CONSTRUCTION AND ANALYSIS OF THE MISCANTHUS GENESPACE

BY

JESSICA R. KIRKPATRICK

THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics with a concentration in Crop Science in the Graduate College of the University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Master's Committee:

Professor Stephen Moose, Chair
Associate Professor Matthew Hudson
Assistant Professor Yoshie Hanzawa
ABSTRACT

*Miscanthus* and energy cane are closely related perennial grasses and candidate bioenergy crops. Both genomes create a challenge for variant discovery and genotyping because of their abundant repeats and the presence of a genome scale duplication with little subsequent divergence. In addition, the *Miscanthus* genome has high heterozygosity due to its self-incompatible breeding system. This complexity necessitates the discovery of sequence variants that distinguish paralogs from alleles, which are most frequent in the low-copy non-coding fraction of the genespace. To enrich for such sequences from *Miscanthus* genomes, we developed a sequence capture array using a solution-based hybridization method. The probe set was designed to capture exons and flanking intronic sequences predicted from alignment of *Miscanthus* transcriptome, fosmid, and genomic reads to the largely syntenic *Sorghum* genome. To facilitate haplotype discovery, we selected for large insert sizes that were subsequently sequenced at high depth as 100x75 bp paired-end reads. Chapter 1 reports the sequence capture of a doubled haploid *Miscanthus sinensis* plant and its parent, and chapter 2 reports the sequence capture of energy cane and two other *Miscanthus* species. This sequence capture design was able to distinguish variation at high depth and it was found that amplification is imperative to achieve this depth. Further analysis of the reads verified that *Miscanthus* has undergone recent whole genome duplication. Additionally, sequence capture serves as a useful tool for comparative genomics and provides strong evidence that *Miscanthus* can serve as a satisfactory reference to the closely related, yet highly complex genomes of *Saccharum spp.*
To Alicia, Steve, and Stephanie
ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Stephen Moose, and my committee members, Dr. Matthew Hudson and Dr. Yoshi Hanzawa for their expert advice and support. I want to thank Dr. Kankshita Swaminathan for her constant and invaluable guidance, for sharing her dessert, and for spoiling my cat. I want to thank Therese Mitros and Dr. Dan Rokhsar for their collaboration and inspiration. Many thanks to Katarzyna Glowaca and Stanislaw Jezowski for providing us with the plant that allowed me to undertake this project. I would also like to thank Dr. Alvaro Hernandez for his protocol advice and wonderful sequencing data. I would like to thank my EBI lab mates, the Moose lab, and the Hudson lab for their assistance and friendship. I want to thank Chris Holmes for all his loving support and encouragement. And last but not least, I would like to thank the Energy Biosciences Institute and the Jonathan Baldwin Turner Graduate Fellowship Program for their gracious financial support.
TABLE OF CONTENTS

CHAPTER 1: *MISCANTHUS SINENSIS* DOUBLED HAPLOID AND PARENT SEQUENCE CAPTURES ................................................................................................................. 1
  1.1 INTRODUCTION ....................................................................................................................... 1
  1.2 MATERIALS AND METHODS .................................................................................................... 8
  1.3 RESULTS .................................................................................................................................. 15
  1.4 DISCUSSION ............................................................................................................................ 22
  1.5 FIGURES ................................................................................................................................. 27
  1.6 TABLES .................................................................................................................................. 36

CHAPTER 2: ENERGY CANE, *MISCANTHUS X GIGANTEUS*, AND *MISCANTHUS SACCHARIFLORUS* SEQUENCE CAPTURES ................................................................. 39
  2.1 INTRODUCTION ....................................................................................................................... 39
  2.2 MATERIALS AND METHODS .................................................................................................... 41
  2.3 RESULTS .................................................................................................................................. 46
  2.4 DISCUSSION ............................................................................................................................ 49
  2.5 FIGURES .................................................................................................................................. 53
  2.6 TABLES .................................................................................................................................. 60

REFERENCES .................................................................................................................................. 61

APPENDIX A: ABBREVIATIONS .......................................................................................................... 67
APPENDIX B: OTHER FIGURES .......................................................................................................... 68
CHAPTER 1: MISCANTHUS SINENSIS DOUBLED HAPLOID AND PARENT SEQUENCE CAPTURES

1.1 Introduction

Miscanthus: Potential Biofuel

In 2007, the US congress passed the Renewable Fuel Standard 2 (RFS2) through the Energy Independence and Security Act, which set a goal of increasing the use of renewable fuels in transportation to 36 billion gallons per year by 2022 (H.R. 6 - U.S. 110th Congress 2007). Sixteen billion gallons of this is forecasted to come from cellulosic biofuels such as crop residues, woody biomass, waste, and perennial grasses. Besides switchgrass, one of the most promising perennial grasses that can fulfill the RFS2’s cellulosic biofuel requirement is Miscanthus. An ideal candidate species of this genus, Miscanthus x giganteus, averages high biomass in comparison to switchgrass and corn, requires relatively low inputs, is sterile and non-invasive, and has chilling tolerance to help it cope with cooler Midwestern weather. All of these traits make it a worthwhile plant to pay attention to.

Miscanthus x giganteus (Mxg), a hybrid of Miscanthus sinensis and (Hodkinson, Chase, Takahashi, et al. 2002) Miscanthus sacchariflorus has many valuable and competitive traits. It is a densely grown, tall grass that attains high biomass. In a 3-year study performed in different regions of Illinois, Mxg achieved an average yield of 30 t ha⁻¹, while switchgrass only averaged 10 t ha⁻¹ (Heaton et al. 2008). In this same study, it was estimated that Mxg would only require 9.3% of US cropland to meet the renewable energy goal and offset 1/5 of U.S. gasoline requirements. For switchgrass and corn stover to meet this same goal, they would require 26.5% and 37.2% of U.S. cropland, respectively. These
findings are corroborated by another study done in Illinois that not only found that Mxg averaged 4 times as much yield as switchgrass, but found that in the long run Mxg would be less expensive to grow in Illinois compared to switchgrass (Khanna et. al 2008). Due to the high biomass that Miscanthus produces, it is likely to be sustainable in terms of land use and long-term economics.

While high biomass is the main reason Mxg makes such a great bioenergy candidate, it also has several other interesting traits. Mxg requires similar amounts of nitrogen as switchgrass, however it needs less fertilizer than most annual food crops like maize (Arundale et al. 2013). Mxg requires less water than other perennial crops, but it does produce better yields with an increase in water supply. As a result, it can also tolerate short periodic flooding (Heaton et al. 2012). Miscanthus x giganteus is also a sterile, non-invasive species. Its rhizomes spread slowly, so it is unlikely to spread too far. Mxg plants have a long life span, and are expected to live for 15 to 20 years after establishment (Brading 2006), so it is not necessary to replant for several years. Finally, Mxg has a cold tolerance that is uncommon among C4 grasses (Beale et al. 1996; Naidu 2003). It was cited that Mxg can continue to photosynthesize at temperatures as low as 47F, and has been shown to survive temperatures as low as -20F (U.S. Department of Energy 2011). As a perennial, many of these additional traits favor its sustainability.

While Miscanthus x giganteus has many positive traits, as with any crop, it also has some disadvantages. Mxg has a high establishment cost, and it is not possible to harvest until at least the second year. Mxg is rhizomatous and so planting is not as easy as it would be with seed. Other challenges in establishing this plant include competition from weeds for the first two years, requiring multiple herbicide applications (Heaton et al. 2012), and
acquiring equipment that can plant the rhizomes efficiently. However, according to the U.S. Department of Energy (2011), there is equipment becoming available that can plant more than 25 acres per day. Since Mxg does have a long lifespan, it is not necessary to replant every year, and these establishment costs will likely only come up once in several years.

In summary, Miscanthus giganteus has numerous beneficial traits that make it one of the front-running candidates for perennial cellulosic biofuels that can help offset petroleum as mandated in the U.S. government’s RFS2. Due to this, it is an important plant to study in depth. However, in comparison to other grasses like maize and sorghum, the Miscanthus x giganteus genome is not well understood. Therefore it is of great importance to develop Miscanthus genomic resources.

**Miscanthus Genomics and Complexity**

While Miscanthus genomics are not as advanced as other major crops, there has been a lot of progress in the past several years. Mxg is determined to be a triploid having 57 chromosomes, with a base number of 19 per subgenome (Linde-Laursen 1993; Lafferty and Lelley 1994). Mxg is highly abundant in repeats, especially transposons and centromeric repeats, which are estimated to make up 80% of the genome (Swaminathan et al. 2010). To further add to this complexity, there is strong evidence that Miscanthus has undergone a whole genome duplication (WGD) event compared to Sorghum bicolor (Ma et al. 2012; Swaminathan et al. 2012) and is thus likely to be an allotetraploid. The Miscanthus x giganteus genome is clearly highly complex and this makes understanding the Miscanthus genome and accelerating genomic resources, which can be used to accelerate domestication, a challenge.
Though *Miscanthus x giganteus* has an innately complex genome, closely related genomes can help deconvolve it. *Miscanthus* is a member of the *Andropogoneae* tribe of grasses that also includes *Sorghum*, maize, and *Saccharum* (Hodkinson, Chase, Lledo et al. 2002). Two previously mentioned studies point to *Sorghum* being a suitable reference genome for *Miscanthus* (Swaminathan et al. 2010; Swaminathan et al. 2012). The later of these studies show that *Miscanthus* demonstrates strong macro-synteny to *Sorghum*. For every chromosome in *Sorghum bicolor*, there are two *Miscanthus sinensis* linkage groups that are syntenic to it, except in one case where it is thought that there was a fusion of two ancestral *Sorghum* chromosomes into one linkage group. Since *Sorghum* is so closely related to *Miscanthus*, and because it has been fully sequenced (Paterson et al. 2009), it is possible to use the *Sorghum bicolor* genome as a reference for *Miscanthus*.

