GENETIC DIVERSITY, POPULATION STRUCTURE AND ASSOCIATION MAPPING OF BIOFUEL TRAITS IN SOUTHERN SWITCHGRASS GERMPLASM

by

ANANTA RAJ ACHARYA

(Under the Direction of E. Charles Brummer and Katrien M. Devos)

ABSTRACT

Switchgrass (*Panicum virgatum* L.), a warm season grass native to North America, is being developed as a biofuel crop. Plant breeding can improve biofuel characteristics further, particularly if the genetic diversity of germplasm resources is clearly understood. The objective of this study was to examine the population structure and relatedness within and among forty-nine switchgrass populations mainly derived from the southern United States and use the information to identify putative QTLs associated with biomass yield, plant height, stem diameter and days to flower. These populations included both upland and lowland ecotypes. A total of 511 genotypes were selected for genotyping and phenotyping. SSR markers developed from switchgrass and well distributed across the switchgrass genome were used to genotype the individuals. We used 35 markers and 365 alleles were discovered for those markers. In addition, we used a Genotyping-by-Sequencing (GBS) protocol to identify and utilize SNPs as genetic markers. With GBS, we identified 65,328 SNP markers. We only used 3,196 SNPs for our analysis, after filtering for read depth of at least 6 reads per locus per genotype and

requiring no more than 20% missing genotypic data for any given locus. In order to investigate the effect of missing data, we also used a second dataset of about 20,000 SNPs allowing up to 50% of individuals to have missing genotypic data for any given locus. We also used nine chloroplast specific markers to identify the cytotype. The data were used to examine the population structure and to perform phylogenetic analysis. Along with measuring dry biomass after harvest, we collected three canonical morphological data; plant height, stem diameter and flowering time on the individuals. We found a population differentiation in the two major groups, upland and lowland ecotypes, with phenotypic, cytotypic and genotypic data. A deeper sub-population structure was identified within the broad lowland and upland population. The sub population structure was correlated with the geographical origins of those accessions. We were able to identify two groups within lowland ecotypes, one of which did not exhibit the typical morphological characteristics of lowland accessions. We also studied the association of markers with above mentioned traits, and within the limitations of the number of environments used, identified several QTL significantly associated with each trait.

INDEX WORDS: Switchgrass, SSR, Genotyping-by-Sequencing, GBS, Association Mapping, Biofuel

GENETIC DIVERSITY, POPULATION STRUCTURE AND ASSOCIATION MAPPING OF BIOFUEL TRAITS IN SOUTHERN SWITCHGRASS GERMPLASM

by

ANANTA RAJ ACHARYA

BS, Tribhuvan University, Nepal, 2005

MS, The University of Florida, 2009

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

GENETIC DIVERSITY, POPULATION STRUCTURE AND ASSOCIATION MAPPING OF BIOFUEL TRAITS IN SOUTHERN SWITCHGRASS GERMPLASM

by

ANANTA RAJ ACHARYA

Major Professor:	Katrien M. Devos
Committee:	E. Charles Brummer
	Joseph H. Bouton
	H. Roger Boerma
	Jeffrey Bennetzen

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
August 2014

DEDICATION

GRANDPARENTS Prajapati Acharya, Draupada Acharya

PARENTS Hari Prasad Acharya, Radha Devi Acharya

BROTHER Kul Mani Acharya

ACKNOWLEDGEMENTS

First and foremost I like to thank Dr. E. Charles Brummer for the opportunity he gave to work for him in an exciting area of research. I would not be here today without his constant help and discussions during the process. I would like to thank Dr. Katrien M. Devos for agreeing to be my major advisor even later in the process and fully supporting my transition. I also like to thank committee members Dr. Joseph Bouton, Dr. Roger Boerma and Dr. Jeff Bennetzen for their continuous push to make me a better student of science and encouraging me to do the best work possible.

I like to thank Dr. Malay Saha of the Noble Foundation for constructive discussions about the project. I cannot thank enough the field crew members Donald, Wesley, Jonathan for all the fieldwork and help in phenotyping. I also like to thank Yanling Wei for her support and teaching regarding laboratory procedures. I also like to thank Dr. Desalegn Serba and field crews at the Noble foundation for all the help. I also like to thank computational support staffs including Yinbing. I also like to thank my lab peers Xuehui, Mohammet, Rafael, Qingzhen for discussions and life beyond lab.

Last but not the least, I like to thank my lovely wife, Smita Sharma, for her constant encouragement when I am feeling discouraged, for her support when I am feeling worn out and the fun beyond the college life.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Switchgrass

Switchgrass is a native North American C4 perennial grass. Historically it has been used as a component of conservation reserve program (CRP) lands and as pasture and hay production (Bouton, 2007). However, the current research focus in switchgrass has been as a major herbaceous feedstock for production of cellulosic biofuel. The Bioenergy Feedstock Development Program (BFDP) in the U.S. Department of Energy began evaluating a wide variety of potential feedstocks in 1978 (McLaughlin and Adams Kszos, 2005). Among herbaceous feedstocks, switchgrass (*Panicum virgatum* L.), was identified as the most promising target to develop as a bioenergy crop. Switchgrass has many characteristics of a desirable biofuel feedstock because it is perennial, has high productivity (Wright, 1994) is adapted to a wide variety of sites, including to poor soil conditions (Sanderson *et al.*, 1996).

Switchgrass is an erect, bunch-type grass with numerous tillers and can grow up to 4.0 m tall (Bouton, 2007). Based on morphological characteristics, switchgrass is mainly divided into two ecotypes, lowland and upland. Lowland plants have taller tillers, thicker stems and later flowering and senescence than upland plants. Lowland switchgrass is predominantly tetraploid (2n = 4× = 36); however, uplands are both tetraploid and octoploid (Costich *et al.*, 2010; Narasimhamoorthy *et al.*, 2008; Zhang *et al.*, 2011a; Zhang *et al.*, 2011b) and aneuploidy (Costich *et al.*, 2010), especially in

octoploids, has been identified. Upland and lowland genotypes can be distinguished based on morphological characteristics, or by DNA-based markers from the nuclear or chloroplast genomes (Missaoui *et al.*, 2005; Narasimhamoorthy *et al.*, 2008; Zhang *et al.*, 2011a; Zhang *et al.*, 2011b). These uplands and lowlands can further be classified into subgroups of genetic pools that can be differentiated with genetic markers. Those include the Gulf Coast and Great Plains lowlands, eastern and western upland and octoploid uplands (Zhang *et al.*, 2011a; Zhang *et al.*, 2011b). The tetraploid switchgrasses have interbred giving rise to some mixed populations. Switchgrass migrated north following the last glaciation of North America, and the upland tetraploids possibly arose from upland octoploids, which emerged after a ploidy level shift from southern lowlands (Lu *et al.*, 2013).

Switchgrass is allogamous, and consequently, more genetic variation is generally observed within than among accessions. Within accession variation ranges from 65-80% of the total observed variation based on molecular marker diversity (Casler, 2012). Significant phenotypic variation also exists in terms of key yield and compositional traits (Lemus *et al.*, 2002).

The main breeding objectives for switchgrass improvement are biomass yield and altered composition to minimize recalcitrance to digestion. Switchgrass is affected by some diseases, including rust and smut, and resistance to these and other pathogens is also a breeding goal. Improving biomass per se is a main breeding goal that could be facilitated by modifying yield components, such as plant height, stem diameter, and tiller number. However, perhaps the easiest way to improve total biomass yield under single

harvest biofuel management is to extend the vegetative growth period – for northern US regions, this can be accomplished by adapting lowland germplasm to survive winter.

Recalcitrance to digestion, either by ruminant livestock or by industrial enzymes, is significantly affected by lignin content, so minimization of lignin is important for digestion-based used. Ultimately, for liquid fuel or animal performance, the real goal is to increase the amount of sugar released from the biomass, so in addition to lignin, total sugars are also of relevance to breeders. Rust (*Puccinia emaculata*) can affect biomass yield in switchgrass (Zale *et al.*, 2008). We have also observed blast (*Magnaporthe grisea*) that appeared to affect plant growth in Watkinsville, GA, although we do not have quantitative data supporting this observation (unpublished).

Genetic markers

Molecular markers can be used in modern breeding programs to assess population diversity and relatedness of germplasm or to apply marker assisted selection. Older types of markers, including Restriction Fragment Length Polymorphism (Botstein *et al.*, 1980), Random Amplified Polymorphic DNA (Williams *et al.*, 1990), Amplified Fragment Length Polymorphism (Vos *et al.*, 1995), and simple sequence repeats (SSR) (Tautz and Renz, 1984) enabled the development of molecular breeding programs but all suffered from various shortcomings that limited their utility for many applications on a routine basis.

Today, most genotyping is based on Single Nucleotide Polymorphisms (SNP). These polymorphisms represent base changes at a specific nucleotide position and are highly abundant in genome. SNP can be identified easily by sequencing a given genomic

region in multiple genotypes (or even in a single heterozygous genotype), and next generation sequencing technologies greatly facilitate comparative sequencing and SNP discovery. Once identified, SNP can be developed into assays for analysis. Numerous genotyping techniques are available to assay SNPs, mostly based on either allele specific hybridization or primer extension (Kim and Misra, 2007; Sobrino *et al.*, 2005), and depending on the technology, from one to hundreds of thousands of SNP loci can be assayed simultaneously. Basically SNP assay technologies can be grouped in six groups based on assay technology; a) physical property based (melting) b) single nucleotide extension based c) 5' exonuclease activity based d) ligation based e) hybridization based and f) sequencing based. Higher Resolution Melting (HRM) analysis based on the difference in melting temperature of two different alleles is mostly used to assess single SNPs. LightScanner® (BioFire Diagnostics) platform utilizes the high resolution melting technology. Single nucleotide extension based technologies such as SNaPshot® (Life Biosciences) and iPlex® (Sequenom) extend the single base from the primer and allelic differences are assessed by capillary electrophoresis (SNaPShot) or mass spectrometry (iPlex). KASPar® (K-Biosciences). TaqMan® (Life Biosciences) uses the 5'exonuclease property of Taq polymerase to detect the allelic differences through Real Time (RT) PCR. The ligation based primer extension technologies are SNPplex® (Life Biosciences) and GoldenGate® (Illumina). Infinium® (Illumina) and Axiom® (Affymetrix) are hybridization based SNP assays. These different methods are suitable for different kind of studies. HRM based technology has been used in alfalfa (Han *et al.*, 2012), potato (De Koeyer *et al.*, 2010), soybean (Ha and Boerma, 2008) etc. The cost is higher per data point and this method has less throughput as compared to technologies that are able to

assay multiple SNPs at a time. However it is best suited for screening a single to few markers in a large number of populations (e.g., marker assisted backcrossing selection). Microarray based technology has been used for genotyping soybean (Hyten *et al.*, 2010), wheat (Wang *et al.*, 2014), and alfalfa (Li, 2014). Even though the cost of developing the initial assay is rather high, subsequent cost per data point is low.

Next generation sequencing technologies can be used to inexpensively identify SNPs. In some cases, SNPs can be determined directly from sequence data on all members of a mapping population, thus combining SNP detection and genotyping into one step. Because the entire genome of each individual cannot be sequenced economically at the current time, sequencing generally is focused on a part of the genome and can be accomplished using a) transcriptome sequencing (Marioni *et al.*, 2008), b) sequence capture (Ng *et al.*, 2009), c) restriction enzyme-based genome reduction (Baird *et al.*, 2008). Transcriptome sequencing (RNA-seq) methods sequence cDNA developed from RNA and is representative of gene space. This method has been used to develop SNP marker arrays in many crops including alfalfa (Li *et al.*, 2012). Transcriptome sequencing, however, is mostly used for identifying the SNPs and developing an array based on the results rather than combining the steps of SNP identification and genotyping in the same step. Sequence capture or exome sequencing targets specific genomic regions of interest with oligonucleotide baits and the captured sequences are sequenced to discover SNPs. Exome capture requires known sequences to generate the baits, and consequently, has only been used in a few crops to date, including wheat (Winfield *et al.*, 2012).

Restriction enzyme-based genome reduction is accomplished by first restricting the genome with one or two enzymes and then sequencing the ends of the resulting fragments. Methylation sensitive restriction enzymes will preferentially target genic regions of the genome rather than repetitive sequences. Enzymatic genome reduction can assess a broader region of the genome. Reduced representation libraries (RRL) can be filtered by a size selection step to further reduce the complexity and to have similarly sized fragments to facilitate consistent sequencing (Altshuler *et al.*, 2000).

Two primary methods have been developed for reduced representation sequencing – Restriction site Associated DNA (RAD-seq) and Genotyping-by-Sequencing (GBS). The RAD-seq method uses restriction enzymes to digest the DNA followed by shearing and size selection to further reduce the complexity (Baird *et al.*, 2008). Double digest RAD-seq (Peterson *et al.*, 2012) eliminated the shearing step by using another frequent cutting restriction enzyme to digest the genome. The GBS procedure (Elshire *et al.*, 2011) removes the size selection step to make the method simpler. Modifying the GBS protocol by using two enzymes improved the sequence representation throughout the genome in wheat (Poland *et al.*, 2012). A novel GBS method was proposed using a restriction enzyme type 2b that produces equal sized fragments flanking a recognition site (Wang *et al.*, 2012). However, because of the small fragment size (28-33 bp), including a 6-8 bp non-polymorphic recognition site, 2b-RAD has limited use.

The goal of a GBS project is to inexpensively generate a large amount of SNP marker data on a large number of individuals. The challenge is generating enough sequence so that most individuals have been sequenced for most of the SNPs to be

assayed. Generally, restriction enzymes with a longer recognition site will generate fewer fragments (and thus loci) and provide a higher marker coverage compared to restriction enzymes with shorter recognition sites, given the same number of sequences generated per individual.

After the libraries for all individuals being assayed have been sequenced, the data need to be manipulated so that SNPs can be identified and marker scores generated for all individuals in the population. This is a computationally intensive process and canned programs have not been developed, in general. Although traditional assembly algorithms can be used, new algorithms have been developed to handle these specific types of sequence data. STACKS (Catchen *et al.*, 2013; Catchen *et al.*, 2011) is a program specially designed for handling RAD-associated loci and has been used extensively. The UNEAK pipeline (Lu *et al.*, 2013), incorporated into the TASSEL software (Bradbury *et al.*, 2007), is also useful. STACKS is more flexible and provides the user with greater control over how the data are analyzed than UNEAK, but the latter is faster.

STACKS basically consists of a series of steps to call the genotype. First, it searches for exactly matching sequencing reads and builds a dictionary. Second, it finds the polymorphism within a single genotype and records the reference sequence of each locus. Third, it builds a catalog of those loci from across the population and merges two or more loci depending upon the among-individual merging parameter and compares the loci to call the genotype and haplotype of the individual. In contrast, UNEAK searches for the tag pairs with exactly one mismatch within the read and a filter is applied to discard complex SNP structures involving non-reciprocal SNPs and only SNPs with reciprocal pairs are retained and used in SNP calling. Further, UNEAK pipeline does not

process the paired end data and does not utilize the fragments of length longer than 64. The suitability of UNEAK and STACKS depends several factors, one of which is complexity of the genome. The availability of a reference genome makes the locus identification process easier, and both STACKS and UNEAK can map reads to the reference sequence using open source programs such as BWA (Li and Durbin, 2010) or BOWTIE2 (Langmead *et al.*, 2009). However, the reference genome of switchgrass is still incomplete, so only partial alignment of GBS reads to known locations can be completed.

At the current time, GBS typically results in a large amount of missing data – that is, individuals for which no or insufficient sequencing reads are available for a given locus. Numerous algorithms have been developed and tested to impute the missing genotype calls such as random forest regression (Stekhoven and Bühlmann 2011), nearest neighborhood (Troyanskaya *et al.,* 2001), singular value decomposition (Troyanskaya *et al.,* 2001), expectation maximization (Dempster *et al.,* 1977) etc. Rutkoski *et al.* (2013) experimented with several algorithms, including k-nearest neighbors, singular value decomposition, random forest regression, and expectation maximization imputation, and concluded that the imputation accuracy depended on factors such as proportion of missing values, heterozygosity, linkage disequilibrium, but that the inclusion of markers that have had genotypes imputed led to the increased genomic selection accuracy (Poland *et al.*, 2012).

Genetic mapping and marker-trait associations

Generally speaking, the use of genetic markers in most plant breeding programs has been based on specific markers associated with specific trait loci. Many agronomically important traits are controlled by multiple quantitative trait loci (QTL). Traditionally, QTL mapping has being done based on biparental, segregating populations (Lander and Botstein, 1989; Zeng, 1994; Li *et al.*, 2007). This approach has been successful in identifying the QTL and in some cases, applying the resulting markers in selection programs (Duvik *et al.*, 2004; Collard and Mackill, 2008; Cooper *et al.*, 2009). Classical biparental mapping has limitations. First, the genetic variance within the population is limited to the two parents (Malosetti *et al.*, 2007), and second, the inference space of loci is also limited to those parents or related population. Third, because of the extended linkage disequilibrium (LD), very few identified loci can be pinpointed to the gene level, and are often located in quite large genomic intervals (Zhu *et al.*, 2008).

Genome wide association study (GWAS) or linkage disequilibrium (LD) mapping can be conducted using markers to saturate the genome of populations with limited LD, thereby identifying markers closely associated with trait loci. In the best case, association mapping can resolve complex trait variation down to the sequence level by exploiting evolutionary or historical recombination events (Nordborg and Tavare, 2002). The key to successful LD mapping is a population with a limited extent of LD and consequently a high density of markers. Association mapping is conducted using a mixed model statistical analysis to associate phenotypes and genotypes (Yu *et al.*, 2006; Stich *et al.*, 2008). The structure of the population due to the presence of subpopulations and the kinship of the individuals being assayed need to be controlled in the analysis to avoid

false positive associations (Yu *et al.*, 2006). The population structure is generally controlled with the fixed effect covariates of either principal components or membership to a given subpopulation. This approach may miss true positive marker-trait associations if the real association is nested within the structure (Brachi *et al.*, 2011). Ideal populations for GWAS have been constructed to introduce new recombination into existing historical recombination to help avoid this problem.

Apart from LD-based QTL mapping with bi-parental families and association mapping using germplasm accessions or breeding populations, bulked segregant analysis (Michelmore, *et al.*, 1991) can be used to map loci with large effects, as has been done in rice (Venuprasad *et al.*, 2009), wheat (Shen *et al.*, 2003), maize (Cai *et al.* 2003), soybean (Hyten *et al.*, 2009), and others. A high density of markers makes this method feasible.

Marker-assisted breeding

Markers can be used to identify different heterotic groups. In maize, the variance of yield described by SSR markers was slightly better (54%) than traditional Specific Combining Ability (SCA) based heterotic group analysis (52%) (Fan *et al.*, 2009), suggesting that markers could replace costly and time-consuming traditional heterotic group identification. In switchgrass, heterosis can be exploited using semi-hybrids between contrasting populations (Brummer, 1999). The upland and lowland switchgrass ecotypes are reported to be different heterotic groups (Martinez-Reyna and Vogel, 2008), and markers could help distinguish populations within these ecotypes for more careful heterotic pairings.

Marker-assisted selection (MAS) can be used to evaluate breeding material for the presence, absence, or allele frequencies of loci associated with important traits. Given a reliable marker or a pair of flanking markers tightly linked to the trait locus, the markers can be used in a selection program. These markers can be used for backcrossing with both "foreground selection" (using markers linked to the trait) and "background selection" (using markers not linked with target locus to select against other donor parent chromatin) (Frisch *et al.* 1999). In addition to the use of markers to select for single loci, multiple loci can be pyramided to increase the trait value or to provide durable disease or pest resistance (Singh *et al.*, 2001). MAS can also be used in a selection program especially in early generations when LD remains extended due to few recombination events (Ribaut and Betran, 1999). Genotype × environment interactions and germplasm × marker interactions can limit the reproducibility of markers across breeding programs (Li *et al.*, 2003). Nevertheless, a number of breeding programs are using MAS successfully, including in rice (Singh *et al.*, 2011), wheat (Collard *et al.*, 2005), barley (Dwivedi *et al.*, 2007), fava bean (Torres *et al.*, 2010), soybean (Walker *et al.*, 2004), and others.

An alternative to identifying and applying selected markers linked to a trait is to use genomic selection (GS) (Heffner *et al.*, 2009; Jannick *et al.*, 2010). In a genomic selection program, genome-wide markers are assayed and the breeding value for each marker is determined. Selection is based on the aggregate breeding value of all markers, regardless of whether those markers are individually associated with the trait or not. The availability of high density SNP markers with reduced cost has made genomic selection practical for selection (Jannink et al., 2010). The breeding values changes as the allele frequencies change but practically it will still be useful for a certain cycles of selection.

Assuming that the breeding values of markers are largely retained across several cycles of selection i.e , genomic selection can outperform phenotypic selection or marker-assisted selection (Jannink *et al.,* 2010). The SNP discovery by Genotyping-By-Sequencing (GBS) is the cheapest and fastest way to get high density markers, thus, enhancing the application of GS in plant breeding.

Switchgrass Breeding

Switchgrass is a cross-pollinating, largely self-incompatible species, so the breeding method is mostly recurrent selection (Vogel and Pederson, 1993), typically using space-plant nurseries. Space-planting can be efficient for high heritability traits, but quantitative traits with low heritability may benefit from controlling for spatial variability (Missaoui *et al.,* (2005), Casler (2010) or using family selection (Casler, 2010) (Bhandari *et al.,* 2010, Rose *et al.,* 2008). Early generation cultivars were simply the seed increases from adapted landraces. Later, regional seed collections at the University of Nebraska were used to initiate a breeding program (Eberhart and Newell, 1959). While switchgrass cultivars are commercialized as synthetics today, the presence of heterosis for yield has shown the potential of hybrid cultivar development (Vogel and Mitchell, 2008; Martinez-Reyna and Vogel, 2008).

In most breeding programs, the breeding emphasis focused on In Vitro Dry Matter Digestibility (IVDMD), a lab method simulating the process in rumen of animal (Hopkins *et al.,* 1993, Vogel *et al.,* 2002). The increased IVDMD was reported to be associated with low lignin content (Casler *et al.,* 2002). With the choice of switchgrass as source of cellulose based ethanol biofuel, breeding has emphasized increased biomass

yield and decreased recalcitrance to degradation by industrial enzymes. Recently, biomass yield has been studied extensively (Bhandari *et al.*, 2010, Rose *et al.*, 2008, Casler, 2010, Vogel and Mitchell, 2008; Martinez-Reyna and Vogel, 2008). Choosing appropriately adapted cultivars can increase biomass yield by approximately 20-25% (Sanderson *et al.*, 2007).

Along with the traditional breeding approaches, switchgrass could greatly benefit from advances in genomics and biotechnology (Bouton, 2007). The use of molecular markers to assess the genetic diversity was already discussed above. Several switchgrass genetic maps have been published (Missaoui *et al.*, 2005, Okada *et al.*, 2010 and Liu *et al.*, 2012, Serba *et al.*, 2013) and linkage map based QTL mapping is in progress (Serba, pers. comm.). These QTL can be used in marker-assisted selection as described above. Further QTL mapping using high density markers can assist in fine mapping of trait loci. A draft reference genome has been built for switchgrass (http://www.phytozome.net/ panicumvirgatum_ er.php). The genetic map can be linked to the sequence assembly to locate genes of interest and to facilitate comparative mapping across species.

Dissertation Objectives

My dissertation research will analyze switchgrass germplasm derived from the southern United States. Our guiding hypotheses are that this germplasm consists of diverse, primarily tetraploid populations, that abundant variability exists among this germplasm for important biofuel-related traits, and that associations between genetic markers and phenotypes in this population can identify QTL useful for breeding programs. To address these hypotheses, first, I will develop SNP markers using GBS.

Then, I will analyze the population structure of our germplasm collection using nuclear SSR markers, SNP markers, and plastid markers to assess the presence of population substructure and to understand the relationships of the accessions in this study. Then, I will assess phenotypic diversity within this collection for agronomic traits. Finally I will use the SNPs to identify quantitative trait loci associated with these traits.

References

Altshuler D., Pollara V.J., Cowles C.R., Van Etten W.J., Baldwin J., Linton L., Lander E.S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513-516.

Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z.A., Selker E.U., Cresko W.A., Johnson E.A. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. Plos One 3. DOI: ARTN e3376 DOI 10.1371/journal.pone.0003376.

Botstein D., White R.L., Skolnick M., Davis R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics 32:314-31.

Bouton J.H. (2007) Molecular breeding of switchgrass for use as a biofuel crop. Current Opinion in Genetics & Development 17:553-558. DOI: Doi 10.1016/J.Gde.2007.08.012.

Bradbury P.J., Zhang Z., Kroon D.E., Casstevens T.M., Ramdoss Y., Buckler E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635. DOI: Doi 10.1093/Bioinformatics/Btm308.

Casler M.D. (2012) Switchgrass Breeding, Genetics, and Genomics, in: A. Monti (Ed.), Switchgrass, Springer, Bologna, Italy. pp. 29-54.

Catchen J., Hohenlohe P.A., Bassham S., Amores A., Cresko W.A. (2013) Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124-3140. DOI: Doi 10.1111/Mec.12354.

Catchen J.M., Amores A., Hohenlohe P., Cresko W., Postlethwait J.H. (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3-Genes Genomes Genetics 1:171-182. DOI: Doi 10.1534/G3.111.000240.

Collard, B. C., & Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1491), 557-572.

Cooper, M., van Eeuwijk, F. A., Hammer, G. L., Podlich, D. W., & Messina, C. (2009). Modeling QTL for complex traits: detection and context for plant breeding. Current opinion in plant biology, 12(2), 231-240.

