first assembled into contigs using GS assembler. SSRs were called using SSR finder. In addition, we obtained 44 markers by mining the 20 sequenced BACs for SSRs.

1.3.3 DNA extraction and PCR amplification

SSR markers were surveyed in a segregating population of an LAP x Molakai 5829 cross. Total genomic DNA was extracted from 12 to 15 ml of powdery lyophilized leaf tissue. The extraction method was described previously (Wang et al. 2008). Amplification and polymorphism surveys of the SSR markers were conducted on four individuals from the segregating population and the LAP parent. The markers showing polymorphisms were then reamplified using 59 different individuals from the population and the LAP parent. SSR markers were amplified in a 10 µl PCR mix containing 5 ng of template DNA, 5 µl 5x GoTaq® Green Master Mix (Promega), and 0.15 mm each of reverse and forward primers.

The PCR reactions were performed using a 2720 thermocycler (Applied Biosystems) in which the reaction mixture was incubated at 94°C for 5 minutes, then for 35 cycles of 45 seconds of denaturing at 94°C, 30 seconds of annealing at 55°C, and 45 seconds of extension at 72°C, with a final extension at 72°C for 7 minutes. PCR products were separated on 4% super fine resolution (SFR) agarose (Amresco, Solon, OH) gels and visualized by ethidium-bromide staining. All marker data was scored by visual inspection and proofread to correct errors.

1.4 RESULTS

From 117 Mbp of EST sequences, 2,640 SSRs were identified with an average of one per 44.3 Kb. The 454 sequences had a similar average of one per 48.8 Kb with 2,991 SSRs being identified from 145.9 Mbp. To produce 44 SSRs with a density of one SSR per 48.8 Kb, 1.9 Mbp of BAC sequences were mined (Table 1). SSR frequency was similar in each dataset, ranging from 44.3 to 48.8 Kb. Ests come from a variety of genotypes while genomic data comes from two interspecific genotypes.

The ranges of repeat length were between 20 and 132 bp in EST; 18 and 128 in 454; and 20 and 73 in BAC sequence datasets. BAC and 454 data were combined and called genomic. The predominant EST repeat length was 24 bp accounting for 32.9% of the total SSRs. This was larger than the genomic length of 20 bp accounting for 30.2% (Fig 1 a).

When looking at the distribution of SSR motif length, the trinucleotide repeats were the most abundant for both ESTs and genomic with both sets having about 40% trinucleotide repeats. The majority of the remaining SSRs were clustered between the dinucleotides and the hexanucleotides. The EST abundances ranged from highest to lowest are as follows: trinucleotide, hexanucleotide, pentanucleotide, dinucleotide, and tetranucleotide. For the genomic, the range

listed from highest to lowest is as follows: trinucleotide, pentanucleotide, tetranucleotide, hexanucleotide, and dinucleotide (Fig 1 b).

Of the EST SSRs, the dinucleotide repeats had the highest polymorphism rate of 30.2%. At 20.4%, the more abundant trinucleotide repeats were below the average EST polymorphism rate of 22.6%. The pentanucleotide repeats were also above average in their polymorphism rate of 23.8%. The dinucleotide repeat lengths of 33.7 bp and 34.2 bp for amplified and polymorphic SSRs were above the average SSR lengths of 25.3 bp and 27.0 bp. The trinucleotides were again lower than average in their lengths for amplified and polymorphic SSRs with lengths of 23.0 and 23.8 respectively. For all types of repeat motifs, the polymorphic SSRs had a longer length compared to the nonpolymorphic SSRs (Table 2).

The genomic SSRs had a polymorphism rate of 11.2%. This is about half of the EST polymorphism rate. In this case, all of the nucleotide groups except trinucleotide repeats had a higher than average polymorphism rate with hexanucleotide repeats being the highest with 16.3%. The average amplified SSR length was 23.4 bp and the average polymorphic length was 28.5 bp. Trinucleotides were also below average in both amplified and polymorphic SSR length with lengths of 20.8 bp and 25.9 bp. Polymorphic dinucleotides were also shorter with an average of 23.0 bp. With the exception of dinucleotides, average polymorphic length was longer than average amplified length (Table 3).

Both genomic and EST dinucleotide sequences were AT rich. Trinucleotide, pentanucleotide, and hexanucleotide repeats were predominantly GC rich in the ESTs. The genomic trinucleotide and hexanucleotide repeats also had a higher GC content, but the difference was less drastic while the pentanucleotide sequences were about evenly distributed among AT and GC. Tetranucleotide sequences were also more evenly spread for both the genomic and EST sequences, but the ESTs were still slightly more GC rich (Fig 2 a, b, c, d, e).

1.5 DISCUSSION

The most significant finding in this study was the low polymorphism rate (16.4%) compared to AFLP (50%) and RFLP (80%) marker studies (Lima et. al. 2002; Lu et. al. 1994). RFLP is based on hybridization using a probe about 1 kb long. Fragments sharing 80% or higher DNA sequence identity would hybridize to the probes, resulting in multiple bands in polyploidy sugarcane (Beckman and Soller 1984). AFLP is based on selective amplification of anonymous fragments matching the three nucleotide extension of the adaptors (Vos et. al. 1995); whereas, SSRs are generated by amplification of target fragments matching the sequence specific primers that are at least 20 bp. An RFLP marker generally has 8 to 10 bands, an AFLP marker has 30 to 60, and an SSR marker usually only has 1 to 3 bands although they sometimes have up to 7 or 8 bands. While the rate of polymorphism of sugarcane is lower than expected, it is higher than the polymorphic rate (13.3%) in Papaya, a diploid system (Chen et. al. 2007). Sugarcane retains a high level of heterozygosity because of polyploidy and vegetative reproduction.

The difference in the polymorphism rate of the ESTs (22.6%) and the genomic (11.0%) could be explained by the method used to create the reduced representation libraries. The enzyme NcoI was used to target start codons because it cuts at the restriction site CATG. It has been shown that 5' DNA regions are less polymorphic than 3' regions. If our technique to target start codons was successful, that would mean that the majority of our sequences would be from 5' regions. The ESTs SSRs were not biased in the same way and therefore might include more 3' regions explaining their higher rate of polymorphism.

This investigation focused on identifying SSRs that are potentially linked with genes. By using sugarcane EST data and a reduced representation genome biased toward low-copy sequences and coding regions, we increased our ability

to locate SSRs that could be used for generating genetic maps that are highly representative of the genic regions. This is especially important in complex polyploid species like sugarcane that has generally lacked in informative marker generation (Parida et al. 2009). In addition, any useful markers generated here can be mapped back to an existing segregating population, accelerating further studies on relating genetic markers/genes to phenotype.

We surveyed 5,675 SSRs, 5,631 of which are biased toward gene sequences. Of these 5,675 SSRs, 4,987 were successfully amplified and 820 were found to be polymorphic in the mapping population derived from LAP x Mol 5829. Though only 16.4% of the total amplified SSRs were polymorphic in the segregating population, other amplified SSRs could be used in studying populations that have greater genetic diversity or were generated from more distantly related parents. The two parents in this study S. officinarum 'LA purple' and S. robustum Molakai 5829 are closely related species with S. robustum postulated to be the progenitor of S. officinarum (Brandes, 1958; Warner, 1962; Grassl, 1977; Daniels et al., 1980). In a segregating population generated from a wider cross between two sugarcane species like *S. officinarum* and *S. spontaneum*, the SSRs identified in this study may have a higher rate of polymorphism. There would be greater genetic diversity within the progeny population, related to the time of divergence between the two parent species. This idea could be further applied to inter-generic crosses like those between sugarcane and *Miscanthus* whose progeny are termed "Miscane" (Jakob et. al. 2009). In addition, these identified SSRs could serve as building blocks for identifying polymorphisms in other closely related species such as *Miscanthus* and *Erianthus*.

The SSR motifs varied according to sequence type, either EST or genomic. As in previous studies, the trinucleotide repeats occurred in the highest frequency in both EST and genomic sequences at a rate of nearly 40% for both types. This type of repeat usually corresponds to protein coding regions of all

taxa and is typically in the codon region of the genes (Toth & Jurka, 2000). This high trinucleotide repeat frequency is counteracted by the relatively low frequency of polymorphisms within this repeat type. Out of the 920 trinucleotide repeat SSRs amplified in EST sequences, only 20% were found to be polymorphic, and an even lower 7.4% polymorphism rate was found in the genomic (Table 2 and 3). The trimers also correspond to an increased GC content, which is expected of gene rich regions and thus regions of low polymorphic rate. These GC rich trimers have been found in high quantity in other SSR studies on sugarcane as well as on other monocots such as maize, wheat, barely and rice (La Rota et al., 2005; Pinto et al., 2004; Chin et al., 1996). AT rich trimers can also be found in non-transcribed regions but were underrepresented in this study indicating our method of enriching for gene coding regions in reduced representation libraries was successful.