*Miscanthus* complexity, especially that resulting from the WGD, necessitates the need to separate homeologs from alleles. Variants can be extracted by using next-generation sequencing (NGS) in combination with variant calling, but high depth is needed for the complex *Miscanthus* genome. Ideally one would choose an inbred or homozygous plant with no allelic variation to identify variants that define the two homeologs, however, because *Miscanthus* is self-incompatible, most lines are highly heterozygous. A doubled haploid line generated from a *Miscanthus sinensis* plant (Glowacka et al. 2012) has been shown to lack allelic variation (Swaminathan et al. 2012). This line and its parent were chosen to quantify paralogous and allelic variation using sequence capture technology, which can generate the read depth required to confidently call variants.
Sequence Capture

Sequence capture is a new technique that allows scientists to pull out specific, usually exonic, regions of a genome and sequence them at higher depth than whole genome sequencing (WGS) allows (Clark et al. 2011). This high depth allows for not only more confidence in the resultant sequence data, but also for better variant calling, which is of great importance for complex plant genomes that can have numerous allelic variants per locus. Sequence capture sequencing differs from WGS because it does not sequence random pieces of DNA from the entire genome; instead it theoretically sequences only the DNA it is designed to capture, which is often genic. Thereby, it reduces the complexity of the genome by avoiding repeats and leaving only the gene space. Sequence capture sequencing differs from transcript sequencing, because it permits the “capture” of flanking intronic regions which often contains more variation. In situations where paralogous copies of a gene are highly conserved, the similarity is usually highest in the coding region making the variations in the flanking non-coding sequences essential for correct identification of the copies. Since sequence capture is done at the genomic level, unlike transcriptome sequencing it is not limited by tissue/time-specific expression; it is able to capture all exons that it is designed for.

Exome capture has been performed extensively in human studies. Only recently has it been applied in plant genomes, which are more of a challenge considering that plants can have high ploidy, abundant repeats, numerous duplication events, and often lack a reference genome. Three of the plant genomes that have used exome capture to date, possess a reference genome and are diploid: maize, soybean, and black cottonwood (Fu et al. 2010; Haun et al. 2011; Zhou and Holliday 2012). It has been found, however, that
exome capture can be done in plant genomes lacking a reference by using other resources such as cDNAs, ESTs, and/or predicted coding regions; this technique has been done in sugarcane, barley, loblolly pine, wheat, and cotton (Saintenac et al. 2011; Bundock et al. 2012; Salmon et al. 2012; Winfield et al. 2012; Mascher et al. 2013; Neves et al. 2013).

There are 3 main exome capture platforms: Agilent’s SureSelect, Roche-Nimblegen’s SeqCap EZ Exome and Illumina’s TruSeq Exome Enrichment. All of these platforms use probes, or baits, that are complementary to the targeted exome regions. These probes are used to hybridize to fragmented genomic DNA, and then they attach to magnetic beads while all other DNA that is not targeted is washed away. The sequences attached to these probes are then sequenced on a NGS platform. The basic protocol is illustrated in Figure B.1.

While the protocols between these platforms are similar, they differ most in their probe design. Nimblegen designs high density probes that cover a relatively smaller portion of the genome and thus requires less sequencing to detect variation (Clark et al. 2011). Agilent and Illumina design less dense probes covering a larger portion of the genome that can detect the same amount of variation as Nimblegen with more sequencing. It is not clear which platform is superior, but the general trend in plant genomics is to implement either Nimblegen or Agilent platforms.

There have been no sequence capture studies done on Miscanthus to date, however a few studies have used genotyping by sequencing (Ma et al. 2012) and transcriptome sequencing to call variants (Chouvarine et al. 2012; Swaminathan et al. 2012). The Ma et al. and Swaminathan et al. studies both generated genetic maps for M. sinensis. The Chouvarine et al. study of variation among Miscanthus x giganteus genotypes using rhizome
mRNA data was handicapped by the lack of a reference genome, inadequate depth and the lack of knowledge of the WGD. In both the Swaminathan et al. and the Chouvarine et al. studies the use of transcriptomic data limited the number of variants observed, because they only capture the exonic regions. This resulted in the inability to distinguish between homeologous variation from true allelic variation. A sequence capture designed to capture gene space and avoid repeats can use exonic probes that capture both the exons as well as the flanking non-coding regions. This is important in polyploids like Miscanthus and Saccharum. Aside from the WGD, Miscanthus can naturally range in ploidy from 2x to 3x to 4x. The homeologs and alleles can only be separated out if enough variation is captured at high depth, and since introns tend to contain more variation than exons, exome capture is more equipped to differentiate these alleles and homeologs.

**Research Justification**

As previously mentioned, the complexities of the Miscanthus genome make it difficult to sort out alleles from homeologs. To do this, it is necessary to achieve high sequence depth to detect variation in genic regions, to know what sequences contain genes, and to be able to fix allelic variation. In this study, sequence capture and subsequent sequencing is implemented to achieve high depth in exonic and flanking intronic regions. To capture genes, the sequence capture probes are designed from previously collected RNA-seq, 454, and fosmid sequences. Fixation of allelic variation was realized by obtaining a doubled haploid Miscanthus plant, and so any observed sequence variants can be attributed to homeologous sequences generated by the whole genome duplication. Exome capture was performed on the doubled haploid and its parent, and non-captured libraries of these were
sequenced as well for comparison. With these resources, it was possible to unravel homeologous variation in *Miscanthus*.

### 1.2 Materials and Methods

**Collection of Materials**

A doubled haploid (DH) *Miscanthus sinensis* plant (IGR-2011-001) was created by Katarzyna Glowaca at the Institute of Plant Genetics at the Polish Academy of Sciences in Poznan, Poland (Figure B.2) (Glowacka et al. 2012). Glowaca collected and mailed stem and leaf tissue from the DH and the parent of the doubled haploid (PDH) (IGR-2011-003) to our laboratory at the Energy Biosciences Institute (EBI) at the University of Illinois in Urbana-Champaign (UIUC), Illinois.

**Probe Design**

To perform sequence capture, it was first necessary to design a probe set that could capture genomic regions of interest. This probe set was developed by Therese Mitros, a collaborator from the University of California, Berkeley (UC Berkeley). The probes were made to be 65bp in length. They were designed from *Miscanthus* RNA-seq assemblies (Barling et al. 2013), a *Miscanthus x giganteus* (*Mxg*) short-read genomic assembly (Swaminathan et al. 2010), and *Mxg* fosmids (an unpublished contribution from Liang Xie) (Figure B.3). The *Miscanthus* RNA-seq assemblies were locally aligned to *Sorghum* using BLAT (BLAST-like Alignment Tool). If there was a match, the boundaries of the match were used to design probes within exons and the probes were tiled across. If there was no match
to *Sorghum*, then probes were densely tiled with a 5bp offset across the entire *Miscanthus* EST contig. The *Mxg* short-read genomic assembly was locally aligned, using BLAT, to the *Miscanthus* RNA-seq assemblies. If there was a match, this region was ignored because probes should have already been designed to it in the first step. If there was no match, probes were densely tiled across the genomic contig. For the *Mxg* fosmids, probes were densely tiled across. Once all of these probes were developed, Mitros filtered sequences that were highly repeated and removed all probes with high-GC content, since these have a lower efficacy in the sequence capture protocol. The remaining 2,036,158 probes were submitted to and validated by Roche Nimblegen for compatibility with their assay.

**Sequence Capture**

*Genomic DNA library preparation*

Genomic DNA was extracted from young leaf tissue of the DH line (IGR-2011-001) and the DH parent line (IGR-2011-003) using a CTAB DNA protocol. To perform this protocol, the young leaf tissue is first ground in liquid nitrogen, and then 25ml of CTAB extraction buffer is added for every 2.5g of tissue. This sits for 1 hour at 65°C, then is cooled to room temperature. An equal volume of Phenol:Chloroform:Isoamyl Alcohol (25:24:1) pH 8 is mixed and spun at 7000-8000 Xg at 10°C for 10 minutes. The aqueous top phase is transferred to a new tube, and then an equal volume of Chloroform:Isoamyl Alcohol (24:1) is added to the aqueous phase. This is mixed and spun at 7000-8000 Xg at 10°C for 10 minutes. The aqueous top phase is transferred to a new tube. Seven tenths of the volume of isopropanol is added to the aqueous phase and mixed gently until threads of DNA forms. This is then spun at 7500 Xg at 20°C for 10 minutes. The supernatant is discarded with care
taken to not disturb the pellet. The pellet is air dried and then re-suspended in 200-300μl of 1X TE or 10mM Tris pH 8.0. Lastly, the DNA is quality checked by running 200ng of 1/100 or 5/100 dilutions on a 0.7% agarose gel.

After DNA was extracted, a portion was taken from both the DH and the parent of the DH library and diluted down to 20 ng/μl for a total of 1,500μl. Then 15 100μl aliquots were made in Diagenode non-stick RNase free 1.5ml tubes. These tubes were placed in a Biorupter (Model # B01010002), 6 at a time, and sheared at low setting for 10 minutes (30 seconds on, 30 seconds off). The sheared DNA was run on a 2% E-gel to size select for 400-500bp fragments. The DNA was extracted from the gel using a MinElute Gel Extraction Kit, and quantified on a High Sensitivity (HS) Qubit assay. For both samples there was approximately 2μg of DNA. The following Illumina protocol required 1μg of DNA per library, so there was enough for 2 libraries for the DH and 2 libraries for the parent of the DH.