Costich D.E., Friebe B., Sheehan M.J., Casler M.D., Buckler E.S. (2010) Genome-size variation in switchgrass (*Panicum virgatum*): flow cytometry and cytology reveal rampant aneuploidy. Plant Genome 3:130-141. DOI: 10.3835/plantgenome2010.04.0010.

De Koeyer D., Douglass K., Murphy A., Whitney S., Nolan L., Song Y., De Jong W. (2010) Application of high-resolution DNA melting for genotyping and variant scanning of diploid and autotetraploid potato. Molecular Breeding 25:67-90. DOI: Doi 10.1007/S11032-009-9309-4.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal statistical Society,39(1), 1-38.

Duvick, D. N., Smith, J. S. C., & Cooper, M. (2004). Long-term selection in a commercial hybrid maize breeding program. Plant breeding reviews, 24(2), 109-152.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Plos One 6. DOI: ARTN e19379 DOI 10.1371/journal.pone.0019379.

Ha B.-K., Boerma H.R. (2008) High-throughput SNP genotyping by melting curve analysis for resistance to southern root-knot nematode and frogeye leaf spot in soybean. J Crop Sci Biotech 11:91-100.

Han Y.H., Khu D.M., Monteros M.J. (2012) High-resolution melting analysis for SNP genotyping and mapping in tetraploid alfalfa (*Medicago sativa* L.). Molecular Breeding 29:489-501. DOI: Doi 10.1007/S11032-011-9566-X.

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. Crop Science, 49(1), 1-12.

Jannink, J. L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics, 9(2), 166-177.

Kim S., Misra A. (2007) SNP genotyping: technologies and biomedical applications. Annual review of biomedical engineering 9:289-320. DOI: 10.1146/annurev.bioeng.9.060906.152037.

Lander, E. S., & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121(1), 185-199.

Lemus R., Brummer E.C., Moore K.J., Molstad N.E., Burras C.L., Barker M.F. (2002) Biomass yield and quality of 20 switchgrass populations in southern Iowa, USA. Biomass & Bioenergy 23:433-442. DOI: Pii S0961-9534(02)00073-9 Doi 10.1016/S0961-9534(02)00073-9.

Li, H., Ye, G., & Wang, J. (2007). A modified algorithm for the improvement of composite interval mapping. Genetics, 175(1), 361-374.

Li X.H., Acharya A., Farmer A.D., Crow J.A., Bharti A.K., Kramer R.S., Wei Y.L., Han Y.H., Gou J.Q., May G.D., Monteros M.J., Brummer E.C. (2012) Prevalence of single nucleotide polymorphism among 27 diverse alfalfa genotypes as assessed by transcriptome sequencing. Bmc Genomics 13. DOI: Artn 568 Doi 10.1186/1471-2164-13-568.

Li, X., Han, Y., Wei, Y., Acharya, A., Farmer, A. D., Ho, J. & Brummer, E. C. (2014). Development of an Alfalfa SNP Array and Its Use to Evaluate Patterns of Population Structure and Linkage Disequilibrium. PloS one, 9(1), e84329.

Lu F., Lipka A.E., Glaubitz J., Elshire R., Cherney J.H., Casler M.D., Buckler E.S., Costich D.E. (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. Plos Genetics 9. DOI: Artn E1003215 Doi 10.1371/Journal.Pgen.1003215.

Malosetti M., van der Linden C.G., Vosman B., van Eeuwijk F.A. (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics 175:879-889. DOI: Doi 10.1534/Genetics.105.054932.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research, 18(9), 1509-1517.

McLaughlin S.B., Adams Kszos L. (2005) Development of switchgrass (*Panicum virgatum*) as a bioenergy feedstock in the United States. Biomass and Bioenergy 28:515-535. DOI: 10.1016/j.biombioe.2004.05.006.

Michelmore, R., Paran I., Keselli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. PNAS 88:9828-9832.

Missaoui A.M., Paterson A.H., Bouton J.H. (2005) Investigation of genomic organization in switchgrass (*Panicum virgatum* L.) using DNA markers. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 110:1372-83. DOI: 10.1007/s00122-005-1935-6.

Mullis K.B., Faloona F.A. (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods in enzymology 155:335-50.

Narasimhamoorthy B., Saha M., Swaller T., Bouton J. (2008) Genetic Diversity in Switchgrass Collections Assessed by EST-SSR Markers. BioEnergy Research 1:136-146. DOI: 10.1007/s12155-008-9011-0.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 461(7261), 272-276.

Nordborg M., Tavare S. (2002) Linkage disequilibrium: what history has to tell us. Trends in Genetics 18:83-90. DOI: Doi 10.1016/S0168-9525(02)02557-X.

Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. Plos One 7. DOI: ARTN e37135 DOI 10.1371/journal.pone.0037135.

Poland J.A., Brown P.J., Sorrells M.E., Jannink J.L. (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. Plos One 7. DOI: ARTN e32253 DOI 10.1371/journal.pone.0032253.

Rutkoski J.E., Poland J., Jannink J.L., Sorrells M.E. (2013) Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy. G3-Genes Genomes Genetics 3:427-439. DOI: Doi 10.1534/G3.112.005363.

Sanderson M., Reed R., Mclaughlin S., Wullschleger S., Conger B., Parrish D., Wolf D., Taliaferro C., Hopkins A., Ocumpaugh W. (1996) Switchgrass as a sustainable bioenergy crop. Bioresource Technology 56:83-93. DOI: 10.1016/0960-8524(95)00176-X.

Schuelke M. (2000) An economic method for the fluorescent labeling of PCR fragments. Nature biotechnology 18:233-4. DOI: 10.1038/72708.

Sobrino B., Brion M., Carracedo A. (2005) SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Science International 154:181-194. DOI: Doi 10.1016/J.Forsciint.2004.10.020.

Stekhoven, D. J., Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.

Tautz D., Renz M. (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Research 12:4127-38.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525.

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

Wang S., Meyer E., McKay J.K., Matz M.V. (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. Nature Methods 9:808-+. DOI: Doi 10.1038/Nmeth.2023.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant biotechnology journal.

Wenz H., Robertson J.M., Menchen S., Oaks F., Demorest D.M., Scheibler D., Rosenblum B.B., Wike C., Gilbert D.A., Efcavitch J.W. (1998) High-precision genotyping by denaturing capillary electrophoresis. Genome research 8:69-80.

Williams J.G., Kubelik A.R., Livak K.J., Rafalski J.A., Tingey S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18:6531-6535.

Winfield M.O., Wilkinson P.A., Allen A.M., Barker G.L.A., Coghill J.A., Burridge A., Hall A., Brenchley R.C., D'Amore R., Hall N., Bevan M.W., Richmond T., Gerhardt D.J., Jeddeloh J.A., Edwards K.J. (2012) Targeted re-sequencing of the

allohexaploid wheat exome. Plant Biotechnology Journal 10:733-742. DOI: Doi 10.1111/J.1467-7652.2012.00713.X.

Wright L. (1994) Production technology status of woody and herbaceous crops. Biomass and Bioenergy 6:191-209. DOI: 10.1016/0961-9534(94)90075-2.

Zale J., Freshour L., Agarwal S., Sorochan J., Ownley B.H., Gwinn K.D., Castlebury L.A. (2008) First Report of Rust on Switchgrass (*Panicum virgatum*) Caused by *Puccinia emaculata* in Tennessee. Plant Disease 92:1710-1710. DOI: Doi 10.1094/Pdis-92-12-1710b.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. Genetics,136(4), 1457-1468.

Zhang Y., Zalapa J., Jakubowski A.R., Price D.L., Acharya A., Wei Y., Brummer E.C., Kaeppler S.M., Casler M.D. (2011a) Natural hybrids and gene flow between upland and lowland switchgrass. Crop Science 51.

Zhang Y.W., Zalapa J.E., Jakubowski A.R., Price D.L., Acharya A., Wei Y.L., Brummer E.C., Kaeppler S.M., Casler M.D. (2011b) Post-glacial evolution of Panicum virgatum: centers of diversity and gene pools revealed by SSR markers and cpDNA sequences. Genetica 139:933-948. DOI: Doi 10.1007/S10709-011-9597-6.

Zhu C.S., Gore M., Buckler E.S., Yu J.M. (2008) Status and Prospects of Association Mapping in Plants. Plant Genome 1:5-20. DOI: Doi 10.3835/Plantgenome2008.02.0089.

CHAPTER 2

GENETIC DIVERSITY AND POPULATION STRUCTURE INFERENCE OF SWITCHGRASS (*PANICUM VIRGATUM* L.) ACCESSIONS USING GENOTYPIC AND CYTOPLASMIC SSR MARKERS, SNPS AND MORPHOLOGICAL TRAITS[1]

---

[1] Acharya, A.R., Y. Wei, M. Saha, K.M. Devos and E.C.Brummer. To be submitted to *Bioenergy Research*

Abstract

Switchgrass (*Panicum virgatum* L.), a warm season grass native to North America, can be cultivated as a forage or biofuel crop. Improved biofuel characteristics can be selected by tapping the genetic diversity within the species. The objective of this study was to examine the population structure and relatedness within and among switchgrass populations mainly derived from the southern United States. A total of 464 individual plants from 49 accessions were selected for genotyping and phenotyping. Genotyping was done using SSR markers and SNP markers well distributed throughout the switchgrass genome. We used 35 SSR markers that generated 365 alleles and Genotyping-by-Sequencing (GBS) to identify 3,196 SNP that were present in at least 80% of the population and that had at least six sequencing reads per locus per genotype. We also used nine chloroplast specific indel markers to identify the cytotype of each individual. We collected data from a field trial on three canonical morphological traits that typically are used to differentiate switchgrass ecotypes: plant height, stem diameter, and flowering time. Based on marker analyses, we identified the expected population differentiation into two major clusters representing upland and lowland ecotypes, with a further clear subdivision of lowland ecotypes into two subgroups. Further subgroups could be discerned in all major groups, and these groupings were largely based on geographical origin of the accessions. Based on the literature, lowland ecotypes are expected to flower later, be taller, and have thicker stems than upland ecotypes. Our analysis showed that one of the lowland clusters had plant height similar to the upland cluster and stem diameter intermediate between the other lowland cluster and upland

accessions. Thus, this experiment shows that few distinguishing traits are not sufficient for clasifying ecotypic status in switchgrass.

Introduction

Bioenergy is a renewable energy source that can help meet the world's increasing demand for energy in an environmentally sensible manner. The Bioenergy Feedstock Development Program in the U.S. Department of Energy began evaluating a wide variety of potential feedstocks in 1978 (McLaughlin and Adams Kszos, 2005). Among herbaceous feedstocks, switchgrass (*Panicum virgatum* L.), a perennial, warm-season (C_4) grass native to North America, was identified as the most promising target to develop as a bioenergy crop. Switchgrass has many characteristics of a desirable biofuel feedstock including high productivity (Wright, 1994), broad adaptation to a wide variety of environments, including to poor soil conditions (Sanderson *et al.*, 1996), and familiarity to farmers as a forage crop.

Switchgrass germplasm has been classified into lowland and upland ecotypes based on their phenotypes (Das *et al.*, 1997; Porter, 1966). Uplands tend to be shorter, have thinner stems, and flower and senesce earlier compared to lowlands. All lowland and some upland ecotypes are tetraploid (2n = 4x = 36), but most upland ecotypes are octoploid (2n = 8x = 72) (Hopkins *et al.*, 1996; Narasimhamoorthy *et al.*, 2008b). Agenetic linkage map has recently been developed from a cross between two tetraploid genotypes, and disomic inheritance has been confirmed (Okada *et al.*, 2010c; Serba *et al.*, 2013). Aneuploidy has been identified in switchgrass (Costich *et al.*, 2010), although the extent to which it exists across all germplasm is not known.

A successful crop improvement program requires genetic variation. Genetic diversity within switchgrass germplasm is extensive, although the amount identified in any given experiment varies depends on the populations evaluated (Cortese *et al.*, 2010;

Gunter *et al.*, 1996; Narasimhamoorthy *et al.*, 2008b; Zalapa *et al.*, 2011; Zhang *et al.*, 2011a; Zhang *et al.*, 2011b). Variation within switchgrass populations is considerably higher than variation among populations (Narasimhamoorthy *et al.*, 2008a; Zalapa *et al.*, 2011), which is a common finding in other outcrossing grass species (Mian *et al.*, 2005; Ubi *et al.*, 2003).

Ecotypic differentiation was initially based on morphological and ecogeographic characteristics, but molecular marker analyses have enabled a more detailed exploration of switchgrass germplasm relationships. Two main germplasm groups, upland and lowland ecotypes, have been easily distinguished using chloroplast markers (Hultquist 1997), nuclear SSR markers (Cortese *et al.*, 2010; Gunter *et al.*, 1996; Narasimhamoorthy *et al.*, 2008; Zalapa *et al.*, 2011; Zhang *et al.*, 2011a; Zhang *et al.*, 2011b), and SNP markers (Lu *et al.*, 2013). Marker-based diversity has been explicitly related to morphological diversity (Cortese *et al.*, 2010). Genotyping-by-sequencing (Elshire *et al.*, 2011) was used to identify over 700,000 SNP markers in a (mostly) northern USA association mapping panel (Lu *et al.*, 2013). A diversity analysis based on 29,000 of these markers indicated clear differentiation between ploidy levels, with two main subpopulations present among lowland accessions, among 4x upland accessions, and among 8x upland accessions (Lu *et al.*, 2013). Fine-grained analysis of switchgrass germplasm using SSR markers has identified numerous sub-populations of both upland and lowland germplasm and apparent hybrids both between ecotypes and across ploidy levels (Zalapa *et al.*, 2011; Zhang *et al.*, 2011a; Zhang *et al.*, 2011b)

As a prelude to association mapping of important traits for bioenergy production, we evaluated a large germplasm collection of accessions from mainly the southern USA,

a complement to the northern germplasm panel described by Lu *et al.* (2013). Although many of these germplasms have been included in previous genetic diversity analyses, we needed to estimate relationships among these genotypes to appropriately adjust our GWAS models. Further, because a number of accessions in our panel were collected in the southeastern US, which appears to be a center of diversity for switchgrass (Zhang *et al.*, 2011b), we were interested in clarifying the levels of hybridity between ecotypes observed previously (Zhang *et al.*, 2011a) and in comparing the marker-based results with canonical morphological traits often used to discriminate ecotypes, especially for the hybrids.

Our hypotheses were (1) that upland and lowland accessions can be identified by marker analyses, (2) that evidence of hybrid origin of at least some genotypes in certain populations could be identified using markers, and (3) that a phenotypic analysis of maturity, plant height, and stem diameter would be able to clarify hybrid status in southern USA switchgrass germplasm. The objective of this study was to test these hypotheses by evaluating diversity of southern USA switchgrass accessions using chloroplast and nuclear SSR markers, SNP markers generated by GBS, and phenotypic analysis of key traits. In addition, because we used both nuclear SSR markers and GBS-generated SNP markers, a further goal of this experiment was to compare genetic diversity estimates generated by the two marker types.

Materials and Methods

Plant materials:

The germplasm we analyzed in this experiment largely derived from the southern half of the US, with a few exceptions, and included both upland and lowland accessions. We included 29 accessions from the National Plant Germplasm System (NPGS), seven populations we collected from Florida, Georgia and South Carolina, and two genotypes that are the parents of a genetic mapping population (Missaoui *et al.*, 2005; Serba *et al.*, 2013), AP13 derived from 'Alamo' and VS16 derived from 'Summer' (Table 2.1, Fig 2.1). Each accession was represented by between one and 16 plants (each having a distinct genotype) for a total of 480 genotypes measured for various phenotypic traits in Watkinesville, GA. An additional 31 genotypes from seven populations recently collected from Florida by the NPGS and six ornamental switchgrass genotypes were included for both nuclear and cytoplasmic SSR genotyping. Of the 511 total genotypes, 372 were sequenced for SNP genotyping. Overall, there were 322 "core" genotypes on which all genotypic and phenotypic data were recorded. The core genotypes were used to compare marker types.

Phenotyping:

The 480 genotypes from the 36 accessions were clonally propagated in the greenhouse and planted into the field at the UGA Plant Sciences farm near Watkinsville, GA in July 2009 with 90 cm spacing on center in a 16 × 30 α-lattice design with 3 replications. Each replication consisted of 16 genotypes in each of 30 blocks. Each genotype was represented by a single clone in each replication. No data were taken during 2009, allowing the plants to fully establish. We collected data for three canonical

traits – height, stem diameter, and flowering date – that define switchgrass ecotypes. Height was measured on three tillers per plant from the ground to the uppermost node of a flowering stem. Stem diameter was measured on three tillers 5 cm from the ground. For flowering date, we recorded the date when at least three tillers showed emergence of the inflorescence. The height and stem diameter were measured after full maturity and before harvest in December 2011 and December 2012.

***SSR Genotyping**:*

We extracted DNA from young leaves of switchgrass using the CTAB method (Doyle and Doyle, 1990). We screened 50 SSR markers developed from switchgrass (Okada *et al.*, 2010b; Serba *et al.*, 2013) covering the genome and selected 35 SSR markers for analysis based on the polymorphism and signal quality (Table 2.2). The M13 tailing method (Schuelke, 2000) was used to label PCR products. Reactions were prepared in a volume of 10 µl with 20 ng of template DNA, 2.5mM of $MgCl_2$, 10× buffer, 0.5 U AmpliTaq Gold® (Applied Biosystems, Foster City, CA, USA), 0.15 mM dNTPs, 0.25 pmol forward primer, 0.5 pmol backward primer and 1.0 pmol M13 universal primer. The M13 universal primer was labeled with one of the blue (FAM), green (HEX) or yellow (NED) fluorescent dyes. We pooled the PCR products of different fragment size and florescent labels. The pooled sample was then mixed with 4µl deionized formamide, 25 µM ROX size standard were added, and the sample was analyzed on an ABI3730 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) at the University of Georgia Genomics Facility (GGF). The genotypic data received from GGF were analyzed with Genemarker 1.85 software. Each allele (fragment) for each marker was scored as a dominant marker, with 1 for presence and 0 for absence of the

specific DNA fragment size in each individual genotype. Individual markers had <5% missing data.

We also designed nine primers specific to the chloroplast, one from Missaoui *et al.* (2005) and eight from Young *et al.* (2011). Each of these markers showed an indel of more than 17 base pairs between the upland and lowland ecotypes. We genotyped these markers using conditions similar to nuclear SSRs described above and scored the fragment sizes. If the genotyping yielded more than two previously identified alleles, at least three genotypes were sequenced with ABI 3730XL (Applied Biosystems) at The Samuel Roberts Noble Foundation genomics core facility for each allele size to identify any further indels.

SNP Genotyping*:*

In addition to SSR genotyping, we also analyzed SNP genotypes on 372 genotypes. We used the same genomic DNA as above. The concentration of DNA was initially measured by nanodrop and DNA concentrations were diluted and normalized to 20 ng/ul based on further quantification using PicoGreen. Single nucleotide polymorphisms (SNP) were identified using the two-enzyme Genotyping-by-Sequencing (GBS) method described by Poland *et al.* (2012) except that we used *Fse*I, a methylation sensitive restriction enzyme with an 8-bp recognition site, instead of *Pst*I, to further reduce the number of fragments generated. Following digestion with *Fse*I and *Msp*I, sample barcodes, *Fse*I adaptors, and a common adapter were ligated and fragments were amplified by PCR using Illumina sequencing primers. We multiplexed 48 genotypes for single end sequencing with a read length of 101 bases in a single lane on an Illumina HiSeq 2000 sequencer at the University of Texas, Austin sequencing facility

(http://www.icmb.utexas.edu/core/DNA/). We obtained about four million raw sequencing reads per genotype.

We analyzed the sequencing reads using the STACKS software (Catchen *et al.*, 2013; Catchen *et al.*, 2011) to assemble reads and to identify SNP marker loci without the use of reference genome. We considered de novo assembly to capture all polymorphic site because only 65% of raw reads aligned to 18 pseudomolecules of *Panicum virgatum* reference genome 1.1 (available at ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/early_release/Pvirgatum_v1.1/). After Illumina's internal quality filtering (chastity filter, www.illumina.com) discarding low quality reads, we demultiplexed the data according to barcode and trimmed the sequencing reads to 64 bases, including the restriction site. We allowed up to three mismatches within a genotype and up to two additional mismatches among genotypes to assemble reads. We required a minimum of six sequencing reads in order to call a given locus as either homozygous or heterozygous, with the further requirement that any given allele be represented by at least two reads. Loci with an allelic ratio less than 0.01 within a given genotype were corrected to be homozygous and loci with allelic ratio between 0.01 and 0.1 were discarded from the analysis because their genotype would be ambiguous (that is, potentially either homozygous or heterozygous). We used loci with a minor allele frequency (MAF) greater than 3% across the entire population for further analysis. We required at least 80% of genotypes to be called with a minimum of six reads per loci (S80-6) in order for them to be included in the analysis of population structure. Additionally, we prepared other datasets with lower number of genotypes called (S50-6) and also lower number of minimum reads (four) for comparison purposes. The scripts to

filter genotype calls are available in supplementary data and online at https://gist.github.com/anantaacharya.

A sample with few or no sequencing reads for a given locus was assigned a genotype based on imputation. We imputed SNP genotypes using the random forest procedure implemented in the MissForest (Stekhoven and Buhlmann, 2012) package in R. The SNP markers were aligned with the *Panicum virgatum* reference genome 1.1 (available at ftp://ftp.jgipsf.org/pub/compgen/phytozome/v9.0/early_release/Pvirgatum_v1.1/) by using a custom BLAST (Altschul et al., 1990). We discarded the SNP's from fragments that mapped to more than one genome position to ensure that the SNP's were not from paralogous or homeologous loci.

Population Structure and Diversity:

The structure of the entire population was evaluated using the program STRUCTURE v2.3.1 (Hubisz *et al.*, 2009; Pritchard *et al.*, 2000). The model was evaluated with 50,000 repetitions of Markov Chain Monte Carlo (MCMC) preceded with 20,000 burn-ins and with a predetermined number of sub-populations ranging from one to twelve with five replication each. We used the change in likelihood method (Evanno *et al.*, 2005) to determine the most likely number of sub-populations within the overall population of genotypes. The program POWERMARKER (Liu and Muse, 2005) was used to calculate genetic distances among genotypes based on shared alleles and to draw a neighbor joining dendrogram (Jin and Chakraborty, 1994). FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/) was used to format the dendrograms. A principal components analysis was conducted to further characterize the population

structure using the R statistical packages (R Development Core Team, 2011). An analysis of molecular variance (AMOVA) was conducted to partition molecular variation within and among populations and ecotypes using the "pegas" software package for R (Paradis, 2010). Graphics were produced with R (R Development Core Team, 2011) with aid of "ggplot2" package (Wickham, 2009).

We calculated summary statistics for the phenotypic data using analysis of variance. For each trait, we performed the mixed model (with lme4 package in R) with genotype as a fixed effect and replication within environment, environment, and genotype × environment interaction as random effects. We computed Pearson correlations (r) among the phenotypic traits based on mean values for each genotype. Statistical significance was assessed at the 5% level unless indicated otherwise.

Results

Nuclear DNA SSR polymorphism:

The 35 SSR markers we evaluated produced 389 alleles, with individual markers having between two and 28 alleles across all genotypes. All markers had a very high Polymorphism Information Content (PIC), ranging from 0.38 to 0.99. Of the 389 alleles, 32 were present in fewer than 5% of the genotypes. However, eight of these 32 were present in genotypes of only one or two accessions and consequently, we included them in further analyses while removing the others. Therefore, overall, 365 alleles were used for the genetic diversity and population structure analysis. Considering only the alleles retained in the analysis, markers averaged 10.4 alleles across the population and individual genotypes averaged two alleles per marker.

SNP polymorphism:

We identified 65,328 SNPs across all genotypes regardless of read depth. We required at least six sequencing reads within a given individual in order to call a genotype, and further required that heterozygotes included at least two reads of each allele. By requiring six reads per locus (assuming bi-allelic loci), we could be 95% sure of identifying at least one read of each allele in a heterozygote, assuming each allele had an equal opportunity to be sequenced (Sedcole, 1977). We removed SNPs with a minor allele frequency (MAF) less than 3% from further analysis. After applying these filters, 3,196 SNPs were present in at least 80% of the population. The minor allele frequency ranged from 0.03 to 0.50 with the average of 0.16. The mean PIC per locus was 0.19, with a maximum of 0.38. If we retained SNP whose genotype could be called in at least 50% of the population (S50-6), Re-filtering to retain SNP loci present in at least 50% of the population resulted in 20,233 SNP markers, with a minor allele frequency ranging from 0.03 to 0.50, and an average of 0.14.

Chloroplast specific indel polymorphism:

We identified 32 alleles from nine chloroplast markers. Fourteen of these alleles had not been identified in previous experiments. For all markers, only one allele was present per genotype, as expected. Two markers (rps4-ndhJ-a and rbcL-psaI) only had two alleles and upon sequencing, they showed the same indels as reported previously (Young *et al.*, 2011). The remaining seven chloroplast markers generated three to seven alleles across the whole population. For each of these markers, sequencing identified the previously reported indel alleles (Missaoui *et al.*, 2005; Young *et al.*, 2011) as well as additional indels (Table 2.3).

Genetic Diversity and Population Structure:

We assessed population structure independently for the nuclear SSR and SNP datasets, and in both cases, the STRUCTURE (Pritchard *et al.*, 2000) analysis suggested that our genotypes were substructured into two groups (K) (Fig. 2.2a). An additional analysis using the change in likelihood (ΔK) method (Evanno *et al.*, 2005) suggested two major subpopulations as well, with possibly additional subpopulations (Fig. 2.2b). The membership plot suggested that the two groups generally corresponded to upland and lowland ecotypes, with some individuals showing mixed parentage (Fig 2.3a). The membership plots from both the SSR and SNP analyses were consistent, with SNP markers showing less admixture of subpopulations (Fig 2.3b). Genetic distances computed between all genotypes for both SNP and SSR were similar when plotted as neighbor joining tree.

