# Analysis of Segregation Ratio Distortion and Linkage Mapping in Two Switchgrass F1 Populations: Lowland-lowland and Lowland-upland Using Genotyping by Sequencing Data

**Rasyidah Razar** ( rm61196@uga.edu )
University of Georgia https://orcid.org/0000-0002-2463-7386
**Katrien Devos**
University of Georgia
**Ali Missaoui**
University of Georgia

**Research article**

**Keywords:** Switchgrass, Genotyping-by-sequencing, Linkage maps, Segregation ratio distortion

**Posted Date:** September 30th, 2019

**DOI:** https://doi.org/10.21203/rs.2.15334/v1

## Abstract

Background: Switchgrass is an emerging bioenergy crop due to its perennial nature, high biomass yield, and ability to grow in marginal land. The high genetic diversity in switchgrass germplasm can be exploited to capture favorable traits that increase the range of adaptation and biomass yield. Genetic diversity can be explored using single nucleotide polymorphisms (SNPs) that next-generation sequencing has made possible for high-throughput genotyping. We used genotyping-by-sequencing (GBS) of genomic fragments resulting from two methylation sensitive restriction enzymes: PstI and MspI . Two bi-parental F1 populations were developed from crosses between lowland B6 and lowland AP13 (AB population), and lowland B6 with upland VS16 genotypes (BV population), with a target number of 298 progenies in each population. Pseudo-testcross strategy was adopted to perform linkage analysis in these populations that are segregating for winter dormancy using single dose markers (SDA): heterozygous in one parent and homozygous in the other parent. We compared the amount of polymorphisms between the two crosses and examined the pattern of segregation distortion based on the SNPs data generated.

Results: Two genetic maps were generated for each population, with 2772 markers in AB and 3766 markers in BV. The higher number of markers in the BV population was expected for since the parents originated from different ecotypes and verified to have the highest genetic distance. More segregation distortion was observed in markers located in the telomeric regions where more genes reside. More markers from the AB population exhibited segregation distortion compared to the BV, and the proportion of heterozygous alleles were significantly higher than homozygous alleles in AB population. The linkage maps showed strong collinearity with P. virgatum V5.1 reference genome with a very minimal number of markers originating from different chromosomes.

Conclusion: Understanding the extent of segregation distortion in switchgrass crosses is important for the correct inclusion of markers based on their segregation ratio when constructing a linkage map. Switchgrass linkage maps should be a useful resource to dissect beneficial biomass traits linked to SNP markers.


## Background

Switchgrass is a C4 perennial warm season grass native to most of North America, spanning southern Canada, most of the United States, and northern Mexico (1). It has traditionally been used for pasture and rangeland grazing since the 1940s (1). It was selected by the U.S. Department of Energy (DOE) Biofuel Feedstock Development Program (BFDP) in 1991 (2) as a herbaceous model species for biomass energy due to its high biomass yield, low nutrient and water requirements and suitability for planting in marginal land unsuitable for grain and forage crops (3, 4).

Switchgrass is divided taxonomically into two major ecotypes, upland and lowland based on their phenotypic divergence caused by the difference in latitudinal adaptation (1). Upland ecotypes are widely adapted to latitudes north of 34°N, extending into much of eastern Canada, while lowland ecotypes are adapted to the south, approximately up to 42°N in the western portion of the grassland range, but can be found as far north as 45°N in eastern North America due to the moderate climate resulting from ocean effects (1). Adaptation of upland ecotypes involves phenology for a short growing season and tolerance to cold winter temperatures (reviewed in Lowry et al. (5)). On the other hand, lowland ecotypes are more adapted to a longer growing season and are less tolerant to cold temperatures. In terms of morphological characters, lowland ecotypes are taller, have fewer and larger tillers, longer and wider leaf blades, and thicker stems than the upland ecotypes (1).

The basic chromosome number of switchgrass is x = 9, with a wide range of somatic chromosome counts, from n = 18 to 108 (6, 7). Lowland ecotypes are mostly tetraploid (2n = 4x = 36 chromosomes) while octoploids (2n = 8x = 72 chromosomes) are very rarely found (8). Upland ecotypes exist for both tetraploids and octoploids, with the octoploids being approximately two to three times more abundant than the tetraploid. Zhang et al. (9) estimated the earliest divergence of upland and lowland ecotypes to be around 1.3 Mybp and the divergence of two similar species in the *Panicum* genus: *P. virgatum and P. hallii* was around 5.3 Mybp. Switchgrass is predominantly cross-pollinated with gametophytic self-incompatiblity (10), however a wide range of successful self-pollination has been reported in literature (11–14). Crossing between the upland and lowland ecotypes is possible at the tetraploid level and several studies showed that the inheritance in tetraploid switchgrass is disomic (15–17). Significant flow of genes occurred between the two diverse ecotypes during the glacial maxima about a million years ago (8, 9).

Prior to the use of genetic marker to group the different accessions of switchgrass, the distinct phenotypic morphologies and natural habitat were used as primary sources of information (18). With the advances in molecular markers for genotypic identification, switchgrass ecotypes were grouped based on genetic marker profiles, including chloroplast (cpDNA) sequences (9, 19–21), simple sequence repeats (SSR) (9, 21), and recently single nucleotide polymorphisms (SNP) (22–24). Not only have these markers successfully grouped switchgrass accessions into lowland and upland ecotypes, SNP markers could cluster them into higher number of groupings that are linked to ecotypes, cytotypes, and geographical origins (22).

In this study we developed two F1 mapping populations derived from crosses between the winter dormant 4x lowland AP13 and a non-dormant 4x lowland B6, and B6 with the winter dormant 4x upland VS16. The objectives of the crosses are to develop two populations that segregate between the two dormancy levels, together with other traits such as cold tolerance, biofuel quality traits, and biomass yield. We used genotyping by sequencing to selectively genotype each individual in the two populations by target sequencing at the potential gene regions. This is done

through DNA cleaving with a combination of two methylation-sensitive restriction enzymes (REs): “rare cutter” PstI and “common cutter” MspI, and DNA barcoding to enable sample multiplexing in one library pool (25). By using methylation-sensitive REs, the repetitive regions of genomes can be avoided and lower copy regions can be targeted more for sequencing (25). The use of a “rare cutter” enzyme has been shown to successfully reduce genome complexity for large, complex, and polyploid genomes (26, 27).

For the construction of linkage maps, the two-way pseudo-testcross mapping strategy is adopted whereby recombinations that occur at each parental side; during egg and pollen formation, are used to construct two linkage maps per chromosome (28, 29). This paper describes the genotyping-by-sequencing methodology and construction of two genetic maps for each population. We then investigate the amount of polymorphisms between the two crosses; lowland-lowland and lowland-upland, and the pattern of segregation distortion based on the SNPs data generated. We used SNPs data from the three parents used in the crossings to calculate Euclidean genetic distance in order to verify the level of diversity within each cross in view of two hypotheses: 1) A population derived from a wide cross such as between different ecotypes will have a higher number of SNPs, and 2) A population with higher diversity should contain more segregation distortion markers due to the existing reproductive barrier that resulted from the cross between two diverse parents (16).

## Results

# a) Genotyping by sequencing

## AB population (AP13 x B6)

The size of raw reads for Pool 1 was 100 GB and Pool 2 was 90 GB. Total number of sequences generated by the Illumina NextSeq platform was 2,200,716,698 (2.2 B) with an average number of 7,335,722 (7.3 M) sequences per individual. The range of sequence length was from 35 to 151 bases, mean percentage of GC content was 51.3%, and average phred score per sequence was 33.6. The average number of sequences for reads that were trimmed for enzyme cut site and low quality bases (Q < 33) were 6,204,226 (6.2 M) sequences per individual with a range of 20—142 bases. Mean percentage of GC content was 50.7% and mean phred score per sequence was 33.9.

## BV population (B6 x VS16)

The size of raw reads for both pools was 100 GB. Total number of sequences generated by the Illumina NextSeq platform was 2,372,553,906 (2.4 B) with average number of 7,908,513 (7.9 M) sequences per individual. The range of sequence length was 35—151 bases, mean percentage of GC content was 49.2%, and average phred score per sequence was 34.1. The average number of sequences for reads that were trimmed for enzyme cut site and low quality bases (Q < 33) were 7,187,731 (7.2 M) sequences per individual with a range of 21—142 bases. Mean percentage of GC content was 48.6% and mean phred score per sequence was 35.0.

# b) Variant calling

## AB population (AP13 x B6)

Mean percentage alignment of reads to the reference genome was 65.12% (3,984,908 bases), suggesting that 34.9% of the sequenced reads cannot be used in the SNPs mining. Out of the 3.98 million aligned reads, 2.18% have mate mapped to a different chromosome in the reference genome. This suggests paralogous regions which are expected to align to multiple locations in the genome (30), and this is common in polyploids where homoeologous chromosomes can have duplicated genomic regions. By considering biallelic loci in SNP calling, the problem of calling paralogs can be significantly reduced (30).

The total number of variants in the raw VCF output after GATK HaplotypeCaller and GenotypeGVCFs processes was 2,539,025 (2.5 M). Removal of loci with quality score less than 20 and genotype quality scores less than 20 resulted in the same number of variants. This means that the GATK pipeline applied a stringent filtering step and gave an output of only high quality SNPs. Removal of the loci that were 20% missing in genotypes and loci that were not biallelics resulted in 44,698 variants. Variants were filtered for minor allele frequency (MAF) < 0.05 and resulted in 19,830 sites. The same number of variants was retained after filtering for alleles with less than 8 sequence depth. Coverage of 8 reads for variants calling means only 0.4% ($0.5^8$) of the heterozygous alleles will be mistakenly called as homozygous, setting a very high confidence level for SNP calling.

The average allele counts per variant site in the final filtered file was 519 alleles, which means that the average SNP coverage for the population is 519X and for individuals is 1.73X. Histogram of allele frequency showed 3 peaks—frequency of 0.25, 0.5, and 0.75 (Figure 1a), which could be

the result of the following parental allele segregations:

### AA x Aa = 1AA: 1Aa (0.75 A, 0.25 a)

### Aa x Aa = 1AA: 2Aa: 1aa (0.50 A, 0.50 a)

### AA x aa = Aa (0.50 A, 0.50 a)

Removal of variants that were not present in both parents resulted in 14,816 variants. Selfed and/or low coverage individuals were identified using markers that are fixed for different alleles between parents—all marker “0” (AA) in AP13 and marker “2” (CC) in B6 were used whereby hybrid progenies were expected to have at least 80% total number of marker “1” (AC). As a result, 9 progenies were discarded from further analysis. Markers were then extracted for single dose alleles (SDA) in each side of the parents. Additional 4 progenies were discarded because of > 50% missing data in SDA for both parents, leaving 285 progenies used in linkage mapping. Final filtering for number “2” allele frequency > 20% and chi-square goodness of fit for 1:1 allele segregation (p-value > $E10^{-15}$) resulted in 3548 (24% of filtered variants) and 3823 (26% of filtered variants) alleles for maternal and paternal SDA, respectively. There were a total of 398 alleles (2.7% of filtered variants) that were heterozygous in both parents (single dose alleles present in both parents).