Though the frequency of the dinucleotide repeats was lower than the tri-, penta- and hexanucleotide repeats of the EST derived SSRs, dinucleotide repeats had the highest polymorphism rate of 30.2% (Fig 1 b). The dinucleotide repeat motif type had the second highest rate of polymorphism in the genomic sequences at 14.2% with the hexa-repeat having the highest rate of 16.3%. In contrast to the trimers, dinucleotide repeat motifs had higher AT content (Fig 2 a and b), which indicated their prevalence in untranscribed regions of the genome.

1.6 TABLES

Table 1 Distribution of SSRs among sequence types in sugarcane

Source Sequences	Size (Mbp)	Mined SSRs	Density (Kb/SSR)	Amplified SSRs	Polymorphic SSRs	% Polymorphic
EST	117	2640	44.3	2301	520	22.6
BAC	1.9	44	42.2	36	8	22.2
454	145.9	2991	48.8	2650	292	11
Total	**264.8**	**5675**	**46.7**	**4987**	**820**	**16.4**

Table 2 SSR amplification and polymorphism distribution in EST sequences

Motif	Amplified	Amplifed average length (bp)	Polymorphic	Polymorphic average length (bp)	Polymorphism rate (%)
Di-	344	33.7	104	34.2	30.2
AT/TA	109	42.8	28	42.2	25.7
AG/GA/CT/TC	170	28.0	60	28.5	35.3
AC/CA/GT/TG	65	33.3	16	38.8	24.6
Tri-	920	23.0	188	23.8	20.4
AAT/ATA/TAA/ATT/TTA/TAT	8	33.5	2	26.0	25.0
AAG/AGA/GAA/CTT/TTC/TCT	28	23.8	6	12.1	21.4
AAC/ACA/CAA/GTT/TTG/TGT	55	24.8	13	11.5	23.6
ATG/TGA/GAT/CAT/ATC/TCA	21	25.0	13	24.4	61.9
AGT/GTA/TAG/ACT/CTA/TAC	4	15.5	1	4.0	25.0
AGG/GGA/GAG/CCT/CTC/TCC	101	22.6	17	22.4	16.8
AGC/GCA/CAG/GCT/CTG/TGC	150	22.3	35	19.0	23.3
ACG/CGA/GAC/CGT/GTC/TCG	53	22.6	11	19.0	20.8
ACC/CCA/CAC/GGT/GTG/TGG	54	22.2	9	18.9	16.7
GGC/GCG/CGG/GCC/CCG/CGC	446	22.4	81	22.4	18.2
Tetra-	205	23.4	39	26.6	19.0
Penta-	353	22.9	84	24.6	23.8
Hexa-	468	25.9	102	27.3	21.8
Total	2301	25.3	520	27.0	22.6

Table 3 SSR amplification and polymorphism distribution in genomic sequences

Motif	Amplified	Amplified Average length (bp)	Polymorphic	Polymorphic Average length (bp)	Polymorphism rate (%)
Di-	212	23.8	30	23.0	14.2
AT/TA	69	12.9	8	23.8	11.6
AG/GA/CT/TC	68	18.3	9	22.7	13.2
AC/CA/GT/TG	75	22.7	13	21.1	17.3
Tri-	1074	20.8	80	25.9	7.4
AAT/ATA/TAA/ATT/TTA/TAT	65	28.0	10	30.2	15.4
AAG/AGA/GAA/CTT/TTC/TCT	145	20.1	7	13.6	4.8
AAC/ACA/CAA/GTT/TTG/TGT	102	24.5	21	28.3	20.6
ATG/TGA/GAT/CAT/ATC/TCA	78	21.2	5	11.3	6.4
AGT/GTA/TAG/ACT/CTA/TAC	35	20.1	3	8.0	8.6
AGG/GGA/GAG/CCT/CTC/TCC	103	19.4	5	11.5	4.9
AGC/GCA/CAG/GCT/CTG/TGC	151	19.9	5	9.5	3.3
ACG/CGA/GAC/CGT/GTC/TCG	100	20.0	7	19.4	7.0
ACC/CCA/CAC/GGT/GTG/TGG	137	19.7	9	18.0	6.6
GGC/GCG/CGG/GCC/CCG/CGC	158	19.6	8	19.9	5.1
Tetra-	469	24.0	59	31.0	12.6
Penta-	563	24.4	71	30.5	12.6
Hexa-	337	27.2	55	29.8	16.3
Total	2686	23.4	300	28.5	11.2

1.7 FIGURES

Fig. 1 Comparison of distributions amplified and polymorphic SSRs in EST and genomic sequences. **a.** Nucleotide length **b.** Motif length

a

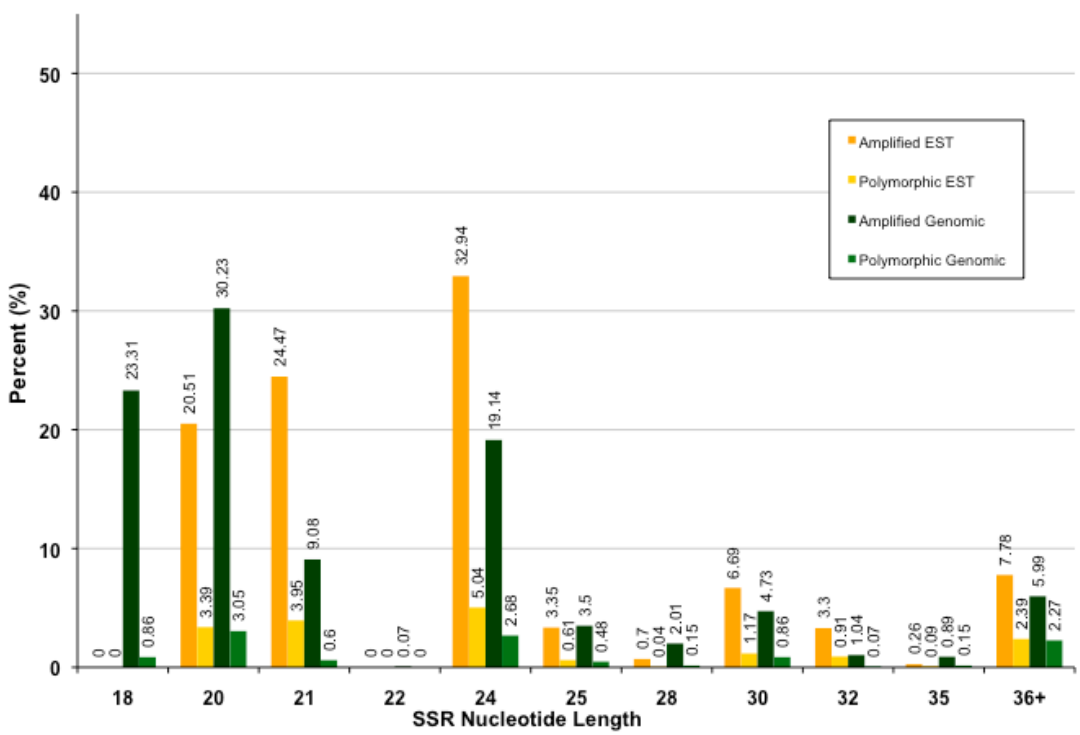


Fig. 1 (cont.)