We computed genetic distances among genotypes based on the SNP data and used these distances to build a phylogenetic tree. This analysis also identified the two broad groups (Fig 2.4), but clearly indicated that the lowland group was further divided into two subgroups, one with mostly southeastern USA accessions and the other containing accessions from other regions (Fig 2.1). Using STRUCTURE to place genotypes into K=3 groups, we similarly identified the two main lowland groups (Fig 2.5). We identify these three groups as Upland, Lowland A and Lowland B, and use these group names as a basis of discussion of cytotype and phenotypic variation below.

On the phylogenetic tree, further clusters of accessions and/or genotypes were present within each of the three groups, suggesting a fine-grained structure similar to that identified by Zhang (2011b). In that study (Zhang *et al.*, 2011b), switchgrass germplasm

was partitioned into ten separate groups. In our experiment, eight of these groups could be identified based on clustering in the neighbor-joining tree (Fig 2.4, labels in second columns from right), but two upland groups were not identified probably because our experiment focused on southern-USA derived accessions. The “Lowland 4x C” was identified as “Lowland A” and “Lowland 4x D” was identified as “Lowland B”. Other lowland groups of Zhang *et al.* (2011b) were hybrids of Lowland A and Lowland B and hybrids of Upland with either or both of Lowlands. We identified more clusters of lowland groups and a new upland group compared to Zhang *et al.* (2011b).

Most accessions belonged to one of the three main groups, and all genotypes of those accessions were members of the same group. We noted several exceptions, however. We identified accessions that appeared to be hybrids between groups and others that had mixtures of genotypes derived from multiple groups (Fig 2.5). Hybrid accessions (PI 315723, PI 317525, PI 422016, PI 476290, PI 476293, PI 422003 and Sprewell Bluff) uniformly included genotypes having genomes with similar proportions derived from two or more groups (e.g., Sprewell Bluff with Upland and Lowland B) based on the Structure analysis (Fig 2.5).

However, some accessions (PI 422006, PI 476291 and Pasco co-FL) included genotypes with strikingly different group memberships. PI 422006 is putatively the cultivar Alamo, but some genotypes showed membership to Lowland A and the rest to the Upland cluster. Because Alamo is known to be a lowland accession, we believe that our seed source was contaminated; in fact, a similar result was found by Narasimhamoorthy *et al.* (2008). Accessions with mixed Upland and Lowland B genotypes included PI 476291 (four and three genotypes in each group, respectively), and

Pasco Co-FL (seven and five, respectively). Two genotypes of accession PI 476293 belonged to Lowland A but six of them had an apparent hybrid origin showing membership to both Lowland A and Upland. Seven out of eight genotypes of PI 422003 showed membership to all three groups (Upland, Lowland A, and Lowland B). The one remaining individual belonged to the Upland group.

Based on the chloroplast specific markers, the genotypes were classified into two broad groups. The Lowland A and Lowland B group were also evident with chloroplast specific SSR markers, but the genetic distances were not as pronounced as with SNP markers. Clustering based on chloroplast marker cytotypes generally reflected the SNP marker groupings. However both Lowland B genotypes of PI 315725 (Coffeeville, MS) had a cytotype typical of Upland genotypes. We also identified some accessions that were a mixture of Lowland A and Upland. Interestingly, these had only one of the cytotypes, either upland or lowland. The Sprewell Bluff accession from Georgia had a Lowland A phenotype except for plant height, but had an upland cytotype. Genetically, it was hybrid of both upland and lowland. PI476293 from New Jersey had a Lowland B but exhibited a mixed phenotype and a hybrid nuclear genetic constitution. One accession from Florida (SWFWMD) had Upland cytotype but clustered with Lowland A genetically and to Lowland B morphologically. The hybrids between Lowland A and Lowland B had either of lowland specific cytotype. For mixture accessions, the cytotype reflected nuclear genotype groups.

Analysis of Molecular Variance:

We used analysis of molecular variance (AMOVA) to estimate the relative amount of SSR and SNP-based genetic variation present within and among accessions.

About half of the SNP-based variance (49.2%) was present within accessions with the remainder among accessions; for SSR markers, nearly three-quarters of the variance (71.5%) was within accession. Adding an ecotypic classification, based on the clusters developed with genetic markers, as a third hierarchical level to the analysis showed that SNP-based among accession variation was split roughly equally among ecotypes (25.1%) and among accessions within ecotypes (25.7%). However, with SSR markers, 28.2% of 28.5% variance was explained by ecotype and virtually none was accorded among accessions within ecotypes.

Phenotypic diversity:

Across all germplasm, the three phenotypic traits often used to distinguish between ecotypes suggested a bimodal distribution (data not shown). Flowering time was positively correlated with both plant height (r=0.49) and stem diameter (r=0.62) and plant height was positively correlated with stem diameter (r=0.85). We compared phenotypes of these traits among the three DNA-based groups and also included a fourth group that included hybrids between any of the groups. Upland genotypes flowered earlier than either of the Lowland groups, which were similar and flowered late; the hybrids were intermediate. The Lowland B group was taller than the other groups; interestingly, the Lowland A group had similar height as the Upland group, suggesting that this trait alone is insufficient to discriminate among ecotypes. Lowland B had the thickest stems, Upland thinnest, and the other groups in between (Table 2.4, Figure 3.6 and Fig 3.7). One particular hybrid is noteworthy; PI 422003 contained the genome of all three major germplasm groups. The mean flowering time of this accession was 199 days, which was later than any other groups, including both lowland groups. Mean plant height was 104

cm and stem diameter was 48 mm, both of which were in between upland and lowland types.

Based on marker profiles, some accessions were mixtures of genotypes with distinctly different genome group memberships (see above). Pasco Co-FL was a mixture of seven Upland and five Lowland B genotypes. The average flowering time for the Upland genotypes was 160 days and for lowlands, 196 days. Similarly, the stem diameter was 22 and 38 mm and plant height 54 and 68 cm, respectively. The two groups differed for these traits based on a t-test. The difference in the flowering time probably enabled these two groups to remain genetically isolated within the same collection location. PI 476291 was a mixture of four Upland genotypes, one Lowland A genotype, and two genotypes that were putative hybrids between the two Lowland groups. The average flowering time for Upland genotypes was 165 days, and for the Lowlands 168 days, a non-significant difference. However, for height the two groups measured 80 and 112 cm, respectively, and for stem diameter measured 26 and 40 mm, respectively; in both cases, these differences were statistically significant.

Discussion

From the results of both SSR and SNP markers, we differentiated the switchgrass into the two well-known switchgrass ecotypic groups – Upland and Lowland (Casler *et al.*, 2007; Cortese *et al.*, 2010; Gunter *et al.*, 1996; Lu *et al.*, 2013; Narasimhamoorthy *et al.*, 2008b; Okada *et al.*, 2010c; Zhang *et al.*, 2011b). Both SNP and SSR markers gave similar results, with SNP markers giving higher resolution. Some accessions appeared to

be mixtures of ecotypes, and others are apparent hybrids, some of which were reported previously by Zhang (2011b).

Our DNA marker results-suggested that the Lowland accessions could be split into two groups – the southern Great Plains lineage (Lowland A) and southeastern lineage (Lowland B). Accessions from Florida belonged to either the Lowland A or Lowland B cluster, potentially because their natural habitat could have been either in or near wetlands or at higher elevations. We identified accessions that were natural hybrids between groups, and these tended to be derived from geographic zones of overlap between ecotypic groups. Across accessions, 20 were Upland, 13 Lowland A, 13 Lowland B, four hybrids and two mixtures.

This study complements a previous switchgrass diversity study using SNP markers (Lu *et al.*, 2013) by including more lowland genotypes. Based on accession that overlapped, our clustering results using fewer SNP markers with less missing data were similar to those of Lu *et al.* (2013). Although we did not present the results, we did similar analyses using SNP datasets we generated based on fewer required reads to call genotypes and with higher percentages of missing data, and we achieved a similar result in terms of clustering relationships. We conclude that for the study of genetic diversity and population classification, lower quality genetic data with less read depth and more missing genotypes can be used but probably does not yield more information than does a smaller amount of high quality data.

All individuals that were clustered as Uplands using DNA markers had the same cytotype. Except for one Lowland A (PI317525, MS) and one lowland B (SWFWMD, FL) accession, which had Upland specific cytotypes, all Lowlands had cytotypes that

matched their marker-assigned group. For mixtures, the cytotypic group of an individual reflected its genotypic group, so accessions with mixtures based on SSR or SNP markers were mixtures of cytotypes as well. For individuals that appeared to be hybrids, their cytotype matched only one of the putative parental groups. The chloroplast specific markers could be useful in classifying upland and lowland ecotypes, but populations from geographic regions in between upland and lowland centers of diversity were often ambiguous in terms of cytotype. Genomic markers are more effective in separating ecotypes.

A higher molecular variance within vs. among populations is expected in cross-pollinated crops (Huff, 1997; Kölliker *et al.*, 1999; Ubi *et al.*, 2003). The high genetic diversity within populations will likely make selection from any accession successful for many traits (Bouton, 2007). The morphological analysis indicated that Lowland A was superior in terms of biofuel-related traits, with Uplands decidedly inferior when grown in this Georgia environment.

We assigned membership of individual plants to ecotypic clusters based solely on DNA marker analysis. Once we made the assignment to clusters, we then evaluated the phenotypes of plants within those clusters. Traditionally, plants with early flowering, short stature, and thin stems have been described as upland ecotypes, whereas lowland ecotypes were described as late flowering, tall, and having thick stems. We discovered in this experiment that one lowland group (Lowland B) did not have the typical lowland characteristics – while it was indeed late flowering as expected for lowland ecotypes, it did not have thick stem or tall stature. This result indicates the value of DNA markers to differentiate germplasm. Divergent germplasm with similar trait values may be useful to

breeders because it could provide additional alleles and/or alternate loci for the control of traits of interest.

Although we did not analyze any hybrids of known accessions, an accession PI422003 that was apparently derived by roughly equal hybridization among all three ecotypic groups had very desirable characteristics for biomass production: taller, thicker stems, late flowering, and high yield. This observation may support the suggestion that different ecotype pools represent heterotic groups for hybrid switchgrass production (Brummer, 1999; Martinez-Reyna and Vogel, 2008). Tropical maize germplasm heterotic groups developed using SSR markers were similar to those based on test cross and hybrid index (Aguiar *et al.*, 2008). Specific Combining Ability (SCA)-based heterotic groups were also consistent with DNA marker-based groupings in other experiments (de MC. Pinto *et al.*, 2003; Fan *et al.*, 2009). This suggests the grouping in our study could possibly represent the heterotic groups, and with high similarity of SSR and SNP markers in our study, any type of DNA based markers can be used.

The presence of population structure among this germplasm collection has implications for genetic mapping and for breeding. For genome-wide association studies, we will incorporate population structure into our statistical models to minimize false associations of markers and traits. For breeding programs, the clearly distinct genetic pools can be used in two main ways. First, while the upland/lowland heterotic grouping has been discussed previously, the presence of a bifurcation of the lowland ecotype into two major groups should be investigated further. Ploidy differences that are often present between the upland and lowland groups will not pose a challenge for inter-lowland group hybrids, which will all be tetraploid. Population hybrids between these groups should be

evaluated. Second, the two different lowland groups may represent distinct reservoirs of alleles for traits of interest, and careful selection and hybridization of germplasm from these two groups may result in significant improvement in some traits. Marker-assisted selection to transfer desired genes or QTL will further help this trait introgression.

Finally, the direct result of this experiment is the identification of individual genotypes that have desirable phenotypes, and these can be added to existing breeding programs directly. Because their genetic profile is known, possibly useful QTL alleles can be monitored, and their effect in hybrid progenies evaluated.

.

References

Aguiar, C. G., Schuster, I., Amaral Júnior, A. T., Scapim, C. A., & Vieira, E. S. N. (2008). Heterotic groups in tropical maize germplasm by test crosses and simple sequence repeat markers. Genetics and Molecular Research, *7*(4), 1233-1244.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403-10. DOI: 10.1016/S0022-2836(05)80360-2.

Bouton J. (2007) The economic benefits of forage improvement in the United States. Euphytica 154:263-270. DOI: 10.1007/s10681-006-9220-6.

Brummer, E. C. (1999). Capturing heterosis in forage crop cultivar development. Crop Science, 39(4), 943-954.

Brummer E.C., Cazcarro P.M., Luth D. (1999) Ploidy Determination of Alfalfa Germplasm Accessions Using Flow Cytometry. Crop Science 39:1202. DOI: 10.2135/cropsci1999.0011183X003900040041x.

Brunken J.N., Estes J.R. (1975) Cytological and Morphological Variation in Panicum virgatum L. The Southwestern Naturalist:379-385.

Casler M.D., Stendal C.A., Kapich L., Vogel K.P. (2007) Genetic diversity, plant adaptation regions, and gene pools for switchgrass. Crop Science 47:2261-2273. DOI: 10.2135/cropsci2006.12.0797.

Catchen J., Hohenlohe P.A., Bassham S., Amores A., Cresko W.A. (2013) Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124-3140. DOI: Doi 10.1111/Mec.12354.

Catchen J.M., Amores A., Hohenlohe P., Cresko W., Postlethwait J.H. (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3-Genes Genomes Genetics 1:171-182. DOI: Doi 10.1534/G3.111.000240.

Cortese L.M., Honig J., Miller C., Bonos S.A. (2010) Genetic Diversity of Twelve Switchgrass Populations Using Molecular and Morphological Markers, BioEnergy Research, Springer New York. pp. 262-271-271.

Costich D.E., Friebe B., Sheehan M.J., Casler M.D., Buckler E.S. (2010) Genome-size variation in switchgrass (*Panicum virgatum*): flow cytometry and cytology reveal rampant aneuploidy. Plant Genome 3:130-141. DOI: 10.3835/plantgenome2010.04.0010.

Das M.K., Fuentes R.G., Taliaferro C.M. (1997) Genetic Variability and Trait Relationships in Switchgrass:443-448.

de MC. Pinto, R., de Souza, C. L., Carlini-Garcia, L. A., Garcia, A. A. F., de SOUZA, A. P. (2003). Comparison between molecular markers and diallel crosses in the assignment of maize lines to heterotic groups. Maydica, *48*(1), 63-74.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Plos One 6. DOI: ARTN e19379

Evanno G., Regnaut S., Goudet J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14:2611-2620.

Fan, X. M., Zhang, Y. M., Yao, W. H., Chen, H. M., Tan, J., Xu, C. X. & Kang, M. S. (2009). Classifying maize inbred lines into heterotic groups using a factorial mating design. Agronomy Journal, *101*(1), 106-112.

Gunter L.E., Tuskan G.A., Wullschleger S.D. (1996) Diversity among populations of switchgrass based on RAPD markers. Crop Science 36:1017-1022.

Hopkins A.A., Taliaferro C.M., Murphy C.D., Christian D.A. (1996) Chromosome Number and Nuclear DNA Content of Several Switchgrass Populations. Crop Sci 36:1192-1195.

Hubisz M.J., Falush D., Stephens M., Pritchard J.K. (2009) Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources 9:1322-1332. DOI: 10.1111/j.1755-0998.2009.02591.x.

Huff D.R. (1997) RAPD Characterization of Heterogenous Perennial Ryegrass Cultivars. Crop Science 37:557.

Jin L., Chakraborty R. (1994) Estimation of genetic distance and coefficient of gene diversity from single-probe multilocus DNA fingerprinting data. Molecular Biology and Evolution 11:120-127.

Kölliker R., Stadelmann F.J., Reidy B., Nösberger J. (1999) Genetic variability of forage grass cultivars: A comparison of *Festuca pratensis* Huds., *Lolium perenne* L., and *Dactylis glomerata* L., Euphytica, Springer Netherlands. pp. 261-270-270.

Liu K., Muse S.V. (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics (Oxford, England) 21:2128-9. DOI: 10.1093/bioinformatics/bti282.

Lu F., Lipka A.E., Glaubitz J., Elshire R., Cherney J.H., Casler M.D., Buckler E.S., Costich D.E. (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution:

Novel Insights from a Network-Based SNP Discovery Protocol. Plos Genetics 9. DOI: Artn E1003215

Martinez-Reyna J.M., Vogel K.P. (2008) Heterosis in switchgrass: Spaced plants. Crop Science 48:1312-1320. DOI: 10.2135/cropsci2007.12.0695.

McLaughlin S.B., Adams Kszos L. (2005) Development of switchgrass (Panicum virgatum) as a bioenergy feedstock in the United States. Biomass and Bioenergy 28:515-535. DOI: 10.1016/j.biombioe.2004.05.006.

Mian M.A.R., Zwonitzer J.C., Chen Y.W., Saha M.C., Hopkins A.A. (2005) AFLP diversity within and among hardinggrass populations. Crop Science 45:2591-2597. DOI: Doi 10.2135/Cropsci2005.04-0029.

Missaoui A.M., Paterson A.H., Bouton J.H. (2005) Investigation of genomic organization in switchgrass (*Panicum virgatum* L.) using DNA markers. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 110:1372-83. DOI: 10.1007/s00122-005-1935-6.

Narasimhamoorthy B., Saha M.C., Swaller T., Bouton J.H. (2008) Genetic Diversity in Switchgrass Collections Assessed by EST-SSR Markers. Bioenergy Research 1:136-146. DOI: 10.1007/s12155-008-9011-0.

Okada M., Lanzatella C., Saha M.C., Bouton J., Wu R., Tobias C.M. (2010a) Complete switchgrass genetic maps reveal subgenome collinearity, preferential pairing and multilocus interactions. Genetics 185:745-60. DOI: 10.1534/genetics.110.113910.

Okada M., Lanzatella C., Tobias C.M. (2010b) Single-locus EST-SSR markers for characterization of population genetic diversity and structure across ploidy levels in switchgrass (*Panicum virgatum* L.). Genetic Resources and Crop Evolution. DOI: 10.1007/s10722-010-9631-z.

Paradis E. (2010) pegas: an R package for population genetics with an integrated, modular approach. Bioinformatics 26:419.

Poland J.A., Brown P.J., Sorrells M.E., Jannink J.L. (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. Plos One 7. DOI: ARTN e32253

Porter C.L. (1966) An Analysis of Variation Between Upland and Lowland Switchgrass, *Panicum Virgatum* L., in Central Oklahoma. Ecology 47:980 - 992. DOI: 10.2307/1935646.

Pritchard J.K., Stephens M., Donnelly P. (2000) Inference of population structure using multilocus genotype data. Genetics 155:945.

R Development Core Team. (2011) R: A Language and Environment for Statistical Computing, Vienna, Austria.

Sanderson M., Reed R., Mclaughlin S., Wullschleger S., Conger B., Parrish D., Wolf D., Taliaferro C., Hopkins A., Ocumpaugh W. (1996) Switchgrass as a sustainable bioenergy crop. Bioresource Technology 56:83-93. DOI: 10.1016/0960-8524(95)00176-X.

Schuelke M. (2000) An economic method for the fluorescent labeling of PCR fragments. Nature biotechnology 18:233-4. DOI: 10.1038/72708.

Sedcole J. (1977) Number of plants necessary to recover a trait. Crop Science 17:667-668.

Serba D., Wu L.M., Daverdin G., Bahri B.A., Wang X.W., Kilian A., Bouton J.H., Brummer E.C., Saha M.C., Devos K.M. (2013) Linkage Maps of Lowland and Upland Tetraploid Switchgrass Ecotypes. Bioenergy Research 6:953-965. DOI: Doi 10.1007/S12155-013-9315-6.

Stekhoven D.J., Buhlmann P. (2012) MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics 28:112-118. DOI: Doi 10.1093/Bioinformatics/Btr597.

Ubi B.E., Kolliker R., Fujimori M., Komatsu T. (2003) Genetic diversity in diploid cultivars of rhodesgrass determined on the basis of amplified fragment length polymorphism markers. Crop Science 43:1516-1522.

Wright L. (1994) Production technology status of woody and herbaceous crops. Biomass and Bioenergy 6:191-209. DOI: 10.1016/0961-9534(94)90075-2.

Young H.A., Lanzatella C.L., Sarath G., Tobias C.M. (2011) Chloroplast Genome Variation in Upland and Lowland Switchgrass. Plos One 6. DOI: ARTN e23980

Zalapa J.E., Price D.L., Kaeppler S.M., Tobias C.M., Okada M., Casler M.D. (2011) Hierarchical classification of switchgrass genotypes using SSR and chloroplast sequences: ecotypes, ploidies, gene pools, and cultivars. Theoretical and Applied Genetics 122:805-817. DOI: 10.1007/s00122-010-1488-1.

Zhang Y., Zalapa J., Jakubowski A.R., Price D.L., Acharya A., Wei Y., Brummer E.C., Kaeppler S.M., Casler M.D. (2011a) Natural hybrids and gene flow between upland and lowland switchgrass. Crop Science 51.

Zhang Y.W., Zalapa J.E., Jakubowski A.R., Price D.L., Acharya A., Wei Y.L., Brummer E.C., Kaeppler S.M., Casler M.D. (2011b) Post-glacial evolution of Panicum virgatum: centers of diversity and gene pools revealed by SSR markers and cpDNA sequences. Genetica 139:933-948. DOI: Doi 10.1007/S10709-011-9597-6

Table 2.1: Accessions used in diversity study with their location, ecotype and morphology.

ID	Name	Location	Latitude	Longitude	Genotypic group[1]	Cytotype[2]	Plant height	Flowering time	Stem diameter
							-cm-	-Julian Days-	-mm-
1	PI 315723	Hoffman, NC	35.04	-79.56	L-A:L-B	L	96	180	43
2	PI 315724	Ellsworth, KS	38.73	-98.23	U	U	67	164	23
3	PI 315725	Coffeeville, MS	33.98	-89.68	L-A	U	99	189	44
4	PI 315727	Apex, NC	35.75	-78.84	L-A	L	57	157	24
5	PI 315728	Scotland County, NC (donated by Maryland)	34.79	-79.55	L-A	L	54	165	29
6	PI 337553	Rafaela Experiment Station, Argentina	-31.18	-61.55	U	U	67	166	24
7	PI 414065	Pangburn, AR	35.42	-91.84	L-A	L	121	178	55
8	PI 414066	Grenville, NM	36.74	-103.46	U	U	62	151	27
9	PI 414067	Soil Conservation Service, NC	35.84	-78.63	U	U	74	170	30
10	PI 414068	Soil Conservation Service, KS	38.83	-97.62	U	U	73	162	24
11	PI 414070	Soil Conservation Service, KS	38.83	-97.62	L-A	L	118	186	52
12	PI 421138	Moore County, NC	35.34	-79.36	L-A	U	79	178	32
13	PI 421520	Kay County, OK	36.8	-97.29	U	U	72	163	24
14	PI 421521	Wetumka, OK (developed in KS)	35.24	-96.24	L-A	L	126	185	51

15	PI 421999	Pangburn, AR	35.42	-91.84	L-A	L	131	175	50
16	PI 422001	Stuart, Martin County, FL	27.2	-80.25	L-B	L	80	191	43
17	PI 422003	FL	28.57	-82.38	L-A:L-B:U	L,U	99	194	45
18	PI 422006	George West, TX	29.29	-98.73	L-A	L	119	188	51
18.1	Unknown	Unknown	NA	NA	U	U	78	162	27
19	PI 422016	FL	28.57	-82.38	L-B	L	96	172	41
20	PI 431575	Raleigh County, WV (from KY)	37.79	-81.19	U	U	61	177	32
21	PI 476290	Wilmington, NC	33.92	-78.13	L-A	L	98	170	41
22	PI 476291	MD	39.04	-76.91	L-A,U	L,U	92	167	32
23	PI 476292	Franklin County, AR	36.23	-91.75	U	L,U	65	168	29
24	PI 476293	Heislerville, NJ	39.25	-74.99	U	L	60	161	26
25	PI 476294	Eads, CO	38.46	-102.65	U	U	54	154	26
26	PI 476295	Colorado Springs, CO	38.79	-104.83	U	U	41	156	21
27	PI 476296	MD	39.04	-76.91	U	U	57	153	23
28	PI 642190	NM	34.78	-106.69	U	U	45	147	20
29	PI 642191	SD	44.08	-103.17	U	U	56	156	23
30	Citrus Co-FL	Citrus County, FL	28.75	-82.53	L-B	L	45	178	23
31	HSP-FL	Hillsborough River State Park, FL	28.06	-82.3	L-B	L	123	187	52
32	OSSP-FL	Oscar Scherer State Park, FL	27.18	-82.49	L-B	L	84	204	40
33	Pasco Co-FL	Pasco County, FL	28.36	-82.21	L-B, U	L,U	60	175	29

34	SNF	Sumter National Forest, SC	34.22	-82.16	L-B	L	126	191	53
35	SPBluff	Sprewell Bluff, GA	32.91	-84.33	L-B:U	U	58	195	35
36	SWFWMD-FL	Southwest Florida Water Management District, FL	28.58	-82.17	L-B	L,U	44	187	22
38	GRIF16964	Florida	30.49	-81.62	U	U	NA	NA	NA
40	GRIF16967	Florida	29.6	-81.11	L-B	L	NA	NA	NA
41	GRIF16968	Florida	29.59	-81.1	L-B	L	NA	NA	NA
42	GRIF16969	Florida	29.28	-81.04	L-B	L	NA	NA	NA
46	GRIF16982	Florida	30.26	-84.03	L-B	L	NA	NA	NA
49	GRIF16991	Florida	29.72	-85.4	U	U	NA	NA	NA
57	JDSPI	NA	NA	NA	L-A	L	NA	NA	NA
58	GRIF16570	NA	NA	NA	L-B	L	NA	NA	NA
59	Rotstrahlbusch	NA	NA	NA	L-B	L	NA	NA	NA
60	KC/SC	NA	NA	NA	L-A:U	L	NA	NA	NA
63	Hanse Herms	NA	NA	NA	U	U	NA	NA	NA
66	Heavy Metal	NA	NA	NA	L-A	L	NA	NA	NA
68	Tyclo/MS	NA	NA	NA	U	U	NA	NA	NA
98	AP13	Derived from Alamo	29.29	-98.73	L-A	L	119	175	47
99	VS16	Derived from Summer	44.08	-103.17	U	U	52	160	24