## BV population (B6 x VS16)

Mean percentage alignment of reads to the reference genome was 73.26% (5,253,577 bases), meaning that 26.74% of the sequenced reads cannot be used in SNP mining. Out of the 5.25 million aligned reads, 2.81% have mate mapped to a different chromosome in the reference genome. The total number of variants in the raw VCF output after GATK HaplotypeCaller and GenotypeGVCFs processes was 6,043,505 (6.0 M). The amount of variants in this population is more than twice the amount in AB population (2.5 M). Removal of loci with quality score less than 20 resulted in 5,016,367 variants. Removal of individual genotypes with genotype quality scores less than 20 resulted in the same number of variants (5.0 M). Removal of loci that are 20% missing across genotypes resulted in 87,281 loci. Removal of loci that are not biallelics resulted in 75,431 variants. Filtering for minor allele frequency (MAF) < 0.05 resulted in 31,374 variants. The same number of variants was retained after filtering for alleles with less than 8 sequences depth.

The average SNP coverage per population is 536X and per individual is 1.79X. Histogram of allele frequency showed 5 peaks—frequency of <0.025, 0.25, 0.5, 0.75, and >0.975 (Figure 1b). These can be the results of the following parental allele segregations:

### AA x AA = AA (1 A, 0 a)

### AA x Aa = 1AA: 1Aa (0.75 A, 0.25 a)

### Aa x Aa = 1AA: 2Aa: 1aa (0.50 A, 0.50 a)

### AA x aa = Aa (0.50 A, 0.50 a)

Removal of variants that were not present in both parents resulted in 27,517 variants. Using markers that were fixed for different alleles between parents and expectation of a minimum of 80% heterozygous genotype in the progenies resulted in the identification of 66 progenies that were potentially a cross between B6 and contaminated pollen source. Markers were then extracted for single dose alleles (SDA) in each side of parents. Additional 5 progenies were discarded because of > 50% missing data in SDA for both parents, leaving 227 progenies used in linkage mapping. Final filtering for number “2” allele frequency > 20% and chi-square goodness of fit for 1:1 allele segregation (p-value > $E10^{-15}$) resulted in 5693 (21% of filtered variants) and 7883 (29% of filtered variants) alleles for maternal and paternal SDA, respectively. There were a total of 758 alleles (2.8% of filtered variants) that were heterozygous in both parents.

# c) Genetic maps

## AB population (AP13 x B6)

For maternal markers, 45 loci were found to be identical and thus discarded. 871 well segregated markers were firstly used in the construction of framework LG, while the remaining segregation distorted markers were added in their respective LG based on their strongest cross-link to the well-segregated loci in the framework LG. This step maintained the 18 LGs for maternal map. There were a total of 1540 markers (43% of total maternal SDA) that were able to be ordered within linkage groups with a total map length of 2475.61 cM and 1.76 cM of average inter-marker distance (Table 1). For paternal markers, there were 65 identical loci that were discarded and a total of 1171 well segregated markers that were used in the construction of framework LG. There were a total of 1232 markers (32% of total paternal SDA) that were able to be ordered within linkage groups with a total map length of 1704.06 cM and 1.78 cM of average inter-marker distance (Table 1). Illustrative comparison of maternal and paternal linkage maps shows that maternal LGs are consistently longer than paternal LGs (Figure 2). Haplotype map shows distinct recombination blocks with minor occurrences of only one marker genotype per recombination block (Figure S1). Distribution of markers along genetic map points to the possible pericentromeric region where the marker density is highest within 20 cM genetic bin sliding window (Figure S2). Distribution of allele frequency for mapped markers showed two peaks for 0.25 and 0.75 allele frequencies signifying AA x Aa cross (Figure S3).

## BV population (B6 x VS16)

For maternal markers, 547 loci were found to be identical and thus discarded. 3867 well segregated markers were used in the construction of framework LG, while the remaining segregation distorted markers were added using strongest cross-link. There were a total of 1824 markers (47% of total maternal SDA) that were able to be ordered within linkage groups with a total map length of 1482.90 cM and 1.07 cM of average inter-marker distance (Table 1). For paternal markers, there were 904 identical loci that were discarded and a total of 4433 well segregated markers that were used in the construction of framework LG. There were a total of 1942 markers (44% of total paternal SDA) that were able to be ordered within linkage groups with a total map length of 1606.79cM and 0.92 cM of average inter-marker distance (Table 1). Comparative illustration of maternal and paternal linkage maps shows no obvious pattern of higher centiMorgan distance between the two maps (Figure 3). The haplotype map for this population still shows distinct recombination blocks, however with higher occurrences of one marker genotype per recombination block (Figure S1). Possible pericentromeric region is shown in the marker frequency along genetic map histogram (Figure S2). Distribution of allele frequency for mapped markers showed two peaks for 0.25 and 0.75 allele frequencies signifying AA x Aa cross (Figure S3).

# d) Segregation ratio distortion

## AB population (AP13 x B6)

Percentage distorted markers was higher than well segregated markers for both maps across linkage groups, with the exception of LG 3N in the maternal map and LG 4K and 5N in the paternal map (Figure 4). Total percentage of distorted alleles is 70%. The percentage distorted allele across linkage groups ranged from 42—89% in maternal map and 33—100% in paternal map. The highest frequency of distorted alleles in maternal map was observed in LG 7K (89%) and in LG 7N (100%) of paternal map. For both maternal and paternal maps, segregation ratio distortion was caused mainly by an excess of heterozygous genotypes to homozygous genotypes from the 1:1 ratio across the linkage groups, with the exception of LG 7N in B6 map where frequency of homozygous alleles was higher (Figure 5, Table S1). Wilcoxon signed rank test comparing heterozygous and homozygous allele frequencies in each linkage group for both AP13 and B6 maps proved that the frequency of heterozygous alleles are significantly higher than homozygous alleles (p-value < 0.0001) (Table S1). There is more prevalence of contiguous severe segregation distortion in the telomeric regions compared to the pericentromeric region (Figure 7). LGs with severe distortion at telomeric regions include 1K, 4K, 5K, 7K, 8K, 9K, 3N, 4N, 5N, 6N, and 8N of AP13 map, and 2K, 3K, 6K, 7K, 8K, 9K, 5N, 8N, and 9N of B6 map. Severe distortion at pericentromeric regions was observed in 2K, 3K, 7N, and 9N of AP13 map, and 1N and 6N of B6 map.

## BV population (B6 x VS16)

Percentage of well segregated markers was higher than distorted markers for both maps across linkage groups, with the exception of LG 5N in maternal map and LG 5K, 5N, and 7N in paternal map (Figure 4). Total percentage of distorted alleles is 20%; percentage of distorted allele across linkage groups ranged from 2—80% in maternal map and 8—100% in paternal map. The highest frequency of distorted alleles in maternal map was observed in LG 5N (80%) and in LG 5K (100%) of paternal map. For both maternal and paternal maps, segregation ratio distortion was caused mainly by an excess of homozygous genotype to heterozygous genotype from the 1:1 ratio across the linkage groups, except for LG 2K and 4K in B6 map and LG 1K, 5N, and 7N in VS16 map (Figure 6, Table S1). Wilcoxon signed rank test comparing heterozygous and homozygous allele frequencies across linkage groups for both B6 and VS16 maps proved that the frequency of homozygous alleles are significantly higher than heterozygous alleles (p-value < 0.0001) in majority of linkage groups (Table S1). Similar with AB population, there is more incidence of contiguous severe segregation distortion at telomeric regions compared to pericentromeric regions (Figure 8). LGs with severe distortion at

telomeres include 2K, 3K, 4K, 3N, 4N, 6N, and 8N of B6 map, and 1K, 5K, 7K, 8K, 4N, 6N, and 7N of VS16 map. Severe distortion at pericentromeric region was observed in 9K and 2N of B6 map, and 1N, 2N, and 9N of VS16 map.

# e) Estimation of genetic distance

The number of SNPs aggregating in the 3 parents after filtering steps was 39,259. After selecting for SNPs that are present in all parents, 34,011 SNPs were used for calculation of genetic similarity and Euclidean distance (Table 2). Parents with the highest genetic similarity were AP13 and VS16, with the fraction of shared loci exceeding half of the total SNPs. Parents with the highest genetic divergence were B6 and VS16.

# f) Collinearity of genetic maps with P. virgatum V5.1 reference genome

## AB population (AP13 x B6)

Collinearity of markers in the genetic position (cM) with the physical position in the reference genome (Mb) suggested high level of collinearity, apart from some local misordering such as in LG 4N of AP13 map (Figure 9). There were only few markers originating from different chromosomes; for the concatenated K and N chromosomes in AP13 map, only 0.88% and 0.40% SNPs, respectively with a total of 0.65% for both maps. For B6 map, a slightly higher number of SNPs originated from different chromosome, 1.62% for K chromosomes and 2.61% for N chromosomes with a total of 2.11% for both maps. For this population, only 0.76% of the markers had no hit with BLASTn with a cutoff of E-value < $1 \times 10^{-5}$. Very few small gaps were observed in the physical map due to the scarcity of SDA alleles in these regions. These included LG 3N of AP13 map, and 4K, 5K, and 4N of B6 map. Small regions of inversion can also be observed at some LGs such as 4K of AP13 map and 3K of B6 map.

## BV population (B6 x VS16)

Collinearity between physical and genetic maps was also observed for BV maps, but not as strong as AB maps (Figure 10). There were few markers originating from different chromosomes; for the concatenated K and N chromosomes in the B6 map, 1.50% and 2.42% SNPs, respectively with a total of 1.92% for both maps. For VS16 map, 1.47% and 1.76% of the SNPs originated from different chromosomes for K and N subgenomes, respectively with a total of 1.60% for both maps. For this population, 1.27% of the markers showed no hit with BLASTn with a cutoff of E-value < $1 \times 10^{-5}$. More gaps were observed in the physical maps including 4K, 5K, 7K, 8K, 4N, 6N, and 7N of BV map, and 1K, 2K, 3K, and 3N of VS16 map. A small region of inversion is observed at 5K of BV map, and 9K of VS16 map.

### Discussion

Several genomic studies in switchgrass have contributed to the ample genomic resources for the species. These include restriction fragment length polymorphism (RFLP) (15), cDNA libraries sequencing (31–35), bacterial artificial chromosome (BAC) libraries sequencing (36), simple sequence repeats (SSR) (14, 16, 37–39), and more recently is genotyping-by-sequencing (GBS) (17, 40). Genetic and genomic studies in switchgrass were greatly improved by the publicly available reference genome in the JGI Phytozome database; *Panicum virgatum* V4.1. The reads were sequenced from AP13 and the overlapped contigs were anchored on the AP13 x VS16 genetic map. The size of genome assembly is approximately 1,165.7 Mb where a total of 89,680 contigs were localized in the main 18 scaffolds (pseudomolecules) that were labeled 1–9 (K or N) based on subgenome specific markers. At the time of preparation of this manuscript, a draft *Panicum virgatum* V5.1 reference genome was just released internally and we were able to use it for estimating the collinearity between genetic and physical maps.