b

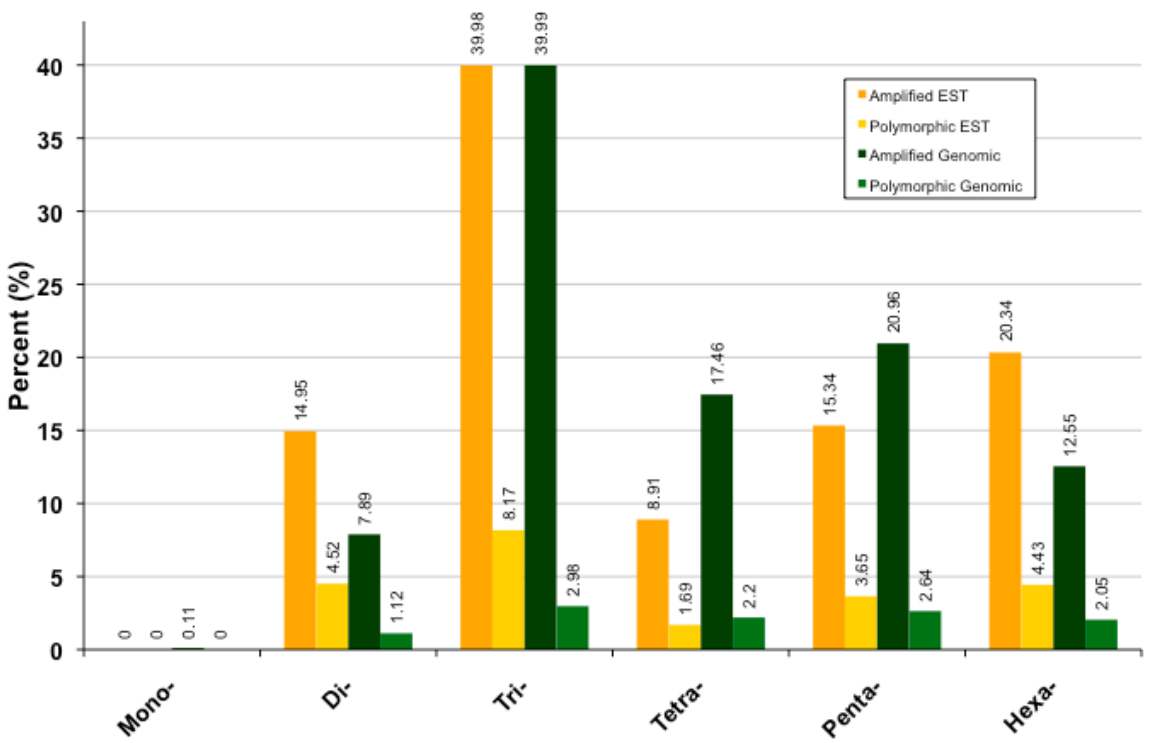

Percent (%)
0
5
10
15
20
25
30
35
40
Amplified EST
Polymorphic EST
Amplified Genomic
Polymorphic Genomic
0
0
0.11
0
14.95
4.52
7.89
1.12
39.98
8.17
39.99
2.98
8.91
1.69
17.46
2.2
15.34
3.65
20.96
2.64
20.34
4.43
12.55
2.05
Mono-
Di-
Tri-
Tetra-
Penta-
Hexa-

Fig. 2 Distribution of SSRs according to nucleotide composition of motif. **a.** Dimer; **b.** Trimer; **c.** Tetramer; **d.** Pentamer; and **e.** Hexamer

a

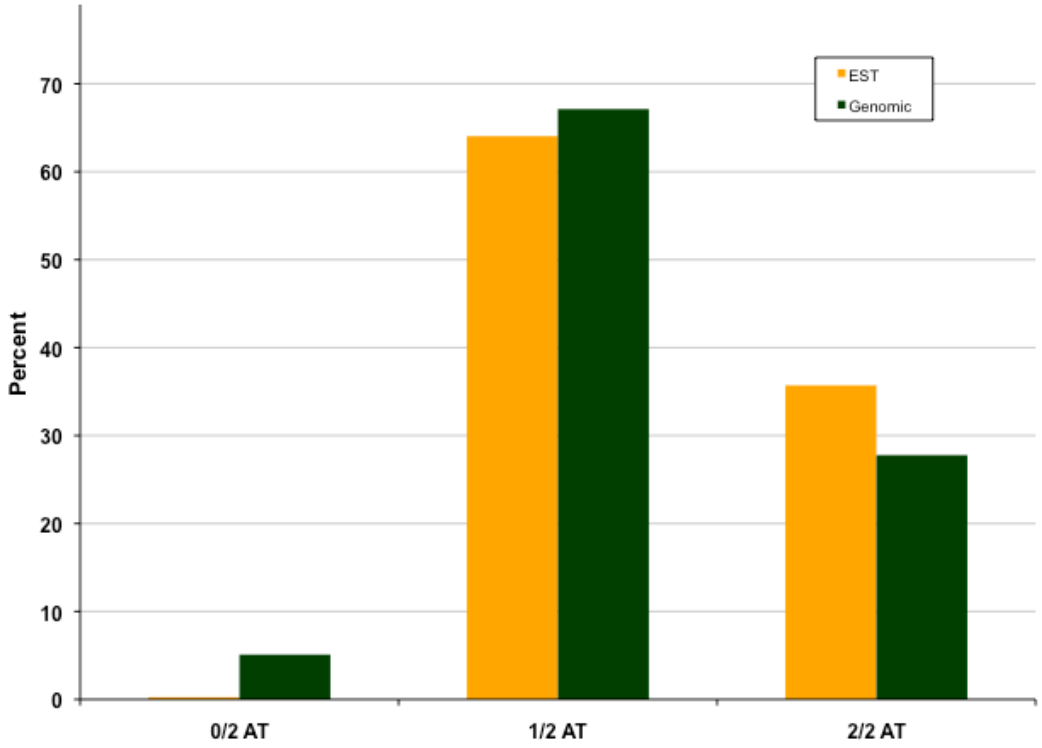


Fig. 2 (cont.)

b

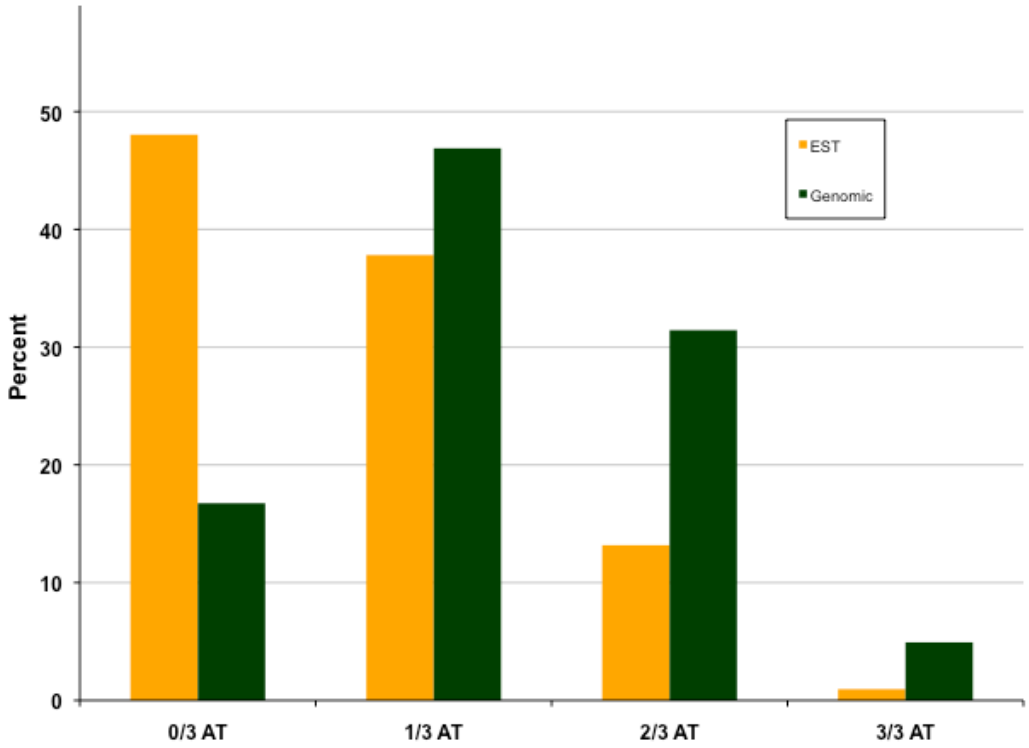


Fig. 2 (cont.)

c

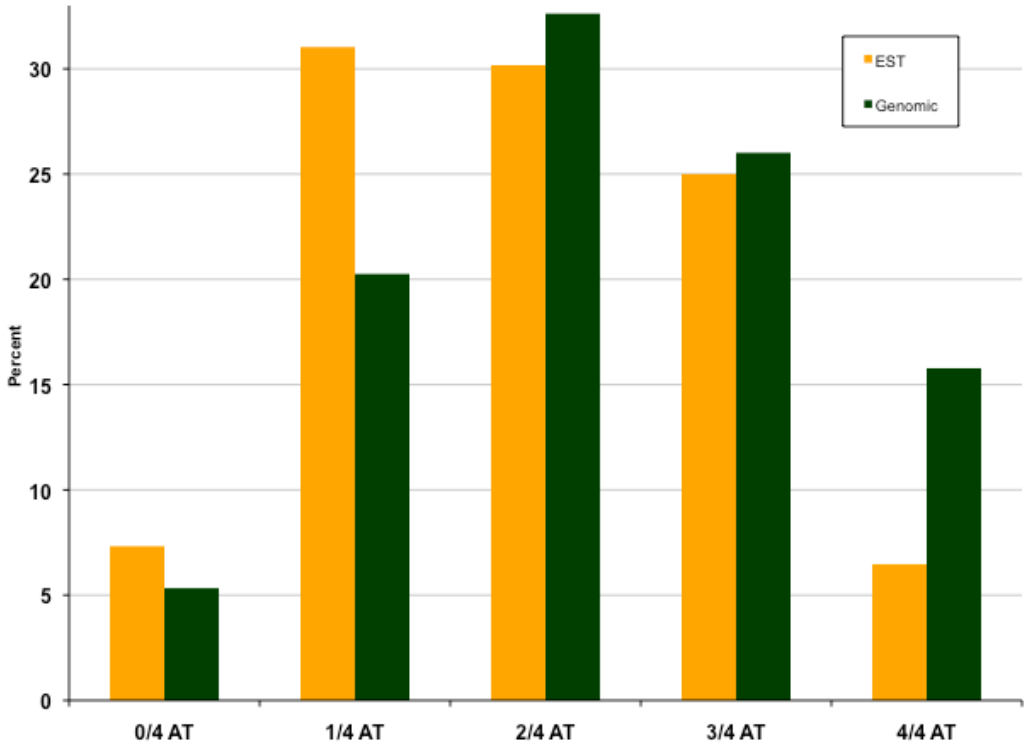


Fig. 2 (cont.)