1. Genotypic group based on SNP markers. L-A, Lowland-A; L-B, Lowland-B; U, Upland. Separated by ":" hybrids, separated by ",", mixture of individuals in accesion
2. Cytotype based on nine chloroplast specific markers. L, Lowland; U, Upland

Table 2.2: List of nuclear SSR markers and their source used in the study

Marker	Forward	Reverse	Source
SWW918	GCAAGATGGGAAGACA	GCTTGACATTGTTGAAG	(Okada *et al.*, 2010a)
SWW620	ACCTCAAGTGGGTGTCC	CTTATCTCCCTCGGTAC	(Okada *et al.*, 2010a)
SWW303	ATCCTCTCCTTCCTCATC	GGTAGTAGTGCTCCACG	(Okada *et al.*, 2010a)
SWW301	CCTCTCCTGCCTTTTAA	CTCTTCTGTGCCATGAA	(Okada *et al.*, 2010a)
SWW298	ACTGAATCATTCGTCTT	AATGGAGTAAGAGCGA	(Okada *et al.*, 2010a)
SWW279	AAAACCAAGGCGTGTG	GCTGTCCGGGTT	(Okada *et al.*, 2010a)
SWW277	TTGAAAATGCACGCCAA	CACGAAGCCCCAACAGT	(Okada *et al.*, 2010a)
SWW271	TTGCACAGCCAACCAAT	CTTGGCTTGAGGCTCTG	(Okada *et al.*, 2010a)
SWW256	CAAATCGCAGTTCGGTA	CTTTGCAGCAAGCAAGA	(Okada *et al.*, 2010a)
SWW240	CAACTAATCGCCACCTC	CTTGGGAGCGGAAGAG	(Okada *et al.*, 2010a)
SWW170	TGGCCTTAGTTTCAGGT	CTTTGGGCTATGTGTGG	(Okada *et al.*, 2010a)
SWW141	AACCAAACCTATGCACA	GATCAAGGACAAGTGC	(Okada *et al.*, 2010a)
SWW114	GAGAAGAACGCTCTCCC	CTTTTCTGATGGTTATTC	(Okada *et al.*, 2010a)
SWW106	ACCTACGGCCCCATCAG	GGCCGTTGATCAGGATG	(Okada *et al.*, 2010a)
SWW127	CTCCCCTACCGCCTCCG	ACTCGGGATGGTGATGA	(Okada *et al.*, 2010a)
UGSW37	GTTTGCTCCTTTTCTCCC	TACTCCCCTCATTCTCAT	(Serba *et al.*, 2013)
UGSW33	CCAAACACTCACCCTCA	CGGTGGTTTACTGGTGA	(Serba *et al.*, 2013)
UGSW31	AAACCCTGGGCTATTCA	GGCTGTTAACACAGGCA	(Serba *et al.*, 2013)
UGSW29	AGAGAAGGAGGGAGTG	GTACTTGTAACCCACGG	(Serba *et al.*, 2013)
UGSW25	TTAAAAACCTCCCCGAA	GAAAGAAAGGCAGTTG	(Serba *et al.*, 2013)
UGSW17	CAGGTCCTGGAAGCTCA	ACAAAAGTTAATTGCCG	(Serba *et al.*, 2013)
UGSW13	ACGTTCGCCATCATCAA	CACCTCAAGCTACCTAC	(Serba *et al.*, 2013)
UGSWP6	CGTGCTGCTCTGTTTTCT	TTGTCTTTATCGACCCG	(Serba *et al.*, 2013)
UGSWP4	GACTTAGCTGTCTCTCG	GTGTAGGGGTGGCGTTG	(Serba *et al.*, 2013)
UGSWP2	GTCCCTTTTCAACACAC	AAGGTGGCGGGTTATAT	(Serba *et al.*, 2013)
UGSW26	AACCGTGGCAAATCAA	TGAAATTTTAACTCCGC	(Serba *et al.*, 2013)
NFSG387	AAGGAATCATTGCTCGC	GCAGCCTTATATTTGAT	(Serba *et al.*, 2013)
NFSG377	GTATCTCTTGCTGCCCA	AATATTGCGACCAAGAT	(Serba *et al.*, 2013)
NFSG293	CAAGCCGCCAAGACAG	ACGGAGTTCTAGAAGCC	(Serba *et al.*, 2013)
NFSG274	TCCTCTCCCATCTTCCTT	ATCAAGAGGGTTTGGAT	(Serba *et al.*, 2013)
NFSG252	TTCACACTCACAAGGAT	CATGTGATGTTGCTCTT	(Serba *et al.*, 2013)
NFSG137	CGTACACCTGATCCAAA	CTTGTCCATTGCTTCATC	(Serba *et al.*, 2013)
NFSG107	ATTCCCTCCCTCTACTC	TTGTACGGAAGGGCGA	(Serba *et al.*, 2013)
NFSG050	CCCTTCTCATAAAAGAA	ACCAGGATTGTCTTTCT	(Serba *et al.*, 2013)
NFSG026	CCTTCATAGTCAAATTG	TGTACCACTATTGAGGC	(Serba *et al.*, 2013)

Table 2.3: List of chloroplast indelmarkers and their source used in the study

Marker	Forward	Reverse	Allele Size[1]	Source
trnL-UAA	CGAAATCGGTAGACGCTACG	GGGGATAGAGGGACTTGAAC	575,619,580	Missaoui, 2005
Rps16-psbk	ACAGATCGAGATCGTTTTTGC	TGAGACCTATCCCTTTATGATCG	277,301,252	Young, 2012
rpoC2	GGGGTTCAGGAATTGTGAAAT	CCTATGTTCATTCTCTGTGCTCTC	146,163,128,133,	Young, 2012
rps4-ndhJ-a	TGCAGAGACTCAATGGAAGC	TCCTCGTTCGATTAATCCACTT	140,191,221 345,366	Young, 2012
rps4-ndhJ-b	GAAAAGGGCTAAAATCTCTGGTT	GTACCGCGCGGATTACTTAG	166,187, 144,208,229	Young, 2012
ndhC-atpE	TGAACCGACTGTTTGTCAGG	CCACAAAAGAAGCCCCATTA	298,348	Young, 2012
rbcL-psaI	GCGATGAGAATGGGAAAAGA	TTGCAATTGCCGGAAATACT	386,360,391	Young, 2012
psbE-petL	TTCCGTAAAAGATGGGATCG	GGGGTTCTATTGATGCCTTG	276,301,252	Young, 2012
ndhF-rpl32	AATAAAGGAGCTCTCTTGTTTCGT	TGGGGGATAAGCCTCCATA	289,308,303,310	Young, 2012

[1] The first and second allele size were lowland and upland type identified in the source papers

Table 2.4: Mean separation of morphological traits based on genetic cluster

Group	Ecotype	Flowering time		Plant Height		Stem Diameter	
		Mean (Julian Days)	Group	Mean (cm)	Group	Mean (cm)	Group
1	Upland	160.87	c	62.93	c	2.48	d
2	Lowland A	185.33	a	63.26	c	3.17	c
3	Lowland B	183.11	a	122.42	a	5.12	a
4	Mixed	177.30	b	85.34	b	3.83	b

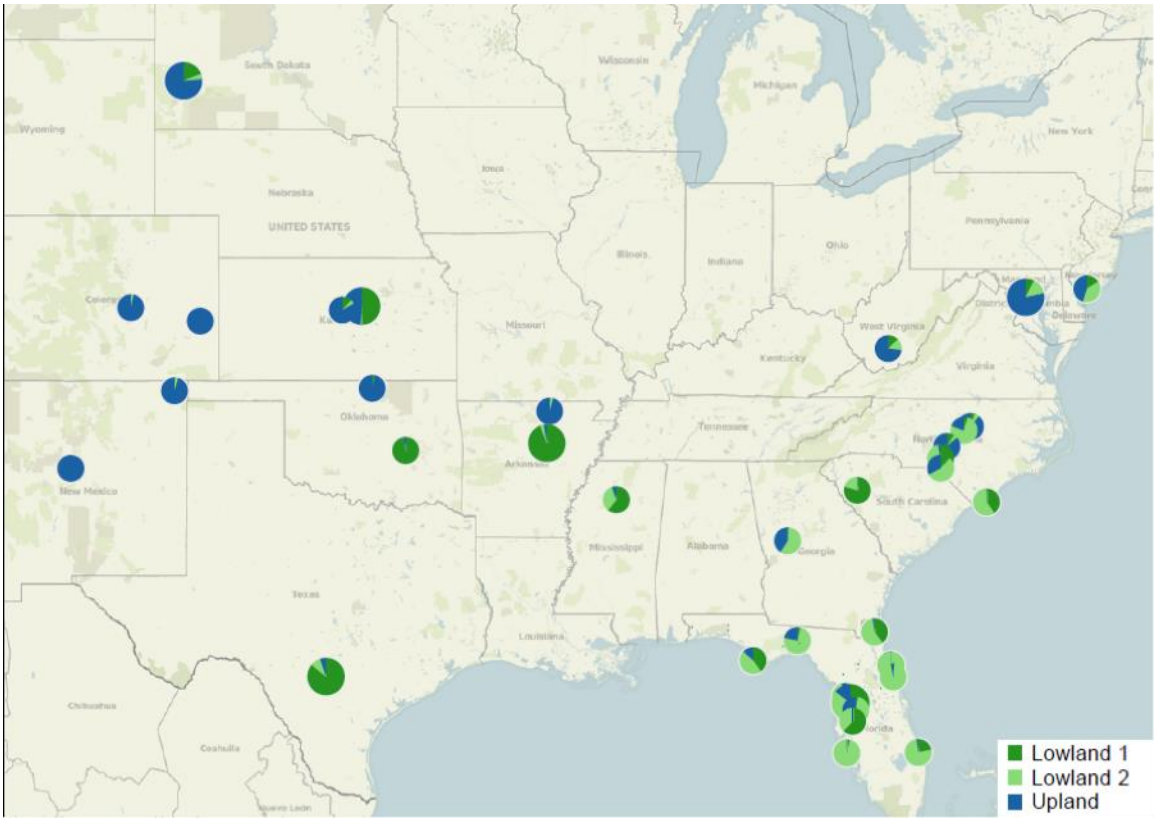
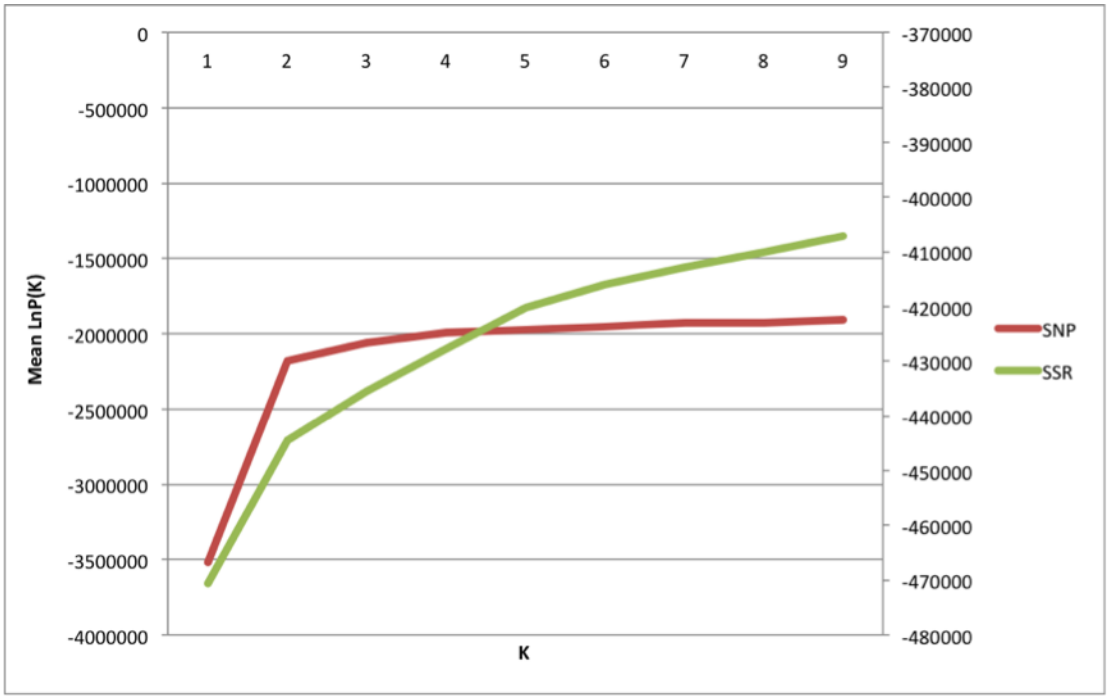
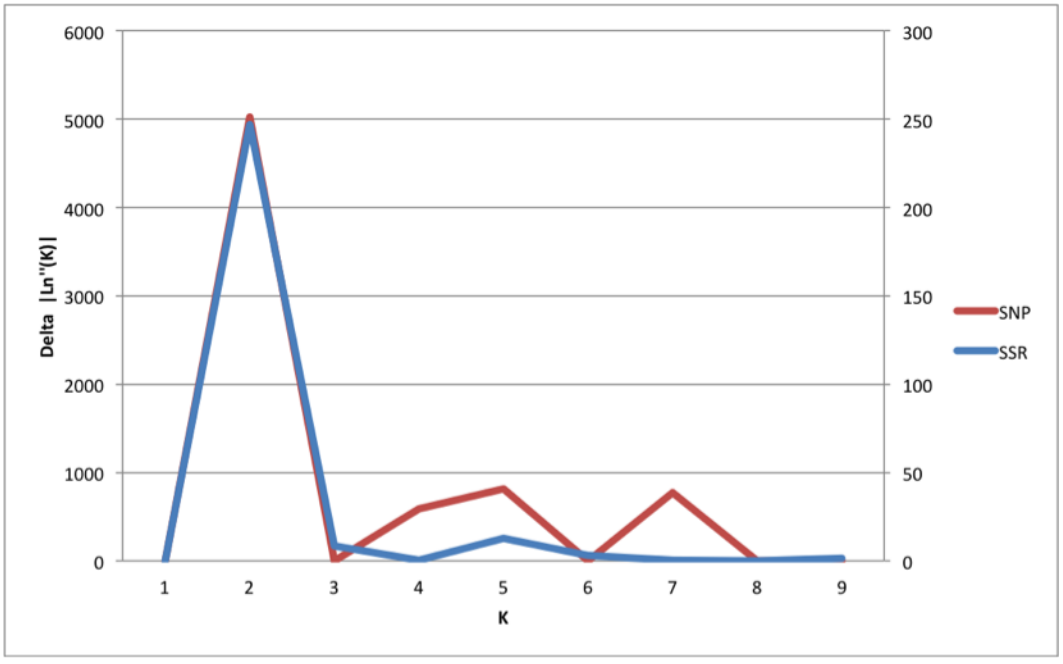


Fig2.1. The accessions used for this study and their origin. The colors represent their membership to the group after the SNP analysis. Green represents southern Lowland A, light green represents Lowland B, and blue represents Upland.

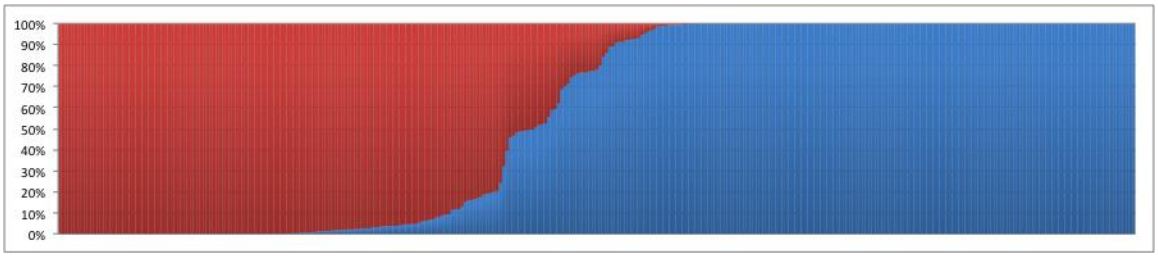


(a)

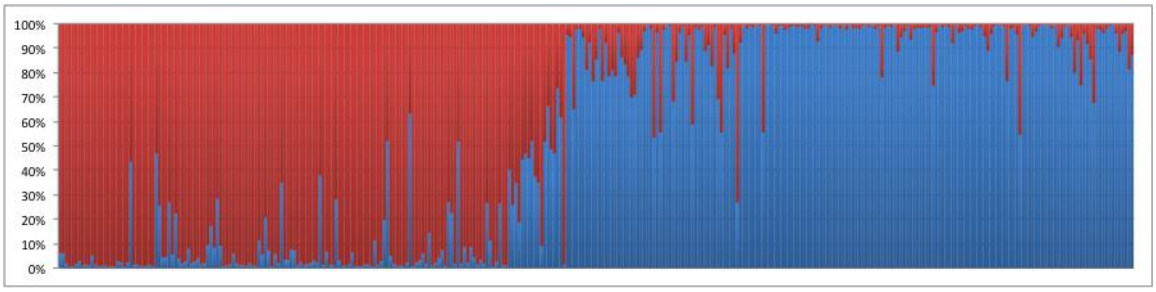


(b)

Fig 2.2. The plot probability of likelihood with subgroups (K) of switchgrass accessions as depicted from structure analysis. a) raw likelihood b) change as described by Evanno *et al.* The best number of cluster seems to be 2, 5 or 7.



a



b

Fig 2.3. The graphical representation of 372 switchgrass genotypes from structure analysis with k=2 a) with SNP b) with SSR. Red color represents upland population origin and blue color represents lowland origin. The genotypes are in same order.

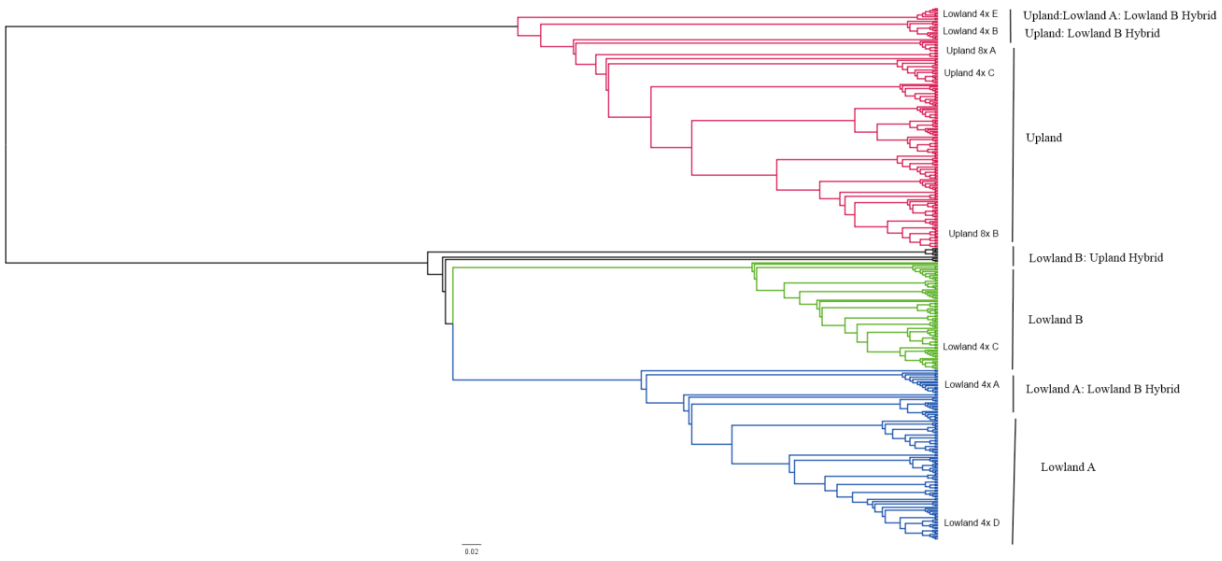
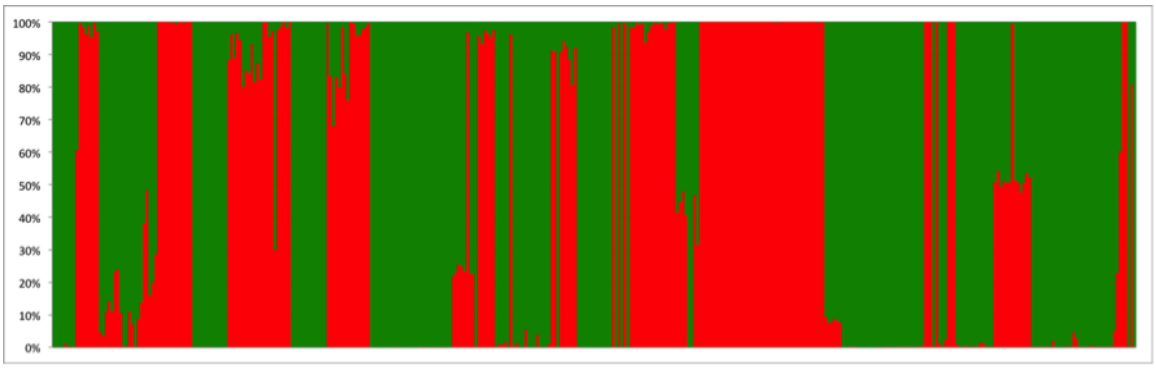
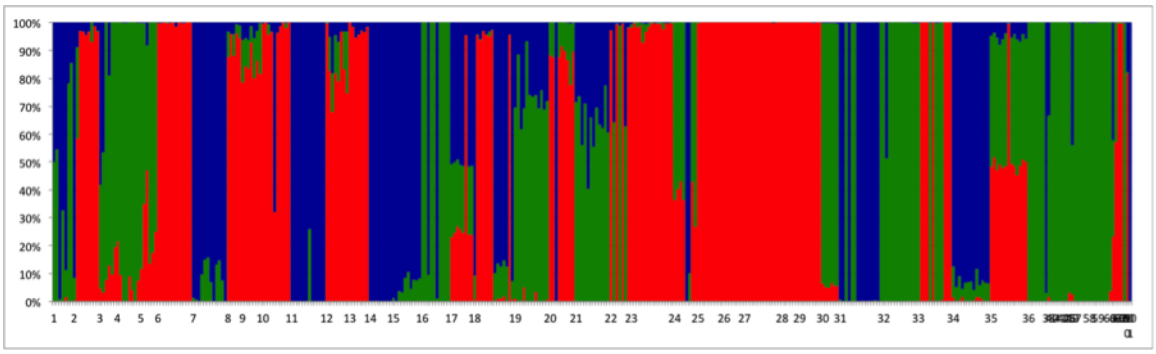


Fig 2.4. The neighbor joining tree of the genotypes used with shared allele genetic distance with SNP markers. Blue represents southern Great Plains lowland (Lowland A), green represents Gulf Coast lowland (Lowland B), and red represents Upland. Black colored were hybrids of Upland and Lowland B. The first set of labels (second column from right) represent the grouping from a previous study (Zhang *et al.*, 2011) and rightmost labels represent our classification.