Both the DH DNA and that of its parent were split into to two tubes each for a total of four genomic libraries. The prepared genomic DNA Illumina paired-end library was constructed by following Illumina’s low-throughput protocol in their TruSeq DNA Sample Preparation Guide. Since there were only four samples, Diagenode non-stick RNase free 1.5 ml tubes were used instead of 96-well PCR plates. When needed, DNA concentrations were quantified using the Qubit BR and HS assays and quantified DNA on an Agilent 2100 Bioanalyzer by using the Agilent DNA 7500 Kit. After the libraries were made, one library of the DH and one library of its parent were saved and stored at -20°C. These were saved as “non-captured” libraries for comparison to the sequence captured libraries.
**Sequence capture protocol**

Sequence captured libraries were constructed by using the NimbleGen SeqCap EZ hybridization and wash kit (#05 634 261 001) and by following Nimblegen’s protocol (NimbleGen SeqCap EZ Choice Library SR User's Guide v1.0, December 2010). This protocol's LM-PCR and hybridization steps were adapted for use of Illumina TruSeq DNA libraries by following the guide provided by Nimblegen (Technical Note Supplement: Targeted Sequencing with NimbleGen SeqCap EZ Libraries and Illumina TruSeq DNA Sample Preparation Kits, Instructions for using Illumina TruSeq DNA Libraries with SeqCap EZ Libraries). Once again, Non-stick RNase free 1.5 ml tubes were used instead of 96-well PCR plates. Five µl of plant capture enhancer, provided by Nimblegen was used instead of 5 µg Cot-1 DNA. After cleaning up the final PCR step, the samples were run on a DNA 7500 Kit on the Agilent bioanalyzer. Aside from expected 300-600bp peak, the samples showed a second smaller but significant peak around 125bp. This appeared to be primer dimers, so an additional clean-up step was applied by following the AMPure bead cleanup outlined in the “TruSeq DNA Sample Preparation Guide”. After this, there was still a significant peak left over, so the libraries were also run through a 2% E-gel and bands from 300bp and up were cut out. Finally, the libraries were quantified using qPCR with a Kapa Library Quantification Kit (Kit #KK4808).

**Sequencing**

The four genomic libraries --captured DH DNA, captured PDH DNA, non-captured DH DNA, and non-captured PDH DNA-- were sequenced on an Illumina HiSeq 2000 system at UIUC’s Roy J. Carver Biotechnology Center. They were sequenced as 100bp paired-end reads.
Before these libraries were fully sequenced, a power fluctuation at the Biotechnology Center caused the sequencing machine to terminate early resulting in reads that were 100bp by 75bps, rather than 100bp by 100bp.

Sequencing coverage was determined by using the Lander/Waterman equation for computing coverage: \( C = \frac{LN}{G} \), where \( C \) stands for coverage, \( G \) is the haploid genome length, \( L \) is the read length, and \( N \) is the number of reads (Lander and Waterman 1988). For the non-captured libraries, \( G = 2.5 \text{Gbp} \) (the estimated length of a Miscanthus haploid genome) and for the captured libraries \( G = 54.66 \text{Mbp} \) (the estimated length that the probes cover in the Miscanthus genome). \( L \) was set to 175 and \( N \) was set to the specific library's raw read count.

**Bioinformatics Analyses**

*Quality control and coverage*

Each library included 6 adapters because of a mistake made during Illumina DNA library preparation. First, the FASTQ files were de-multiplexed using a custom Perl script, resulting in 6 smaller pairs of FASTQ files per library. The files needed to be de-multiplexed because only one 3-prime adapter could be removed from paired-end FASTQ files at a time, given our computational resources. For each smaller pair of FASTQ files, the reads were trimmed to a minimum PHRED quality score of 25 and the 3-prime adapter sequences were removed simultaneously using a custom Python script. Then the small pairs of FASTQ files were concatenated back into one whole pair of FASTQ files. 5-prime adapters were masked using a NovoalignMPI parameter in the alignment step (see the following section).
Alignments

Reads were aligned using NovoalignMPI (v. 3.00.05) (http://www.novocraft.com/) on the Energy Biosciences Institute’s computer cluster. The *Sorghum bicolor* genome sequence (release number 79) was used as a reference genome. Parameter ‘-c 1’ paired with ‘-n #’ was used to restrict the number of threads NovoalignMPI could use on the computer cluster to a specified number #. The ‘-F ILM1.8’ parameter indicated which Illumina quality scores the FASTQ files were encoded with, and ‘-i PE 500 150’ specified that the libraries were paired-end with a mean insert length of 500 bp and 150 bp standard deviation. All alignment locations were reported in SAM (Sequence Alignment/Map) format by using options ‘-r All’ and ‘-o SAM’ respectively. The gap opening penalty was set to 25 and the gap extension penalty to 5 using ‘-g 25 -x 5’ options to allow for gaps, since Miscanthus was being aligned to a different genome, Sorghum. 5-prime adapter sequences were removed by denoting the ‘-a’ parameter.

The alignment to Sorghum reported an oddity that was found in the paired-end fragment length distribution that NovoalignMPI returned for the captured libraries, but not the non-captured libraries. These fragment lengths are virtual lengths predicted by NovoalignMPI, and do not represent actual physical fragment lengths. We expected the distribution to be unimodal around the mean insert size, but there was an extra peak centered on 195-224bp. It is not completely clear why this peak is present, but possibilities are explained in the following Results section. Since an insert size of 400-500 bp was originally selected for, all paired-end reads with a fragment length below 300 bp in size were removed from all four libraries. This step essentially discarded all paired-end reads.
within the anomalous peak, and was performed using ‘grep’ and ‘cat’ UNIX commands on alignment SAM files.

The probes were aligned to *S. bicolor* using the same parameters used for the DH and PDH alignments, with a few exceptions. The parameters ‘-i PE 500 150’ and ‘-a’ were removed. The parameter ‘-F ILM1.8’ was changed to ‘-F FA’ since the probes were in FASTA format, not FASTQ format. Parameter ‘-r All’ was changed to ‘r Random’ so that only a maximum of one alignment, chosen randomly, was returned per probe sequence.

*Variant calling*

Before using variant calling software, SAMtools (Li et al. 2009) was implemented to convert SAM files to BAM (Binary Alignment/Map) format, to sort the BAM files, and finally, to convert BAM files to pileup format. Duplicate reads were not removed so as to avoid the risk of deleting nearly identical homeologous reads. The variant caller, VarScan (Koboldt et al. 2009) was employed to call variants from pileup files by implementing the pileup2snp command. A minimum coverage of 10 (--min-coverage 10) and a minimum variant coverage of 5 (--min-reads2 5) was specified. All other parameters were left at default.

*Assembly*

Assemblies were made *de novo* for the DH library by running ABySS software (Simpson et al. 2009). The abyss-pe command was implemented with the parameter ‘b=0’ to avoid “bubble popping” which might collapse homeologs and parameter ‘E=0’ to prevent base erosion due to low coverage. Several assemblies were constructed to determine the most successful k-mer length. Assemblies were made with the following k-mer lengths: 40, 50,
60, 70, and 80. An attempt was made to make an assembly with k-mer 90, but the computer cluster did not have enough memory for this. Results showed that a k-mer length of 60 resulted in the largest contig N50 score, so another round of assemblies were made at k-mer lengths 55, 57, 59, 61, 63, and 65.

Assembly alignment
Contigs from the k-mer 61 assembly were aligned to *Sorghum bicolor* using Novoalign (v. 3.00.05) and the same parameters that were indicated for the probe alignment previously described, with two exceptions. The ‘-k’ option was removed because it caused the program to crash, and the ‘-r’ parameter was set to ‘-r All’ to retrieve all possible alignments.

Additional analyses
Analyses performed on the output of the alignment and variant files were done using custom Perl scripts to mine the data and Microsoft Excel was used to create charts and graphs of the data. Visualization of the alignments was done using IGV (Integrative Genomics Viewer) software (Robinson et al. 2011).

1.3 Results
Sequence Capture Design
Normally the capture platform vendor, Roche-Nimblegen in our case, designs the probe set, but since there is little publically available data for *Miscanthus* and the sequences in hand came from multiple resources--both transcriptomic and genomic--we developed the probe
set in house by creating densely tiled probes from *Miscanthus* transcripts (Barling et al. 2013), *Mxg* genomic contigs (made from 454 sequencing data) (Swaminathan et al. 2010), and *Mxg* fosmids (Figure B.3). This resulted in approximately 2 million probes covering 54.66 Mbp of the *Miscanthus* genome (Table 1.1). These probes were aligned to *Sorghum bicolor* (release 79), being the closest fully sequenced genome available, to determine where the probes were distributed based on *S. bicolor* annotations (Figure 1.1). Only 1,226,946 probes actually mapped to sorghum. Around 73% of these alignments are in known sorghum exons, while approximately 9% of these alignments are completely inside of or partially inside and partially outside of a gene. Approximately 18% of these alignments are in “other” un-annotated regions. These “other” probes could have been designed from non-genic regions covered by the long 454 and fosmid reads.

**Performance of Sequence Capture**

The following 4 libraries were sequenced on an Illumina HiSeq 2000 platform, resulting in 100x75bp paired-end reads: sequence-captured doubled haploid (DH), non-captured DH, sequence-captured parent of the doubled haploid (PDH), and non-captured PDH. The raw read counts returned from sequencing can be found on Table 1.2, as well as the read counts after quality control (QC). Quality control includes removing 3-prime adapter sequences and trimming to a minimum PHRED quality score of 25. This table also depicts the coverage of each of these libraries, where the non-captured libraries are calculated based on the size of the estimated haploid genome size (2.5Gbp) and the captured libraries are based on the length of the genome the probes are estimated to cover (54.66Mbp). The
captured libraries had more than 50 times more coverage than the non-captured libraries when calculated this way.