d

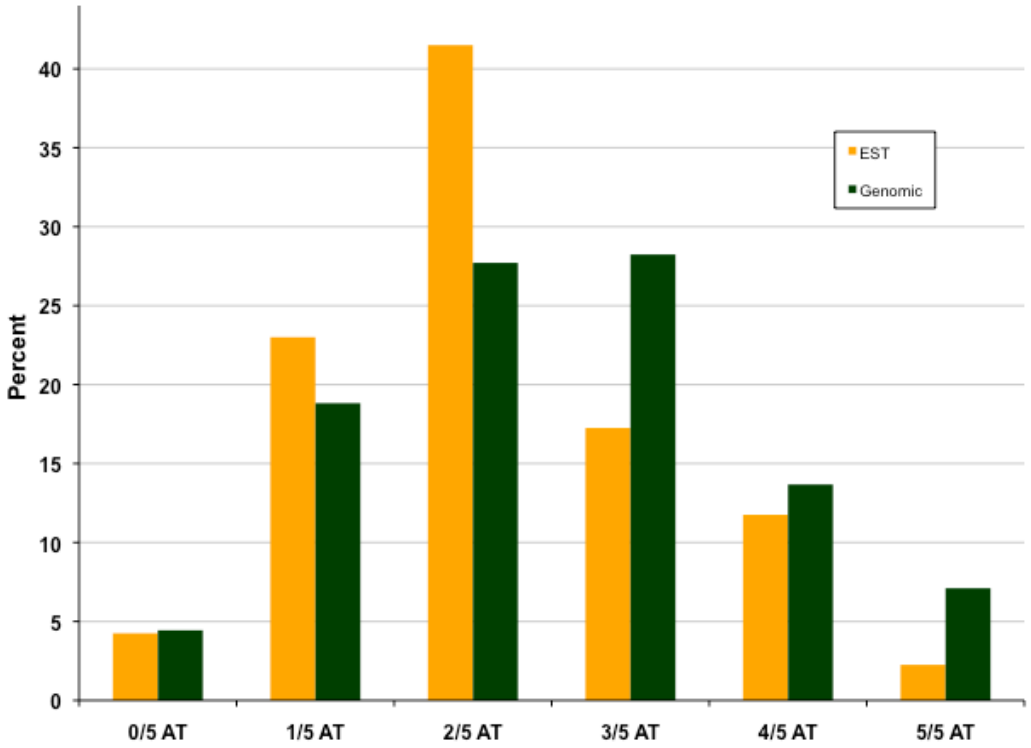
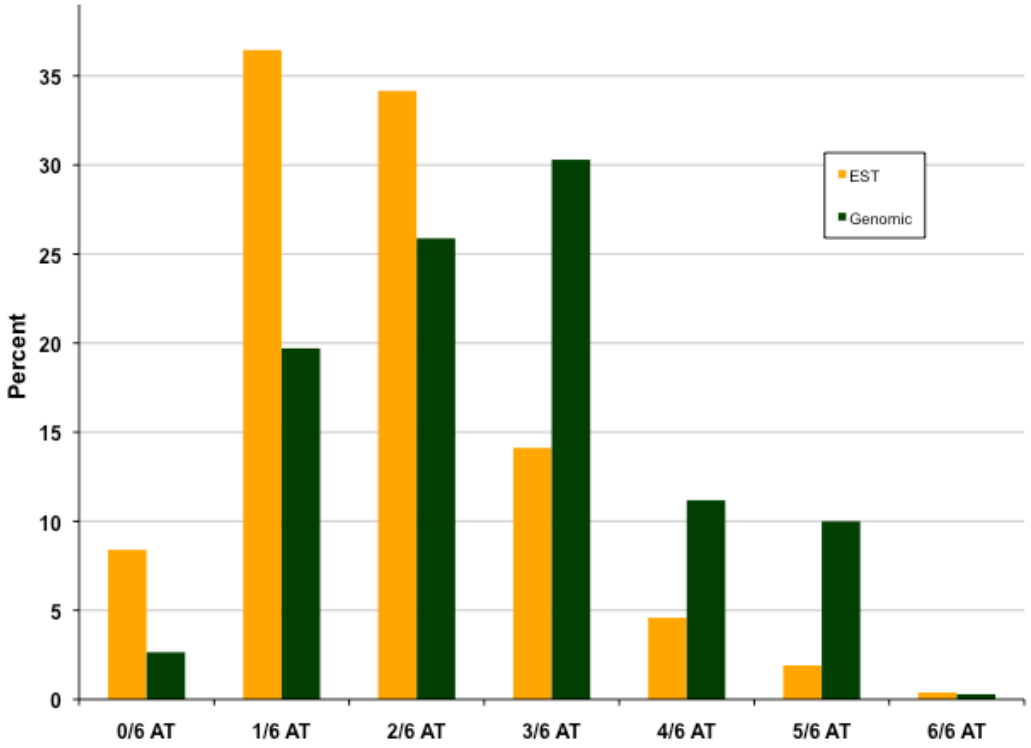

Percent
40
35
30
25
20
15
10
5
0
EST
Genomic
0/5 AT
1/5 AT
2/5 AT
3/5 AT
4/5 AT
5/5 AT

Fig. 2 (cont.)

e


Percent
35
30
25
20
15
10
5
0
EST
Genomic
0/6 AT
1/6 AT
2/6 AT
3/6 AT
4/6 AT
5/6 AT
6/6 AT

CHAPTER II
SNP DETECTION AND ALLELIC VARIATION IN TWO CLOSELY RELATED GENERA, *MISCANTHUS* AND *SACCHARUM*

2.1 ABSTRACT

Miscanthus and sugarcane both have complex polyploid genomes with *Miscanthus x giganteous* (Mxg) being a triploid allopolyploid and *Saccharum officinarum* 'LA Purple' (LAP) being an octoploid autopolyploid. Single nucleotide polymorphisms (SNPs) are more commonly being used as a marker type due to their high frequency in the genome and multiple detection methods. The objective of this study was to discover SNPs within and between *Miscanthus* and *Saccharum*. Primers were designed from 454 FLX reads of Mxg. The PCR products from Mxg and LAP were cloned and then sequenced using Sanger sequencing methods. The sequences were then assembled and analyzed. A total of 166 contigs were constructed of which 90 were Mxg and 76 were LAP. The consensus strand for each of these contigs was blasted to various databases to annotate the sequences. The SNP frequencies were one in every 74 bases in LAP and one in every 39 bases in Mxg. The higher frequency in Mxg reflected its interspecific hybrid origin; whereas, LAP is a pure S. officinarum. The abundant SNPs in *Miscanthus* and sugarcane will facilitate the genomic research for feedstock development in these two biofuel crops.

2.2 INTRODUCTION

Sugarcane (*Saccharum* spp. L.) and species of *Miscanthus* belong to the Andropogoneae tribe along with other important crops such as maize and *Sorghum bicolor*. Andropogoneae are C_4 grasses that are known for their high efficiency of photosynthesis and biomass production. Cultivated forms of sugarcane are most notably used for sugar and alcohol production and more

recently biofuel production. *Miscanthus* has also recently emerged as a top candidate for biofuel production.

Miscanthus x giganteous (Mxg) and *Saccharum officinarum* 'LA Purple' (LAP) are both polyploid with large, complex genomes. Mxg is an allopolyploid that is believed to be a hybrid of *M. sinesis* and *M. sacchariflorus*. It has a triploid genome with 57 chromosomes and a basic chromosome number of x = 19 (Hodkison et al. 2002; Linde-Laursen 1993; Lafferty and Lelly 1994). LAP is an autopolyploid with 80 chromosomes and a basic chromosome number of x = 10 making it an octoploid (Daniels and Roach 1987; Sreenivasan et al. 1987).

Many types of markers have been used to assess genetic diversity in sugarcane: restriction fragment length polymorphisms (RFLP) (da Silva et al. 1993; Al-Janabi et al. 1993; da Silva et al. 1995; Grivet et al. 1996); amplified fragment length polymorphism (AFLP) (Hoarau et al 2001, Aitken et al. 2005); and simple sequence repeats (SSRs) (Pinto et al. 2004; Pinto et al. 2006; Aitken et al. 2005; Parida et al 2009; da Silva et al. 2001). Random amplification of polymorphic DNA (RAPD) and AFLP studies have also been conducted on *Miscanthus* (Ateinza et al. 2002; Greef et al 1997).