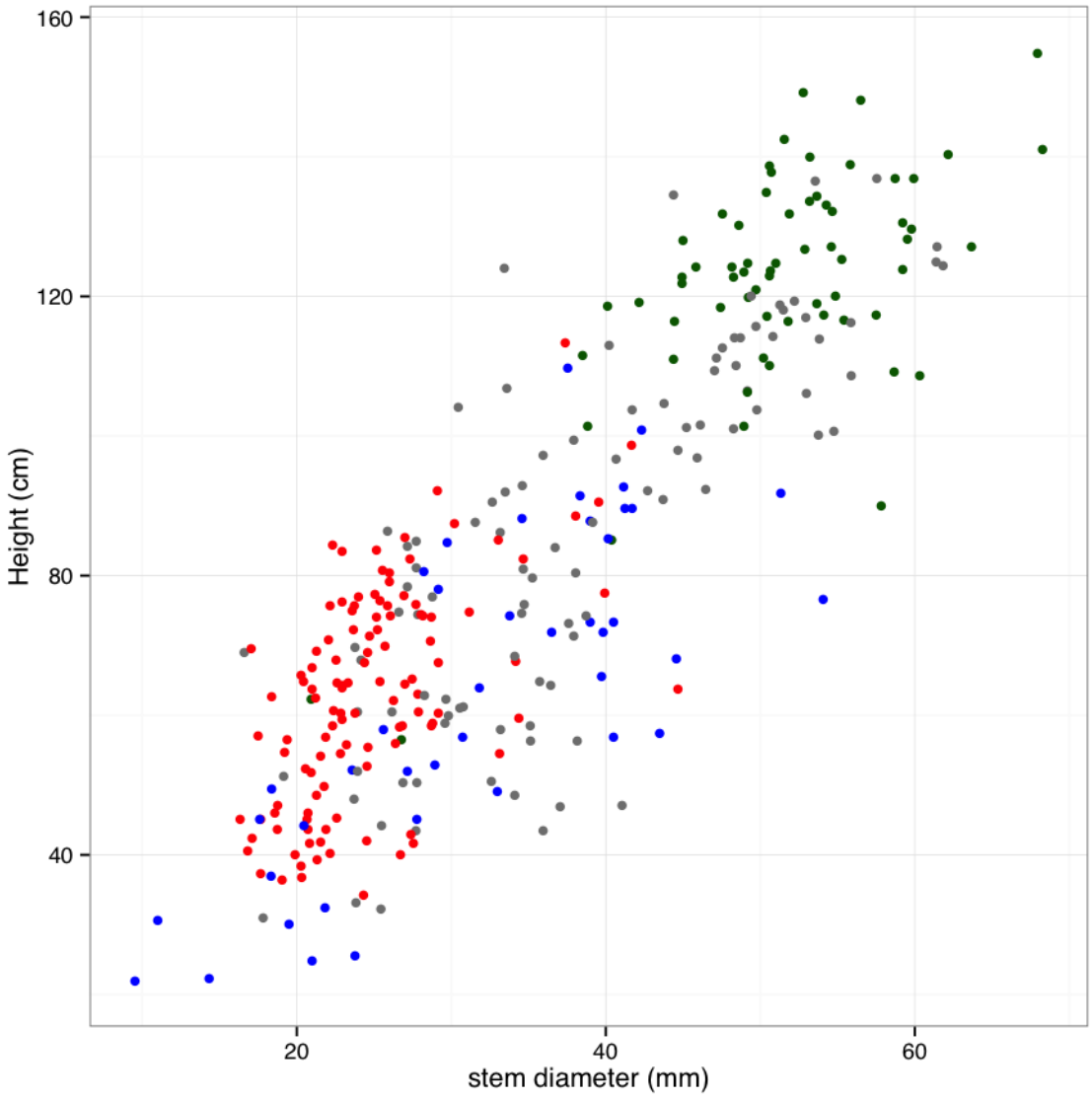


(a)

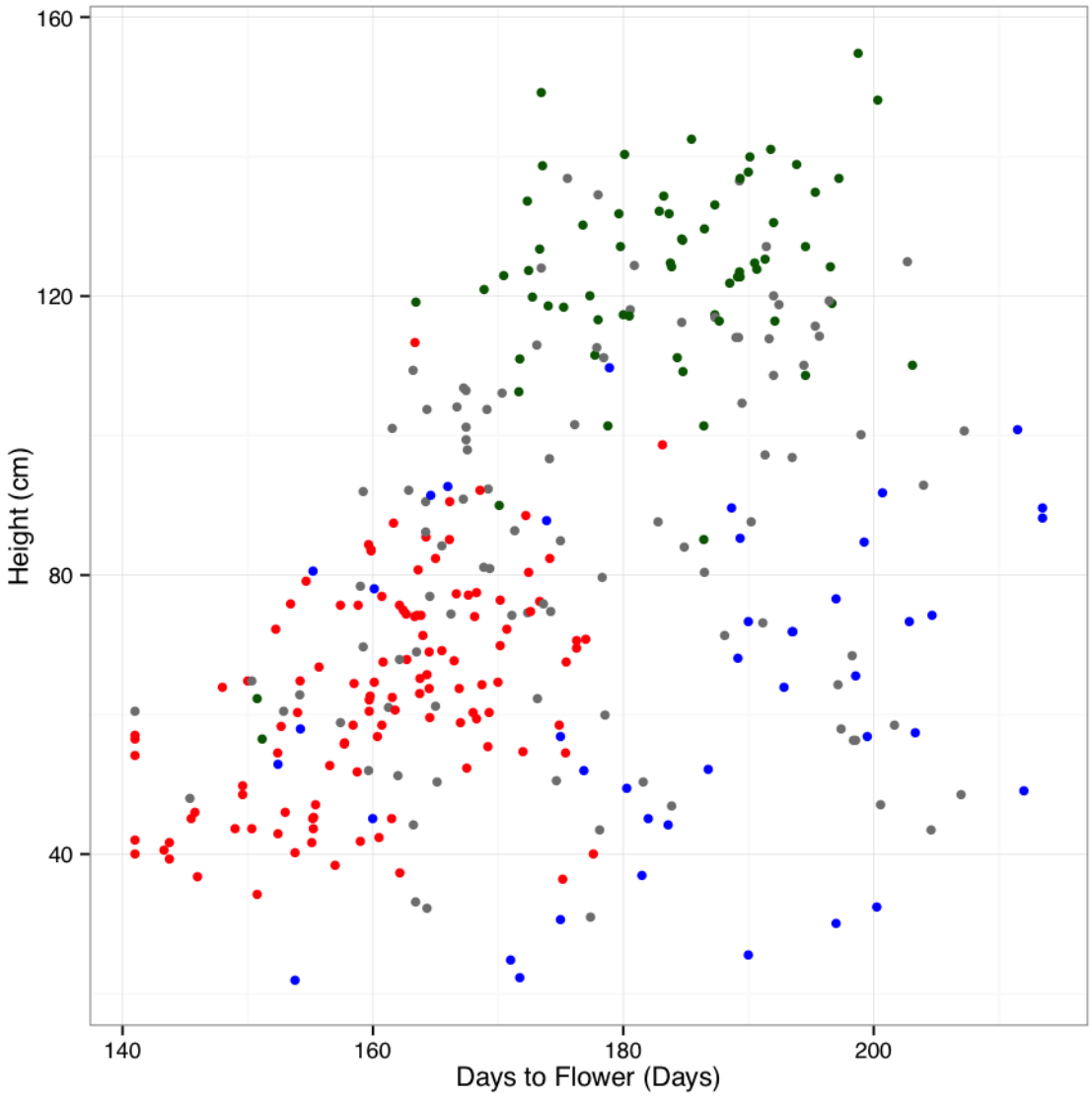


(b)

Fig 2.5. The graphical representation of 372 switchgrass genotypes from structure analysis with nuclear SNP. Numbers in the x-axis represent the accessions as in Table 1. a) k=2 and Red color represents upland population origin and green color represents lowland origin. b) when k=3, the lowland type is further divided into two lowland subgroups blue (Lowland A) and green (Lowland B).



(a)



(b)

Fig 2.6. Distribution of individuals from 36 accessions with relation to phenotype. a) height in y-axis and stem diameter on x-axis grouped by genetic group. Red: Upland, Blue: Lowland A, Green: Lowland B and Grey: hybrids of any two groups. b) height in y-axis and flowering time on x-axis grouped by genetic group. Red: Upland, Blue: Lowland A, Green: Lowland B and Grey: hybrids of any two groups.

CHAPTER 3

GENOME WIDE ASSOCIATION ANALYSIS OF BIOMASS YIELD IN SWITCHGRASS (*PANICUM VIRGATUM* L.) [2]

[2]Acharya, A.R,, Y. Wei, M. Saha, K.M. Devos and E.C. Brummer. To be submitted to *Theoretical and Applied Genetics*

Abstract

Biomass yield is one of the most important traits for an energy crop such as switchgrass (*Panicum virgatum* L.). Understanding the genetic architecture of biomass yield and yield components, such as plant height, stem diameter, or flowering time, could accelerate breeding progress. Genome-wide association mapping studies (GWAS) are one of the methods to identify genetic markers associated with complex traits. In this study, we identified over 65,000 SNPs from genotyping-by-sequencing in a population of 352 diverse switchgrass genotypes. We used a subset of 3,196 SNPs that could be genotyped in at least 80% of the populations examined and another subset of about 20,000 SNPs to identify marker-trait associations for biomass yield, plant height, stem thickness, and days to flower, based on phenotypic data generated in Georgia and Oklahoma over three years. The repeatability estimates were 0.72 for yield, 0.88 for height, 0.90 for stem diameter, and 0.68 for days to flower. We used a mixed model to account for both structure and kinship in our population. After correction for false discovery, we identified more than 50 SNPs associated with the four investigated traits. Some SNPs showed the association only one of the locations or years suggesting genotype-by-environment interaction. Using the larger SNP dataset, which included more missing data, we were able to identify all the associations that were identified with the smaller dataset as well as additional marker-trait associations. Some of the loci we identified as associated with these traits showed similarity to genes linked to similar traits identified in other related species. This is the first association study in switchgrass, and although the number of environments examined was limited, it offers a starting point for the application of markers to switchgrass breeding programs.

Introduction

Dedicated bioenergy feedstock will contribute a renewable energy source to help meet the world's increasing demand for energy in an environmentally sensible manner. The Bioenergy Feedstock Development Program (BFDP) in the U.S. Department of Energy began evaluating a wide variety of potential feedstock in 1978 (McLaughlin and Adams Kszos, 2005). Among herbaceous feedstock, switchgrass (*Panicum virgatum* L.) was identified as the most promising target to develop as a bioenergy crop (Wright, 1994). Switchgrass has many characteristics of a desirable biofuel feedstock because it is perennial, has high productivity is adapted to a wide variety of sites, including to poor soil conditions (Sanderson *et al.*, 2006) and is familiar to farmers as it has been developed as forage species for a long time. In addition, switchgrass is a native grass of North America.

A successful crop improvement program requires genetic variation. Genetic diversity within switchgrass germplasm varies depending on the populations studied (Cortese *et al.*, 2010; Gunter *et al.*, 1996; Narasimhamoorthy *et al.*, 2008; Zhang *et al.*, 2011). In our previous study (Acharya *et al*., previous chapter), there is more variation within population than among which is also true for the above mentioned studies. All of these studies have identified distinct clusters of upland and lowland ecotypes with variation within populations higher than the variation among populations. However we identified two distinct groups within lowland ecotypes too. With high number of markers, apart from three groups (two lowlands and one upland), the accessions were divided in respect to geographical origin. Such a clustering of accessions helps in the plant breeding. The parents of breeding program can be diversified to include the gene pool of different groups. Or in the contrary, for a breeding program focused in a geo-ecological region, the accession adapted in that region can be selected as a start of a selection program.

The true value of this diversity lies in being able to identify important alleles for key traits in the germplasm and then using them effectively in the breeding program. Identifying quantitative trait loci (QTL) for important traits will enable breeders to manipulate specific chromosome segments during the breeding process. Traditionally, QTL were identified using genetic linkage maps constructed in bi-parental populations. In recent years, association mapping (also known as linkage disequilibrium (LD) mapping) has been done using diverse germplasm or breeding populations (Gupta *et al.*, 2005; Honsdorf *et al.*, 2010). Association mapping enables mapping QTL with high resolution because of the higher amount of historical recombination that has occurred in these populations compared to an F_1 or F_2 bi-parental population (Ewens and Spielman, 2001; Jannink *et al.*, 2001). The key to successful LD mapping is having a high density of genetic markers to adequately cover the genome. With the accessibility of large amounts of DNA sequence from multiple individuals within a species, single nucleotide polymorphisms (SNP) are easily identified and can be developed into high-throughput marker assays. Genotyping-by-sequencing further enables the detection of SNPs without pre-identification and development of assay (Poland and Rife, 2012). With their high resolution, low mutation rate, and suitability for high-throughput systems (Zhu *et al.*, 2008), SNP are the marker of choice for most mapping applications.

Association mapping is typically conducted using a mixed model statistical analysis to associate phenotypes and genotypes (Stich *et al.*, 2008;Yu *et al.*, 2006). The structure of the population due to the presence of subpopulations and the kinship of the individuals being assayed need to be controlled in the analysis to avoid false positive associations (Yu *et al.*, 2006). Association mapping in a highly structured population will probably limit the detection of

rare variants fixed within subpopulations (Breseghello and Sorrells, 2006) and misses some true positives where the real association is nested within the structure (Brachi *et al.*, 2011).

Yield is one of the major traits of any biomass breeding program. Switchgrass shows a high variation in yield especially between ecotypes (Cassida *et al.*, 2005; Lemus *et al.*, 2002; Boe and Lee, 2007). Tulbare *et al.* (2012) estimated the genetics being 5th most important variable contributing to the yield after nitrogen fertilizer, age, climate and soil types. Within the genetic contribution, the contribution due to cytotype (or ecotype as they are correlated), is larger than the accession within a cytotype. Similar result was found in the study by Wullschleger *et al.* (2010), where the yield between ecotypes was significant with lowland yielding higher. The accessions within each groups also showed the significant variation. In both lowlands and uplands, the distribution of yield was skewed to left with a long tail. Understanding the genetic architecture of quantitative traits like yield will have a huge impact in marker assisted selection and genomic selection. The heritability estimates of yield are varied from very low to moderately high (Bhandari *et al.*, 2010, Rose *et al.*, 2008). For a low heritable trait, selection based on secondary trait has been successful (Hansen *et al.*, 2005). Several traits affecting the biomass yield in switchgrass has been identified both in spaced plant nurseries (Boe and Deck, 2008; Das *et al.*, 2004) and swards (Price and Casler, 2014). Few morphological traits directly affecting the biomass (or yield) are plant height, tiller diameter and tiller density and the length of vegetative growth period with height showing the highest correlation in all studies. Elongated leaf height and canopy height explained >91% and >82% of variation in switchgrass biomass (Schmer *et al.*, 2010). Bhandari *et al.* (Bhandari *et al.*, 2011; Bhandari *et al.*, 2010) also reported the high positive correlation of biomass yield with plant height, tiller diameter and days to flowering. All of which showed higher heritability than biomass yield itself. The objective of this study is to

identify QTLs associated with dry biomass yield and related vegetative traits (plant height, stem diameter and days to flowering time of switchgrass

Materials and Methods

Plant materials:

The germplasm we analyzed in this experiment largely derived from the southern half of the US, with a few exceptions (Table 3.1). We included 29 accessions from the National Plant Germplasm System (NPGS), seven populations we collected from Florida, Georgia and South Carolina, six populations recently collected from Florida by the NPGS, and two genotypes that are the parents of a genetic mapping population (Missaoui *et al.*, 2005), AP13 derived from 'Alamo' and VS16 derived from 'Summer'. The populations included both upland and lowland ecotypes and some accessions with intermediate phenotypes. Each accession was represented by between one and 16 plants (i.e., each having a distinct genotype) for a total of 511 genotypes evaluated in at least one field location. Of the 511 genotypes, 413 were included at both locations, 67 only in Athens and 67, including 31 newly collected genotypes, only in Ardmore. A subset of these genotypes was ultimately used for association analysis because SNP data were only available for 352 individuals (see below).

Field planting and phenotype data collection:

Genotypes were clonally propagated from individual ramets in the greenhouse. Field plots were established at the University of Georgia Plant Sciences Farm near Watkinsville, GA in July 2009 and at the Noble Foundation Research Park Farm near Ardmore, OK in May 2011. At each location, a 16 × 30 α-lattice design with 3 replications was planted. Each replication consisted of 16 genotypes in each of 30 incomplete blocks for a total of 480 genotypes at each

location. In Ardmore, we divided the genotypes into two sets based on the height data from Watkinsville, and planted a sets within reps design. The clones were separated by 90 cm within and between rows. Each genotype was represented by a single clone in each replication, for a total of 1440 plants (plots) at each location. At Watkinsville, weeds were controlled by spraying Atrazine at 4.6 liters/hectare in February and voluntary switchgrass in following years were removed by manually hoeing the plot area. Nitrogen was applied at 56 kg/hectare in April of each year. At Ardmore, Prowl H_2O at10.2 liters/hectare, Charger Max at 1.5 liters/hectare and 2,4 D amine at 4.6 liters/hectare was applied to control weeds and voluntary switchgrass in following years were removed by manually hoeing. Nitrogen was applied at 112 kg/hectare for each year and phosphorus was applied at 168 kg/hectare on first year.

No data were taken during the first year to allow the plants to fully establish. At the end of the establishment year, when plants had fully senesced, we removed all above ground biomass at a height of 10 cm. Beginning in the second year, we measured biomass yield on each plant after all plants in the experiment had fully senesced and/or been killed by freezing temperatures. We harvested each plant at 10 cm from the soil surface using a sickle bar harvester, measured the fresh weight of the entire plant in field, and subsampled two to three entire tillers from each plant. Subsamples were weighed immediately after harvest, dried for a week at 50 °C, and weighed dry. The dry weight of each plant in the field was computed based on the dry matter percentage of the sample. Biomass was harvested in Watkinsville on Feb 10, 2011; Feb 7, 2012 and Dec 20, 2012 and in Ardmore on Jan 7, 2013. The height was measured after full maturity and before harvest each year. Stem diameter was measured after full maturity and before harvest at Watkinsville location in 2011 and 2012. Height was measured on three tillers per plant from the ground to the uppermost node of a flowering stem. Stem diameter was measured on three

tillers 5 cm from the ground. For flowering date, we recorded the date when at least three tillers showed emergence of the inflorescence.

Genotyping:

We obtained SNP data from 352 of the genotypes grown in the field as described previously (Acharya *et al.*, Chapter 2). Briefly, we extracted the DNA from young leaves of switchgrass using the CTAB method (Doyle and Doyle, 1990) and quantified the concentration using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Inc., (http://www.nanodrop.com). DNA concentrations for all samples were normalized to 20 ng/ul and concentrations confirmed with PicoGreen. Single nucleotide polymoprhisms (SNP) were identified using the two-enzyme Genotyping-by-Sequencing (GBS) method described by Poland *et al.* (Poland *et al.*, 2012) except that we used *Fse*I instead of *Pst*I to reduce the number of fragments generated. Following digestion, sample barcodes, *Fse*I adaptors, and the common adapter were ligated prior to sequencing on an Illumina HiSeq 2000 sequencer at the University of Texas, Austin sequencing facility (http://www.icmb.utexas.edu/core/DNA/), with 48 genotypes multiplexed per sequencer lane. Sequencing resulted in an average of about four million raw sequencing reads per genotype.

We first analyzed the sequence data using the STACKS software (Catchen *et al.*, 2013; Catchen *et al.*, 2011) as described previously (Acharya *et al.*, 2014). We identified 65,328 SNP across all genotypes using the STACKS pipeline. Using the SNP identified in this overall analysis, we then developed the marker dataset used for association analysis. We truncated the total set of SNP by requiring at least six sequencing reads within a given individual in order to call a genotype, and further required that heterozygotes include at least two reads of each allele. By requiring six reads per locus (assuming bi-allelic loci), we could be 95% sure of identifying

at least one read of each allele in a heterozygote (Sedcole, 1977). For this analysis, we assumed that we are not detecting homoeologous loci with a given sequence, and thus, each locus resides within a single genome and is effectively diploid (Liu and Wu, 2012; Lu *et al.*, 2013). We then required that at least 80% of genotypes in the population could be assigned a genotype – that is, not more than 20% of the genotypes were missing data for any given locus due to few or no sequencing reads. Loci with an allelic ratio less than 0.01 within a given genotype were corrected to be homozygous and the loci with allelic ratio between 0.01 and 0.1 were discarded from the analysis because of the uncertainty of genotype calls. We removed SNPs with a minor allele frequency (MAF) less than 3% from further analysis. The custom scripts used for filtering the datasets along with parameters used to run both pipelines and to filter genotype calls are available in supplementary data and online at https://gist.github.com/anantaacharya.

Because this is our first analysis using GBS in switchgrass, in addition to the procedure outlined above, we also developed additional SNP datasets under different conditions. The main issue to be managed with GBS is the amount of missing data present in the dataset (Poland and Rife, 2012). More markers can be identified if the amount of missing data that can be accepted is increased. So, we developed SNP marker datasets across a range of missing values and using different stringencies to call genotypes in order to determine whether more markers with more missing data are better than fewer markers with less missing values. We varied the number of individuals with missing data for a given locus, and developed SNP marker datasets considering that up to 10, 50, and 90% of the population had missing data for that locus. For each of these datasets we made the same requirements as our core set in terms of read numbers in order to call a genotype. We also developed SNP marker datasets using the same four levels of missing data but only requiring a single sequence read to call a genotype.

In addition, we also developed a similar set of SNP datasets using the UNEAK pipeline (Lu *et al.*, 2013). With the UNEAK pipeline, we identified 25,341 SNPs among all genotypes. However UNEAK missed those SNPs with 2 or more SNPS within the same fragment, which describes the fewer number of SNPs. However, 91.8% of all those SNPs were also identified by STACKS. The remaining SNPs could not be identified with STACKS because of the read depth requirement. For simplicity, we refer to the different datasets using a code of U or S for UNEAK or STACKS, respectively, a number between 0 and 100 to indicate the percentage of plants in the population for which a genotype could be assigned, and a number indicating the number of sequencing reads required in order to make a genotype call. For example, S80-6 represents the SNPs identified using the STACKS pipeline for which 80% of genotypes were identified based on at least six sequencing reads. This particular dataset is the one we use for most of our subsequent analyses and serves as the comparator for other datasets we created.

If we could not assign a genotype for a given locus to a given individual due to inadequate (or no) sequencing reads, then we imputed a genotype using the random forest procedure implemented in MissForest (Stekhoven and Buhlmann, 2012) package in R. We used the imputed dataset, which had no missing values, for further analysis. The out-of-bag (OOB) accuracy reported by program, which is the measure of accuracy based on averaging the accuracy when a sample was left out in particular run was noted.

The names of SNP markers were prefixed with S (STACKS) or U (UNEAK) denoting pipeline that was used. The alphabet was followed by catalog number and position of SNP within that catalog separated by underscore; i.e S1000_10. For UNEAK pipeline SNP, there is no position within 64 basepair as it only reports a single SNP per tagpair (catalog). The SNP markers were aligned with the *Panicum virgatum* reference genome 1.1 (available at ftp://ftp.jgi-

psf.org/pub/compgen/phytozome/v9.0/early_release/Pvirgatum_v1.1/) by using a custom BLAST (Altschul *et al.*, 1990). Based on the alignment, the position of SNP was determined by adding the position of that SNP in that catalog to the start of alignment.

Linkage Disequilibrium:

We calculated linkage disequilibrium (LD) between the markers with TASSEL v4 (Bradbury *et al.*, 2007) and also with the LDcorSV software package (Mangin *et al.*, 2012) implemented in R, which corrects the bias due to population structure.

Statistical analysis of morphological traits:

We evaluated yield, plant height and flowering time in 2010, 2011, and 2012 at Watkinsville, GA and in 2012 at Ardmore, OK. We evaluated stem diameter in 2010 and 2011 at Watkinsville. For each trait, we performed an analysis of variance with R package "lme4" with genotype as a fixed effect and replication within year by location, location, year, genotype × location interaction and genotype × year interaction as random effects. We estimated the Best Linear Unbiased Predictor (BLUP) of each genotype for each trait across all environments. Depending upon the result of genotype × location interaction and genotype × year interaction, we estimated additional BLUPs for each location and/or each year and subsequently used for association analysis. We calculated repeatability of each trait across all environments as the ratio of genotypic variance to the sum of the variance components of genotype, genotype × location interaction, genotype × year interaction and error (residual). We calculated the phenotypic correlations among all traits based on mean values for individual genotypes in each environment. We assessed statistical significance as the 5% probability level, unless stated otherwise.

Association Analysis:

We conducted association mapping using single marker regression to identify SNP markers associated with biomass yield, plant height, stem diameter, and flowering time. First, we analyzed the data using the GLM procedure implemented in TASSEL v4 (Bradbury *et al.*, 2007) based on a naïve model that did not correct for either population structure or for the relationships among individuals. We next computed principal components (PC) of the genotypic data in order to correct for population structure, and reanalyzed the trait data for associations. Next, we estimated kinship using the Identity-by-State (IBS) allele sharing matrix implemented in the EMMA package in R (Kang *et al.*, 2008) and used this matrix in a mixed linear model (MLM) to account for kinship (Yu *et al.*, 2006) in the association analysis. Finally, we evaluated a fourth model that controlled both population structure and kinship. The significance probabilities (*P*-values) for association between markers and traits were corrected to reduce the false discovery rate (at P=0.05) with the Benjamini & Hochberg algorithm (Benjamini and Hochberg, 1995) and with the Bonferroni correction (Bland and Altman, 1995) using the R statistical package. For traits where no significant markers were identified after correction, results using raw *P*-values were reported. Quantile-Quantile (QQ) plots were used to visualize the distribution of *P*-values against the null hypothesis expectation of no association of markers with traits.

Results

Trait analysis:

The analysis of variance showed significant effects for genotype, year, location, genotype × year interaction and genotype × location interaction on biomass yield. The proportion of variance explained by genotype was larger than for any other component, ranging from 0.29 for

days to flower to 0.68 for plant height (Table 3.2). Across traits, the genotype × location interaction effect explained more of the variation than genotype × year interaction, but none of these effects was larger than 0.16 for any trait (Table 3.2). The repeatability estimates were high for all traits, ranging from 0.68 for days to flowering to 0.90 for stem diameter (Table 3.2). For the association analysis, we estimated BLUPs of each genotype across locations and years. In addition, we estimated BLUPs for each location for biomass yield and for each year and location for days to flowering because the genotype × location interaction effect and, in the case of flowering, the genotype × year interaction effect accounted for more than 10% of the trait variance.

All traits varied substantially among genotypes, among years, and between locations (Table 3.3). Yield was higher in Ardmore in 2012 than in any year in Watkinsville; yield increased every year in Watkinsville (Table 3.3). Plant height differed each year in Watkinsville; however, the average height in Ardmore was not different from 2010 and 2012 in Watkinsville. Stem diameter was different between years. The start of flowering ranged across 80 days in Watkinsville, but was about 118 days in Ardmore.

Biomass yield was positively correlated with plant height (r=0.86), stem diameter (r=0.81), and flowering time (r=0.49). Similarly, height was positively correlated with stem diameter (r=0.86) and flowering time (r=0.61). The correlation between stem diameter and flowering time was also positive (r=0.43). The correlations within one environment were similar to the overall correlations (Table 3.4).