The trimmed reads were then aligned to the *Sorghum bicolor* assembly (release 79) and all possible alignments were returned. About 35% of the non-captured libraries’ single reads, meaning that each paired-end read is considered to be two single reads, aligned to sorghum, while nearly twice as many aligned for the captured libraries. Table 1.3 shows these exact percentages and the total number of alignments that were returned. As mentioned in section 1.2, there was an oddity found in the paired-end fragment distribution when aligning the two captured libraries to *S. bicolor* (Figure 1.2). Again, these fragment lengths are virtual fragment lengths predicted by NovoalignMPI, and do not represent actual physical fragment lengths. The smaller peak in this distribution is abnormal, and must have been caused by the capture, since it is not seen in non-captured libraries. Further investigation into this revealed that the small fragments were partially accounted for by telomeric repeats and sub-telomeric repeats. The other fragments could be caused by tandem repeats. When reads contain repeats like this, and some did in our data, this can cause them to map to multiple locations on multiple chromosomes that contain this repeat, and so they become highly represented in the fragment distribution. Furthermore, if two paired reads contain this repeat, then they are likely to map very close to each other, causing their virtual fragment length to be 175bp or more, which is close to the size at the apex of this small peak. Because this small peak caused noise in our data and we originally selected fragment sizes ranging from 300-700bp, we decided to remove all paired reads that had a fragment length of 300bp or less. The number of alignments left after this removal is listed in the right-most column of Table 1.3.
When making the physical sequence capture libraries, genomic fragments of 300-700bp were selected. Depending on where a probe attaches, this allows for extra sequence, usually intronic, to be captured where there are no probes designed. Figure 1.3 illustrates this well. At the top in purple and pink are the probes aligned to *S. bicolor* and there is a space in between them of about 250bp. Below that is the captured DH library alignment depicting many reads completely covering that approximately 250bp region where there are no probes. This demonstrates that more genomic region is being captured than what is targeted by the probes.

**Depth Comparison**

Depth can be visually compared in Figure 1.4. The top panel of this figure illustrates the C4-PPDK gene model in *S. bicolor*. Below that are all the probes that align to this region; probes cover most of the gene model. In the third panel, the read depth of each base is represented for the non-captured DH library, and in the fourth panel the same is depicted for the captured DH library. Both the third and fourth panels are on the same scale, so it is clear that the captured library has substantially higher depth than the non-captured library.

In Figure 1.5, depth is depicted in a different way. A “window” is defined as being a series of continuous bases of at least 75bp from an alignment with base read depths of at least five. In Figure 1.5, the non-captured DH’s window depths are compared to the captured DH’s window depths. On the x-axis are binned read depths, and on the y-axis are the window counts, which are normalized to their number of alignments. The circled portion of the graph shows that the majority of non-captured windows have depths of 5-15, and there are more non-captured windows in these bins than captured windows. In all
other higher-depth bins there are more captured windows than non-captured windows. This demonstrates that the captured DH library has more windows at higher read depths than the non-captured DH library does.

When visualizing the library alignments on IGV, the differences between captured and non-captured are obvious. Figure 1.6 depicts an alignment to a portion of the C4-PPDK gene in *S. bicolor*. In the top panel is the non-captured DH alignment, in the second panel is the captured DH alignment, and below that are the reference sequence and an illustration of the C4-PPDK gene model. Again it is demonstrated that the captured library has higher read depth than the non-captured, but more importantly it demonstrates that the homeologous reads, denoted by the labels ‘I’ and ‘II’, have higher read depth in the captured library as well. This is important for the success of variant calling, which will be discussed in the next section. In the case of a doubled haploid genome, which should theoretically only contain homeologous variation, it may be acceptable to only have a read depth of five to call each variant. In that case, the non-captured library would satisfy this, however in natural *Miscanthus* genomes there can be more than two variants per locus, and so a higher depth is needed.

**Variant Calling**

Variant calling was performed on all libraries using VarScan. A minimum read depth of ten and a minimum variant depth of five was required. The number of total base pairs per library and the number of single nucleotide variants (SNVs) that were predicted from these bases are shown in Table 1.4. In the captured libraries approximately 3% of the total bases
are predicted to be SNVs, while in the non-captured libraries only ~1% of the total bases are predicted to be SNVs.

The distribution of these SNVs was determined based on *S. bicolor* annotations. Panel A of Figure 1.7 illustrates how many SNVs are in sorghum exons, introns and other regions for each library. The captured PDH library has a higher number of called SNVs in comparison to the captured DH. This difference is likely to be the alleles that the DH did not and could not have. In the non-captured libraries the opposite trend is found. In this case, VarScan may not have been able to call as many SNVs in loci where there were three or more variants (only found in the PDH) due to low depth, since there needs to be a depth of at least five for each variant. For example, if at a specific base the DH and the PDH had a depth of 20 and the DH had two variants and the PDH had three variants, all of these variants would be called if each variant had a depth of at least five. If you cut this depth in half to ten, like what could be the case for the non-captured libraries, then in the DH there is still just enough to call two variants at a depth of five, but not enough to call the three variants in PDH, since not all three can have a depth of five. Also important to note, is that the captured libraries have more SNVs in both exons and introns than the non-captured libraries do. Panel B also shows that the captured libraries have a higher average SNV depth than the non-captured libraries. This indicates that the sequence capture returned high depth in targeted exons and introns, allowing for more variant calls than the non-captured libraries could achieve. The “other” category includes SNVs that do not lie in any known *Sorghum* gene annotations. These could include SNVs in repeats, or *Miscanthus*-specific genes, or non-coding regions that were captured. The average depth of the “other”
SNVs is higher in the non-captured libraries because, unlike the captured libraries, coding regions were not targeted.

Lastly, the SNVs in the captured DH library were defined as being either ‘fixed’ or ‘50%’ variants. A fixed variant is represented 90% - 100% of the time at a single base pair locus, and so is likely to be a fixed SNV that is different from *S. bicolor*. A 50% variant is represented 40 – 60% of the time, and is likely to be one of two homeologous SNVs at that position. We did not include SNVs represented 0 – 39% or 61-89% of the time to reduce doubt that these are true fixed and homeologous SNVs. For every *Sorghum* chromosome, the count of fixed and 50% SNVs were found for each known gene model. We found that fixed SNVs tended to outnumber 50% SNVs 2:1 (Figure B.5). These counts, along with the average depth across each gene model, were graphed into an area plot. Figure 1.8 depicts one of these graphs for chromosome 4, where a fusion is thought to have occurred in *Miscanthus*. Above this graph is a 3-D area plot that zooms into a smaller section of chromosome 4. There were two situations we looked for. First, regions with significant depth that have fixed SNVs but no 50% SNVs suggest that one homeolog was deleted. Secondly, regions with significant depth that had no fixed or 50% SNVs indicated that either this region had no variants originally or both homeologs had been deleted. The close-up 3-D area plot shows what a typical region of the graph looked like. There are no situations like the two described in any of the chromosomes.

**Assembly**

Assemblies of the captured DH reads were made to acquire contigs that truly represented each of the two homeologs. They were made with ABySS software at varying k-mer lengths
to find the best assembly. The assemblies were not allowed to collapse reads at a base pair that was majority represented. Rather, the contigs were allowed to break apart into two if a confident continuous path was not obtainable. The ability to preserve both homeologs was one of the main reasons we chose ABYSS as our assembler. The results of a subset of these assemblies are shown in Table 1.5. For each k-mer length, the contig N50 score and the scaffold N50 score is given. The two k-mers that are highlighted in gray had the highest combined scores. The k-mer 61 contig assembly was randomly chosen to be aligned to the *S. bicolor* genome. Of these contigs, 65.81% of them aligned to *S. bicolor*. Figure B.6 shows that both homeologs were preserved in the contigs, and these same homeologs are seen in the capture DH alignment.

Figure 1.9 shows a larger section of the assembly in comparison to the captured DH alignment. It is clear from the top panel, which depicts the entire C4-PPDK gene, that many sections are covered by contigs, but there are also many areas that appear to have high depth and yet there are no contigs covering them. The second panel shows a close-up of this gene. In this panel, it is even more obvious that there is high read depth in these sections, but no contigs covering those areas. This may, in part, be due to missing contigs that did not align to *S. bicolor*. Or perhaps, different parameters or a different assembler should be tried to achieve better results.

**1.4 Discussion**

Exome capture has been successful in several plant genomes, but no sequence capture has been previously attempted in *Miscanthus* (Fu et al. 2010; Haun et al. 2011; Saintenac et al.
2011; Bundock et al. 2012; Winfield et al. 2012; Zhou and Holliday 2012; Mascher et al. 2013; Neves et al. 2013). Our goal was to successfully perform sequence capture and develop a robust method to enhance variant calling and haplotype detection in complex polyploids. The most apparent result of this study is that our sequence capture design was successful. It resulted in a high number of reads (Table 1.2) and alignments to *S. bicolor* (Table 1.3), and high read depth in comparison to non-captured libraries (Figure 1.5).