In this study we compare the outcome of genotyping-by-sequencing on two switchgrass F1 populations: lowland-lowland (AP13 x B6) and lowland-upland (B6 x VS16).Both populations share a common parent, B6, which is a winter non-dormant lowland genotype. The number of progenies that were initially intended to be used for SNPs generation was 298, however due to low reads coverage for some individuals, 285 progenies for AB population and 227 progenies for BV population were used in the analysis. The size of mapping population is important to determine the correct ordering of markers that are closer than ≈2–5 cM (41) and to improve map resolution (26). The moderate to large number of progenies used in this study are comparable to numbers of progenies used in many linkage mapping studies that were around 80 - 200 progenies ((14, 26, 37, 41, 42)).

The number of sequences generated was 2.2 B for AB and 2.4 B for BV population. The higher number of sequences generated for BV population resulted in higher number of sequences per individual for reads that were already trimmed for enzyme cut site and poor quality; 7.2 M sequences/individual for BV compared to 6.2 M sequences/individual for AB might be due to difference in genomic DNA concentrations. By increasing genomic DNA concentration for sequencing we were able to increase the mean SNPs coverage per population from 519 alleles per site

for AB population to 536 alleles per site for BV population. The higher percentage of reads alignment to V4.1 reference genome for BV (73.26%) compared to AB (65.12%) can be explained by the improved genome coverage in this population. The high SNP coverage in our study was also due to the use of two enzymes combination: PstI-MspI. In this strategy, fragment amplification occurs for DNA strands containing the ligated forward (PstI cut-site) and common Y (MspI cut-site) adapters (26). Since PstI is a six base rare cutter and MspI is a four base common cutter, this enzymes combination could lead to higher sequencing depth. Previous GBS study in switchgrass utilizing only one enzyme; ApeKI (17) and PstI (40), reported lower loci coverage. The choice of RE in GBS is crucial since it determines the tradeoff between high number of fragments and sequencing depth of fragments (17, 41). Several studies showed that ApeKI can generate large numbers of SNPs because it is a frequent cutter with partial sensitivity to methylation, but PstI can give at least 8 times higher coverage and its methylation sensitive nature can target more gene-rich regions (17, 27). In this study both aspects were covered, the large number of variants generated due to the high output sequencing platform, and deep loci coverage from the use of two enzymes combination.

The number of variants in the raw VCF output was also higher for the BV population, 6M compared to 2.5M for the AB population. This can be explained by the higher sequence coverage and higher number of polymorphisms in BV resulting from a cross between lowland and upland ecotypes. It was interesting to note that five peaks were observed in the histogram of allele frequency after reads were retained for biallelics in BV population. This suggests that there were more hybridization of the same homozygous loci for B6 and VS16 parents. To ensure that only high quality markers were included in the final map, allele frequency distributions for mapped markers in both populations were analyzed (Figure S3). The observed two peaks at 0.25 and 0.75 frequencies confirm the crosses between heterozygous and homozygous loci (SDA), and allele frequencies showed MAF $\geq$ 0.05 for both populations. These results confirm that we have included only true markers with no sequencing errors.

We chose to include only single dose alleles (SDA) present in either parent for linkage maps construction because there were too little number of SDA x SDA alleles, 398 in AB and 758 in BV. We have successfully produced four linkage maps with a total number of 2772 SNPs in AB population and 3766 SNPs in BV population. It was expected that more markers can be mapped in BV population since there is a higher number of SNPs and Indels. Even though more markers were included in the map, the total map length was lower for both parental maps in BV compared to AB. This is due to the low inter-marker distance for BV map; 1.07 and 0.92 for B6 and VS16 maps, respectively, compared to AB, 1.76 and 1.78 for AP13 and B6 maps, respectively. A possible pericentromeric region for each linkage group is suggested based on the highest marker density found within 20 cM bin sliding window. The high frequency of markers in the likely pericentromeric region of the genetic map may have lower recombination resulting in lower amount of recombination events observed (26, 27, 29).

Inter- and intraspecific crosses can often lead to a distortion of segregation ratios in the hybrid progenies (43–46). Segregation distortion is a deviation of segregation ratios from the expected Mendelian fractions (47, 48) that may result from competition among gametes or from abortion of the gamete or zygote. Competition among gametes may occur because of gametophyte genes expressed during postmeiosis of the microspore and pollen development in angiosperms (48, 49). Genetic differences among pollen may lead to gametophyte competition and selection, which result in nonrandom fertilization, while hybrid sterility genes can cause abortion of a specific gamete or zygote genotypes (49). Several linkage mapping studies in switchgrass reported the existence of segregation distortion with level of distortion that was dependent on population. Okada et al. (16) reported 3% distorted single dose alleles (SDA) in female map and 14 % in male map, in a population that was derived from lowlands Kanlow and Alamo. Higher amount of marker distortion in male map was hypothesized to be due to the presence of rye's self-incompatibility (SI) Z locus ortholog that present in LG VIIb. SI locus can cause failure of fertilization by the affected pollen and hence more marker distortion was observed for male maps. Marker distortion in other LG not containing S-Z locus was proposed to be due to post-zygotic mechanisms resulting from two-locus interactions between parents, by which the author suggested that geographically distant parental origin may contribute to a reproductive barrier leading to segregation distortion. Other suggested cause for segregation distortion in the study was the presence of transmission distorter loci.

Serba et al. (37) reported 12 % of distorted markers in the linkage map of the an F1 population derived from lowland AP13 and upland VS16.The majority of the markers formed three clusters in the male map LG Ia, IIIb, and VIIb. The study was in agreement with the SI loci hypothesis since rye's Z and S orthologs were found in LG VIIb and IIIb, respectively, by which both of these LG were found to contain clusters of distorted markers. In another study conducted by Li et al. (14), a total of 17.1 % distorted markers were observed in the final integrated map of selfed and hybrid third generation populations. The first generation populations in this study were derived from crosses developed by Okada et al. (16) and Serba et al. (37). The distorted markers were however found distributed throughout the maps and did not form any major cluster. A post-zygotic interaction was suggested as the cause of distortion due to the significant interaction between markers in the male and female maps. Fiedler et al. (40) used the same mapping population developed by Okada et al. (16) and produced a comparative genetic map using SNP marker, and found higher amounts of distorted alleles in the mild-TRD and severe-TRD maps with 21% and 33% distorted markers, respectively. The inclusion of distorted alleles had increased the total map length by 10% for mild TRD and 20% for severe TRD, when compared to the well-segregated SDA map. The study showed a distribution of distorted markers in some instances in half of the length of LGs, at pericentromeric or telomeric regions, and on majority of the markers in the LGs.

In our study, the percentage of distorted markers in the final map of AB population is higher than BV population - 70% in AB and 20% in BV. We also found longer length of genetic maps for AB population (2475.61 cM for AP13 and 1704.06 cM for B6) compared to BV population (1482.90

cM for B6 and 1606.79 cM for VS16). Longer map length could indicates greater rates of recombination (16, 42), or simply a result of map inflation caused by transmission ratio distortion that could introduce spurious linkage, biased estimates of recombination fractions, or incorrect marker order (reviewed in Okada et al. (16)). Because of a significantly higher amount of distorted markers was observed for AB population and distortion was evidently seen in both AP13 and B6 parental maps, we can suggest postzygotic interactions as the cause of segregation ratio distortion. Postzygotic interaction can be caused by the presence of hybrid sterility genes favoring certain types of genotype formation which lead to abortion of zygote. In our case, it favored the formation of heterozygous loci. For polyploid outcrossers, it is known that heterozygosity can increase colonization ability via improved plant fitness through heterosis and masking of deleterious alleles by the presence of extra genome (50, 51). In our study, the allele state was analyzed within subgenome and not across subgenomes; thereby the theory of fixed heterozygosity through hybridization of diverged subgenomes may not be the case. From our first year field phenotypic data, there was a high mid-parent heterosis (MPH) for these three traits: fall regrowth height, normalized difference vegetation index, and spring emergence date, with 44%, 56%, and 15.3%, MPH respectively (unpublished data). These three traits were showed to be significantly correlated with dry biomass yield in a switchgrass diversity panel consisting of both upland and lowland ecotypes (52). Therefore, we can attribute the overall increase in plant fitness for AB population was due to high level of heterozygosity.

For BV population, since the two parents were different ecotypes with high degree of divergence (high Euclidean distance calculated between B6 and VS16 parents), crossing them could potentially lead to accumulation of genes favoring survival in both adapted regions. Minimal segregation distortion was observed, and across LGs, most markers presented a 1:1 segregation between heterozygous and homozygous genotypes, with inclination towards a Mendelian segregation model. In this population, we observed more regions of distortion in the paternal map especially on chromosomes 5K, 5N, and 7N, where 100%, 52%, and 96% of the markers on each respective LG were distorted. Since more distortion was observed in paternal map and involved chromosome 7N that contains rye's self-incompatibility (SI) Z locus ortholog, we can infer that the mechanism of distortion might be at least partly due to male self-incompatibility. We observed more instances of contiguous severe segregation distortion at telomeric regions compared to pericentromeric regions for both AB and BV populations. Daverdin et al. (29) reported a significantly higher number of genic markers located at the distal chromosome regions compared to pericentromeric regions, while the opposite was true for genomic markers. With this in mind, we can deduce that more functional segregation distortion occurred at genic regions in the distal locations compared to non-genic regions in the pericentromeric location.

A very high percentage of SNPs were successfully aligned through BLASTn to the correct chromosomes in the switchgrass reference genome V5.1. In B6 population, 99.35% for AP13 map and 97.89% for B6 map while 98.09% for B6 map and 98.40% for VS16 map from BV population aligned correctly. Fiedler et al. (40) found approximately 60% of markers in their severe TRD map to align to the same chromosome when *P. virgatum* V1.1 was used for alignment. This suggested a tremendous improvement of switchgrass sequence assembly in the V5.1 since we only have 0.65% and 1.92% of markers in AB and BV population, respectively, that were not aligned to the same chromosomes. We found a stronger collinearity between physical and genetic maps for AB population compared to BV population, and the presence of more marker gaps in the physical map of BV population. One obvious reason is due to the construction of switchgrass reference genome from the AP13 genotype, the same genotype that we used as the maternal parent in AB population. According to Lu et al. (17) the use of reference genome for SNP discovery may not represent the whole genome of a species because some genomic regions might be technically missing and the presence and absence variation (PAV) that is unique to the referenced genotype.