SNP polymorphism:

Out of the total 65,328 SNP loci identified with STACKS, 59,288 were able to be assigned a genotype in at least 10% of the population based on the requirement of at least six

sequence reads to call the genotype. We designated this dataset with the name S10-6, based on the percentage of individuals genotyped and the number of reads required to assign a genotype. Other datasets were named in an analogous fashion. A total of 20,233 SNP were genotyped in at least 50% of the population (S50-6), 3,194 in at least 80% (S80-6), and only 457 in at least 90% of genotypes in the population (S90-6). Out of the total 25,341 SNP loci identified by UNEAK, 4,947 were present in at least 10% of the population (U10-6), 1471 in at least 50% (U50-6), 820 in at least 80% (U80-6), and only 280 (~1%) in at least 90% of genotypes in the population (U90-6). If we simply required a single read per individual, then 1931 SNP loci could be scored in at least 80% of the population (U80-1) or 3601 SNPs in at least 50% of the population (U50-1). Because STACKS requires at least two identical reads to identify a locus, genotyping based on a single read is not possible, and hence we did not create datasets with this criterion for STACKS. For all datasets, we imputed missing data for each locus using random forest imputation with imputation accuracy of 0.72 for the S80-6 dataset and 0.64 for the S50-6 dataset. For most of the analysis described below, we focus on the S80-6 SNP dataset, and make comparisons to other datasets as warranted.

For S80-6, 62.6% (1999) of SNPs aligned to the 18 pseudomolecules of the reference sequence (Fig 3.1), 35.3% (1134) aligned to the other contigs of the reference that have not yet been assembled, and only 2.1% (61) did not align to the reference sequence. The number was similar for all other datasets from either pipeline, with a range of 58 to 64% of SNP loci aligning to the 18 pseudomolecules.

Linkage Disequilibrium:

We calculated LD between markers aligned to the 18 pseudomolecules of the reference sequence. Because our previous analysis (Chapter 2) indicated that our population had

significant substructure – separated into upland and lowland clusters, with the lowland cluster further divided into two subclusters – as well as kinship among individual genotypes, we evaluated LD accounting for both structure and kinship. After adjustment, only 0.37% of SNP pairs had LD greater than 0.1 (Fig 3.2). The structure and kinship corrected LD coefficient was 0.13 for loci within 10kbp, 0.09 for loci within 100kbp, 0.03 for loci within 1Mbp, and 0.008 for loci within 10mbp.

Association Mapping:

We performed the single marker association analysis using four models: (1) a naïve model without correction for population structure or relatedness among individuals, (2) a linear model with structure (P) correction, (3) a mixed model with kinship (K) correction, and (4) a mixed model correcting for both P and K. We compared the distribution of *P*-values for the four models to the expected distribution under the null hypothesis of no association using a QQ plot. The naïve model showed substantial deviation from expectations; correcting for population structure, kinship, and both improved the distribution. In particular, models including kinship showed little deviation from the expected *P*-value distribution (Fig 3.3). In the discussion below, we only report SNP loci associated with traits after raw *P*-values have been corrected for multiple testing.

Under the naïve model, we identified 2,527 loci out of 3194 controlling average biomass yield across years and locations. Correcting for population structure reduced this to 370 loci, and correction for kinship only resulted in identification of four loci. Adjusting for both structure and kinship, we identified eight loci, including the four identified from a kinship-only correction. Five of the eight loci were in the assembled part of the 18 pseudomolecules and three were in unidentified locations (Table 3.5). The variance explained by individual marker loci was small, ranging from 5-7%. For several markers, the minor allele frequency was less than 10%, but for

one marker, the minor allele frequency of 0.41 (Table 3.5). The sum of the variance explained by all eight SNPs was 0.45, but this is an over-estimate based only on single marker analysis.

We performed an association analysis using yield BLUPs from Watkinsville only (WAT) and Ardmore only (ARD). Based on the P+K model, only one locus (S170007_27) on pseudomolecule 14 was identified in common between locations and also in the overall analysis (Table 3.4); this locus also had the highest marker R^2 (0.07) in the combined analysis. Several additional QTL were detected in only one of the locations (Table 3.5). The variance explained by markers ranged from 0.05 to 0.08. The combined variance was 0.59 (10 SNPs) for Watkinsville and 0.43 (6 SNPs) for Ardmore. The two SNPs that were 21 base pairs apart in the same 64 bp tag in pseudomolecule 18 both showed significant associations with yield for Watkinsville. Both of them explained the same amount of variation and also had same minor allele frequency. Out of the remaining seven loci, three did not have other SNPs in the same tag but four had other SNP in the same tag, which did not have the same likelihood of association. In those cases, the other SNP had a different minor allele frequency.

In addition to the S80-6 dataset, for which we required at least six sequencing reads to call a genotype and no more than 20% of genotypes with missing data, we also evaluated several other SNP datasets, including S50-6 with up to 50% missing genotypes for a given locus and those derived from the conservative SNP-calling model UNEAK.

Using the S50-6 SNP dataset with the P+K model, we identified 100 loci associated with biomass yield, compared to eight with S80-6 (Supplemental Table 3.1). Thus, a 6.4 fold increase in SNP markers resulted in a 12.5 fold increase in loci associated with yield. We identified all the loci, except one, that were identified using S80-6; the missing locus was on pseudomolecule 17 and had a *P*-value of 0.0001, just above the threshold for false discovery rate correction. Some

loci that were identified with S80-6 had nearby loci in S50-6. For example, Locus S18231_10 was within 8 bp of locus S96112_6, which was identified with both S50-6 and S80-6. The locus S63822_57 was within 100kbp of S170007_2, which was, again, identified by both datasets. All the loci from Watkinsville previously identified with S80-6 were identified with S50-6 and all except one for Ardmore. The locus that was missed had other marker loci with significant associations within 10kbp.

We also analyzed the association with the SNPs identified using UNEAK. For comparison purposes, we used four datasets described earlier. U80-6 was the database directly comparable to S80-6 with same number of read depth and same percentage of individuals for which a genotypic call had to be imputed. From this data set, we identified only two loci as compared to eight for S80-6, perhaps not surprisingly since this dataset consisted of only 802 SNPs as compared to 3,196 SNPs from S80-6. One locus was in pseudomolecule 15 and had nearby loci identified with S80-6 and the other locus was in the unmapped region. When comparing among the datasets from UNEAK, both the SNPs identified with U80-6 were also identified with U50-6 and two additional loci were identified by the latter. No loci associated with biomass yield were identified using U80-1, but seven were found using U50-1. Out of four loci identified by U50-6, two were also identified with U50-1 (data not shown). The variance explained by a single marker was up to 14%, but most were in the range of 5-8%. To better visualize the results of associated SNPs from these different combinations of datasets (S80-6, S50-6, U80-6, U50-6, U80-1, U80-6), locations (Both, WAT, ARD) along with other criteria (such as population structure and kinship derived from different metrics, not discussed here), we have implemented a web application (available at http://spark.rstudio.com/antu/GLMMLM_yield).

Some SNP markers from S80-6 that were associated with yield were in potential candidate genes (Table 3.6). Candidate genes based on SNP markers associated with yield from S50-6 are provided as supplemental file. SNP S17790_25 was in pseudomolecule 14, which has a similarity to G2-like transcription factors, which are important to chloroplast development (Fitter *et al.*, 2002). This marker had a minor allele frequency of 0.04, with five homozygous individuals and 18 heterozygotes carrying the "G" allele, which had a positive yield effect. These genotypes were high yielding genotypes of accessions PI 414065 and PI 414070, both of which are lowland ecotypes. This SNP was also identified by all models, all datasets and in all environments. SNP S145171_59 showed similarity to *nec*1 gene of barley (Rostoks *et al.* 2006).

We are only presenting results from the P+K mixed model using the S80-6 and S50-6 datasets for other morphological traits. For plant height, no markers were associated with plant height when using the FDR correction at $p = 0.05$ with either dataset. Therefore, in order to identify loci potentially involved with the trait, we identified markers associated with height having a raw *P*-value < 0.001. We identified 60 SNPs in the S50-6 dataset and 17 in the S80-6 dataset as having an association with height. All the SNPs identified with S80-6 (Table 3.7, Fig 3.5) were also identified with S50-6 (Supplemental Table 3.2, Supplemental Fig. 3.1). Three of 17 SNPs were associated with both height and yield (S17790_25 on pseudomolecule 14 and S145171_38, S145171_59 on pseudomolecule 18). Because of high repeatability and little evidence for genotype × year or genotype × location interaction effects, we only reported the analysis from combined phenotype.

Six SNP markers were associated with stem diameter (Fig 3.6), including the two markers on pseudomolecule 18 that were associated with both biomass yield and height (Table

3.8). We identified 20 SNPs associated with stem diameter with the S50-6 dataset (Supplemental Table 3.3, Supplemental Fig 3.2).

Because both genotype × year and genotype × location interaction effects were relatively large for days to flower, we analyzed the marker trait association for all four environments individually as well as across all four environments. Nine SNP markers were associated with days to flower but none were in common between environments and only one was identified in the combined data, which was also identified in Ardmore (Table 3.9, Fig 3.7). Using the S50-6 dataset, 26 SNP markers, including those found with S80-6, were associated with days to flower across different environments (Supplemental Table 3.4; Supplemental Fig 3.3). SNP S172069 showed sequence similarity to an anthocyanidin 5,3-O-glucosyltransferase-like gene affecting flower color. For all the traits, candidate SNPs from S50-6 are not discussed here but sequences are provided as supplemental file.

Three of the SNPs associated with plant height were also associated with biomass yield (Fig 3.8). Similarly, one of six SNPs from stem diameter was also identified for yield. However, none of the SNPs that was associated with flowering time was also associated with yield. For the dataset S50-6, eighteen of the SNPs associated with plant height were also associated with biomass yield (Fig 3.8). Similarly, nine out of 21 SNPs from stem diameter was also identified for yield. However, only one of the SNP that was associated with flowering time was also associated with yield (Supplemental Tables 3.1, 3.2, 3.3, 3.4). Six height QTL that were not the same as yield QTLs, however, were within 1 kbp of other yield QTLs. All QTL associated with stem diameter, except one, that were not directly associated with yield QTL were within 1kbp of yield QTL. About one quarter of the QTLs associated with flowering time were near loci associated with yield.

Discussion

The phenotypic data were robust for all four traits being analyzed, with a larger percentage of phenotypic variance as due to variance among genotypes than to any other factor. In general, the variance due to the interactions of genotypes with years or locations was relatively small. As a consequence, repeatability for all traits was high. Stem diameter showed the highest repeatability, which may be due to use of data from only one location. Surprisingly, for flowering time, more phenotypic variance was explained by genotype × year and genotype × location interactions than for the other traits. Although temperature and rainfall affect flowering time, photoperiod plays the major role (Parrish and Fike, 2005). The photoperiod was not only invariant year-to-year, but also nearly identical between Watkinsville, GA and Ardmore, OK.

Genotyping-by-sequencing (Elshire *et al.,* 2011) has led to rapid and inexpensive marker discovery provided sufficient bioinformatics support and computing resources are available to process the data into meaningful genotypic information. A fully automated bioinformatics pipeline is not yet in place for GBS, but several methods can be applied to GBS data, including STACKS (Catchen *et al.,* 2011; Catchen *et al.*, 2013) and UNEAK (Lu *et al.*, 2013), which work well in the absence of a robust reference genome. Various parameters can be modulated in these programs, which alter the ultimate number of SNP markers that are identified. Generally, genotypes cannot be called for all individuals for a given locus, making missing data a concern with GBS (Poland and Rife, 2012). Using STACKS, we only used about 10% of the potential SNPs identified across the entire population after filtering for lower read depth and allowing for up to 50% missing genotype calls for any individual. We only used 1.5% of potential SNPs when restricting missing data to 20% of the population for a given locus. Increasing the number of sequencing reads per genotype and/or reducing the number of sites in the genome generated by

the restriction enzymes could minimize missing data. However, imputation methods can be used to predict the genotype of given individual at a given locus. Of the various methods, random forest showed generally high imputation accuracy in an experiment comparing different datasets and imputation algorithms (Rutkoski *et al.*, 2013). Ultimately, the question for breeders is whether the amount of marker data that is generated for a given cost fulfills the breeding goal, not how much of the data generated is discarded.

Switchgrass is an open pollinating species, and consequently, we showed that linkage disequilibrium quickly decays as expected. Assuming that the switchgrass genome is 1.3 Gbp (www.phytozome.org), we expected to see approximately 115,000 restriction enzyme cut sites throughout the genome, based on an *in silico* computation using the initial switchgrass genome assembly (www.phytozome.org). Each enzyme cut site generates potentially two tags, so the expected genome coverage was about two tags per 113kbp. In our SNP dataset S50-6, we identified 20,000 SNP, which would correspond to one SNP per 65kbp, assuming even coverage across the genome. We know that a number of those SNP occur in the same 64bp tag. However, for the primary dataset S80-6, we identified 3,196 SNP, which is equivalent to one SNP per 406kbp. The *Fse*I enzyme is methylation sensitive and has a GC rich recognition site; therefore, it will preferentially cut in genic regions rather than regions with repetitive DNA elements. The distribution of SNPs in this experiment shows fewer SNPs in centromeric regions, which consist of highly repetitive DNA sequences. Therefore, the average distance between SNP is unrealistic, and given that we are skewing our sequenced sites toward genic regions suggests that our effective genome coverage is better than we may otherwise expect. Nevertheless, in a population where linkage disequilibrium decayed within 170 bp on average, the number of SNPs we

identified is very small and an increased number of markers would help identify more marker-trait associations.

The two SNP detection methods we used here – STACKS and UNEAK – identified very different numbers of SNP for a given set of criteria. UNEAK algorithm utilizes only those reciprocal fragments with one mismatch per tag to call the SNPs but with STACKs, it can be customized and we allowed up to two mismatches within genotype and one additional mismatch among genotypes. Because most of the SNPs identified using UNEAK were also identified using STACKS, we suggest using STACKS because it generates more markers. For a given pipeline, multiple SNP datasets can be generated by altering the amount of missing data accepted, the number of reads required to call genotypes, and so on, and each dataset can produce different QTL results due to the correction for multiple testing. This can help explain why a few QTL associated with yield based on UNEAK were not identified with STACKS.

We previously (Chapter 2) reported that this population of genotypes had a strong population substructure, which was evident from analysis with either SSR or SNP markers. For this experiment, we used a GWAS model that accounted for both population structure and kinship to help avoid false QTL detection (Yu *et al.*, 2006, Bradbury *et al.*, 2011; Korte and Farlow, 2013).

Although most studies remove the markers below a MAF threshold of 5 to 10%, false discovery of marker-trait associations because of low MAF (5%) was not significantly different from that seen with a higher MAF threshold (Moskvina *et al.*, 2006). Thus, removing minor alleles because of fear of false discovery is not recommended (Tabangin *et al.*, 2009). Our inclusion of markers above 3% MAF ensured that at least ten genotypes, carried the minor alleles. This number is approximately the number of individuals within a given accession, so

inclusion of this level of minor alleles ensures that one accession within our diverse set of populations is not discarded from analysis. Other experiments that included few individuals from many diverse populations also included low MAF (Eckert *et al.*, 2012). Based on this literature, we believe that although power was decreased using low MAF, the false discovery rate was not significantly different from those SNPs with higher MAF.

Several of the identified SNP markers were in tags that had sequence similarity to annotated genes from related species. Two tags containing candidate SNPs showed similarity to a G2-like MYB transcription factor. Because these transcription factors are responsible for a variety of plant growth responses, like development, metabolism, and biotic and abiotic stress response in Arabidopsis (Dubos *et al.*, 2010), they could be involved in biomass yield of switchgrass. Another candidate SNP showed sequence similarity to the necrogenic phenotype (*nec*1) gene conferring the resistance to biotic stress (Keisa, 2011), and it may have played role in increasing the biomass of switchgrass through improved disease resistance (Shavannor Smith, pers. comm.). Markers associated with plant height showed sequence similarity to an ADP/ATP carrier-like protein, a zinc finger-like protein and a NAC domain containing protein. The capacity to import ATP is related to plant growth (Reiser *et al.*, 2004), and the NAC domain containing protein affects plant meristem growth (Wang and Li, 2008), which could relate to the role of this SNP in plant height. The anthocyanidin 5,3-O-glucosyltransferase-like gene showing sequence similarity to one of our flowering time SNP markers, has a role in flower color, but is not known to be directly related to flowering time in other species. Of course, other genes likely exist within the vicinity of the SNP markers, even with rapid LD decay, and these other genes may be the true candidate locus controlling the trait.

This is the first genome wide association mapping experiment in switchgrass. Because biomass yield is a key characteristic necessary for economical biofuel production, identifying loci putatively associated with yield could accelerate breeding new cultivars by using marker assisted selection. In this experiment, we identified 19 QTL associated with biomass yield across multiple environments. These loci could form the basis of a switchgrass marker-assisted breeding program to increase biomass yield. Components of yield, including the other three traits used in this experiment of stem diameter, plant height, and flowering time may be under less complex genetic control and more amenable to marker-based selection. Improving them could indirectly improve yield. These correlations between linked SNP genes and traits find some interesting cases, but there are thousands of genes that might be logically associated with yield. And we have not enriched for such genes in our SNP correlations. Several of the QTL we identified for yield were also important for these traits.

Some genotypes and or populations in this experiment showed superior phenotypes and desirable marker profiles at useful QTL. These individuals can be used in breeding programs, and QTL can be selected based on marker alleles identified here after the validation. The markers can be used to both the selection of desirable genotypes and to discard of inferior material before taking it to the field. The use of markers can continue with the marker assisted selection or marker-assisted backcrossing to increase QTL allele frequency and to stack QTL for a given trait. Further research is needed to determine if the markers identified here are associated with these traits in other genetic backgrounds including breeding programs and in other environments. More fully saturating this population with additional markers is underway using exome capture. The improved resolution will undoubtedly assist us in identifying other QTL for these and other

biofuel-related traits. Additional QTL experiments using other populations will help validate the results here.

Reference:

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403-10. DOI: 10.1016/S0022-2836(05)80360-2.

Benjamini Y., Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological):289-300.

Bhandari H.S., Saha M.C., Fasoula V.A., Bouton J.H. (2011) Estimation of Genetic Parameters for Biomass Yield in Lowland Switchgrass (*Panicum virgatum* L.). Crop Science 51:1525. DOI: 10.2135/cropsci2010.10.0588.

Bhandari H.S., Saha M.C., Mascia P.N., Fasoula V.A., Bouton J.H. (2010) Variation among Half-Sib Families and Heritability for Biomass Yield and Other Traits in Lowland Switchgrass (*Panicum virgatum* L.). Crop Science 50:2355-2363. DOI: 10.2135/cropsci2010.02.0109.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. Bmj, 310(6973), 170.

Boe, A., Beck, D. L. (2008). Yield components of biomass in switchgrass. Crop Science, 48(4), 1306-1311.

Bouwmeester K., de Sain M., Weide R., Gouget A., Klamer S., Canut H., Govers F. (2011) The Lectin Receptor Kinase LecRK-I.9 Is a Novel Phytophthora Resistance Component and a Potential Host Target for a RXLR Effector. Plos Pathogens 7. DOI: Artn E1001327

Bradbury P., Parker T., Hamblin M.T., Jannink J.L. (2011) Assessment of Power and False Discovery Rate in Genome-Wide Association Studies using the BarleyCAP Germplasm. Crop Science 51:52-59. DOI: Doi 10.2135/Cropsci2010.02.0064.

Bradbury P.J., Zhang Z., Kroon D.E., Casstevens T.M., Ramdoss Y., Buckler E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635. DOI: Doi 10.1093/Bioinformatics/Btm308.

Breseghello F., Sorrells M.E. (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165-1177. DOI: Doi 10.1534/Genetics.105.044586.

Cassida K.A., Muir J.P., Hussey M.A., Read J.C., Venuto B.C., Ocumpaugh W.R. (2005) Biomass yield and stand characteristics of switchgrass in south central US environments. Crop Science 45:673-681.

Catchen J., Hohenlohe P.A., Bassham S., Amores A., Cresko W.A. (2013) Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124-3140. DOI: Doi 10.1111/Mec.12354.

Catchen J.M., Amores A., Hohenlohe P., Cresko W., Postlethwait J.H. (2011) Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3-Genes Genomes Genetics 1:171-182. DOI: Doi 10.1534/G3.111.000240.

Cortese L., Honig J., Miller C., Bonos S. (2010) Genetic Diversity of Twelve Switchgrass Populations Using Molecular and Morphological Markers, BioEnergy Research, Springer New York. pp. 262-271-271.

Das, M. K., Fuentes, R. G., Taliaferro, C. M. (2004). Genetic variability and trait relationships in switchgrass. Crop science, 44(2), 443-448.

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., Lepiniec, L. (2010). MYB transcription factors in Arabidopsis Trends in plant science, 15(10), 573-581.

Eckert, A. J., Wegrzyn, J. L., Cumbie, W. P., Goldfarb, B., Huber, D. A., Tolstikov, V., ... & Neale, D. B. (2012). Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. new phytologist, 193(4), 890-902.

Ewens W.J., Spielman R.S. (2001) Locating genes by linkage and association. Theoretical population biology 60:135-9. DOI: 10.1006/tpbi.2001.1547.

Fitter D.W., Martin D.J., Copley M.J., Scotland R.W., Langdale J.A. (2002) GLK gene pairs regulate chloroplast development in diverse plant species. Plant Journal 31:713-727. DOI: Doi 10.1046/J.1365-313x.2002.01390.X.

Giorno F., Wolters-Arts M., Mariani C., Rieu I. (2013) Ensuring Reproduction at High Temperatures: The Heat Stress Response during Anther and Pollen Development. Plants 2:489-506.

Gunter L.E., Tuskan G.A., Wullschleger S.D. (1996) Diversity among populations of switchgrass based on RAPD markers. Crop Science 36:1017-1022.

Gupta P.K., Rustgi S., Kulwal P.L. (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant molecular biology 57:461-85. DOI: 10.1007/s11103-005-0257-z.

Hansen, K. A., Martin, J. M., Lanning, S. P., Talbert, L. E. (2005). Correlation of genotype performance for agronomic and physiological traits in space-planted versus densely seeded conditions. Crop science, 45(3), 1023-1028.

Hill W.G., Weir B.S. (1988) Variances and covariances of squared linkage disequilibria in finite populations. Theoretical population biology 33:54-78.

Himelblau E., Amasino R.M. (2000) Delivering copper within plant cells. Current Opinion in Plant Biology 3:205-210.

Honsdorf N., Becker H.C., Ecke W. (2010) Association mapping for phenological, morphological, and quality traits in canola quality winter rapeseed (*Brassica napus* L.). Genome 53:9. DOI: 10.1139/G10-049.

Jannink J.-L., Bink M.C.A.M., Jansen R.C. (2001) Using complex plant pedigrees to map valuable genes. Trends in Plant Science 6:337-342. DOI: 10.1016/S1360-1385(01)02017-9.

Kang H.M., Zaitlen N.A., Wade C.M., Kirby A., Heckerman D., Daly M.J., Eskin E. (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709-1723. DOI: Doi 10.1534/Genetics.107.080101.

Keisa, A., Kanberga-Silina, K., Nakurte, I., Kunga, L., Rostoks, N. (2011). Differential disease resistance response in the barley necrotic mutant *nec*1.BMC plant biology, 11(1), 66.

Korte A., Farlow A. (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant methods 9:29. DOI: 10.1186/1746-4811-9-29.

Lemus R., Brummer E.C., Moore K.J., Molstad N.E., Burras C.L., Barker M.F. (2002) Biomass yield and quality of 20 switchgrass populations in southern Iowa, USA. Biomass & Bioenergy 23:433-442. DOI: Pii S0961-9534(02)00073-9

Liu L.L., Wu Y.Q. (2012) Identification of a Selfing Compatible Genotype and Mode of Inheritance in Switchgrass. Bioenergy Research 5:662-668. DOI: Doi 10.1007/S12155-011-9173-Z.

Lu F., Lipka A.E., Glaubitz J., Elshire R., Cherney J.H., Casler M.D., Buckler E.S., Costich D.E. (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. Plos Genetics 9. DOI: Artn E1003215

Mangin B., Siberchicot A., Nicolas S., Doligez A., This P., Cierco-Ayrolles C. (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity 108:285-291. DOI: Doi 10.1038/Hdy.2011.73.

McLaughlin S.B., Adams Kszos L. (2005) Development of switchgrass (*Panicum virgatum*) as a bioenergy feedstock in the United States. Biomass and Bioenergy 28:515-535. DOI: 10.1016/j.biombioe.2004.05.006.

Missaoui A.M., Paterson A.H., Bouton J.H. (2005) Investigation of genomic organization in switchgrass (*Panicum virgatum* L.) using DNA markers. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 110:1372-83. DOI: 10.1007/s00122-005-1935-6.

Mohan S., Ma P.W.K., Williams W.P., Luthe D.S. (2008) A Naturally Occurring Plant Cysteine Protease Possesses Remarkable Toxicity against Insect Pests and Synergizes Bacillus thuringiensis Toxin. Plos One 3. DOI: ARTN e1786

Moskvina, V., Craddock, N., Holmans, P., Owen, M. J., O'Donovan, M. C. (2006). Effects of differential genotyping error rate on the type I error probability of case-control studies. Human heredity, 61(1), 55-64.

Narasimhamoorthy B., Saha M., Swaller T., Bouton J. (2008) Genetic Diversity in Switchgrass Collections Assessed by EST-SSR Markers. BioEnergy Research 1:136-146. DOI: 10.1007/s12155-008-9011-0.

Poland J.A., Brown P.J., Sorrells M.E., Jannink J.L. (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. Plos One 7. DOI: ARTN e32253

Poland J.A., Rife T.W. (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome 5:92-102. DOI: Doi 10.3835/Plantgenome2012.05.0005.