Single nucleotide polymorphisms (SNPs) are emerging as a new marker tool to evaluate genetic diversity in organisms. SNPs are one of the most common types of genetic variation and can show differences both within and among genomes in related species. They can occur throughout the genome in genic, intergenic, and non-coding regions. A SNP is a single base change in a polynucleotide sequence and can be the result of insertions, deletions, or substitutions. For those species with no genome sequence available, SNP discovery can be accomplished by using expressed sequence tags (ESTs) (Kota et al. 2004; Batley et al. 2003; Picoult-Newberg et al. 1999). The limited genomic sequences available in sugarcane have also led to the mining of EST sequences for SNP markers (Bundock et al. 2009; Grivet et al. 2001, 2003; Cordeiro et al.

2006). In plant species where genome sequence information is available, a great deal of SNP discovery has taken place, including work on the model species *Arabidopsis thaliana* (Jander et al. 2002) and in other members of the Poaceae family such as rice (Hayashi et al. 2004; Shen et al. 2004; McNally et al. 2006), barley (Kanazin et al. 2002; Rostoks et al. 2005), and maize (Ching et al. 2002). Recent sequencing of the Mxg genome has provided genomic sequences for SNP discovery (Swaminathan et al. 2010). *Miscanthus* is most closely related to *Saccharum* among the Andropogoneae. It has also been shown that markers from related species have cross-transferability (Selvi et al. 2003; Hernández et al. 2001; Kuleung et al. 2004). The objective of this study was to use markers designed from *Miscanthus* genome sequences to analyze SNP frequency within and between Mxg and LAP.

2.3 MATERIALS AND METHODS

2.3.1 Primer design

LAP and Mxg genomic DNA was extracted from 12 to 15 ml of powdery lyophilized leaf tissue. The extraction method was described previously (Wang et al. 2008). Primers were designed from low copy reads obtained from a single plate of 454 FLX on Mxg, and 96 primers were generated and used to amplify both Mxg and LAP (Swaminathan et al. 2010).

2.3.2 DNA extraction and PCR amplification

Samples from both species along with a water negative control were amplified in a 50 µl PCR mix containing 5 ng of template DNA, 25 µl 5x GoTaq® Colorless Master Mix (Promega), and 2 µM each of reverse and forward primers. The PCR reactions were performed using a 2720 thermocycler (Applied Biosystems) in which the reaction mixture was incubated at 94°C for 5 minutes, then for 35 cycles of 30 seconds of denaturing at 94°C, 30 seconds of annealing at 52°C for LAP and 58°C for Mxg, and 50 seconds of extension at 72°C, with a final

extension at 72°C for 10 minutes. PCR products were separated on 2% agarose gels and visualized by ethidium-bromide staining. PCR purifications were done using QIAquick 96 PCR Purification Kit and QIAvac 96 (QIAGEN).

2.3.3 Ligation and transformation

Ligation and transformation for each species were done using the pGEM-T Easy Vector System II (Promega). For each species, two plates were made for each primer. Four white colonies were picked from each plate for a total of eight clones per primer per species. If a plate failed to produce enough colonies, additional colonies were chosen from other plates to fill the 96-well block. Colonies were grown, and DNA was isolated using the procedures in the QIAprep 96 Turbo Miniprep Kit (QIAGEN).

2.3.4 Sequencing

A 12 µl sequencing mix containing 50 ng of template DNA, 2 pmol of SP6 or T7 primer (each clone was sequenced from both ends), 2.5-µl 5x Sequencing Buffer, 0.5-µl 5M Betaine, and 0.5-µl BigDye v3.1 was used. The sequencing reactions were performed using a 2720 thermocycler (Applied Biosystems), in which the reaction mixture was incubated at 94° for 3 min, then for 35 cycles of 10 sec of denaturing at 96°, 15 sec of annealing at 52°, and 3 min of extension at 60°, with a final extension at 72° for 7 min. Samples were then given to the University of Illinois Core Sequencing Facility for Sanger sequencing.

2.3.5 Sequence analysis and annotation

Sequence results were assembled using the Sequencher 4.7 program with 90% identity cut off. SNPs were called manually by looking at the consensus and comparing the sequences for those contigs that had at least four unique cloned strands. Consensus strands from contigs that had two or more reads

from both Mxg and LAP were assembled with a 90% identity cut off and SNPs were manually detected for these contigs as well.

All consensus sequences from both species were then put into a blastn program and compared against the Sugarcane EST database (http://sucest.lad.ic.unicamp.br/en/), the Saccharum officinarum Gene Index (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=s_officinarum), Sorghum cds (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v5.0/Sbicolor/annotation/Sbi1.4/), Sorghum Genome (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v5.0/Sbicolor/assembly/Sbi1/), Nucleotide database from NCBI (nt) (ftp://ftp.ncbi.nih.gov/blast/db/), TIGR repeat database (http://plantrepeats.plantbiology.msu.edu/), TAIR database (http://www.arabidopsis.org/), Miscanthus ESTs (not published), and the Mxg High Copy Reads database (Swaminathan et al. 2010).

2.4 RESULTS

Ninety-six pairs of primers were designed from single or low copy *Miscanthus* genome sequences (Swaminathan et al. 2010). These primers were used for amplifying LAP and Mxg and resulted in a 96.9% amplification rate with three primers not producing bands in either species. These amplified sequences were then transformed and cloned. There was a 78.5% transformation rate for both species; however, a different set of 20 out of the possible 93 amplified sequences failed for each species to arrive at this percentage. Six 96-well plates of DNA were isolated for each species. Eight colonies were chosen for each primer, but in the cases where the transformation was not as successful or failed, colonies from different primers were used to fill the empty wells of the plate. Each plate was sequenced from both ends, resulting in 1,152 sequences for both Mxg and LAP.

The sequenced fragments were assembled into 166 contigs. There were 90 Mxg contigs and 76 LAP contigs. Of the 90 Mxg contigs, 59 matched the original

454 reads that were used in the primer design; whereas, only 27 of the 76 LAP contigs matched the original reads. This means that of the 96 primers 61.5% and 28.1% were successfully sequenced in Mxg and LAP respectively. However, some primers amplified regions that did not match the original reads for both species.

Of the 90 Mxg contigs, only 52 produced four or more reads for SNP discovery. Of these 52 contigs, the average length of sequenced reads was 123 bp with an average SNP count of 3.2 per contig. The contigs ranged from 0 to 12 for the amounts of SNPs recorded with 28.8% of contigs having no detectable SNPs. The SNP per base pair average was 0.03 or one SNP in every 39 base pairs (Table 4, Fig 3).

In LAP, 76 contigs were produced, but only 27 contigs had four or more reads. The LAP contigs had a slightly longer average read length of 128.8 bp. The average SNP count was lower, 1.7 per contig, and they had a range of 0 to 13 SNPs per contig. There were no detectable SNPs in 51.9% of the contigs, and the SNP per base pair average was 0.01 or one SNP in every 74 base pairs (Table 5, Fig 3).

The Mxg and LAP contigs previously described were condensed into consensus strands and then assembled to produce 25 contigs. Consensus strand SNPs were counted two ways. The first method was to count similar SNPs. A SNP was considered similar if it was completely identical or if it was a partially identical SNP where one strand had a standard base and the other strand had a SNP at that position that included this base. For example, a SNP would be considered identical if both the Mxg and LAP original contigs had a C or a T at a certain location, thus making both their consensus strands a Y. However, it could also be considered similar if one strand had a G and the other had an R (an A or a G SNP) at a particular location. The second way SNPs were counted was the standard method where a position has completely unrelated

bases. The average length of 129.3 bp for the consensus contigs was slightly longer than both individual species. The average number of similar SNPs was 5.2 per contig and the average for different SNPs was 1.6 per contig. This makes the similar count much higher than the individual species and the different count lower than both species. The combined count was 6.8 SNPs per contig, meaning there is about one SNP in every 19 base pairs or 0.05 SNPs per base pair between *Miscanthus* and *Saccharum* (Table 6, Fig 3).

All 166 contigs were used in the blast, but only the 86 Mxg and LAP contigs that matched the original 454 reads were considered when annotating the sequences. A total of 59 of these sequences had a hit in at least one database, 40 Mxg and 19 LAP. However, 22 of these sequences only hit the Mxg High Copy Read database while 12 others matched this database but also had a hit in one of the other databases. These high copy hits accounted for 57.6% of the total reads. They accounted for 62.5% of the Mxg reads and 47.4% in LAP. None of the 166 reads matched the TIGR plant repeat database because they were prescreened against it. Of the reads, 39.0% had a hit to one of the EST databases, 47.4% of which were LAP and 35.0% in Mxg (Tables 4 and 5).