We designed our own sequence capture probe set from RNA-seq assemblies, genomic assemblies, and fosmids. It is not possible to determine exactly how many probes were successful at capturing their intended targets, but at least 24.1% of these probes exhibited a read depth of at least five, and were found to share perfect sequence identity with corresponding sequences from the doubled haploid aligned to *S. bicolor*. The proportion of perfectly matching sequences increased only slightly to 24.9% when captured read depth was relaxed to one. These numbers seem low, but many of the probes were originally designed from *Mxg*, not *M. sinensis*, which makes them less likely to have exact matches, because *Mxg* can theoretically have up to six allele types and the DH can only have up to two homeologs. Additionally, since probes were matched to a DH alignment to *Sorghum*, sequences that were represented in the *Miscanthus* genome but not the *Sorghum* genome were not able to be used in this calculation, so there were less places for probes to have exact matches to. When given these circumstances, a number of 24.1% is quite substantial, and again it is only a percentage that we can say, with little doubt, did successfully work. Many more could have and are likely to have worked considering how much of the genome was covered by our sequencing reads.
To date, no other plant sequence capture study has utilized a doubled haploid to sort out paralogous variation. The doubled haploid *Miscanthus* plant was used as the baseline since it lacks allelic variation, and its parent was used to determine whether we could deconvolve the homeologs and alleles. Genomic evidence that this is a true doubled haploid and parent has been shown before (Swaminathan et al. 2012) and can be seen from the exome capture data set in Figure B.4. The sequencing and alignment results of this study demonstrate that it is possible to sort out homeologous variation and at a much higher depth than a non-captured library (Figure 1.6). Table 1.4, makes it clear that many more SNVs can be called with a captured library in our region of interest. In fact, at least 3% of total base pairs returned from the DH and PDH captured alignments were called SNVs, while in the non-captured alignments, only 1% of the base pairs were determined to be variants. Even though the non-captured libraries cover more base pairs in *Miscanthus*, they do not have as many predicted SNVs as the captured libraries, most likely because the non-captured libraries do not have the necessary read or variant depth required by the VarScan parameters for many SNVs. A doubled haploid sequence capture could be a beneficial application in other large genomes such as sugarcane, cotton or wheat. All of these genomes have likely undergone large duplications, which necessitate the ability to distinguish paralogous variants.

Figure 1.2, which illustrates the odd small peak found in the virtual fragment size distribution of DH, may be more than just noise. As mentioned in the Results, a portion of the paired-end read alignments mapped to telomeric repeats (‘AAATGGG’). Interestingly, the majority of these mapped to chromosomes 4 and 7, the two chromosomes thought to be associated with the chromosomal fusion in *Miscanthus* (Swaminathan et al. 2012). It is
possible that because of this fusion, there are telomeric repeats at the middle of this fused *Miscanthus* chromosome. The non-genic probes that were inadvertently designed may have been close by these interstitial repeats, allowing for their capture. Interstitial repeats resulting from a chromosomal fusion has been found in many plants and animals (Fuchs et al. 1995; Meyne et al. 1990), so this would not be an unexpected scenario.

Now that homeologous variation is known in the *Miscanthus* genespace, it would be worthwhile to compare these results to our transcriptome reads (Barling et al. 2013) in order to sort out alleles from homeologs. This knowledge would also allow for determination of homeolog specific expression. In some cases a homeolog may be expressed more often than the other, demonstrating that they did not retain the exact same function. Paralog-specific expression has already been investigated in the fully sequenced *Andropogoneae* grasses, *Sorghum* and maize (Emrich et al. 2007; Mizuno et al. 2012; Sekhon et al. 2011), but this could also be possible in more complex genomes by implementing sequence capture on a doubled haploid.

Recent whole genome duplication has been previously evidenced in *Miscanthus* (Swaminathan et al. 2012). Figure 1.8 corroborates that this duplication was recent, by providing support that there have been no large deletions of variants in any of the chromosomes. Chromosome 4 was specifically chosen for Figure 1.8, because it is the suspected site of a chromosome fusion in *Miscanthus*, and thus more likely to contain deletions because this activity. Genomes that have underwent recent WGDs, like the Brassicas (Town et al. 2006), are more complicated than older WGDs because it is difficult to differentiate between fixed alleles and paralogs, given that there is less disparity
between them. Sequence capture coupled with a doubled haploid genome proves to be one method of getting around this complication.
1.5 Figures

Figure 1.1  **Probe distribution in *Sorghum bicolor* annotations.** Describes where the designed probes lie in reference to *S. bicolor* annotations. This chart only consists of the 1,226,946 probes that successfully mapped to the *S. bicolor* reference sequence. Gene border includes probes that lie partially in a gene and partially outside of a gene. Exon/intron border includes probes that lie partially in an exon and partially in an intron. “Other” probes do not lie in any known *S. bicolor* annotation.
**Figure 1.2** Paired-end fragment size distribution of captured doubled haploid library. This distribution is expected to be unimodal, but both this distribution and the captured PDH library’s distribution were bimodal. The non-captured libraries did not demonstrate this bimodal distribution. It is uncertain why this smaller second peak is present, but it is assumed to be an artifact and has been removed from this study since we selected for 400-500 bp fragment size originally.
Figure 1.3 Example of extra coverage allowed by large fragment sizes. Using IGV software, displayed the alignment of probes and captured DH reads to the C4-PPDK gene-model in *S. bicolor* chromosome 9. Each long rectangle represents a read, and the arrows at the ends of them indicate forward and reverse reads. In the top panel, with colored reads, is the probe alignment. In the second panel is the DH alignment, and in the third panel is the C4-PPDK gene-model. In between the ends of the probes is a space of nearly 250bp that is not directly selected for. However, large fragment sizes (300-700bp) in the DH library covered this 250bp region with high depth.

**Aligned to C4-PPDK region of *Sorghum* Chromosome 9**

~250 bp

**Probe Alignment**

**DH Alignment**
Figure 1.4 Example of read depth comparison between captured and non-captured DH libraries. In the top panel is an illustration of the C4-PPDK gene-model. In the second panel is the alignment of probes to this gene. The third and fourth panels depict non-captured and captured DH library read depths for each base pair on an equivalent scale of 200. It is clear that the captured library has much higher depth across this gene than the non-captured library does.
Figure 1.5 Window depth for captured and non-captured DH libraries. “Window” is defined as a sequence of bases that has coverage of at least five per base. On the x-axis is the average depth across each window, binned. Along the y-axis is the number of windows normalized to the number of alignments returned for each library that fit into each bin. The encircled bars demonstrate that the majority of non-captured windows have an average depth of 5-15, while the majority of captured windows lie in bins of higher average depths.
Figure 1.6  Example of homeolog depth comparison between non-captured and captured libraries. This illustrates alignments of non-captured DH (top panel) and captured DH (second panel) to C4-PPDK region of S. bicolor chromosome 9. In each alignment, two homeologs ('I' and 'II') are represented. The non-captured library has a depth of four to six for either homeolog, while the captured library has a higher depth of more than ten for both. Pink indicates forward reads, and purple indicates reverse reads. The bottom panel is an illustration of the C4-PPDK gene-model.
Figure 1.7  SNV distribution and depth. Panel A illustrates the counts of SNVs found in *S. bicolor* annotations for each library. “Other” includes SNVs that do not lie in any known *S. bicolor* annotations. Panel B, for comparison, depicts the average SNV depth in each *S. bicolor* annotation. Trends in average depth tend to follow trends seen in panel A, with a couple of expected deviations.
Figure 1.8 Comparison of fixed and "50%" aligned to S. bicolor chromosome 4. The bottom graph shows the number of fixed and "50%" SNVs found in each S. bicolor gene-model for chromosome 4. It also shows, in green, the average depth across these gene-models. The top graph shows a close-up of the bottom graph. This represents what is typically found in regions with either no "50%" and/or fixed SNVs, and that is insignificant average depth.
Figure 1.9 Comparison of k-mer 61 contig and captured DH alignments. The top figure shows the alignment of the DH assembly and the captured DH library to the entire C4-PPDK gene-model of *S. bicolor* chromosome 9. The bottom figure shows a close up of this to better illustrate that there are regions with high read depth in the captured DH library, but no contigs covering it.
### 1.6 Tables

#### Table 1.1  **Probe design summary.** Probe type describes the sequences that the probes were designed from. For each probe type, the count and coverage in base pairs of the *Miscanthus* genome are given.

<table>
<thead>
<tr>
<th>Probe Type</th>
<th>Count</th>
<th>bp in <em>Miscanthus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Miscanthus</em> transcripts, mapped to sorghum</td>
<td>1,053,525</td>
<td>33.8 Mbp</td>
</tr>
<tr>
<td><em>Miscanthus</em> transcripts, not mapped to sorghum</td>
<td>804,250</td>
<td>10.0 Mbp</td>
</tr>
<tr>
<td>Mxg genomic contigs</td>
<td>175,845</td>
<td>10.7 Mbp</td>
</tr>
<tr>
<td>Fosmids (non repetitive regions)</td>
<td>2,538</td>
<td>0.16 Mbp</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>2,036,158</td>
<td>54.66 Mbp</td>
</tr>
</tbody>
</table>

#### Table 1.2  **Sequencing results.** Paired-end read counts before and after quality control (QC), and estimated coverage is given for each library.

*For non-captured libraries, coverage referred to entire estimated size of the *Miscanthus* genome, 2.5 Gbp. For the captured libraries, this referred to the estimated 54.66 Mbp that the probes are estimated to cover.*

<table>
<thead>
<tr>
<th>Library</th>
<th>Raw read count</th>
<th>Read count after QC</th>
<th>Coverage*</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH captured</td>
<td>73,167,138</td>
<td>71,687,998</td>
<td>290x</td>
</tr>
<tr>
<td>DH non-captured</td>
<td>90,665,478</td>
<td>89,037,293</td>
<td>5x</td>
</tr>
<tr>
<td>PDH captured</td>
<td>79,438,050</td>
<td>77,837,750</td>
<td>307x</td>
</tr>
<tr>
<td>PDH non-captured</td>
<td>95,885,882</td>
<td>93,884,347</td>
<td>6x</td>
</tr>
</tbody>
</table>
**Table 1.3 Alignment to Sorghum bicolor results.** For each library, the percentage of single reads that aligned to *S. bicolor* are listed, as well as the total number of possible alignments before and after quality control (QC). Quality control, in this case, includes removing all paired-end read alignments with a fragment size below 300 bp. This was done to remove the unexpected peak (Figure 1.2) that was found in the paired-end fragment distribution.