Price, D. L., Casler, M. D. (2014). Predictive relationships between plant morphological traits and biomass yield in switchgrass. Crop Science, 54(2), 637-645.

Reiser, J., Linka, N., Lemke, L., Jeblick, W., & Neuhaus, H. E. (2004). Molecular physiological analysis of the two plastidic ATP/ADP transporters from Arabidopsis. Plant physiology, 136(3), 3524-3536.

Rose, L. W., Das, M. K., Taliaferro, C. M. (2008). Estimation of genetic variability and heritability for biofuel feedstock yield in several populations of switchgrass.Annals of applied biology, 152(1), 11-17.

Rostoks, N., Schmierer, D., Mudie, S., Drader, T., Brueggeman, R., Caldwell, D. G., Kleinhofs, A. (2006). Barley necrotic locus nec1 encodes the cyclic nucleotide-gated ion channel 4 homologous to the Arabidopsis HLM1. Molecular Genetics and Genomics, 275(2), 159-168.

Rutkoski J.E., Poland J., Jannink J.L., Sorrells M.E. (2013) Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy. G3-Genes Genomes Genetics 3:427-439. DOI: Doi 10.1534/G3.112.005363.

Sanderson M.A., Adler P.R., Boateng A.A., Casler M.D. (2006) Switchgrass as a biofuels feedstock in the USA. Canadian Journal of Plant Science.

Schmer M.R., Mitchell R.B., Vogel K.P., Schacht W.H., Marx D.B. (2010) Efficient Methods of Estimating Switchgrass Biomass Supplies. Bioenergy Research 3:243-250. DOI: Doi 10.1007/S12155-009-9070-X.

Sedcole J. (1977) Number of plants necessary to recover a trait. Crop Science 17:667-668.

Stekhoven D.J., Buhlmann P. (2012) MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics 28:112-118. DOI: Doi 10.1093/Bioinformatics/Btr597.

Stich B., Möhring J., Piepho H.-P., Heckenberger M., Buckler E.S., Melchinger A.E. (2008) Comparison of mixed-model approaches for association mapping. Genetics 178:1745-54.

Tabangin, M. E., Woo, J. G., Martin, L. J. (2009). The effect of minor allele frequency on the likelihood of obtaining false positives. In BMC proceedings (Vol. 3, No. Suppl 7, p. S41). BioMed Central Ltd.

Tulbure, M. G., Wimberly, M. C., Boe, A., Owens, V. N. (2012). Climatic and genetic controls of yields of switchgrass, a model bioenergy species. Agriculture, Ecosystems & Environment, 146(1), 121-129.

Wang, Y., & Li, J. (2008). Molecular basis of plant architecture. Annu. Rev. Plant Biol., 59, 253-279.

Wright L. (1994) Production technology status of woody and herbaceous crops. Biomass and Bioenergy 6:191-209. DOI: 10.1016/0961-9534(94)90075-2.

Wullschleger, S. D., Davis, E. B., Borsuk, M. E., Gunderson, C. A., Lynd, L. R. (2010). Biomass production in switchgrass across the United States: database description and determinants of yield. Agronomy Journal, 102(4), 1158-1168.

Yu J., Pressoir G., Briggs W.H., Bi I.V., Yamasaki M., Doebley J.F., Mcmullen M.D., Gaut B.S., Nielsen D.M., Holland J.B., Kresovich S., Buckler E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203-208. DOI: 10.1038/ng1702.

Zhang Y.W., Zalapa J.E., Jakubowski A.R., Price D.L., Acharya A., Wei Y.L., Brummer E.C., Kaeppler S.M., Casler M.D. (2011) Post-glacial evolution of P*anicum virgatum*: centers

of diversity and gene pools revealed by SSR markers and cpDNA sequences. Genetica 139:933-948. DOI: Doi 10.1007/S10709-011-9597-6.

Zhu C., Gore M., Buckler E.S., Yu J. (2008) Status and Prospects of Association Mapping in Plants. The Plant Genome 1:5-20. DOI: 10.3835/plantgenome2008.02.0089.

Table 3.1: Switchgrass accessions used in this experiment and their geographic origin with putative ecotype classification.

Accession	Geographic Origin	Ecotype	No of genotypes per accession
PI 315723	Hoffman, NC	L	7
PI 315724	Ellsworth, KS	U	9
PI 315725	Coffeeville, MS	L	2
PI 315727	Apex, NC	L	12
PI 315728	Scotland County, NC (donated by Maryland)	I	6
PI 337553	Rafaela Experiment Station, Argentina	U	12
PI 414065	Pangburn, AR	L	12
PI 414066	Grenville, NM	U	5
PI 414067	Soil Conservation Service, NC	U	7
PI 414068	Soil Conservation Service, KS	U	10
PI 414070	Soil Conservation Service, KS	L	12
PI 421138	Moore County, NC	U	8
PI 421520	Kay County, OK	U	7
PI 421521	Wetumka, OK (developed in KS)	L	8
PI 421999	Pangburn, AR	L	10
PI 422001	Stuart, Martin County, FL	L	10
PI 422003	FL	L	8
PI 422006	George West, TX	L	14
NA[†]	NA	U	12
PI 422016	FL	L	9
PI 431575	Raleigh County, WV (from KY)	U	12
PI 476290	Wilmington, NC	L	7
PI 476291	MD	I	15
PI 476292	Franklin County, AR	U	8
PI 476293	Heislerville, NJ	I	9
PI 476294	Eads, CO	U	7
PI 476295	Colorado Springs, CO	U	13
PI 476296	MD	U	6
PI 642190	NM	U	8
PI 642191	SD	U	6
Citrus Co-FL	Citrus County, FL	L	15
HSP-FL	Hillsborough River State Park, FL	L	12
OSSP-FL	Oscar Scherer State Park, FL	L	12
Pasco Co-FL	Pasco County, FL	I	13
SNF	Sumter National Forest, SC	L	13
SPBluff	Sprewell Bluff, GA	I	7
SWFWMD-FL	Southwest Florida Water Management	L	7

	District, FL		
GRIF16964	FL	L	1
GRIF16967	FL	L	1
GRIF16968	FL	L	1
GRIF16969	FL	L	2
GRIF16982	FL	L	1
GRIF16991	FL	L	1
AP13	Derived from PI 422006	L	1
VS16	Derived from PI 642191	U	1

[†]These populations were thought to be Alamo (PI 422006), but genetically and phenotypically they are very different.

Ecotypes were designated based on population structure analysis with nuclear SNP markers

Table 3.2. Variance components for sources of variation derived from an analysis of variance for four morphological traits of switchgrass.

Source	Yield		Height		Flowering time		Stem diameter	
	Variance	Proportion	Variance	Proportion	Variance	Proportion	Variance	Proportion
Genotype	208***	0.43	835.1***	0.68	175.0***	0.29	15.7***	0.67
Replication	0	0.00	0.4	0.00	0.3	0.00	0.0	0.00
Year	30***	0.06	14.7***	0.01	39.7***	0.07	1.2***	0.05
Location	12***	0.02	0.8*	0.00	171.3***	0.28		
Genotype × Year	23***	0.05	25.0***	0.02	71.7***	0.12	0.3	0.01
Genotype × Location	79***	0.16	63.5***	0.05	90.0***	0.15		
Residual	132	0.27	290.2	0.24	58.8	0.10	6.4	0.27
Repeatability	0.72		0.88		0.68		0.90	

*, **, *** Significant at 0.05, 0.01, <0.001 level

Table 3.3. Means, standard deviations, and ranges of biomass yield, plant height, flowering time, and stem diameter of switchgrass genotypes grown in Athens, GA and Ardmore, OK across three years.

	Yield				**Height**				**Flowering time**				**Stem Diameter**			
	Mean -gm-	St. Dev. -gm-	Min -gm-	Max -gm-	Mean -cm-	St. Dev. -cm-	Min -cm-	Max -cm-	Mean -days-	St. Dev. -days-	Min -days-	Max -days-	Mean -mm-	St. Dev. -mm-	Min -mm-	Max -mm-
Athens																
2010	327d[†]	340	1	3417	84b	32	5	179	170d	14	141	218				
2011	581c	541	1	3085	79c	30	8	166	181b	18	152	234	38a	13	6	89
2012	702b	733	1	3807	87a	39	5	187	173c	23	144	212	33b	16	2	90
Ardmore																
2012	859a	806	3	5498	86ab	33	3	182	187a	21	149	267				

[†]Means within columns followed by different letters were significantly different at $p<0.05$.

Table 3.4: Correlation among switchgrass traits over years and location. The overall correlation was calculated by averaging the traits over all year and location for each genotype.

Trait	Location	Year	Height	Stem diameter	Flowering time
Yield	Watkinsville	2010	0.77	NA	0.43
	Watkinsville	2011	0.80	0.75	0.40
	Watkinsville	2012	0.82	0.81	0.49
	Ardmore	2012	0.81	NA	0.45
	Overall		0.86	0.81	0.49
Height	Watkinsville	2010		NA	0.48
	Watkinsville	2011		0.73	0.25
	Watkinsville	2012		0.88	0.39
	Ardmore	2012		NA	0.39
	Overall			0.86	0.41
Stem diameter	Watkinsville	2010			NA
	Watkinsville	2011			0.56
	Watkinsville	2012			0.46
	Ardmore	2012			NA
	Overall				0.63

Table 3.5: The SNP markers associated with biomass yield in switchgrass ($p<0.05$) after false discovery rate correction, the minor allele frequency at the marker locus, the likely genomic location based on switchgrass reference sequence 1.0, and the amount of phenotypic variation explained by the marker.

Marker	Pseudomolecule	Position	MAF[†]	Marker R^2		
				Both	Ardmore	Watkinsville
174224_36	1	28,145,266	0.09			0.07
75170_33	2	26,561,693	0.15		0.08	
55919_34	5	16,459,056	0.15	0.06		0.05
163375_6	10	8,736,442	0.12		0.07	
17790_25	14	33,296,072	0.04	0.07	0.08	0.06
96112_6	14	49,512,709	0.08	0.06		
170007_27	15	724,329	0.19	0.05	0.06	
169502_6	15	9,111,853	0.03			0.06
168152_38	16	45,925,147	0.05		0.06	
91480_8	17	16,277,093	0.41	0.05		
145171_38	18	14,796,918	0.04			0.05
145171_59	18	14,796,939	0.04			0.05
46741_51	18	22,729,999	0.07			0.07
45124_26	Contig		0.18		0.08	
129676_24	Contig		0.14	0.05		
133630_61	Contig		0.06			0.06
17689_32	Contig		0.4	0.05		
171069_7	Contig		0.06	0.06		
162119_44	Contig		0.08			0.07

[†]MAF=Minor Allele Frequency

Table 3.6: The SNPs significantly associated with traits and their similarity to annotated gene in other species.

SNP	Associated trait	Annotation
S17790	yield, height	G2-like trascription factor
S145171	yield, height, stem diameter	Necrotic (nec1) gene
S91480	yield	G2-like trascription factor
S172069	flowering time	Anthocyanidin 5,3-O-glucosyltransferase-like
S46074	height	ADP/ATP CARRIER 3 family protein
S48579	height	Zinc finger protein 1-like
S92451	height	NAC domain-containing protein 18-like

Table 3.7: The SNP markers associated with plant height in switchgrass ($p<0.0001$) after false discovery rate correction, the minor allele frequency at the marker locus, the likely genomic location based on switchgrass reference sequence 1.0, and the amount of phenotypic variation explained by the marker.

Marker	MAF[†]	Pseudomolecule	Position	Marker R^2
85177_12	0.03	1	70,949,740	0.05
46074_20	0.03	3	2,986,042	0.04
98731_42	0.04	3	25,723,483	0.04
48579_35	0.04	3	77,933,511	0.04
157075_52	0.26	9	20,250,257	0.04
92451_34	0.04	14	2,832,983	0.05
17790_25	0.04	14	33,296,072	0.04
180053_9	0.06	15	25,523,375	0.04
145171_38	0.04	18	14,796,897	0.04
145171_59	0.04	18	14,796,918	0.04
158887_22	0.5	Contig		0.04
158887_26	0.5	Contig		0.04
41690_18	0.05	Contig		0.04
19931_32	0.17	Contig		0.04
208454_14	0.31			0.04
208454_9	0.32			0.05

[†]MAF=Minor Allele Frequency

Table 3.8: The SNP markers associated with stem diameter in switchgrass ($p<0.05$) after false discovery rate correction, the minor allele frequency at the marker locus, the likely genomic location based on switchgrass reference sequence 1.0, and the amount of phenotypic variation explained by the marker.

SNP	MAF†	Pseudomolecule	Position	Marker R^2
152890_33	0.09	4	40,221,596	0.06
138040_10	0.08	6	71,235,845	0.06
145171_38	0.04	18	14,796,897	0.06
145171_59	0.04	18	14,796,918	0.06
46741_51	0.07	18	22,729,999	0.07
18831_12	0.03	Contig		0.06

†MAF=Minor Allele Frequency

Table 3.9: The SNP markers associated with flowering time in switchgrass ($p<0.0001$), the minor allele frequency at the marker locus, the likely genomic location based on switchgrass reference sequence 1.0, and the amount of phenotypic variation explained by the marker.

Marker	Pseudomolecule	Position	MAF[†]	Marker R^2				
				Both	Watkinsville			Ardmore
					2010	2011	2012	2012
25104_57	1	15,537,383	0.04	NA	0.06	NA	NA	NA
113324_54	6	3,426,581	0.07	0.06	NA	NA	NA	0.11
172069_30	18	12,363,103	0.05	NA	NA	0.09	NA	NA
172437_17	18	48,991,459	0.04	NA	NA	NA	NA	0.08
172437_45	18	48,991,487	0.04	NA	NA	NA	NA	0.08
78475_49	Contig		0.46	NA	NA	NA	0.07	NA
93553_33			0.05	NA	NA	NA	0.06	NA
93553_53			0.05	NA	NA	NA	0.06	NA
96836_48			0.05	NA	NA	NA	0.06	NA

[†]MAF=Minor Allele Frequency

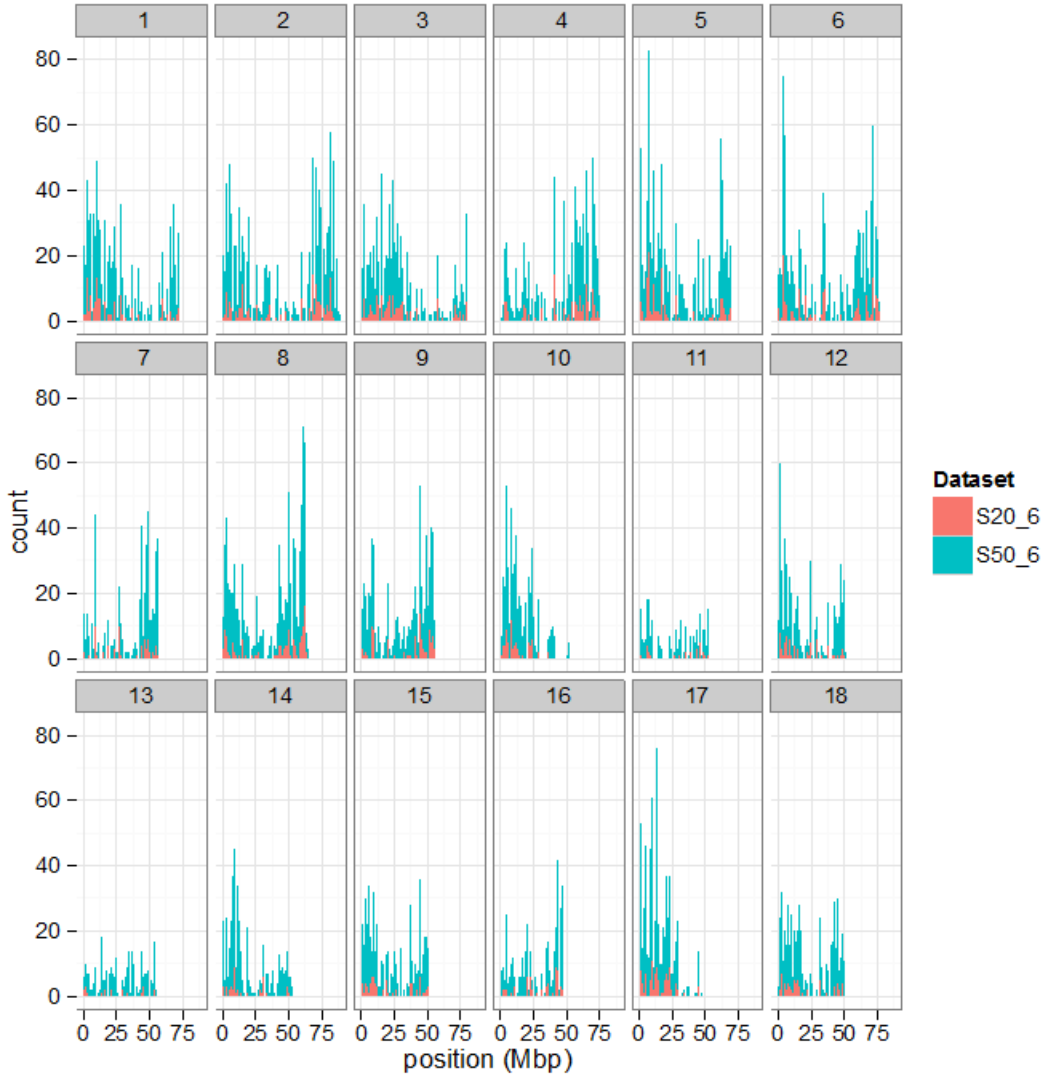


Fig 3.1: Distribution of SNPs in 18 pseudomolecules of AP13 reference genome grouped with 18 pseudomolecules. The bars are count per 1Mbp. Red color up to 20% missing genotype call (S80-6) and blue up to 50% missing genotype call (S50-6).

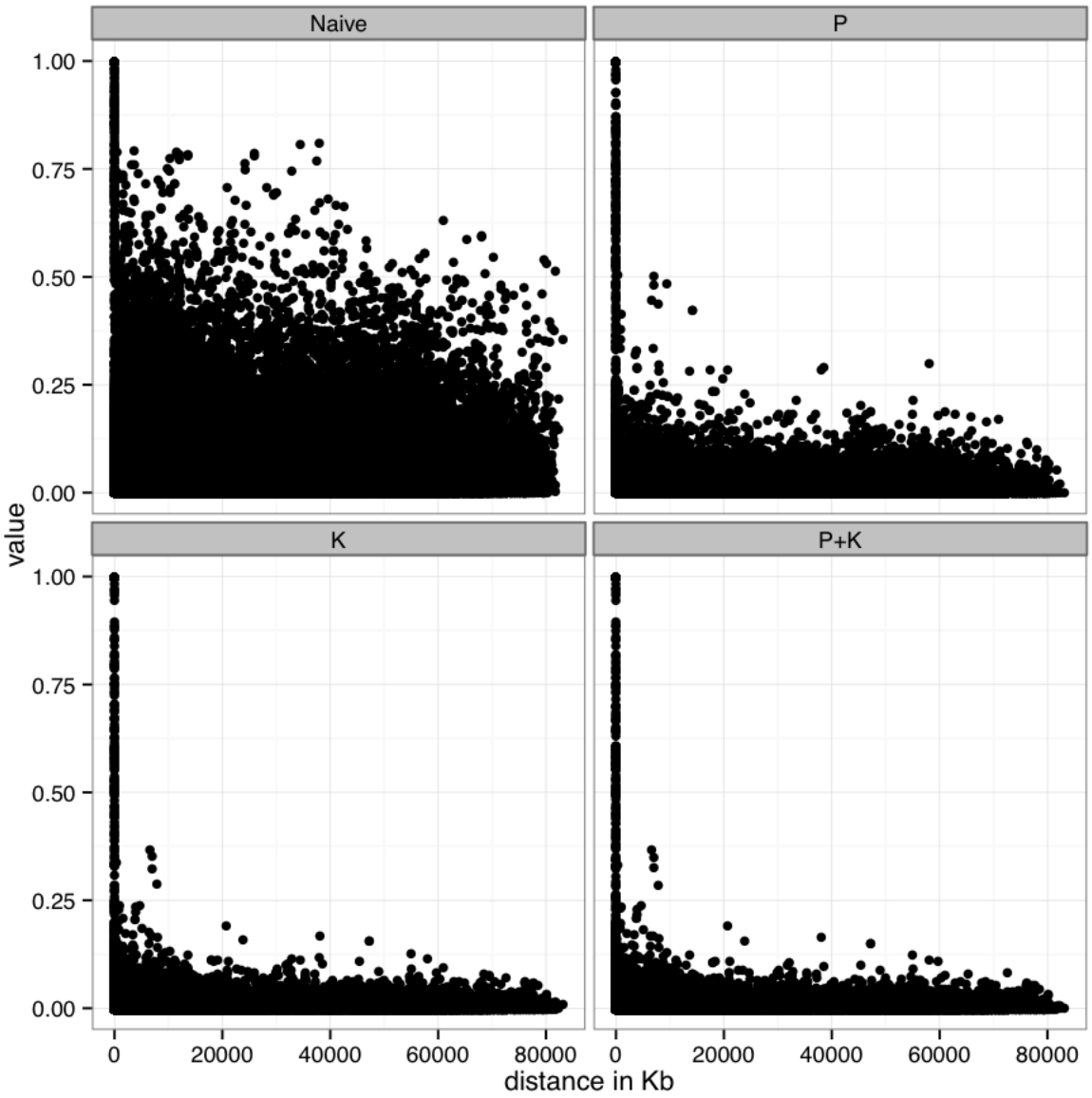


Fig 3.2. Linkage disequilibrium coefficients (r^2) plotted against the distance between the SNP pairs. a) without correcting for structure or population. b) after correcting for kinship c) after correcting for structure and d) after correcting for both structure and kinship. Note the high number of SNP pairs in LD across long distances that is likely due to population structure corrected later by the correction measures.