2.5 DISCUSSION

The complex polyploidy sugarcane genome has not been completely mapped despite numerous mapping projects in the past 20 years due to the large number of chromosomes (often > 100) to be mapped and the lack of a large number of high-throughput DNA markers. No linkage mapping of the *Miscanthus* genome has been attempted to date. This project was initiated to assess the level of heterozygosity within the sugarcane and *Miscanthus* genome for developing a high-throughput SNP genotyping platform. We found that S. officinarum 'LA Purple' had one SNP in every 74 base pairs, comparable to the previously reported 1 SNP every 58 base pairs in pure S. officinarum (Bundock et al. 2009). The decreased amount of SNPs discovered in this study could be due to the very

small sample size that we were able to obtain and the depth of the sequences that have not covered all possible allelic variations. Nonetheless, our results confirmed the high frequency of SNPs in S. officinarum, and abundant SNPs could be mapped to generate a saturated genetic map for genome assembly. The results could also be used for mapping quantitative trait loci (QTL) controlling biomass, sugar yield, and other agronomic traits.

It was also observed that Mxg had one SNP in every 39 base pairs, nearly doubling the SNP frequency of S. officinarum 'LA Purple'. Bundock et al. (2009) used a sugarcane commercial clone that was an interspecific hybrid of S. officinarum and S. spontaneum and reported that there was one SNP in every 35 bases between the two species, similar to that of Mxg. Since sugarcane and *Miscanthus* are so closely related and Mxg is a hybrid of M. sinesis and M. sacchariflorus, it could be expected that their SNP rates would be similar (Bundock et al. 2009). Also, the increased SNP rate in Mxg compared to LAP could be explained by Mxg being an allopolyploid and LAP being an autopolyploid. The chromosomes from both parents in Mxg are likely to have more variation than the doubled chromosomes of LAP.

Although LAP had a lower rate of polymorphism when looking at the consensus strands, it is apparent that the two species are highly related because there was only one different SNP in every 81 base pairs while there was one similar SNP in every 25 base pairs. Sugarcane could serve as a genomic model for *Miscanthus*, which has comparatively less genetic information and fewer genomic resources available. Another study currently is underway to use sugarcane SSR markers to fingerprint *Miscanthus* lines.

The high percentage (57.6%) of sequences that matched the Mxg High Copy Read database is the consequence of primer design. The initial 454 reads were screened against the TIGR plant repeat database. This is the reason that there was no hit to the TIGR plant repeat database. However, *Miscanthus* specific

repeats are not in this database (not available when the database was built), and some primers were not the low copy reads for which we aimed. This could explain some of the discrepancies in the reads obtained from sugarcane and *Miscanthus*. These high copy regions could be rapidly obtaining mutations or be transposable element sequences; therefore, even if there were similar areas in sugarcane, they would most likely have obtained mutations at a different rate via different mechanisms. This could be a possible reason there are only 1.2 times more total *Miscanthus* sequences but 1.9 times more *Miscanthus* sequences that match their original read. This is further supported by the decreased (-15.1%) percentage of sugarcane reads that hit the Mxg high copy database and the increased (+12.4%) amount of LAP reads that hit an EST database. The ESTs are likely to be genic regions that are most likely highly conserved.

2.6 TABLES

Table 4 *Miscanthus* SNP count and sequence annotation

Contig Name	Clones Sequenced	Size (bps)	SNPs	Haplotype	Function	Best Hit Class
Mxg001	4	99	3	2	No Hit	
Mxg001.2	5	99	7	5	giganteus.fas.Contig132636	Mxg High Copy Reads
Mxg002	5	101	6	3	giganteus.fas.Contig128420	Mxg High Copy Reads
Mxg003	4	101	4	2	giganteus.fas.Contig118686	Mxg High Copy Reads
Mxg005	4	101	3	4	giganteus.fas.Contig134478	Mxg High Copy Reads
Mxg007	2	103			chr_1	Sorghum Genome
Mxg009	10	103	0	1	chr_10	Sorghum Genome
Mxg011	7	104	3	2	chr_6	Sorghum Genome
Mxg012	15	104	0	1	Sb04g006100.1 |similar to Putative multidrug resistance protein	Sorghum cds
Mxg015	7	106	6	7	giganteus.fas.Contig126044	Mxg High Copy Reads
Mxg018	9	105	1	2	No Hit	
Mxg028	5	107	0	1	No Hit	
Mxg030	5	112	6	4	chr_6	Sorghum Genome
Mxg050	4	108	0	1	giganteus.fas.Contig131568	Mxg High Copy Reads
Mxg067	11	122	2	3	giganteus.fas.Contig132696	Mxg High Copy Reads
Mxg070	8	127	8	4	No Hit	
Mxg070.2	5	118	5	3	No Hit	
Mxg082	2	132			gb|EC325253.1| SUS31-010-G07-A-054.g mRNA sequence	Sugarcane ESTs
Mxg083	3	131			gb|CA166030.1| SCUTRZ3102C10.g RZ3 mRNA sequence	Sugarcane ESTs
Mxg084	18	113	0	1	chr_1	Sorghum Genome
Mxg085	7	132	7	5	giganteus.fas.Contig133895	Mxg High Copy Reads
Mxg090	12	104	9	11	No Hit	

Table 4 (cont.)

Contig Name	Clones Sequenced	Size (bps)	SNPs	Haplotype	Function	Best Hit Class
Mxg093	15	134	2	3	chr_9	Sorghum Genome
Mxg094	14	135	1	2	No Hit	
Mxg095	9	136	12	8	chr_8	Sorghum Genome
Mxg097	8	111	0	1	No Hit	
Mxg098	6	105	4	2	Mxg_only_AllPooled.fa.screen.Contig21883	Mxg ESTs
Mxg111	2	111			giganteus.fas.Contig118841	Mxg High Copy Reads
Mxg113	18	138	1	2	chr_6	Sorghum Genome
Mxg114	4	110	2	2	giganteus.fas.Contig130896	Mxg High Copy Reads
Mxg114.3	4	110	1	2	giganteus.fas.Contig130896	Mxg High Copy Reads
Mxg118	10	118	2	3	gb|BT067604.1| Zea mays mRNA, complete cds	nt
Mxg122	9	104	1	2	No Hit	
Mxg125	32	146	5	3	No Hit	
Mxg133	6	134	0	1	No Hit	
Mxg136	5	111	7	5	giganteus.fas.Contig130068	Mxg High Copy Reads
Mxg141	28	116	4	5	No Hit	
Mxg144	9	121	12	8	Mxg_only_AllPooled.fa.screen.Contig41310	Mxg ESTs
Mxg146	5	144	5	3	No Hit	
Mxg159	6	153	0	1	No Hit	
Mxg171	2	135			chr_2	Sorghum Genome
Mxg184	3	159			giganteus.fas.Contig128856	Mxg High Copy Reads
Mxg191	7	165	6	6	giganteus.fas.Contig121626	Mxg High Copy Reads
Mxg192	6	138	0	1	Mxg_only_AllPooled.fa.screen.Contig38760	Mxg ESTs
Mxg193	2	142			giganteus.fas.Contig122988	Mxg High Copy Reads
Mxg193.2	4	141	7	4	No Hit	
Mxg196	13	161	0	1	No Hit	
Mxg198	11	127	0	1	giganteus.fas.Contig125729	Mxg High Copy Reads

Table 4 (cont.)

Contig Name	Clones Sequenced	Size (bps)	SNPs	Haplotype	Function	Best Hit Class
Mxg199	10	158	1	2	gb|CA170831.1| SCRUSB1078F12.g SB1 mRNA sequence	Sugarcane ESTs
Mxg200	6	156	2	2	giganteus.fas.Contig119793	Mxg High Copy Reads
Mxg201	12	151	3	4	TC1009 similar to GP|29125030|emb|CAD52890.1||AJ510211 ARIADNE-like protein ARI8	SoGI
Mxg202	7	162	6	7	giganteus.fas.Contig132628	Mxg High Copy Reads
Mxg204	9	142	0	1	chr_6	Sorghum Genome
Mxg215	13	131	11	5	Mxg_only_AllPooled.fa.screen.Contig43459	Mxg ESTs
Mxg219	7	112	1	2	Sb07g025790.1 |similar to Receptor protein kinase PERK1-like protein	Sorghum cds
Mxg220	7	144	0	1	AT5G20890.1 | chaperonin, putative | chr5:7087020-7089906	TAIR
Mxg221	5	102	0	1	gb|CA122113.1| SCJFLR1074A04.g LR1 mRNA sequence	Sugarcane ESTs
Mxg222	6	129	3	2	No Hit	
Mxg223	5	110	0	1	giganteus.fas.Contig125849	Mxg High Copy Reads
Average		124.31	3.25			

Table 5 Sugarcane SNP count and sequence annotation

Contig Name	Clones Sequenced	Size (bps)	SNPs	Haplotype	Function	Best Hit Class
LAP001	3	99			giganteus.fas.Contig119624	Mxg High Copy Reads
LAP005.2	5	233	0	1	giganteus.fas.Contig134478	Mxg High Copy Reads
LAP009	8	103	3	3	chr_10	Sorghum Genome
LAP011	5	104	6	2	chr_6	Sorghum Genome
LAP012	13	104	0	1	Sb04g006100.1 |similar to Putative multidrug resistance protein	Sorghum cds
LAP015	3	107			giganteus.fas.Contig127326	Mxg High Copy Reads
LAP028.2	7	70	0	1	No Hit	
LAP030	3	114			chr_5	Sorghum Genome
LAP050	3	109			giganteus.fas.Contig131568	Mxg High Copy Reads
LAP052.2	4	114	0	1	chr_6	Sorghum Genome
LAP070	6	128	8	5	No Hit	
LAP073	4	129	0	1	No Hit	
LAP073.2	4	113	1	2	No Hit	
LAP083	12	131	0	1	gb|CA166030.1| SCUTRZ3102C10.g RZ3 mRNA sequence	Sugarcane ESTs
LAP085	4	131	0	1	giganteus.fas.Contig133895	Mxg High Copy Reads
LAP090	6	129	7	3	No Hit	
LAP093	9	132	1	2	chr_9	Sorghum Genome
LAP095.2	9	122	3	2	No Hit	
LAP149	4	148	0	1	No Hit	
LAP159	6	120	0	1	No Hit	
LAP160	1	153			chr_5	Sorghum Genome

Table 5 (cont.)