<table>
<thead>
<tr>
<th>Library</th>
<th>% of single reads aligned</th>
<th>Total alignments</th>
<th>Alignments after QC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH captured</td>
<td>69.09%</td>
<td>155,419,704</td>
<td>144,868,561</td>
</tr>
<tr>
<td>DH non-captured</td>
<td>36.36%</td>
<td>272,094,184</td>
<td>268,200,806</td>
</tr>
<tr>
<td>PDH captured</td>
<td>68.44%</td>
<td>162,417,157</td>
<td>154,002,335</td>
</tr>
<tr>
<td>PDH non-captured</td>
<td>35.15%</td>
<td>331,791,556</td>
<td>329,108,579</td>
</tr>
</tbody>
</table>

**Table 1.4 Variant calling results.** The number of base pairs that are aligned to *S. bicolor* and the number of those that are predicted SNVs are listed for each library. Even though the non-captured libraries cover more base pairs, less SNVs are predicted than in the captured libraries.

<table>
<thead>
<tr>
<th>Library</th>
<th># of bp aligned</th>
<th>SNVs predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH captured</td>
<td>110,543,003</td>
<td>3,649,999</td>
</tr>
<tr>
<td>DH non-captured</td>
<td>148,830,194</td>
<td>1,872,490</td>
</tr>
<tr>
<td>PDH captured</td>
<td>114,369,132</td>
<td>3,933,989</td>
</tr>
<tr>
<td>PDH non-captured</td>
<td>154,625,448</td>
<td>1,619,601</td>
</tr>
</tbody>
</table>
Table 1.5 Assembly results. Lists contig and scaffold N50 scores for captured DH assemblies made with ABySS for different k-mer lengths. Lines highlighted in gray specify the two k-mer lengths that produced nearly equal N50 scores.

<table>
<thead>
<tr>
<th>k-mer length</th>
<th>Contig N50</th>
<th>Scaffold N50</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>857</td>
<td>1058</td>
</tr>
<tr>
<td>57</td>
<td>863</td>
<td>1065</td>
</tr>
<tr>
<td>59</td>
<td>869</td>
<td>1080</td>
</tr>
<tr>
<td>61 (gray)</td>
<td>873</td>
<td>1083</td>
</tr>
<tr>
<td>63</td>
<td>874</td>
<td>1082</td>
</tr>
<tr>
<td>65</td>
<td>873</td>
<td>1073</td>
</tr>
</tbody>
</table>
CHAPTER 2: ENERGY CANE, MISCANTHUS X GIGANTEUS, AND MISCANTHUS SACCHARIFLORUS SEQUENCE CAPTURES

2.1 Introduction

Energy Cane and Sugarcane

Sugarcane, of the genus *Saccharum*, is produced primarily for sugar production and more recently for ethanol production, especially in Brazil. Modern sugarcane cultivars are hybrids of *Saccharum officinarum* (N=10, 80 chromosomes) and *Saccharum spontaneum* (N=10, 64 chromosomes), where the *S. officinarum* contributes sugar while the *S. spontaneum* provides resistance to biotic and abiotic stresses and brings in some diversity. In the F1 generation, these often result in what are called energy canes because of their high lignocellulose to sugar ratios. The highest sugar progeny of this F1 generation are backcrossed to commercial high-sugar sugarcanes for multiple generations to achieve high sugar content. It is thus, important to note that energy cane is not a unique species, but is a modern sugarcane cultivar that is selected for high fiber and low sugar content. Energy cane is a potential ligno-cellulosic bioenergy grass like *Miscanthus*, however it and sugarcane can only be planted in limited warm areas of the U.S such as Louisiana and Hawaii. Since sugarcane and energy cane are so vital for food and biofuel in many parts of the world, they are a worthwhile plant to study, especially to help fulfill plant-breeding goals.

There is little molecular biology and genomics used in sugarcane breeding largely because the genomes are highly complex and it is difficult to get a handle on polyploidy and rampant copy number variation seen in sugarcane populations. Energy cane and
sugarcane cultivars have large polyploid genomes with frequent aneuploidy. It is unknown exactly how many subgenomes sugarcane or energy cane have, but it is estimated that *S. officinarum* has 80 chromosomes and *S. spontaneum* has anywhere from 48 to 128 (Bremer 1961). So modern cultivars, which are interspecific hybrids of these two *Saccharum* species could potentially range in chromosome numbers of 64 to 104, or even higher if one of these hybrids were crossed with *S. spontaneum*, and then backcrossed to a commercial hybrid. These are clearly very complex genomes, and the ability to separate and quantify the allelic copies within this genome would be extremely difficult.

*Andropogoneae* Comparative Genomics

As mentioned in chapter 1, the *Andropogoneae* tribe consists of several grasses, including closely-related sorghum, *Miscanthus*, and *Saccharum* grasses (Hodkinson, Chase, Lledo et al. 2002). Like *Miscanthus*, sugarcane demonstrates synteny to *Sorghum* (Grivet et al. 1994; Dufour et al. 1997; Ming et al. 1998; Wang et al. 2010). It is thought that *Miscanthus* and sugarcane may be even more closely related to each other (Hodkinson, Chase, Lledo et al. 2002). Sorghum is a diploid genome, *Miscanthus* can have diploid to tetraploid genomes, although even the diploid has a WGD compared to Sorghum (Swaminathan et al. 2012). The *Saccharum* spp. are octaploids. The relatedness of these 3 species in the *Andropogoneae* tribe creates the potential for these to be genetic stepping-stones for each other. It is possible to use the well-known and relatively simple *Sorghum* genome to shed light on the *Miscanthus* genome, and then, once fully sequenced, use the *Miscanthus* genome to shed light on the even more complex sugarcane and energy cane genomes.
Research Justification

The following study performs sequence capture on energy cane, *Miscanthus x giganteus*, and *Miscanthus sacchariflorus var. Golf Course*. The purpose of this is to observe the utility of the sequence capture design described in chapter 1, as a variant calling and comparative genomics tool. By capturing these 2 *Miscanthus* species, it is possible to compare unique alleles among these and the 2 plants captured in chapter 1. Capturing energy cane will demonstrate whether *Miscanthus*-based technology can be effectively used in *Saccharum* genomes. There is the possibility that *Miscanthus*, once fully sequenced, could be used as a reference to sugarcane. Since sugarcane and energy cane are so highly polyploid, there is a necessity to have an intermediately polyploid genome like *Miscanthus* to help unravel its variation.

2.2 Materials and Methods

Collection of Materials and Probe Design

Brandon James gathered energy cane (Ho02 113), *Miscanthus x giganteus*, and *Miscanthus sacchariflorus var. Golf Course* from the greenhouses at the University of Illinois at Urbana-Champaign (UIUC), and from UIUC’s SoyFACE facility. James performed CTAB DNA extraction on these samples in the same way described in the previous chapter. The same *Miscanthus*-derived probe set described in the previous chapter was used to capture Illumina libraries from these three plants.
Sequence Capture

Genomic DNA library preparation

A portion of the extracted DNA from energy cane (EC), Miscanthus x giganteus (Mxg) and Miscanthus sacchariflorus var. Golf Course (MS) was not diluted down to a specific concentration this time. We had performed experiments with varying concentrations and times on the Biorupter (Model # B01010002) before starting exome capture, and the results showed that concentration did not affect sonication in a significant way. The DNA was diluted down enough to make 600μl. In the 600μl of Mxg there was 55.5μg of DNA. In the 600μl of EC there was 18.5μg of DNA, and in the 600μl of GC there was 17.5μg of DNA. Six 100μl aliquots were made into Diagenode non-stick RNase free 1.5ml tubes. These tubes were placed in the Biorupter and sheared at low setting for 20 minutes (30 seconds on, 30 seconds off). The Biorupter was set to 20 minutes instead of 10 minutes, because Dr. Alvaro Hernández, the director of UIUC's DNA Services, suggested that we would get more consistent results for larger genomes like these for the fragment size we wanted (400-500bp). In addition, our experimental Biorupter testing results showed 20 minutes still returned high amounts of our target fragment size.

The sheared DNA was diluted to 20ng/μl, and DNA was ran in a 2% E-gel to size select for 400-500bp fragments. DNA was extracted from gel by using a MinElute Gel Extraction Kit. A High Sensitivity (HS) Qubit assay was run on the resulting DNA. Energy cane had approximately 2.5μg of DNA, the MS had about 0.8μg, and Mxg had 3.3μg of DNA. MS was lower, because there was less starting material, and more was not easily obtainable at the time.
For the following Illumina protocol, 1μg of DNA was used per library. Genomic DNA was prepared by following Illumina’s low-throughput protocol in their TruSeq DNA Sample Preparation Guide. Diagenode non-stick RNase free 1.5 ml tubes were used instead of 96-well PCR plates, since there were only three samples. When needed, DNA concentrations were measured using Qubit BR and HS assays and quantified DNA on Agilent 2100 Bioanalyzer using the Agilent DNA 7500 Kit.