## Conclusions

Comparison of polymorphic markers between two crosses, lowland x lowland and lowland x upland showed that the population derived from different ecotypes contained more variants. This was supported by the higher genetic distance between B6 and VS16 parents compared to AP13 and B6 or AP13 and VS16 parents. Our hypothesis that a population derived from more divergent parents (BV population) would have more distorted markers due to possible reproductive barriers, was proven inaccurate, instead the less divergent AB population contained more distorted markers. A possible cause for the segregation distortion in AB is post-zygotic interactions favoring heterozygous genotypes which results in high level of heterosis in F1 progenies. Lower extent of segregation distortion found in the population derived from the more divergent parents B6 and VS16 can be explained by the presence of more polymorphic markers in this population suggesting accumulation of alleles for adaptation in different geographical regions, and thus the need for specific zygotic formation is not as crucial. Understanding the extent of segregation distortion in switchgrass crosses is important for the correct inclusion of markers based on their segregation ratio when constructing a linkage map. The linkage maps produced in this study will be used in future studies involving QTL mapping for extended production, biomass yield, and biomass quality traits.

## Materials And Methods

# a) Population development

Two F1 populations consisting of 298 progenies derived from two crosses, AP13 x B6 and B6 x VS16 were produced in the greenhouse in 2015 and 2016. A non-dormant genotype B6 was crossed to the dormant lowland genotype AP13. B6 was also crossed to the dormant upland genotype VS16. The parents of the mapping populations are tetraploid (2n = 4x = 36). Both crosses were made in isolation in separate greenhouse sections to prevent unintentional cross-pollination from unidentified source of switchgrass pollen. The plants were cross-pollinated by placing the parents close together. Seeds produced from cross pollinated plants were harvested at maturity and were dried at room temperature before undergoing pre-chilling treatment to break seed dormancy. The pre-chilling treatment consisted of placing the seed on a wet filter paper laid in a petri dish. Petri dish was closed and sealed with parafilm then placed in a 4°C refrigerator for two weeks. After that, the seeds were planted in flats for germination.

The seedlings were first genotyped using one SSR marker before transferring them into bigger pots to ensure they are hybrids. The plants were later divided into three clones for replications. The parents used in the crosses were also divided into 3 clones and included in the field experiment. The parents and progenies were transplanted at the Iron horse farm in Greene County, GA (33.73° N, –83.30° W) in April—May 2017. The experimental design in the field trial was a randomized complete block design with three replications. Plants were spaced apart by 3 feet (91.4 cm).

# b) Genotyping by sequencing

Leaf tissue samples were collected from all the progenies and kept in an ice box. The samples were dried in a freeze drier for at least two days. Genomic DNA was extracted using 2% hexadecyltrimethyl ammonium bromide (CTAB) method of Doyle and Doyle (53), then purified using RNase A (Thermo Scientific™, Thermo Fisher Scientific, Waltham, MA, USA, EN0531) and proteinase K (Invitrogen™, Thermo Fisher Scientific, Waltham, MA, USA, AM2548). Genomic DNA quality was confirmed by visualizing them in 1% agarose gel. Good quality genomic DNA gives one high molecular weight band with very minimal smears. The DNA was then cleaved using restriction enzymes per the GBS protocol developed by the Devos lab (pers. Communication) that was adapted from Poland et al. (26). In brief, genomic DNA was digested with two restriction enzymes, a “rare cutter” *PstI* (New England Biolabs®, Ipswich, MA, USA, R0140S) and a “common cutter” *MspI* (New England Biolabs®, Ipswich, MA, USA, R0106S). They were then ligated to their respective barcoded forward adapter containing *PstI* cut site and common reverse adapter containing *MspI* cut site. The DNA fragments were size selected for 300 bp and higher using modified magnetic beads (54) Sera-Mag SpeedBeads™ (Fisher Scientific™, Thermo Fisher Scientific, Waltham, MA, USA, 09–981–123) at 1X volume.

The size selected DNA were later PCR amplified for 16 cycles using Illumina forward and reverse primers; forward primer includes Illumina forward primer combined with the first few bases of the barcode adapter and reverse primer consists of Illumina reverse primer attached to the first few bases from the common adapters. DNA concentration for each sample was then measured using Qubit® 3 Fluorometer (Life Technologies, Thermo Fisher Scientific, Waltham, MA, USA) and pooled in equal amounts (about 10 ng/µl concentrations for every sample for a total of 150 samples for each pool). Pooled samples were sent to Georgia Genomic Facilities for fraction analysis, primer dimers filtration, and sequencing on Illumina NextSeq Paired-Ends for 30 cycles on four lanes of a flow cell.

# c) Reads filtering and variant calling

Sequencing reads were first checked for their quality using FastQC. The reads for each lane were de-multiplexed according to their unique barcode sequences using ‘process_radtags’ from Stacks (55) and merged as a single file for each genotype and type of read (read 1 and read 2). After that, reads were trimmed to remove enzyme cut sites and low quality bases where reads with phred score below 33 were discarded. Switchgrass reference genome version 4.1 from JGI Phytozome (56) was used to align the trimmed reads using bowtie2 version 2.2.9 (57) and reads were processed for GATK compatible format using samtools version 1.3.1 (58) and picard version 2.4.1 (59). Alignment of reads to a reference genome is theoretically the best practice for variant callings because the reference genome can separate a real variant and a paralog, thus reducing false SNP calls from paralogs (22). SNP calling were done on each genotype’s output file using “HaplotypeCaller” and pooled into one file using “GenotypeGVCFs” from GATK version 3.4.0 (60).

The final output file generated from “GenotypeGVCFs” from GATK 3.4.0 contains a very large number of variants and these must be filtered to remove the false positive variants that might be caused by sequencing errors. In addition, the filtered variants must meet these criteria to be accepted as true variants: 1) Quality score > 20, 2) Genotype quality score > 20, 3) Less than 20% genotype missing value, 4) Biallelics, 5) Minor allele frequency > 5%, and 6) Sequence depth ≥ 8. SNP with a MAF < 5% were removed because they cannot be distinguished from sequencing errors (22). The raw variant files were filtered using VCFtools (61). The filtered variants in VCF file was then converted to binary format using PLINK 1.07 (62), which assuming C is the minor allele recoded genotypes as follows: 0 (AA), 1 (AC), 2 (CC), and NA (00). The binary file was further filtered for variants that are present in both parents. Besides alleles filtering, progeny sequences were analyzed to ensure that they are true hybrids. For hybrid verification only homozygous markers were used, for example marker “0” in female parent and marker “2” in male parent. For hybrid progenies, they must contain heterozygous allele (marker “1”) at this locus since they are F1. A subset of markers, i.e. 4000 markers were

used to screen out non-hybrids and poorly sequenced progenies—true hybrids have at least 80% heterozygous alleles for these subsets of markers.

# d) Construction of linkage maps

Variants were pulled out for single dose alleles (SDA)—heterozygous in one parent and homozygous in the other parent. Chi-square goodness of fit (p-value > 0.05) was used to identify well-segregating alleles with 1:1 ratio. Only SDAs with < 10% missing data were retained in this category. The rest of the alleles with p-value > $1 \times 10^{-15}$ and < 20% missing data were considered distorted markers, following the classification used by Fiedler (40). Construction of linkage maps was carried out using JoinMap 5.0 with the following settings: 1) Outcrosser population type (CP), 2) Independence LOD $\geq 7$ for markers grouping, 3) Regression algorithm, 4) Kosambi function, 5) Recombination frequency < 0.40, and 6) Jump threshold of 5.0 for removal of loci.

The first step in linkage map construction was to exclude identical alleles, and then only the SDA alleles were used for initial marker grouping (28, 40). The next step was to add the ungrouped distorted alleles to their strongest cross-link (SCL) well-segregated loci using the LOD threshold value of 10.0. This step is important to maintain the integrity of linkage groups (LGs) by giving the strongest link LGs in the initial step and to maximize the number of linkage groups (40). For the parents used in the mapping population, the accurate number of LGs should be 18, representing the nine “a” and “b” subgenomes of switchgrass. Markers that were not ordered by JoinMap were taken out and designated as accessory markers (28, 40). These markers were forced to be included in the alternative map using JoinMap’s third round marker placement function and kept for our reference. LGs were labeled according to the pseudomolecule name in switchgrass reference genome 4.1 (K and N subgenomes) where each LG should have at least 85% alleles belonging to the same pseudomolecule. Finally, linkage maps were drawn using MapChart 2.30 (63).

Marker distribution was plotted along genetic distance to point the regions with high density markers as this might be an indication of a possible pericentromeric region. Marker frequency was shown for 1 cM genetic distance bin for every linkage group.

# e) Analysis of segregation distortion

Chi-square goodness-of-fit test for the expected 1:1 segregation ratio that resulted from the cross between heterozygous allele (H) to homozygous allele (A) was calculated for all markers in the map. A marker is significantly deviated from the expected segregation ratio at p-value < 0.05 and severe distortion is detected at p-value < 0.001. Percentage of distorted and well-segregated markers in each linkage group and number of heterozygous and homozygous genotypes for each marker were calculated and presented in graphs. Chi-square values for each marker across concatenated K and N subgenomes in genetic distance are presented in graph to illustrate region with severe segregation distortion. Region with at least three consecutive markers with p-value < 0.001 was highlighted as severely distorted (64).

# f) Estimation of genetic distance between parents

Euclidean genetic distance was calculated between the parents to estimate the degree of genetic dissimilarity in order to make inference on genetic divergence. SNP aggregation for all three parents were firstly conducted using “GenotypeGVCFs” from GATK 3.4.0 using switchgrass V4.1 reference genome for reads alignment. Raw variants were then filtered in similar steps described previously. Only the sites having genotype calls for all three parents were used for the calculation of pairwise distance. This is to avoid calculation bias due to differential amounts of missing data between pairs in comparison. Similarity between two parents was first calculated using the formula: [Due to technical limitations, this equation is only available as a download in the supplemental files section]. Euclidean genetic distance between two parents was then calculated using the formula (65): [Due to technical limitations, this equation is only available as a download in the supplemental files section.]

# g) Markers collinearity with switchgrass reference genome V5.1

Variant calling was initially done using switchgrass reference genome V4.1 for reads alignment. To see the collinearity between the genetic map distances (cM) with the physical map distances (Mb), genetic distance versus physical map distance was plotted for each chromosome. At the time when the analysis was finalized, a draft of V5.1 was just released and we conducted BLASTn for every SNP and Indel sequence in V4.1 to V5.1 using E-value < $1 \times 10^{-5}$ as a matching hit threshold. The sequences used as queries for BLASTn include 50 bp upstream and downstream of the variant physical position. The best sequence alignment in V5.1 corresponded to the expected chromosome with E-value < $1 \times 10^{-30}$ was selected for the new physical position. In the case where no alignment match was found to have an E-value < $1 \times 10^{-30}$, the next best alignment with E-value < $1 \times 10^{-5}$ was used.