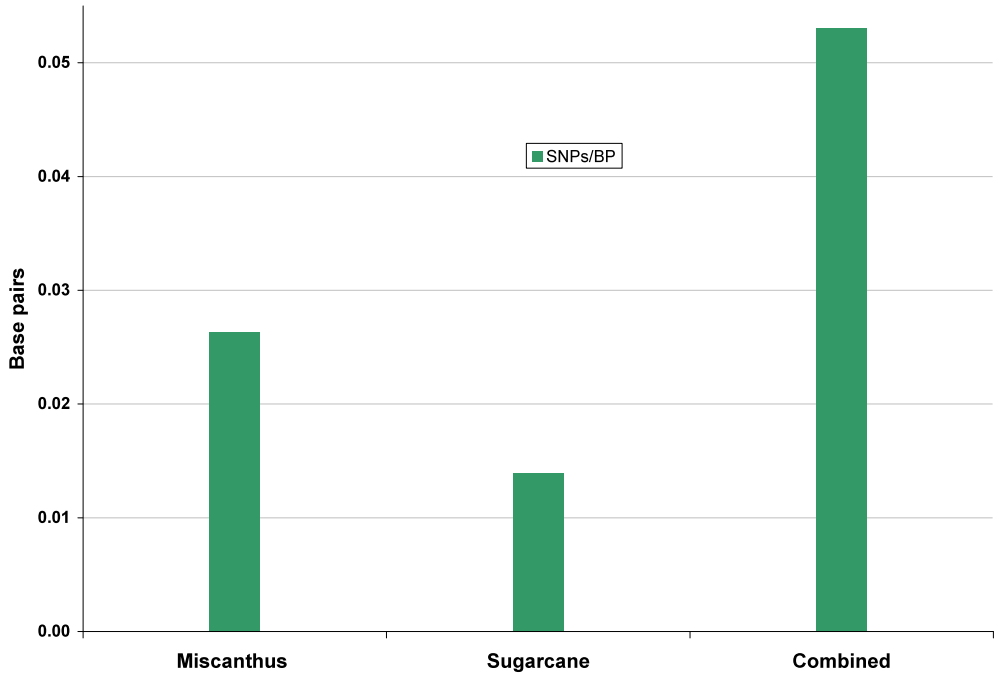
Contig Name	Clones Sequenced	Size (bps)	SNPs	Haplotype	Function	Best Hit Class
LAP171	4	134	1	2	chr_2	Sorghum Genome
LAP180.2	7	120	0	1	No Hit	
LAP191	2	164			giganteus.fas.Contig121626	Mxg High Copy Reads
LAP191.2	4	102	1	2	giganteus.fas.Contig121626	Mxg High Copy Reads
LAP193	5	141	13	4	No Hit	
LAP196	11	161	1	2	No Hit	
LAP199	10	158	0	1	gb|CA170831.1| SCRUSB1078F12.g SB1 mRNA sequence	Sugarcane ESTs
LAP201	5	151	1	2	TC1009 similar to GP|29125030|emb|CAD52890.1||AJ510211 ARIADNE-like protein ARI8	SoGI
LAP204	2	142			chr_6	Sorghum Genome
LAP215	8	131	0	1	ref|XM_002442894.1| Sorghum bicolor hypothetical protein, mRNA	nt
LAP217.2	5	113	0	1	No Hit	
LAP219	11	111	1	2	Sb07g025790.1 |similar to Receptor protein kinase PERK1-like protein	Sorghum cds
LAP220	8	144	0	1	AT5G20890.1 | chaperonin, putative | chr5:7087020-7089906	TAIR
Average		129.27	1.74			

Table 6 *Miscanthus* and sugarcane alignment SNP count and distribution

Contig Name	% Identity	Size (bps)	SNPs (same)	SNPs (different)	Total SNP
Mxg/LAP03	96	100	0	4	4
Mxg/LAP09	99	103	5	0	5
Mxg/LAP11	98	104	7	1	8
Mxg/LAP12	100	104	0	0	0
Mxg/LAP15	99	107	7	0	7
Mxg/LAP30	97	114	8	2	10
Mxg/LAP50	97	108	0	3	3
Mxg/LAP70	96	127	15	0	15
Mxg/LAP83	95	131	0	6	6
Mxg/LAP85	96	132	7	3	10
Mxg/LAP90	94	130	13	3	16
Mxg/LAP93	97	134	3	2	5
Mxg/LAP95	94	136	12	4	16
Mxg/LAP125	98	145	4	2	6
Mxg/LAP159	100	120	0	0	0
Mxg/LAP171	99	134	1	0	1
Mxg/LAP191	98	164	7	0	7
Mxg/LAP193	95	141	16	3	19
Mxg/LAP196	98	161	5	2	7
Mxg/LAP199	99	158	2	1	3
Mxg/LAP201	99	151	4	0	4
Mxg/LAP204	98	142	0	2	2
Mxg/LAP215	96	130	11	2	13
Mxg/LAP219	99	112	3	0	3
Mxg/LAP220	99	144	0	1	1
Average	97.44	129.28	5.20	1.64	6.84

2.7 FIGURES

Fig. 3 Comparison of SNPs per base pair averages for the two species and contigs that were made by aligning the consensus strands for each species



REFERENCES

Aitken KS, Jackson PA, McIntyre CL (2005) A combination of AFLP and SSR markers provides extensive map coverage and identification of hom(oe)ologous linkage groups in sugarcane. Theor Appl Genet 110:789-801.

Al-Janabi SM, Honeycutt RJ, Peterson C, Sobral BWS (1993) A genetic link-age map of Saccharum spontaneum L. 'SES 208'. Genetics 134:1249-1260.

Atienza SG, Satovic Z, Petersen KK, Dolstra O, Martín A (2002) Preliminary genetic linkage map of Miscanthus sinensis with RAPD markers. Theor Appl Genet 105:946–952.

Batley J, Barker G, O'Sullivan, Edwards K, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84-91.

Beckman, J. S. and M. Soller (1984) Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping and costs. Theor Appl Genet 67: 35-43.

Brandes EW (1958) Origin, classification and characteristics. In: Artschwager, E. & E.W. Brandes (Eds.), Sugarcane (Saccharum officinarum L). Agric. Handbk. (U.S. Dept. Agric.) 122, pp. 1– 35.

Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnology Journal 7:347-354.

Chen C, Yu Q, Hou S, Li Y, Eustice M, Skelton RL, Veatch O, Herdes RE, Diebold L, Saw J, Feng Y, Qian W, Bynum L, Wang L, Moore PH, Paull RE, Alam M, Ming R (2007) Construction of a sequence-tagged high-density genetic map of papaya for comparative structural and evolutionary genomics in Brassicales. Genetics 177:2481-2491.

Chin EC, Senior ML, Shu H, Smith JS (1996) Maize simple repetitive DNA sequences: abundance and allele variation. Genome 39:866-873.

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski A (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics 3:19.

Cordeiro GM, Eliott F, McIntyre CL, Casu RE, Henry RJ (2006) Characterization of single nucleotide polymorphisms in sugarcane ESTs. Theor Appl Genet 113:331–343.

da Silva JA, Sorrells ME, Burnquist W, Tanksley SD (1993) RFLP linkage map and genome analysis of Saccharum spontaneum. Genome 36:782-791.

da Silva JA, Honeycutt RJ, Burnquist W, Al-Janabi SM, Sorrells ME, Tanksley SD, Sorbral WS. (1995) Saccharum spontaneum L. 'SES 208' genetic linkage map combining RFLP- and CR-based markers. Mol. Breeding 1:165-179.

da Silva JA (2001) Preliminary analysis of microsatellite markers derived from sugarcane ESTs. Genetics and Mol. Biol. 24:155-159.