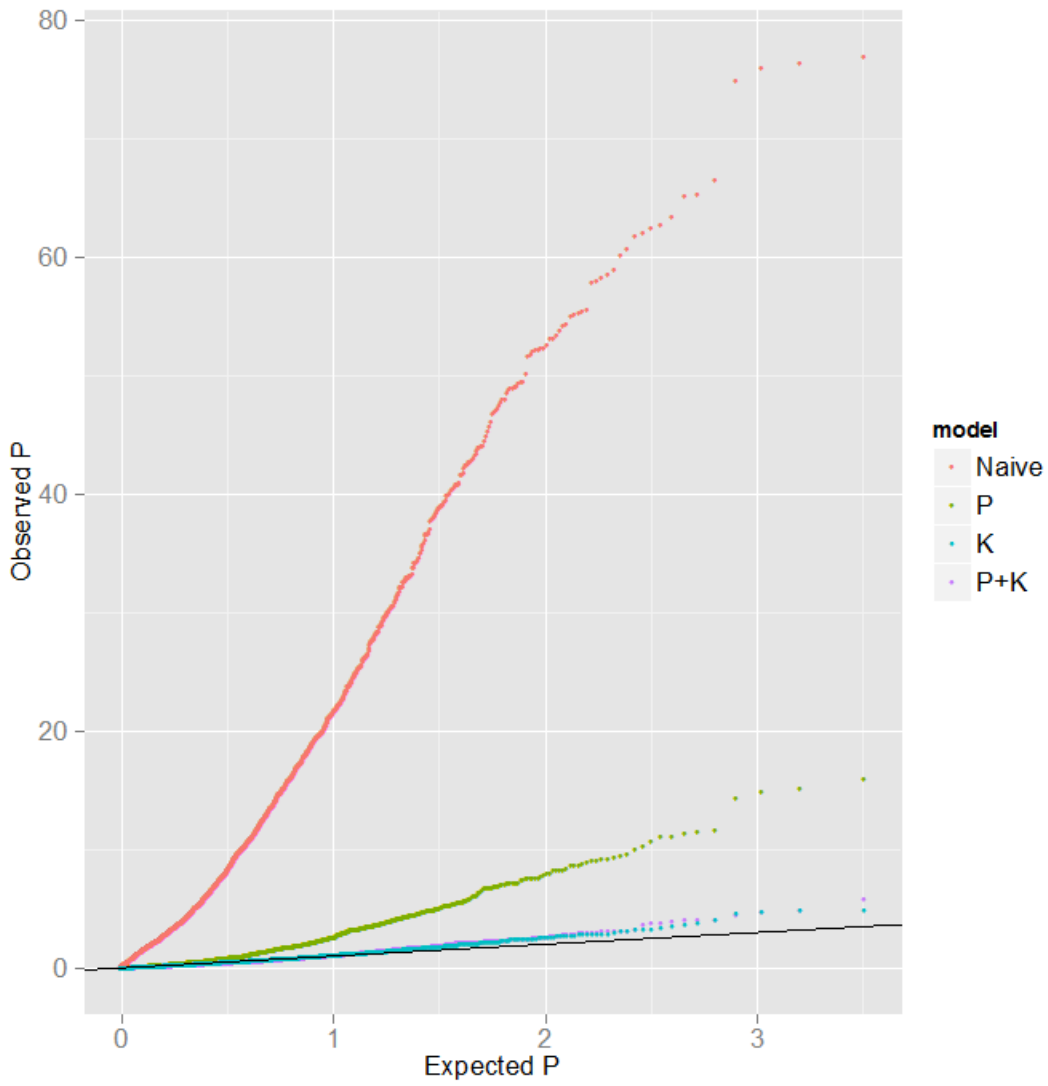
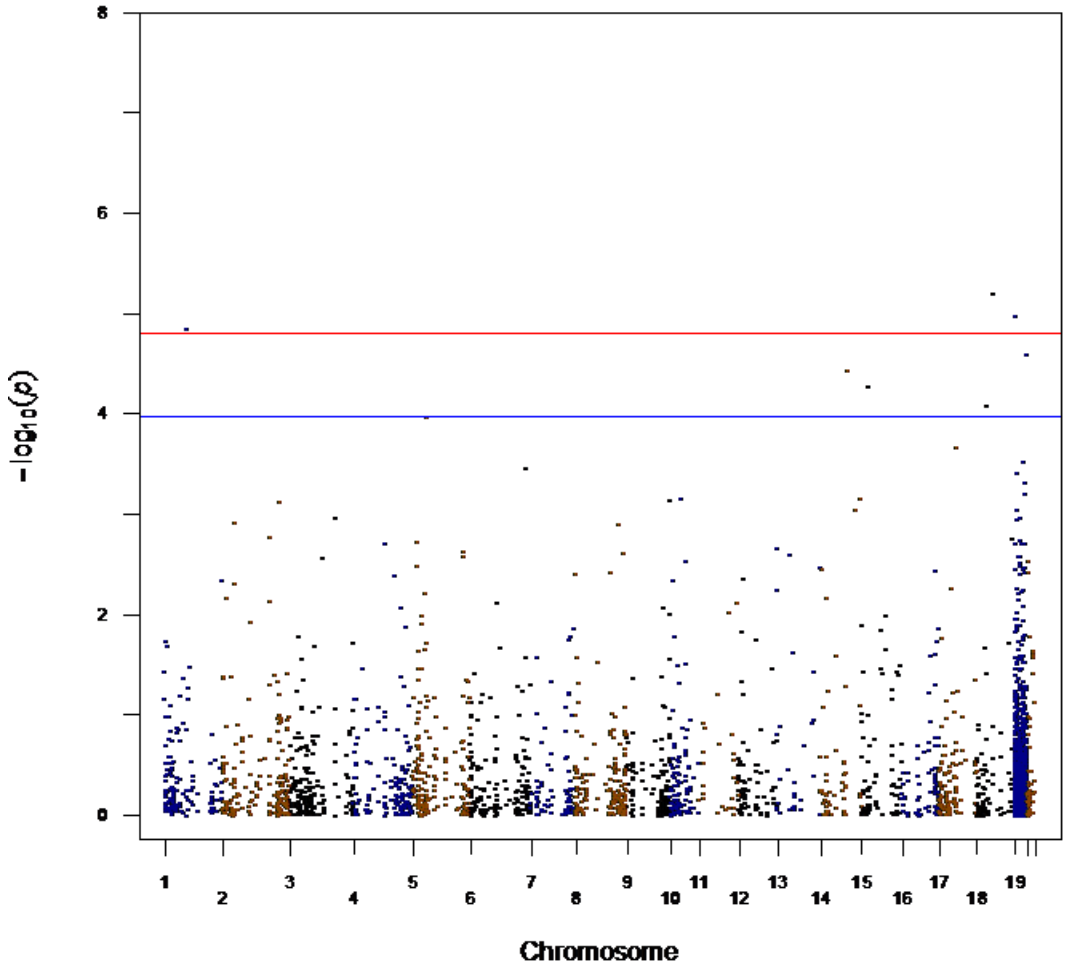
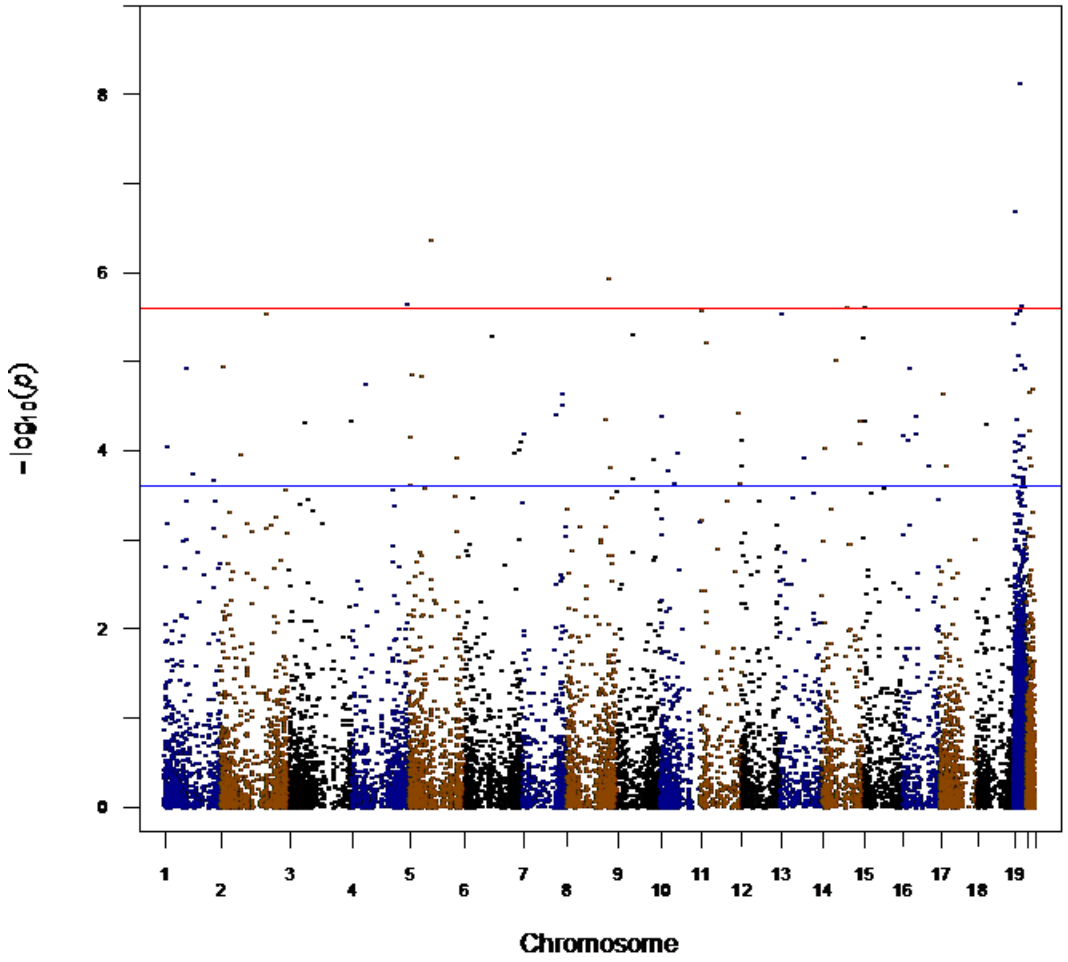


Fig 3.3. Distribution of *P* values for different models. Here x-axis represents the negative $\log_{10}$ of expected *P* values assuming no association of marker and trait and y-axis represents the observed. The black line represents the null expectation.


8
6
4
2
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Chromosome

a)



b)

Fig 3.4. Manhattan plot of association mapping of biomass yield with mixed model with principal component and kinship (P+K). The X axis are 18 pseudomolecules, the 19th position denotes SNPs aligned to contigs that were not assembled in any pseudomolecules and SNPs in last position with orange color are SNPs not aligned to switchgrass reference genome. Y axis is negative log10 of P values. Red line is cutoff using conservative Bonferroni FDR correction at 0.05 and blue line is Benjamini & Hochberg FDR correction at 0.05. a) With the dataset where each individual had at least 80% SNPs genotyped (S80-6) b) at least 50% SNPs genotyped (S50-6)

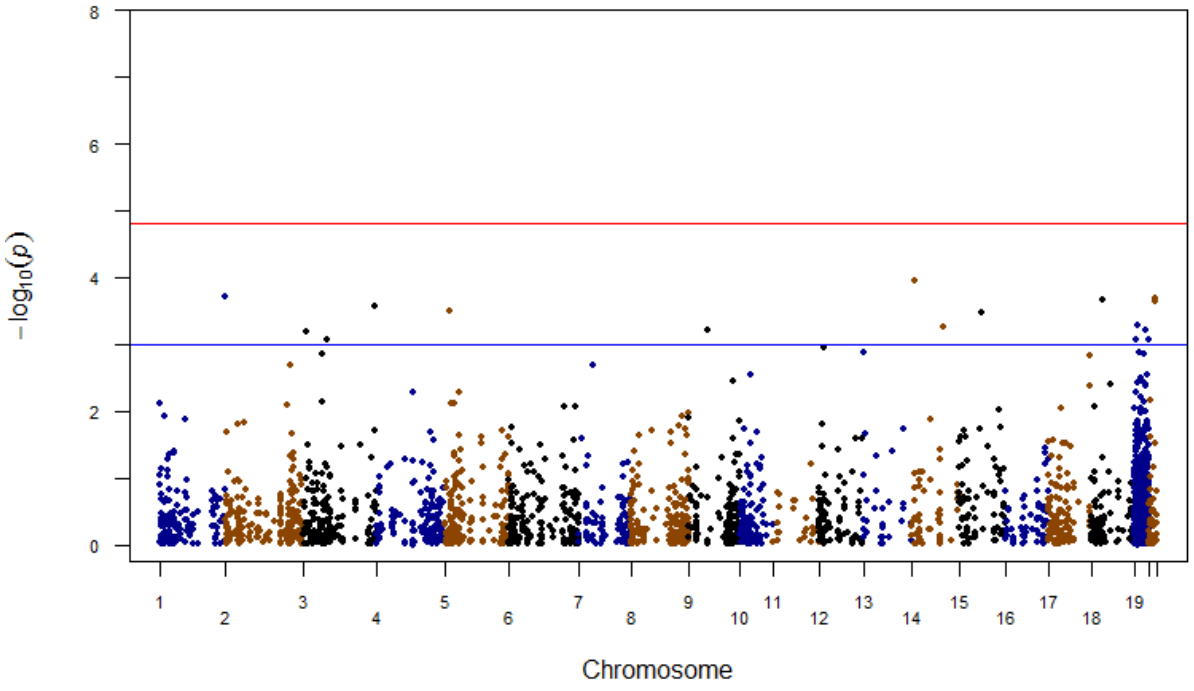


Fig 3.5. Manhattan plot of association mapping plant height with mixed model with principal component and kinship (P+K). The X axis are 18 pseudomolecules, the 19th position denotes SNPs aligned to contigs that were not assembled in any pseudomolecules and SNPs in last position with orange color are SNPs not aligned to switchgrass reference genome. Y axis is negative log10 of P values. Red line is cutoff using conservative Bonferroni FDR correction at 0.05. Blue line is cutoff of $P = 0.001$.In this dataset each individual had at least 80% SNPs genotyped (S80-6)

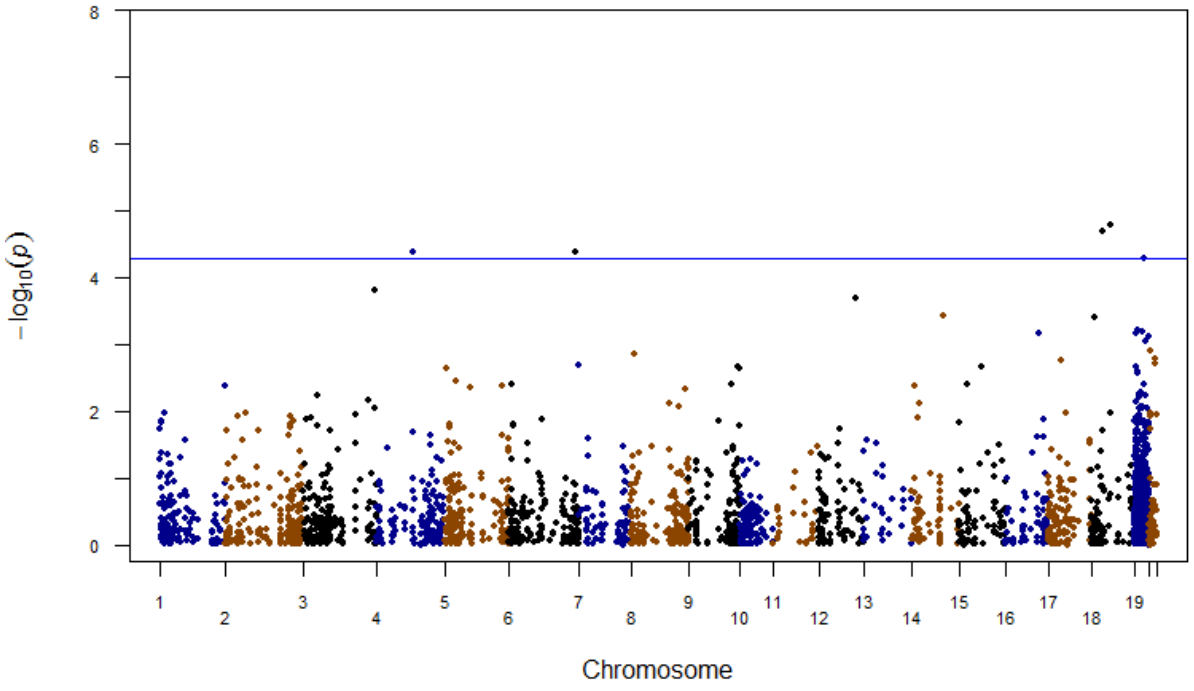


Fig 3.6. Manhattan plot of association mapping stem diameter with mixed model with principal component and kinship (P+K). The X axis are 18 pseudomolecules, the 19th position denotes SNPs aligned to contigs that were not assembled in any pseudomolecules and SNPs in last position with orange color are SNPs not aligned to switchgrass reference genome. Y axis is negative log10 of P values. Blue line is Benjamini & Hochberg FDR correction at 0.05. All were with the dataset where each individual had at least at least 80% SNPs genotyped (S80-6)

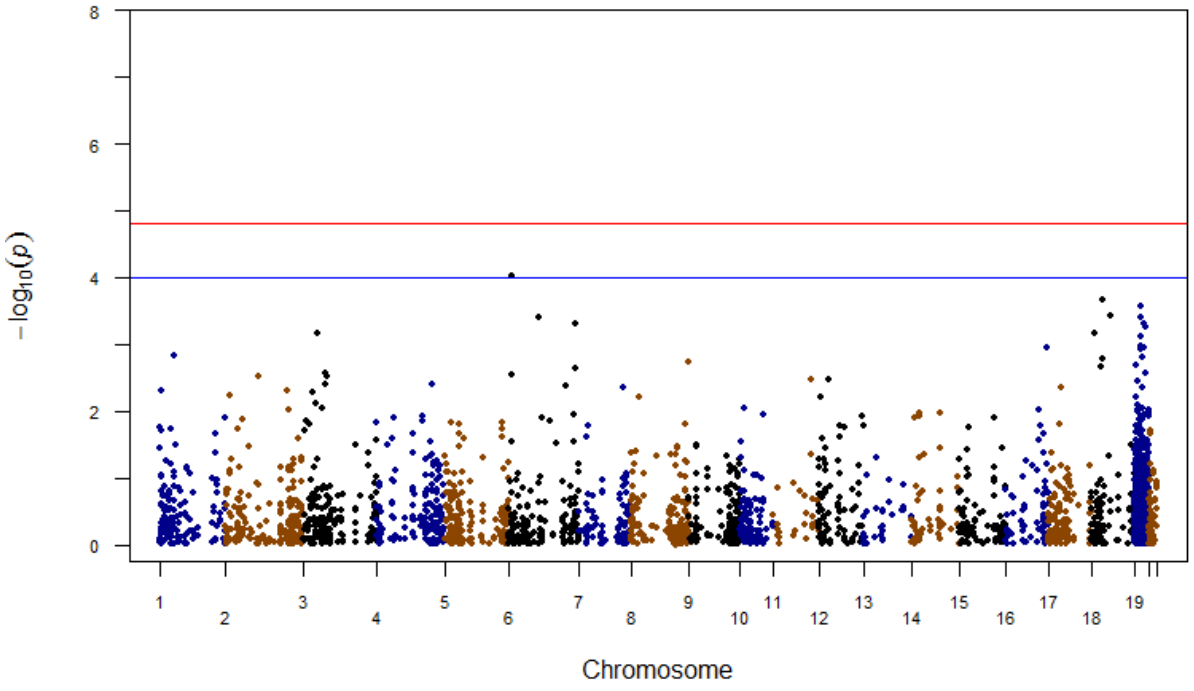


Fig 3.7. Manhattan plot of association mapping flowering time with mixed model with principal component and kinship (P+K). The X axis are 18 pseudomolecules, the 19th position denotes SNPs aligned to contigs that were not assembled in any pseudomolecules and SNPs in last position with orange color are SNPs not aligned to switchgrass reference genome. Y axis is negative log10 of P values. Red line is cutoff using conservative Bonferroni FDR correction at 0.05 and blue line is Benjamini & Hochberg FDR correction at 0.05. All were with the dataset where each individual had at least at least 80% SNPs genotyped (S80-6)

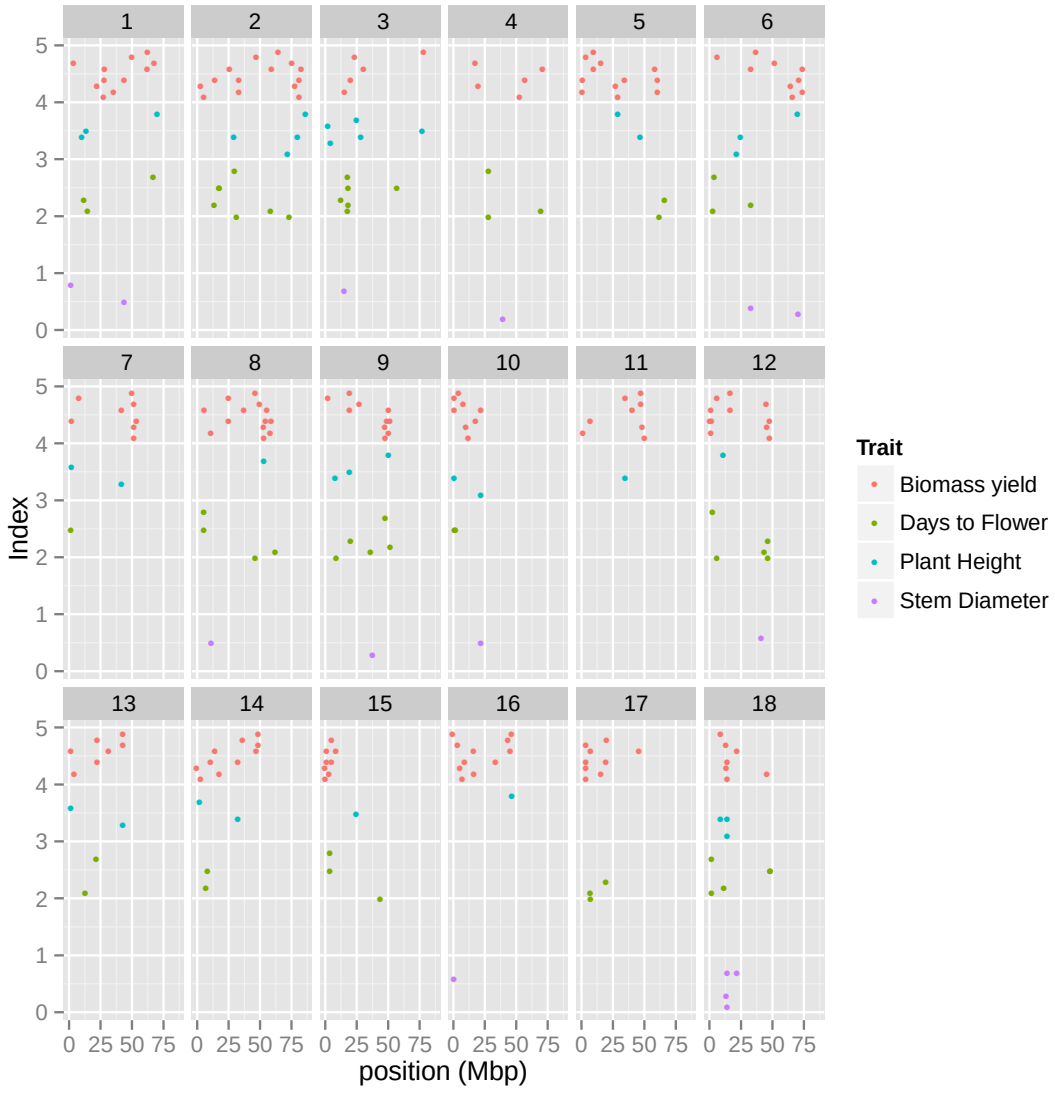


Fig 3.8. All the loci associated with any of the traits (biomass yield, plant height, stem diameter and flowering time) to show the relative position of QTLs. Eighteen boxes represent eighteen pseudomolecules when aligned to switchgrass reference genome. All were with the dataset where each individual had at least at least 80% SNPs genotyped (S80-6)

CONCLUSIONS

Progress from plant breeding depends on the existence of genetic variation, on the accurate identification and selection of superior genotypes, and on the generation of new segregants through recombination. Understanding the extent of genetic diversity within crop germplasm and the way that diversity is structured among and within populations can assist a plant breeding program in using germplasm most effectively. Using genetic markers, plant breeders can theoretically manipulate genes controlling important traits more precisely than previously possible using conventional breeding technologies, thereby accessing previously untapped genetic variants to aid cultivar development.

In the research described in this dissertation, we investigated switchgrass germplasm, both to assess genetic diversity and to identify loci for biomass yield and associated biofuel traits. We developed SNP markers using Genotyping-by-Sequencing (GBS) in order to conduct the major segment of the research.

First, we developed a GBS protocol based on previous methods but with the use of novel enzymes suitable for the switchgrass genome and to improve data quality. We prepared several different SNP datasets based on the read depth and the missing genotype calls at a given locus and used two different software packages to identify SNP markers. We imputed the missing genotypic data.

Second, we used the SNP markers together with SSR and plastid markers to classify switchgrass germplasm largely derived from the southern USA. In agreement with other experiments, we identified the clear split between upland and lowland ecotypes, but in addition,

we showed that the lowland group can be distinctly divided into two subgroups. Generally, upland plants are shorter, have thinner stems, and flower earlier than lowland ecotypes. However, one of our two main lowland groups had similar height as upland accessions, had stem diameter intermediate between upland and the other lowland group, but was still late flowering like other lowland accessions. Some accessions appeared to be hybrids in which each individual was composed of two or three sub populations, depending on the hybrid. In general, these hybrids showed morphological characteristics in between the groups they were composed of. Similarly, some accessions were mixtures of different subgroups, where each individual showed membership to one or another group. The morphological characteristics of these individuals were similar to the group to which they belonged. Apart from the major groups, the accessions could be classified based on their geographic region of origin. The identification of different germplasm pools could guide a plant breeding program to select suitable germplasm. A plant breeder can potentially exploit heterosis by using different subgroups, which are could indicate different heterotic group.

Third, we collected data for important biofuel traits of switchgrass and used the previously identified SNP markers to identify QTL associated with those traits. The repeatability estimates of biomass yield, stem diameter, plant height and flowering time all were very high. There was significant genotype × location interaction and genotype × year interaction for the yield, height and flowering time. The stem diameter did not have significant interaction, which may also be due to the fact that the data were from fewer environments than for the other traits. We applied genome wide association mapping to these traits with the SNP identified earlier. We identified several QTL associated with yield, height, stem diameter and flowering time. The number of QTL increased when using a larger marker dataset even though these SNP had a

larger proportion of missed genotype calls. However, QTL identified by the smaller set were included among those identified with the larger set. A few QTL were identified for multiple traits, especially for biomass yield, height and/or stem diameter. The SNP markers associated with these QTL showed sequence similarity to annotated genes from other organisms, some of which may be potential candidate genes. Along with those QTL with known annotation, we identified several novel loci associated with the traits.

This study has two major limitations. First, the genotypes used in this study were very diverse, so that very limited linkage disequilibrium was present. The numbers of markers we developed was not sufficient to saturate the genome and consequently, we undoubtedly did not detect important QTL that were segregating in the population. Second, the population we used showed strong population structure, potentially fixing the SNP and/or QTL between subpopulations, which would prevent their detection. However, we identified some QTL and the SNP associated with them should be close to or even at the gene controlling the trait. We expect to link the findings of this study to future QTL identifying studies such as that using a nested association mapping population, and to further saturate the genome of this population using other methods, including exome capture.

The results from this study can be used for future switchgrass improvement programs. The two different components of this study; a) identification of population structure b) identification of QTL associated with traits of breeding interest; both contribute to a plant breeding program differently. The population structure helps identifying the germplasm resources for breeding cultivars. A local germplasm pool already adapted to a certain geographic reason is more suitable for breeding program in that locale. The use of genetic markers to identify the genomic group helps evaluate new germplasm. The different heterotic groups, as

assessed by genetic markers, can be used to exploit the heterosis and develop a superior cultivar. The identification of QTL associated with several traits of interest helps a breeding program to select or discard the plant materials. The markers can be utilized in early selection and removal of plant materials. These markers, assuming their validity, can be utilized in a marker assisted selection program for mass screening of progenies and using only the individuals with desired marker signal for further evaluation. These markers are especially useful in early generations when there are higher numbers of progenies to be screened. In addition to the screening, these markers can be utilized in introgression of a single QTL or stacking multiple QTL through a backcrossing procedure.

This study used a diverse population set, with some individuals and populations exhibiting superior phenotypes and desired alleles from the identified QTL. Not only the markers identified in this study were associated with QTL of desired traits, but also the plant material itself can be a useful resource for future breeding programs. The individuals that had desirable traits can also be used as a parental source. Using the parents with different QTL for the same traits will increase richness of desired alleles.

This is a first study in switchgrass identifying QTL from a diverse population set; especially within the Southern germplasm. We believe that the results of this study will be useful in the switchgrass community to exploit its potential as a biofuel crop.