Sequence capture protocol

Sequence captured libraries were constructed by using the NimbleGen SeqCap EZ hybridization and wash kit (#05 634 261 001) and by following Nimblegen’s protocol (NimbleGen SeqCap EZ Choice Library SR User’s Guide v1.0, December 2010). This protocol’s LM-PCR and hybridization steps were adapted for use of Illumina TruSeq DNA libraries by following the guide provided by Nimblegen (Technical Note Supplement: Targeted Sequencing with NimbleGen SeqCap EZ Libraries and Illumina TruSeq DNA Sample Preparation Kits, Instructions for using Illumina TruSeq DNA Libraries with SeqCap EZ Libraries). The pre-capture PCR step was skipped in order to avoid PCR bias, but found it necessary to complete the post-capture PCR step, due to low amounts of DNA. Non-stick RNase free 1.5 ml tubes were used instead of 96-well PCR plates. Five µl of plant capture enhancer was used instead of 5 µg Cot-1 DNA. After cleanup of the final PCR step, the bioanalyzer showed a significant peak around 125bp. As mentioned in the last chapter, these appeared to be primer dimers, so we used an additional clean-up step by running the libraries through a 2% E-gel and bands from 300bp and up were cut out. The three
libraries, each made with a different index, were then pooled together since their DNA content was so low without the pre-capture PCR step.

**Sequencing**

The pooled library was quantified and underwent quality control at UIUC's Roy J. Carver Biotechnology Center. This library was then sequenced on one lane on an Illumina HiSeq 2000 system at the same center. The pooled library was sequenced as 100bp paired-end reads, and the resultant sequence data was de-multiplexed back into three libraries.

Sequencing coverage was determined by using the Lander/Waterman equation for computing coverage: $C = \frac{LN}{G}$, as described in the chapter 1 methods section. For all libraries, $G = 54.66\text{Mbp}$, which is the estimated length that the probes covered in the Miscanthus genome. L was set to 200 and N was set to the specific library’s raw read count.

**Bioinformatics Analyses**

**Quality control**

For each pair of FASTQ files, the reads were trimmed to a minimum quality score of 25 and the 3-prime adapter sequences were removed simultaneously using a custom Python script. 5-prime adapters were masked using a NovoalignMPI parameter in the alignment step.

**Alignment**

Reads were aligned using NovoalignMPI (v. 3.00.02) on the Energy Biosciences Institute’s computer cluster. The Sorghum bicolor genome sequence (release number 79) was used as
a reference genome. Parameters used were the same as those used for alignments of the DH and parent of DH to *S. bicolor*.

EC, *Mgx*, and MS reads were also aligned to a doubled haploid draft assembly using Novoalign (v. 3.00.05). This assembly was created by Therese Mitros and is currently unpublished. Mitros used meraculous software (Chapman et al. 2011) to assemble doubled haploid 100bp paired-end reads with varying selected insert sizes (300-800bp). All libraries’ reads were aligned to this draft assembly by using the same parameters used for their alignments to *S. bicolor*.

**Variant calling**

Before using variant calling software, SAMtools (Li et al. 2009) software was implemented to convert SAM files to BAM format, to sort the BAM files, to remove duplicates and finally to convert BAM files to pileup format. Duplicate reads were removed this time, because visualization of the alignments showed evidence of clonal reads. This is likely because skipping the pre-capture amplification resulted in low yields from the capture and then this small amount of DNA was multiplied by the second amplification step. The variant caller, VarScan (Koboldt et al. 2009) was employed to call variants from the pileup files, using same parameters as in the last chapter.

**Additional analyses**

Analyses performed on the output of the alignment and variant files were done using custom Perl scripts to mine the data and Microsoft Excel to create charts and graphs of the
data. Visualization of the alignments was done by running IGV software (Robinson et al. 2011).

2.3 Results

Performance of Sequence Capture

We performed the same sequence capture protocol as described in chapter 1 on Miscanthus x giganteus (Mxg), energy cane (EC), and Miscanthus sacchariflorus var. Golf Course (MS). Since these genomes have higher ploidy than the DH and its parent, we attempted to skip both amplification steps to avoid PCR bias. After omitting the first amplification step, there was very little DNA, so it was necessary to perform the second amplification step. This resulted in many clonal reads, which we removed using SAMtools.

The raw read counts returned from sequencing can be found on Table 2.1, as well as the read counts after quality control (QC) was performed. Quality control included removing 3-prime adapter sequences and trimming to a minimum PHRED quality score of 25. This table also depicts the coverage of each of these libraries based on the length of the genome that the probes are estimated to cover (54.66Mbp). The MS library had more than twice as many reads and resultantly, twice the coverage as the Mxg and the EC libraries. Quality control did not result in a large loss of reads (less than 1% loss). Overall, the number of sequences returned for these libraries were much lower than those returned for the DH and PDH libraries. This is due to omission of the first amplification step and resultantly, fewer bioinformatics analyses were performed since these libraries lacked the depth required for robust variant calling, especially when considering their high ploidy.
These reads were aligned to the *Sorghum bicolor* assembly (release 79) and all possible alignments were returned. Approximately 72% to 81% of these libraries’ single reads aligned to the *S. bicolor* reference genome, which is higher than the captured libraries achieved in Chapter 1. Again, single reads are not single-end reads, but paired-end reads being defined as two reads rather than one. Table 2.2 holds the exact percentages of single reads aligned and the total number of alignments that were returned. For the captured libraries from Chapter 1 there was an oddity found in the paired-end fragment distribution when aligning the two captured libraries to *S. bicolor*. This abnormal small peak was not seen in any of these three alignments, so there was no need to remove any alignments.

Figures 2.1, 2.2, and 2.3 depict alignments of *Mxg, EC* and *MS* to the genic regions of *S. bicolor* chromosome 7. In each figure, three alleles with at least a read depth of four are pointed out with the white pointed blocks and dotted lines. Note that black lines represent deletions, and purple right brackets represent insertions. These figures demonstrate that it is possible to sort out at least 3 alleles at required depth. If that extra amplification step had been performed, perhaps more of these instances would have been found and with greater confidence.

**Capture Window Depth**

In Figure 2.4, “window” depth is depicted in the same way as it was in Figure 1.5, where a “window” is defined as being a series of continuous bases of at least 75bp from an alignment with base read depths of at least five. In Figure 2.4, the window depths for each of the three libraries are compared. On the x-axis are binned read depths, and on the y-axis are the window counts, which are normalized to their number of alignments. Note that the
y-axis is log base 10 scaled, but the data has not been transformed. This graph clearly illustrates that read depth is low for all of these libraries. The majority of windows have an average depth of 5-15, and window counts decrease rapidly along the x-axis at a more than exponential rate. This again demonstrates that a low number of raw reads results in low read depth.

Multiple Genome Comparison

The doubled haploid theoretically has two homeologs and no allelic variation and thus two possible variants per Sorghum locus. The parent of the doubled haploid, a Miscanthus sinensis plant, has allelic variation as well as the variants that define the homeologs; one of the two alleles for each homeolog is identical to the DH, and so for the parent there are four possible variants per Sorghum locus. Miscanthus x giganteus has three genomes, one of which is theoretically very similar to a M. sinensis genome, and the other two are similar to M. sacchariflorus. Each genome has two homeologs and therefore up to six possible variants is possible per Sorghum locus. The M. sacchariflorus line used here is a diploid and so, like the M. sinensis PDH, can have two alleles and two homeologs, for a possible four variants per Sorghum locus. Theoretically, there will be variants that distinguish the M. sacchariflorus species from M. sinensis and so both types of variants should be seen in Mxg. Since all of these genomes are inter-related, comparisons between their alignments were made to see if this relatedness could be observed. It should be noted that the PDH and DH plants are directly related, however the MS and PDH plants are not directly related nor necessarily parents to Mxg.
Figures 2.5, 2.6, and 2.7 depict alignments of the DH, PDH, Mxg, and MS libraries to the C4-PPDK region of *S. bicolor* chromosome 9. The black oval in Figure 2.5 points out a differing variant from the reference genome, ‘G’, that is unique to only the MS genome. This illustrates that the MS plant is genetically different from the other three libraries. In this same figure, there is a fixed variant indicated by the blue ‘C’s that is shared by all of the genomes. Figure 2.6 highlights a variant, ‘A’, that is unique to only the PDH genome (indicated by the black oval). In this same figure, a variant ‘G’, highlighted by the red ovals, is unique to only the *Mxg* and MS genomes. In the blue oval, a ‘C’ variant is found to be unique to only the *Mxg* genome. If the PDH and MS plants were direct parents of this *Mxg*, then this ‘C’ variant would be extremely unlikely to occur. It is also possible that this ‘C’ variant would be seen in MS if this were a tetraploid rather than a diploid. Figure 2.7 portrays a variant ‘G’ that is only found in the DH and PDH genomes. All of these variations follow the trends that one would expect to find among these four genomes, given their relatedness.

2.4 Discussion

One of the most obvious results from this study is that the Roche-Nimblegen sequence capture protocol does not provide satisfying read depth when even one amplification step is avoided. This is corroborated by the low read counts seen in Table 2.1. It is desirable to eliminate amplification, especially in large and polyploid genomes, because of PCR bias. PCR bias results in chimeric and multi-clonal reads (Kanagawa 2003), which can further
convolute already complex genomes. A sequence capture technology that excluded amplification altogether would be of great benefit to plant genome studies.

Table 2.2 indicates that the percentages of single reads that aligned to *S. bicolor* are appreciably higher for *Mxg*, EC, and MS than they are for DH and PDH. This is expected in the *Mxg* and MS alignments because the probes are mainly designed from *Mxg* sequences. *Mxg* contains two subgenomes donated from a *M. sacchariflorus* plant and only one subgenome donated from a *M. sinensis* plant. Therefore, the probes are likely to be more similar to *Mxg* and MS than they are to DH or PDH, and thus may pull out more targeted DNA in the former libraries. The fact that EC aligns even better than *Mxg* and MS, is a bit more unexpected, but not unreasonable. One explanation is that EC contains less repeats and its genic regions are highly identical to *Sorghum* (Wang et al. 2010), so the hybridization kinetics during sequence capture may favor more efficient recovery of *Saccharum* sequences shared with exonic sequences from *Sorghum* compared to *Miscanthus*.