## Abbreviations

GBS: Genotyping by sequencing; SNP: Single nucleotide polymorphism; SDA: Single dose alleles; LG: linkage group; cM: Centimorgan; Mb: Megabase pair; MYBP: Million years before present; RE: Restriction enzyme; MAF: Minor allele frequency; MPH: Mid-parent heterosis; SI: Self-incompatibility

## Declarations

## Ethics approval and consent to participate

Not applicable

## Consent for publication

Not applicable

## Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Competing interests

The authors declare that they have no competing interests.

## Funding

Funding was provided by Malaysian Rubber Board and The Center for Bioenergy Innovation, a U.S. Department of Energy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

## Authors' contributions

RR conducted all experiments and wrote the manuscript, KD supervised the GBS and linkage mapping works, AM supervised the research, edited and revised the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

We thank Jonathan Markham for greenhouse and field experiment, April Legg for lab and chemical supplies, and Peng Qi for GBS assistance.

## References

1. Casler M. Switchgrass Breeding, Genetics, and Genomics. In: Monti A, editor. Switchgrass. Green Energy and Technology: Springer London; 2012. p. 29–53.

2. Kszos LA, Downing ME, Wright LL, Cushman JH, McLaughlin SB, Tolbert VR, et al. Bioenergy Feedstock Development Program Status Report. Technical Report. 2000.

3. Parrish DJ, Fike JH. The biology and agronomy of switchgrass for biofuels. Critical Reviews in Plant Sciences. 2005;24(5/6):423–59.

4. Sanderson MA, Reed RL, McLaughlin SB, Wullschleger SD, Conger BV, Parrish DJ, et al. Switchgrass as a sustainable bioenergy crop. 1996:83.

5. Lowry DB, Behrman KD, Grabowski P, Morris GP, Kiniry JR, Juenger TE. Adaptations between ecotypes and along environmental gradients in Panicum virgatum. American Naturalist. 2014;183(5):682–92.

6. Barnett FL, Carver RF. Meiosis and Pollen Stainability in Switchgrass, Panicum virgatum L1. Crop Science. 1967;7(4):301–4.

7. Nielsen E. Analysis of variation in *Panicum virgatum.* . J Agric Res. 1944;69:327–53.

8. Zhang Y, Zalapa J, Jakubowski AR, Price DL, Acharya A, Wei Y, et al. Natural Hybrids and Gene Flow between Upland and Lowland Switchgrass. Crop Science. 2011(6).

9. Zhang Y, Zalapa JE, Jakubowski AR, Price DL, Acharya A, Wei Y, et al. Post-glacial evolution of Panicum virgatum: centers of diversity and gene pools revealed by SSR markers and cpDNA sequences. Genetica. 2011;139:933–48.

10. Martinez-Reyna JM, Vogel KP. Incompatibility systems in switchgrass2002.

11. Adhikari L, Anderson MP, Klatt A, Wu Y. Testing the Efficacy of a Polyester Bagging Method for Selfing Switchgrass. Bioenerg Res. 2015;8(1):380–7.

12. Dong H, Thames S, Liu L, Smith M, Yan L, Wu Y. QTL Mapping for Reproductive Maturity in Lowland Switchgrass Populations. Bioenerg Res. 2015;8(4):1925–37.

13. Liu L, Wu Y. Identification of a Selfing Compatible Genotype and Mode of Inheritance in Switchgrass. Bioenerg Res. 2012;5(3):662–8.

14. Li G, Serba DD, Saha MC, Bouton JH, Lanzatella CL, Tobias CM. Genetic Linkage Mapping and Transmission Ratio Distortion in a Three-Generation Four-Founder Population of Panicum virgatum (L. ). 2014. p. 913–23.

15. Missaoui AM, Paterson AH, Bouton JH. Investigation of genomic organization in switchgrass ( Panicum virgatum L. ) using DNA markers. Theoretical & Applied Genetics. 2005;110(8):1372–83.

16. Okada M, Lanzatella C, Saha MC, Bouton J, Wu RL, Tobias CM. Complete Switchgrass Genetic Maps Reveal Subgenome Collinearity, Preferential Pairing and Multilocus Interactions. Genetics 2010;185:745–60.

17. Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, et al. Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. PLoS Genet. 2013;9(1):e1003215.

18. Clyde L. Porter J. An Analysis of Variation Between Upland and Lowland Switchgrass, Panicum Virgatum L. , in Central Oklahoma. Ecology. 1966(6):980.

19. Missaoui A, Paterson A, Bouton J. Molecular markers for the classification of switchgrass (Panicum virgatum L. ) germplasm and to assess genetic diversity in three synthetic switchgrass populations. Genetic Resources and Crop Evolution. 2006;53:1291–302.

20. Hultquist SJ, Vogel KP, Lee DJ, Arumuganathan K, Kaeppler S. Chloroplast DNA and Nuclear DNA Content Variations among Cultivars of Switchgrass, Panicum virgatum L. 1996:1049.

21. Zalapa JE, Price DL, Kaeppler SM, Tobias CM, Okada M, Casler MD. Hierarchical classification of switchgrass genotypes using SSR and chloroplast sequences: ecotypes, ploidies, gene pools, and cultivars. Theoretical and Applied Genetics. 2011(4):805.

22. Fei L, Alexander EL, Jeff G, Rob E, Jerome HC, Michael DC, et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genetics, Vol 9, Iss 1, p e1003215 (2013). (1):e1003215.

23. Acharya AR. Genetic diversity, population structure and association mapping of biofuel traits in southern switchgrass germplasm. [electronic resource]: 2014. ; 2014.

24. Bahri BA, Daverdin G, Xu X, Cheng J-F, Barry KW, Brummer EC, et al. Natural variation in genes potentially involved in plant architecture and adaptation in switchgrass (Panicum virgatum L. ). BMC Evolutionary Biology. 2018;18(1):91-.

25. Elshire RJ, Glaubitz JC, Qi S, Poland JA, Kawamoto K, Buckler ES, et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE. 2011;6(5):1–10.

26. Poland JA, Brown PJ, Sorrells ME, Jannink J-L. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLOS ONE. 2012;7(2):e32253.

27. Russell J, Hackett C, Hedley P, Liu H, Milne L, Bayer M, et al. The use of genotyping by sequencing in blackcurrant ( Ribes nigrum): developing high-resolution linkage maps in species without reference genome sequences. Molecular Breeding. 2014;33(4):835–49.

28. Grattapaglia D, Sederoff R. Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. Genetics. 1994;137(4):1121–37.

29. Daverdin G, Bahri BA, Wu X, Serba DD, Tobias C, Saha MC, et al. Comparative Relationships and Chromosome Evolution in Switchgrass (Panicum virgatum) and Its Genomic Model, Foxtail Millet (Setaria italica). BioEnergy Research. 2015(1):137.

30. Berthouly-Salazar C, Mariac C, Couderc M, Pouzadoux J, Floc'h J-B, Vigouroux Y. Genotyping-by-Sequencing SNP Identification for Crops without a Reference Genome: Using Transcriptome Based Mapping as an Alternative Strategy.

31. Tobias CM, Sarath G, Twigg P, Lindquist E, Pangilinan J, Penning BW, et al. Comparative Genomics in Switchgrass Using 61,585 High-Quality Expressed Sequence Tags. The Plant Genome Journal. 2008;1:111–24.

32. Wang Y, Zeng X, Iyer NJ, Bryant DW, Mockler TC, Mahalingam R. Exploring the switchgrass transcriptome using second-generation sequencing technology. PLoS ONE. 2012;7(3):e34225-e.

33. Palmer NA, Donze-Reiner T, Horvath D, Heng-Moss T, Waters B, Tobias C, et al. Switchgrass (Panicum virgatum L) flag leaf transcriptomes reveal molecular signatures of leaf development, senescence, and mineral dynamics. Functional & Integrative Genomics. 2015;15(1):1–16.

34. Palmer NA, Saathoff AJ, Scully ED, Tobias CM, Twigg P, Madhavan S, et al. Seasonal below-ground metabolism in switchgrass. Plant Journal. 2017(6):1059.

35. Tornqvist C-E, Vaillancourt B, Kim J, Buell CR, Kaeppler SM, Casler MD. Transcriptional Analysis of Flowering Time in Switchgrass. BioEnergy Research. 2017(3):700.

36. Sharma MK, Sharma R, Peijian C, Jenkins J, Bartley LE, Qualls M, et al. A Genome-Wide Survey of Switchgrass Genome Structure and Organization. PLoS ONE. 2012;7(4):1–13.

37. Serba D, Wu L, Daverdin G, Bahri BA, Wang X, Kilian A, et al. Linkage Maps of Lowland and Upland Tetraploid Switchgrass Ecotypes. Bioenerg Res. 2013;6(3):953–65.

38. Liu L, Huang Y, Punnuri S, Samuels T, Wu Y, Mahalingam R. Development and integration of EST–SSR markers into an established linkage map in switchgrass. Molecular Breeding. 2013;32(4):923–31.

39. Liu L, Wu Y, Wang Y, Samuels T. A High-Density Simple Sequence Repeat-Based Genetic Linkage Map of Switchgrass. 2012. p. 357–70.

40. Fiedler JD, Lanzatella C, Okada M, Jenkins J, Schmutz J, Tobias CM. High-Density Single Nucleotide Polymorphism Linkage Maps of Lowland Switchgrass using Genotyping-by-Sequencing. The Plant Genome. 2015;8(2):1–14.

41. Gardner KM, Brown P, Cooke TF, Cann S, Costa F, Bustamante C, et al. Fast and Cost-Effective Genetic Mapping in Apple Using Next-Generation Sequencing. 2014. p. 1681–7.

42. Adhikari L, Lindstrom OM, Markham J, Missaoui AM. Dissecting Key Adaptation Traits in the Polyploid Perennial Medicago sativa Using GBS-SNP Mapping. Frontiers in Plant Science. 2018;9(934).

43. Faris JD, Laddomada B, Gill BS. Molecular Mapping of Segregation Distortion Loci in Aegilops tauschii. 1998:319.

44. Virk PS, Ford-Lloyd BV, Newbury HJ. Mapping AFLP markers associated with subspecific differentiation of Oryza sativa (rice) and an investigation of segregation distortion. Heredity. 1998;81(6):613–20.

45. Mano Y, Muraki M, Fujimori M, Takamizo T, Kindiger B. AFLP–SSR maps of maize × teosinte and maize × maize: comparison of map length and segregation distortion. Plant Breeding. 2005;124(5):432–9.

46. Törjék O, Witucka-Wall H, Meyer RC, Korff Mv, Kusterer B, Rautengarten C, et al. Segregation distortion in Arabidopsis C24/Col–0 and Col–0/C24 recombinant inbred line populations is due to reduced fertility caused by epistatic interaction of two loci. Theoretical & Applied Genetics. 2006;113(8):1551–61.