Daniels J, Ashley JB, Daniels C (1980) The 1914 sugarcane collecting expedition to New Guinea. Sugarcane Breed. Newsl. 43: 24–38.

Daniels J, and Roach BT (1987) Taxonomy and evolution. In sugarcane improvement through breeding. Edited by D.J. Heinz. Elsevier Press, New York. pp. 7–84.

Goldstein DB, Schlotterer C (1999) Microsatellites: evolution and applications. Oxford University Pres, Oxford.

Grassl CO (1977) The origin of the sugar producing cultivars of Saccharum. Sugarcane Breed. Newsl. 39: 8–33.

Greef JM, Deuter M, Jung C, Schondelmaier J (1997) Genetic diversity of European Miscanthus species revealed by AFLP fingerprinting. Genetic Resources and Crop Evolution 44: 185–195.

Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C, Glaszmann J (1996) RFLP mapping in cultivated ugarcane (Saccharum spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 142:987-1000.

Grivet L, Glaszmann JC, Arruda P (2001) Sequence polymorphism from EST data in sugarcane: a fine analysis of 6-phosphogluconate dehydrogenase genes. Genet. Mol. Biol. 24:161–167.

Grivet L, Glaszmann JC, Vincentz M, da Silva F, Arruda P (2003) ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. Theor. Appl. Genet. 106, 190–197.

Gupta PK, Balyan HS, Sharma PC, Ramesh B (1996) Microsatellites in plants: a new class of molecular markers. Curr Sci 70:45-53.

Hayashi K, Hashimoto N, Daigen M, Ashikawa I (2004) Development of PCR-based SNP markers for rice blast resistance genes at the Piz locus. Theor Appl Genet 108:1212-1220.

Heichel GH: Comparative efficiency of energy use in crop production. Bull Conn Ag Expt Sta, 1974, 739:1-26.

Hernández P, Laurie DA, Martín A, Snape JW (2002) Utility of barley and wheat simple sequence repeat (SSR) markers for genetic analysis of Hordeum chilense and tritordeum. Theor Appl Genet 104:735–739.

Hoarau JY, Offmann B, D'Hont A, Risterucci AM, Roques D, Glaszmann J, Grivet L (2001) Genetic dissection of a modern sugarcane cultivar (Saccharum spp.). I. Genome mapping with AFLP markers. Theor Appl Genet 103:84-97.

Hodkinson TR, Chase MW, Takahashi C, Leitch IJ, Bennet MD, Renvoize SA: The use of DNA sequencing (ITS and trnL-F), AFLP, and fluorescent in situ hybridization to study allopolyploid Miscanthus (Poaceae). Am J Bot 2002, 89:279-286.

Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL (2002) Arabidopsis map-based cloning in the post-genome era. Plant Physiol 129:440-450.

Jakob K, Zhou F, Paterson A (2009) Genetic improvement of C4 grasses as cellulosic biofuel feedstocks. In Vitro Cell.Dev.Biol.—Plant 45:291–305.

Kanazin V, Talbert H, See D, DeCamp P, Nevo E, Blake T (2002) Discovery and assay of single-nucleotide polymorphisms in barley (Hordeum vulgare). Plant Molecular Biology 48: 529–537.

Kota R, Varshney RK, Thiel T, Dehmer KJ, Andreas G (2004) Generation and comparison of EST-Derived SSRs and SNPs in barley (Hordeum Vulgare L.). Hereditas 135:145 – 151.

Kuleung C, Baenziger PS, Dweikat I (2004) Transferability of SSR markers among wheat, rye, and triticale. Theor Appl Genet 108:1147–1150.

La Rota M, Kantety RV, Yu JK, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics 6:23.

Lafferty J, Lelly T: Cytogenetic studies of different Miscanthus species with potential for agricultural use. Plant Breeding 1994, 113:246-249.

Lam E, Shine Jr J, da Silva J, Lawton M, Bonos S, Calvino M, Carrer H, Silva-Filho MC, Glynn N, Helsel Z, Ma J, Richard Jr. E, Souza G, Ming R: Improving sugarcane for biofuel: engineering for an even better feedstock. Global Change Biology Bioenergy 2009, 1:251-255.

Lima ML, Garcia AA, Oliveira KM, Matsuoka S, Arizono H, De Souza CL Jr, De Souza AP (2002) Analysis of genetic similarity detected by AFLP and coefficient of parentage among genotypes of sugar cane (*Saccharum* spp.). Theor Appl Genet. 104(1):30-8.

Linde-Laursen IB: Cytogenetic analysis of Miscanthus 'Giganteus', an interspecific hybrid. Hereditas 1993, 119:297-300.

Lu YH, D'Hont A, Walker DIT, Rao PS, Feldmann P, Glaszmann JC (1994) Relationships among ancestral species of sugarcane revealed with RFLP using single copy maize nuclear probes. Euphytica 78:7-18.

McNally KL, Bruskiewich R, Mackill D, Buell CR, Leach JE, Leung H (2006) Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. Plant Physiology141:26-31.

Ming R, Liu SC, Lin YR, Braga D, da Silva J, van Deynze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH (1998) Alignment of *Sorghum* and *Saccharum* chromosomes: comparative organization of closely-related diploid and polyploidy genomes. Genetics 150:1663-1682.

Ming R, Moore PH, WU K, D'Hont A, Glaszmann JC, Tew TL (2006) Sugarcane improvement through breeding and biotechnology. Plant Breeding Reviews 27: 15-119.

Parida SK, Kalia SK, Kaul S, Dalal V, Hemaprabha G, Selvi A, Pandit A, Singh A, Gaikwad K, Sharma T, Srivastava PS, Singh NK, Mohapatra T (2009) Informative genomic microsatellite markers for efficient genotyping applications in sugarcane. Theor Appl Genet 118:327-338.

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M (1999) Mining SNPs from EST databases. Genome Res. 9:167-174.

Pinto LR, Oliveira KM, Ulian EC, Garcia AA, de Souza AP (2004) Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. Genome 47:795-804.

Pinto LR, Oliverira KM, Marconi T, Garcia AAF, Ulian EC, de Souza AP (2006) Characterization of novel sugarcane expressed sequence tag microsatellites and their comparison with genomic SSRs. Plant Breed 125:378-384.

Rosen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana, Totowa, NJ, pp 365–386.

Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Booth A, Svensson JT, Wanamaker SI, Walia H, Rodriguez EM, Hedley PE, Liu H, Morris J, Close TJ, Marshall DF, Waugh R (2005) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Gen Genomics 274: 515–527.

Selvi A, Nair NV, Balasundaram N, Mohapatra T (2003) Evaluation of maize microsatellite markers for genetic diversity analysis and fingerprinting in sugarcane. Genome 46:394–403.

Shen YJ, Jiang H, Jin JP, Zhang ZB, Xi B, He YY, Wang G, Wang C, Qian L, Li X, Yu QB, Liu HJ, Chen DH, Gao JH, Huang H, Shi TL, Yang ZN (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiology 135:1198-1205.

Sreenivasan TV, Ahloowalia BS, Heinz DJ (1987) Cytogenetics. In: DJ Heinz (ed) Sugarcane improvement through breeding. Elsevier, New York. pp. 211–253.

Swaminathan K, Alabady MS, Varala K, de Paoli E, Ho I, Rokhsar DS, Arumuganathan AK, Ming R, Green PJ, Meyers BC, Moose SP, Hudson ME (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.

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10:967-981.

Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MIT, Henrique-Silva F, Giglioti EA, Lemos MVF, Coutinho LL, Nobrega MP, Carrer H, Franca SC, Bacci M, Goldman MHS, Gomes SL, Nunes LR, Camargo LEA, Siqueira WJ, Van Sluys M, Thiemann OH, Kuramae EE, Santelli RV, Marino CL, Targon MLPN, Ferro JA, Silveira HCS, Marini DC, Lemos EGM, Monteiro-Vitorello CB, Tambor JHM, Carraro DM, Roberto PG, Martins VG, Goldman GH, de Oliveira RC, Truffi D, Colombo

CA, Rossi M, de Araujo PG, Sculaccio SA, Angella A, Lima MMA, de Rosa VE., Siviero F, Coscrato VE, Machado MA, Grivet L, Di Mauro SMZ, Nobrega FG, Menck CFM, Braga MDV, Telles GP, Cara FAA, Pedrosa G, Meidanis J, and Arruda P (2003) Analysis and Functional Annotation of an Expressed Sequence Tag Collection for Tropical Crop Sugarcane. Genome Res. 13: 2725-2735.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 23:4407-4414.

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

Wang J, Chen C, Na J, Yu Q, Hou S, Paull RE, Moore PH, Alam M, Ming R (2008) Genome-wide comparative analysis of microsatellites in papaya. Tropical Plant Biol 1:278-292.

Warner JN (1962) Sugarcane: An indigenous Papuan cultigen. Ethnology 1: 405–411.