This study demonstrates the utility of the lower ploidy *Miscanthus* genome to the high ploidy *Saccharum spp.* genomes. The sequence capture probe set that was designed to capture *Miscanthus* sequences was able to capture similar sequences in energy cane. A similar sequence capture study was done in barley; it was not identical to this study, but they did use a Nimblegen platform, designed their probes from RNA-seq and other transcript data, and they had just under 2 million overlapping probes (Mascher et al. 2013). Barley is a diploid grass that is closely related to wheat, a more complex grass ranging from 6x to 8x ploidy, much like *Saccharum spp.* (Catalán et al. 1997). In the barley study, they attempted to show the utility of a simpler barley genome to the more complex wheat
genome and so they performed exome capture on wheat using probes designed from barley sequences. Nearly 50% of the sequenced wheat reads could be mapped onto the barley assembly. In our study, we mapped our energy cane reads to a preliminary Miscanthus assembly (not shown). We found that 90.20% of the energy cane single reads aligned to this doubled haploid assembly. This suggests that Miscanthus is even more closely related to energy cane, than barley and wheat are to each other. Furthermore, it suggests that Miscanthus can serve as a genomic stepping-stone and reference for Saccharum spp.

In addition to EC, Mxg and MS were also aligned to the Miscanthus doubled haploid assembly. The alignments resulted in 92.92% and 94.33% of single reads aligning to the assembly for Mxg and MS, respectively. These are higher percentages than the EC alignment to the assembly, which is expected considering that EC is a different genus. All three of these libraries aligned considerably better to the Miscanthus assembly than to Sorghum. This indicates that Sorghum is not the most adequate reference genome for either Miscanthus spp. or Saccharum spp.

Doing comparative genomics like the kind described in Figures 2.5, 2.6, and 2.7 provides certainty of the nature of each genome and can reveal potentially important variations. For example, if this kind of comparative genomics were done with four directly related plants, it would be possible to not only sort out the homeologs using a doubled haploid plant, but it would be possible to discover which genome each allelic variant came from. A couple exome capture studies have already attempted to compare related genomes by using to their reads to compare different types of variants and to make a phylogenetic tree (Winfield et al. 2012; Mascher et al. 2013). It would be enlightening to perform
comparative genomics on an exome capture study on species belonging to the 
*Andropogoneae* tribe.
2.5 Figures

Figure 2.1 Alignment of *M. x giganteus* to *S. bicolor*. An alignment of *Mxg* to a genic region in *S. bicolor* chromosome 7. As indicated by ‘I’, ‘II’, and ‘III’, there are three alleles with significant amount of depth, proving that this sequence capture can separate out more than two alleles in this genome, and possibly more with higher depth.
Figure 2.2 Alignment of energy cane to *S. bicolor*. An alignment of EC to a genic region in *S. bicolor* chromosome 7. As indicated by ‘I’, ‘II’, and ‘III’, there are three alleles with significant amount of depth, proving that this sequence capture can separate out more than two alleles in this genome, and possibly more with higher depth.
Figure 2.3  Alignment of *M. sacchariflorus* to *S. bicolor*. An alignment of MS to a genic region in *S. bicolor* chromosome 7. As indicated by 'I','II', and 'III', there are three alleles with significant amount of depth, proving that this sequence capture can separate out more than two alleles in this genome, and possibly more with higher depth.
Figure 2.4 Window depth for non-DH libraries. “Window” is defined as a sequence of bases that has coverage of at least five per base. On the x-axis is the average depth across each window, binned. Along the y-axis is the number of windows normalized to the number of alignments returned for each library that fit into each bin. Note that the y-axis is log$_{10}$ scaled but the data itself not log$_{10}$ scaled.
Figure 2.5  Example of unique allele in *M. sacchariflorus* library. This figure shows four alignments of DH, PDH, Mxg, and MS libraries to C4-PPDK of *S. bicolor* chromosome 9. In the fourth panel (MS alignment) a unique allele that has a ‘G’ instead of the reference ‘A’ base pair is circled. This variant is not seen in any alignment other than MS.
Figure 2.6  Examples of unique alleles among 4 libraries. This figure shows four alignments of DH, PDH, Mxg, and MS libraries to C4-PPDK of *S. bicolor* chromosome 9. In the second panel, a unique allele that has a “T” instead of the reference ‘A’ base pair is circled in black. This variant is not seen in any alignment other than PDH, so it must have been in the subgenome that the DH inherited. In the third and fourth panels there is another unique allele represented by a ‘G’ variant instead of the reference ‘T’ base pair (circled in red). This allele is not seen in any alignments other than *Mxg* and MS. Finally, circled in blue, there is an allele unique to *Mxg* that is represented by a ‘C’ variant instead of the reference ‘T’ base pair. This allele is not seen in any alignments other than *Mxg*. 
Figure 2.7  Example of unique allele found in DH and PDH libraries. This figure shows four alignments of DH, PDH, Mxg, and MS libraries to C4-PPDK of *S. bicolor* chromosome 9. In the first and second panels, a unique allele that has a ‘G’ instead of the reference ‘A’ base pair is circled. This variant is not seen in any alignments other than DH and PDH.
2.6 Tables

Table 2.1 Sequencing results. Paired-end read counts before and after quality control (QC), and estimated coverage is given for each library. Coverage is based on the estimated 54.66 Mbp that the probes cover.

<table>
<thead>
<tr>
<th>Library</th>
<th>Raw read count</th>
<th>Read count after QC</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miscanthus x giganteus</td>
<td>6,034,105</td>
<td>6,001,167</td>
<td>22x</td>
</tr>
<tr>
<td>Energy Cane</td>
<td>6,656,154</td>
<td>6,626,538</td>
<td>24x</td>
</tr>
<tr>
<td>Miscanthus sacchariflorus</td>
<td>17,393,400</td>
<td>17,333,372</td>
<td>64x</td>
</tr>
</tbody>
</table>

Table 2.2 Alignment to S. bicolor results. For each library, the percentage of single reads that aligned to S. bicolor are listed, as well as the total number of possible alignments.

<table>
<thead>
<tr>
<th>Library</th>
<th>% of single reads aligned</th>
<th>Total alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miscanthus giganteus</td>
<td>72.02%</td>
<td>10,206,466</td>
</tr>
<tr>
<td>Energy Cane</td>
<td>80.99%</td>
<td>13,403,997</td>
</tr>
<tr>
<td>Miscanthus sacchariflorus</td>
<td>77.55%</td>
<td>29,767,767</td>
</tr>
</tbody>
</table>
REFERENCES


Brading MD. 2006. Miscanthus has the potential to become popular ethanol feedstock. Grainnet - News Inf. Grain, Milling, Feed Seed Ind. [Internet]. Available from: http://www.grainnet.com/info/articles_print.html?ID=35767


**APPENDIX A: ABBREVIATIONS**

BAM: Binary Alignment/Map

BLAT: BLAST-Like Alignment Tool

DH: Doubled Haploid

EBI: Energy Biosciences Institute

EC: Energy Cane

IGV: Integrative Genomics Viewer

MS: Miscanthus sinensis var. Golf Course

Mxg: Miscanthus x giganteus

PDH: Parent of Doubled Haploid

SAM: Sequence Alignment/Map

UIUC: University of Illinois at Urbana-Champaign

WGD: Whole Genome Duplication

WGS: Whole Genome Sequencing
APPENDIX B: OTHER FIGURES

Figure B.1 Basic sequence capture protocol. Illustrated is the basic Roche-Nimblegen sequence capture protocol. This protocol is carried out entirely in solution, no array chip. Genomic DNA is fragmented, and then adapters are attached. After this, the DNA is denatured to allow for hybridization to probes. Magnetic beads are added to the solution and the probes attach to them along with the targeted DNA. All other DNA is washed away, and the DNA attached to probes is eluted in TE buffer. DNA is then amplified and sequenced. Note, that there is also an amplification step immediately before hybridization that was not included in this figure.

![Diagram showing basic sequence capture protocol](image-url)
Figure B.2  Picture of doubled haploid *M. sinensis* plant.
Figure B.3  Probe design pipeline. This figure shows the pipeline that was used to design the sequence capture probes. An in depth description of this can be found in the 1.2 Methods.
Figure B.4 Comparison of PDH and DH alignments. This figure illustrates the DH (first panel) and PDH (second panel) alignments to the C4-PPDK gene-model in *S. bicolor* chromosome 9. The white boxes point to a variant. There are 2 types of variant reads or homeologs in the DH library and 3 types of alleles in the PDH library. Many examples like this can be found across the entire *S. bicolor* genome, providing evidence that this a true doubled haploid and its parent. These reads were colored according to their paired reads (forward = pink, reverse = purple).
Figure B.5  Comparison of the average number of 50% SNVs to fixed SNVs per chromosome. For each *S. bicolor* chromosome, the average number of 50% and fixed SNVs were found per gene-model. This analysis was done on the DH alignment.
Figure B.6 Preservation of both homeologs in DH assembly. This depicts the DH assembly and captured DH library alignments to the C4-PPDK gene-model of *S. bicolor* chromosome 9. ‘I’ and ‘II’ white boxes point to 2 homeologs found in the captured DH library as well as the DH assembly. This shows that the assembly did not collapse reads into one consensus read, but instead preserved both homeologs.