47. Daniel Z, Yaakov T. Unequal Segregation of Nuclear Genes in Plants. Botanical Gazette. 1986(3):355.

48. Lyttle TW. Segregation Distorters. Annual Review of Genetics. 1991;25(1):511–81.

49. Mascarenhas JP. Pollen gene expression: molecular evidence. International Review Of Cytology. 1992;140:3–18.

50. Soltis PS, Soltis DE. The role of genetic and genomic attributes in the success of polyploids. Proc Natl Acad Sci U S A. 2000;97(13):7051–7.

51. te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, Kubesová M, et al. The more the better? The role of polyploidy in facilitating plant invasions. Annals of botany. 2012;109(1):19–45.

52. Razar RM, Missaoui A. Phenotyping Winter Dormancy in Switchgrass to Extend the Growing Season and Improve Biomass Yield. Journal of Sustainable Bioenergy Systems. 2018;Vol. 08No. 01:22.

53. Doyle J, Doyle J. Isolation of plant DNA from fresh tissue. Focus. 1990;12:13–5.

54. Rohland N, Reich D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome research. 2012;22(5):939–46.

55. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences.

56. JGI Phytozome 12. 2017. https://phytozome. jgi. doe. gov/pz/portal. html. Accessed 1 July 2017.

57. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;10(3):R25.

58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078–9.

59. Picard. 2017. http://broadinstitute. github. io/picard/. Accessed 1 August 2017.

60. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. 2010:1297.

61. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.

62. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81(3):559–75.

63. Voorrips RE. MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. Journal of Heredity. 2002;93(1):77–8.

64. Li X, Wei Y, Acharya A, Jiang Q, Kang J, Brummer EC. A Saturated Genetic Linkage Map of Autotetraploid Alfalfa (*Medicago sativa* L. ) Developed Using Genotyping-by-Sequencing Is Highly Syntenous with the *Medicago truncatula* Genome. G3: Genes|Genomes|Genetics. 2014;4(10):1971–9.

65. Gower J, Legendre P. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification. 1986;3(1):5.

## Tables

**Table 1 Comparison of marker numbers, map length (cM), and average inter-marker distance for each parental map in AP13 x B6 and B6 x VS16 populations**

| Linkage groups | AP13 x B6 | | | | | | Linkage groups | B6 x VS16 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AP13 | | | B6 | | | | B6 | | | VS16 | | |
| | Total no. of markers | Total map length (cM) | Average inter-marker distance (cM) | Total no. of markers | Total map length (cM) | Average inter-marker distance (cM) | | Total no. of markers | Total map length (cM) | Average inter-marker distance (cM) | Total no. of markers | Total map length (cM) | Average inter-marker distance (cM) |
| 1K | 107 | 144.76 | 1.37 | 73 | 98.85 | 1.37 | 1K | 60 | 70.82 | 1.2 | 162 | 63.9 | 0.4 |
| 1N | 94 | 152.07 | 1.64 | 122 | 96.81 | 0.8 | 1N | 86 | 110.66 | 1.3 | 106 | 97.07 | 0.92 |
| 2K | 109 | 176.64 | 1.64 | 76 | 92.49 | 1.23 | 2K | 221 | 72.4 | 0.33 | 151 | 84.76 | 0.57 |
| 2N | 87 | 160.88 | 1.87 | 77 | 107.07 | 1.41 | 2N | 226 | 83.38 | 0.37 | 98 | 102.65 | 1.06 |
| 3K | 76 | 167.61 | 2.23 | 108 | 112.89 | 1.06 | 3K | 155 | 135.8 | 0.88 | 183 | 140.07 | 0.77 |
| 3N | 52 | 156.24 | 3.06 | 123 | 99.95 | 0.82 | 3N | 138 | 104.1 | 0.76 | 76 | 140.73 | 1.88 |
| 4K | 72 | 101.83 | 1.43 | 27 | 85.39 | 3.28 | 4K | 47 | 64.74 | 1.41 | 120 | 79.49 | 0.67 |
| 4N | 63 | 120.09 | 1.94 | 16 | 80.06 | 5.34 | 4N | 37 | 54.19 | 1.51 | 68 | 81.3 | 1.21 |
| 5K | 104 | 177.49 | 1.72 | 77 | 119.45 | 1.57 | 5K | 147 | 85.83 | 0.59 | 75 | 45.12 | 0.61 |
| 5N | 125 | 151.04 | 1.22 | 93 | 116.47 | 1.27 | 5N | 69 | 63.41 | 0.93 | 180 | 105.78 | 0.59 |
| 6K | 71 | 98.93 | 1.41 | 47 | 79.08 | 1.72 | 6K | 65 | 120.73 | 1.89 | 88 | 100.19 | 1.15 |
| 6N | 58 | 96.74 | 1.7 | 63 | 91.2 | 1.47 | 6N | 50 | 77.15 | 1.57 | 47 | 73.48 | 1.6 |
| 7K | 82 | 107.21 | 1.32 | 44 | 99.69 | 2.32 | 7K | 36 | 45.43 | 1.3 | 103 | 109.53 | 1.07 |
| 7N | 97 | 121.19 | 1.26 | 16 | 32.04 | 2.14 | 7N | 100 | 77.80 | 0.79 | 67 | 53.61 | 0.81 |
| 8K | 37 | 93.21 | 2.59 | 59 | 92.2 | 1.59 | 8K | 67 | 76.93 | 1.17 | 60 | 52.4 | 0.89 |
| 8N | 42 | 110.2 | 2.69 | 51 | 86.49 | 1.73 | 8N | 72 | 74.26 | 1.05 | 89 | 63.43 | 0.72 |
| 9K | 133 | 162.47 | 1.23 | 108 | 127.36 | 1.19 | 9K | 200 | 77.2 | 0.39 | 147 | 87.59 | 0.6 |
| 9N | 131 | 177 | 1.36 | 52 | 86.59 | 1.7 | 9N | 48 | 88.03 | 1.87 | 122 | 125.67 | 1.04 |
| | **1540** | **2475.61** | **1.76** | **1232** | **1704.06** | **1.78** | | **1824** | **1482.90** | **1.07** | **1942** | **1606.79** | **0.92** |

**Table 2 Genetic similarity and Euclidean distances calculated between parents**

| | AP13 and B6 | B6 and VS16 | AP13 and VS16 |
|---|---|---|---|
| Genetic similarity | 0.225 | 0.159 | 0.545 |
| Euclidean distance | 0.880 | 0.917 | 0.674 |

## Figures

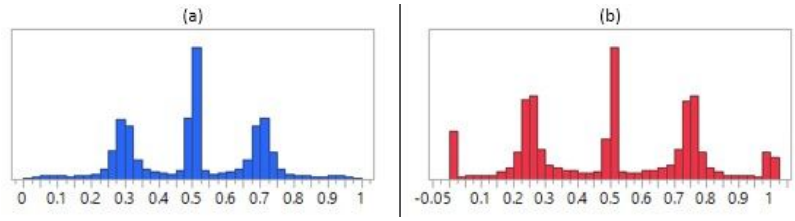


**Figure 1**

Histogram of allele frequency for (a) AB population and (b) BV population

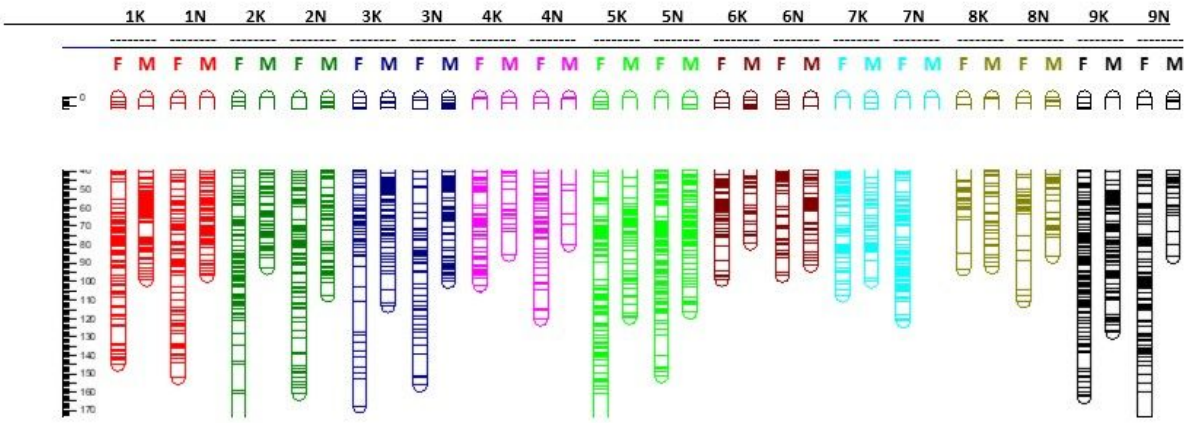

**Figure 2**

Linkage and homology groups of maternal (F) and paternal (M) maps for AB population. Individual bars represent loci positions with thicker bars indicating two or more loci are nearby.

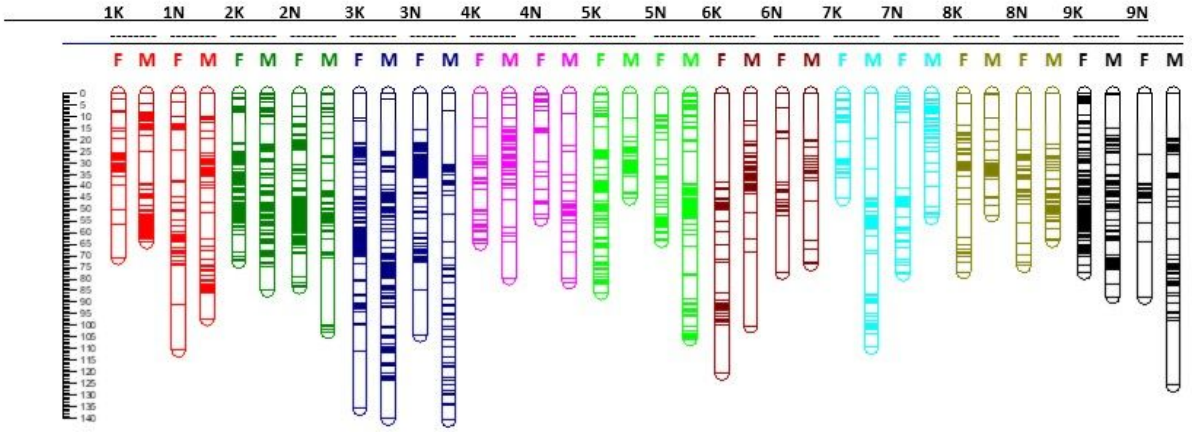


**Figure 3**

Linkage and homology groups of maternal (F) and paternal (M) map for BV population. Each horizontal bar represents a locus position, with thicker bars indicating two or more loci are nearby.

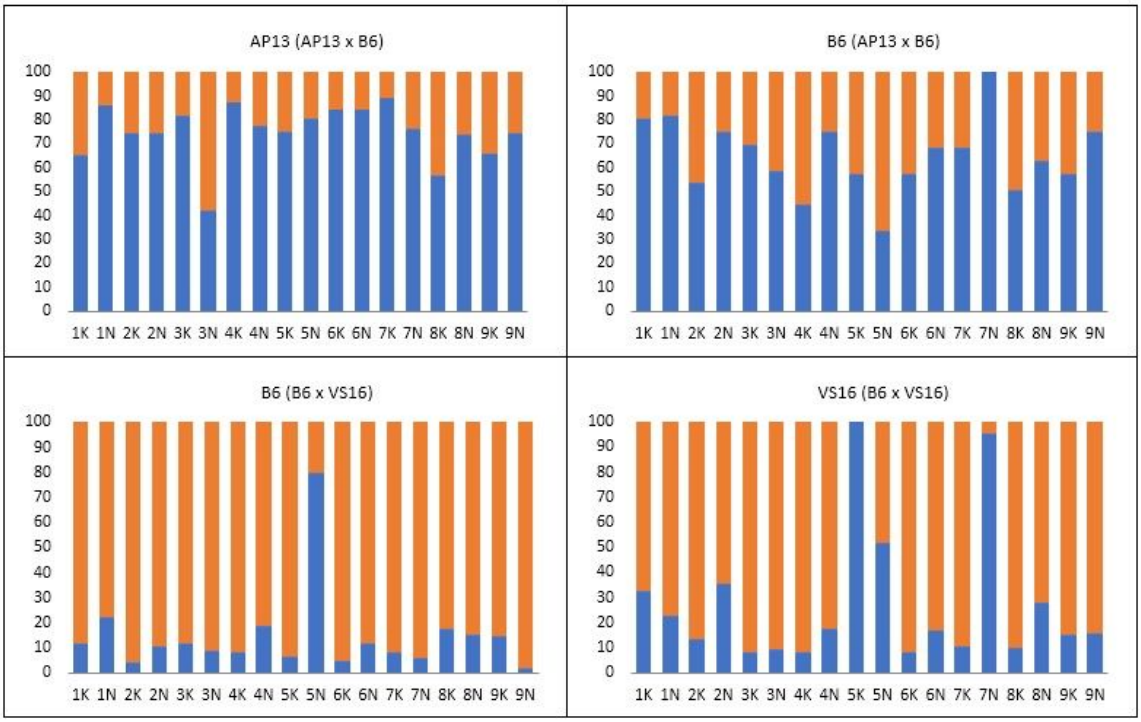


**Figure 4**

Proportion of distorted and well segregated markers across linkage groups. Blue bars = Distorted loci; Orange bars = Well segregated loci; y-axis = Percentage; x-axis = Linkage groups

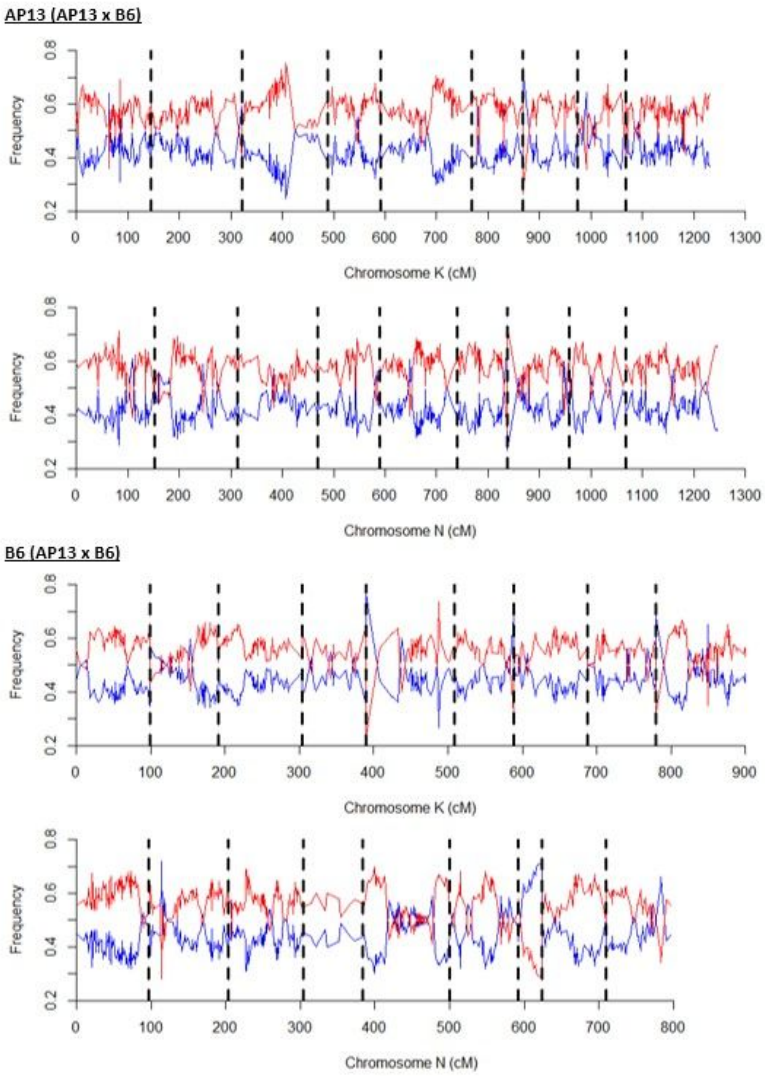


**Figure 5**

Frequency of homozygous and heterozygous genotypes across K and N subgenomes in AP13 (top) and B6 (bottom) maps. Disconnected vertical lines demarcate each linkage group in the concatenated chromosomes. Red line = Heterozygous; Blue line = Homozygous.

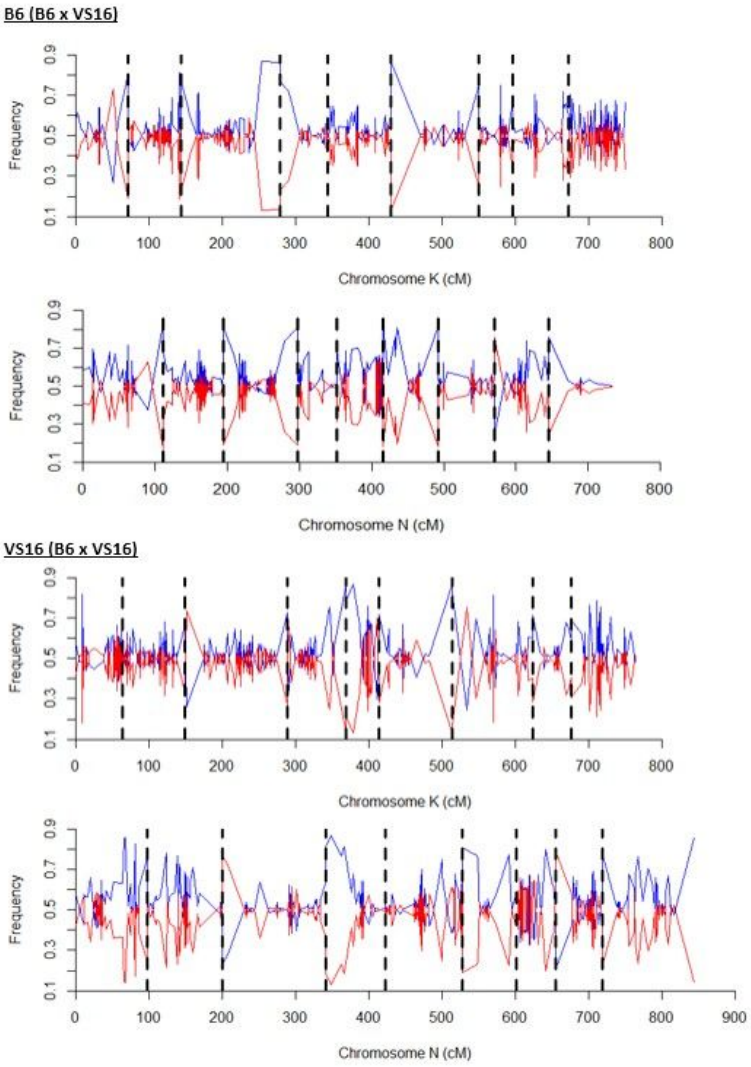


**Figure 6**

Frequency of homozygous and heterozygous genotypes across K and N subgenomes in B6 (top) and VS16 (bottom) maps. Disconnected vertical lines demarcate each linkage group in the concatenated chromosomes. Blue line = Homozygous; Red line = Heterozygous

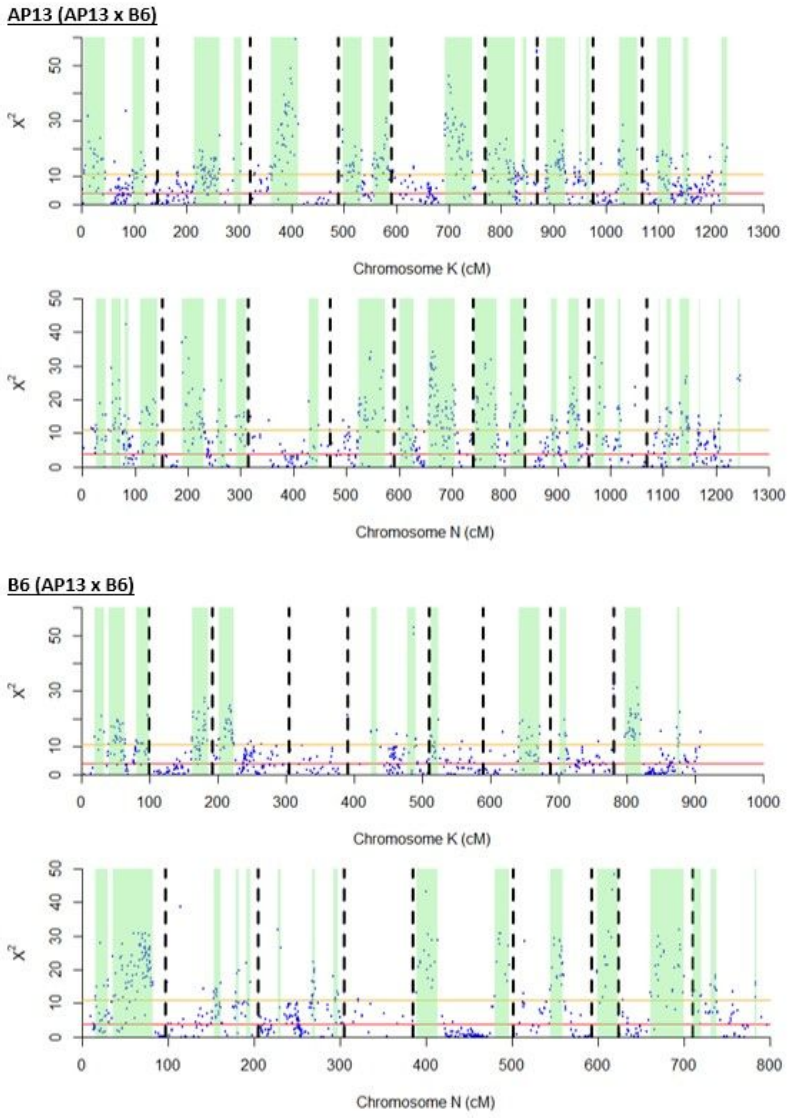


**Figure 7**

Chi square (X2) values for goodness-of-fit test with 1:1 expected segregation ratios across K and N subgenomes in AP13 (top) and B6 (bottom) maps. Red and orange lines mark the critical value of $p < 0.05$ ($X2 < 3.841$) and $p < 0.001$ ($X2 < 10.8276$), respectively. Chromosome regions highlighted in green indicate areas of contiguous severe segregation distortion ($p < 0.001$)

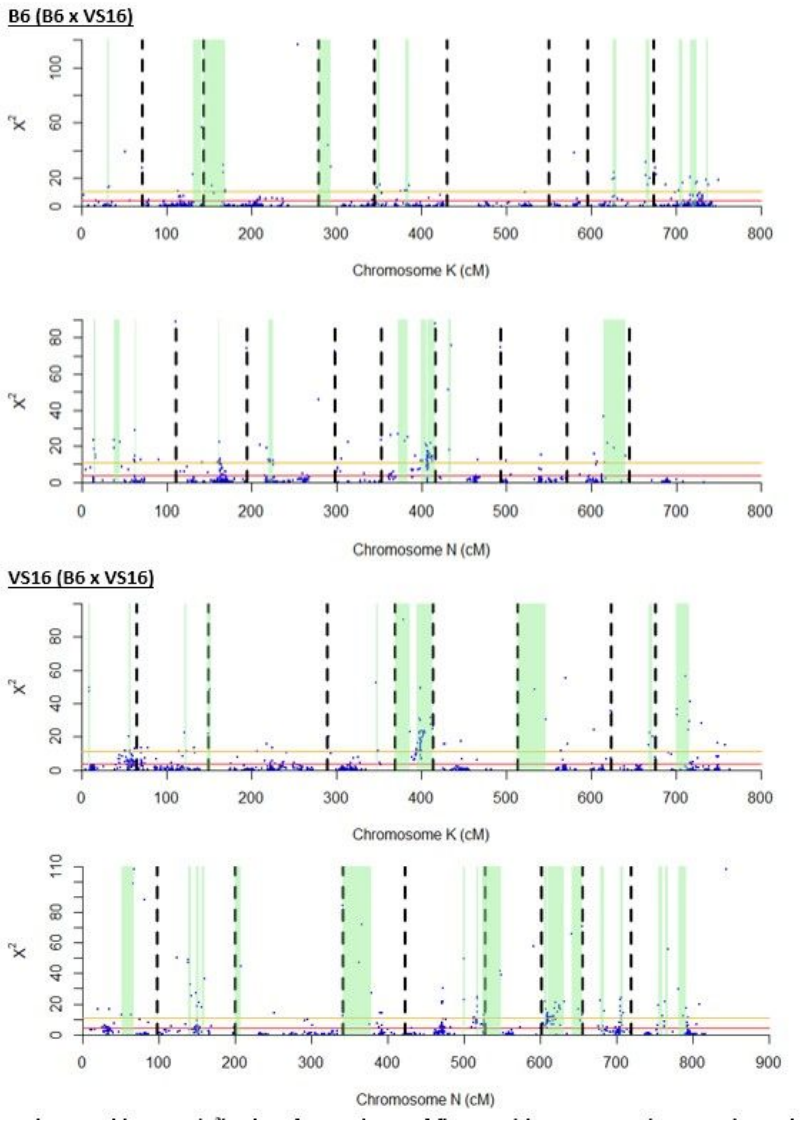


**Figure 8**

Chi square (X2) values for goodness-of-fit test with 1:1 expected segregation ratio across K and N subgenomes in B6 (top) and VS16 (bottom) maps. Red and orange lines mark the critical values of $p < 0.05$ ($X2 < 3.841$) and $p < 0.001$ ($X2 < 10.8276$), respectively. Chromosome regions highlighted in green indicate areas of contiguous severe segregation distortion ($p < 0.001$)

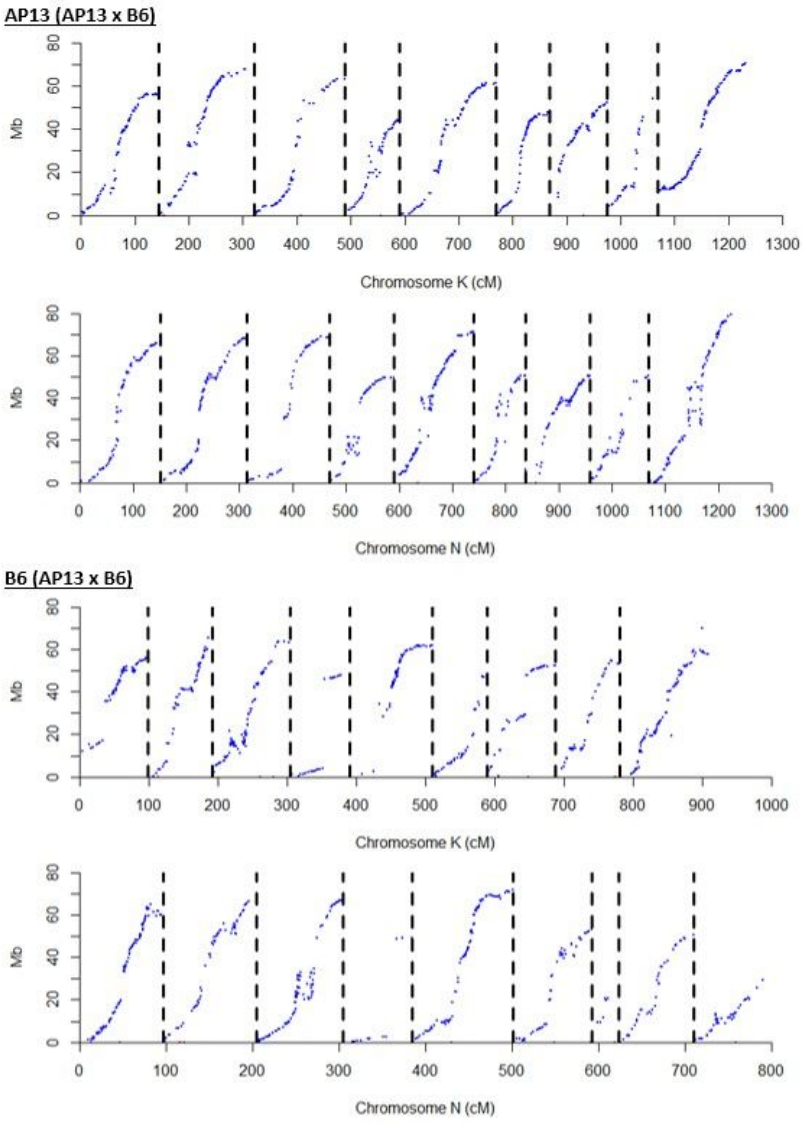


**Figure 9**

Collinearity of marker positions in genetic maps (x-axis) and switchgrass reference genome V5.1 (y-axis) across K and N subgenomes in AP13 (top) and B6 (bottom) maps for AP13 x B6 population. Red dots (y = 0) represent markers that are not from the same chromosome.

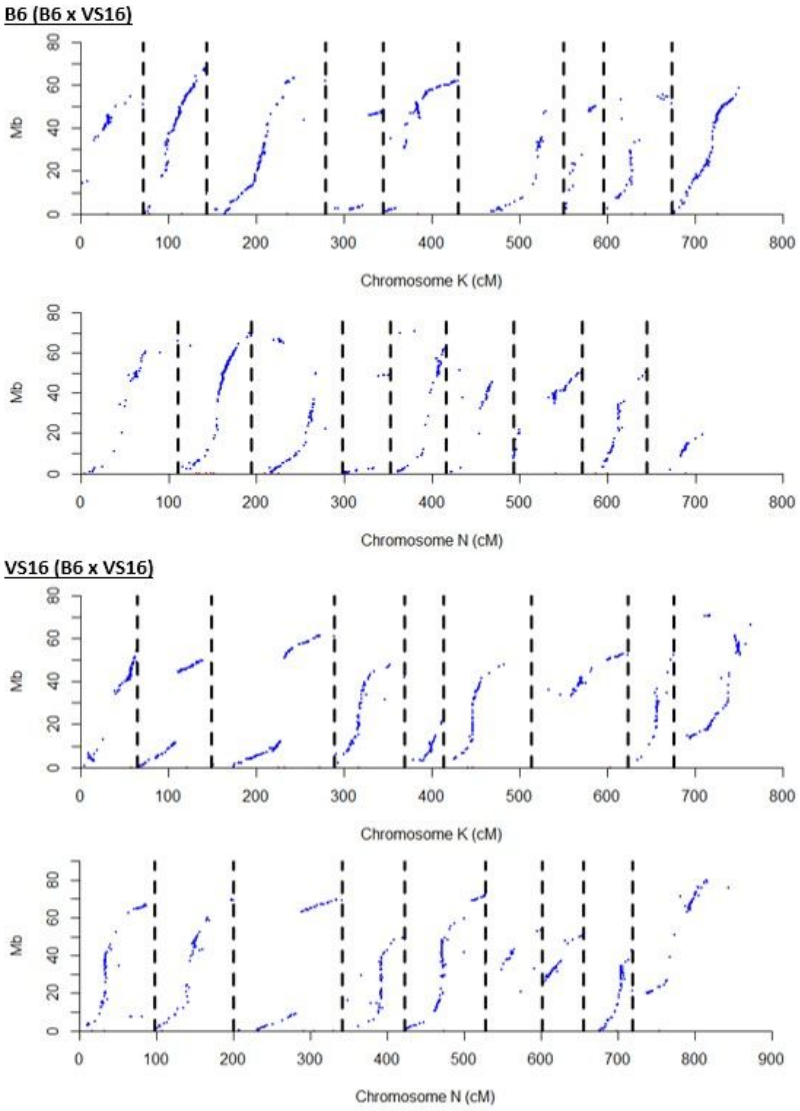


**Figure 10**

Collinearity of marker positions in the genetic maps (x-axis) and switchgrass reference genome V5.1 (y-axis) across K and N subgenomes in B6 (top) and VS16 (bottom) maps for B6 x VS16 population. Red dots (y = 0) represent markers that are not from the same chromosome.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- eq2.jpg
- eq1.jpg
- FigureS1TableS1.xlsx
- FigureS2.docx
- FigureS3.docx