Genomic Variation Analysis of Switchgrass (*Panicum virgatum* L.) NAM (Nested Association Mapping) Parents

Shahjahan Ali[1], Junil Chang[2], Desalegn D. Serba[3], Hem S. Bhandari[4], Laura Bartley[5] and Malay C. Saha[1]

[1]The Samuel Roberts Noble Foundation, Plant Biology Division, 2510 Sam Noble Parkway, Ardmore, Oklahoma 73401 USA
[2]The Samuel Roberts Noble Foundation, Computing Services, 2510 Sam Noble Parkway, Ardmore, Oklahoma 73401 USA
[3]Kansas State University, AG Research Centers-Hays, Kansas 67601 USA
[4]Department of Plant Sciences, University of Tennessee, 2431 Joe Johnson Drive, Knoxville, Tennessee 37996 USA
[5]The University of Oklahoma, 660 Parrington Oval, Norman, Oklahoma 73019 USA

Abstract

NAM (Nested Association Mapping) has been established as an efficient and powerful method for association mapping and QTL analysis. It offers benefits of both bi-parental and association mapping to dissect complex traits. We have generated 2,000 pseudo F_2 NAM population by crossing 15 switchgrass low land ecotypes with a common parent, AP13. This population has been evaluated in Ardmore, OK and Knoxville, TN locations along with pseudo F_1 and parents. We used whole genome sequencing approach to delineate allelic variations within NAM parental genomes. Genomic shotgun sequencing of NAM parental genomes produced 28-66 Gb high-quality sequence data. Alignment of these sequences with the reference genome, AP13 (v3.1), revealed that up to 99.00% of the genomic sequences mapped to the AP13 genome. The parent, NFGA16_05, produced the highest number (9.94 million) of polymorphic loci whereas, the least polymorphic loci (6.43 million) were observed in NFGA09_02. We cataloged 27.78 million bi-allelic SNPs in the 18 chromosomes of a tetraploid switchgrass genome. On an average one SNP was identified in every 48 to 64 bp of chromosome sequence of the NAM parental genomes. The ration of intronic to exonic SNPs was 1.72. We have identified 1.09 million nonsynonymous SNPs in the exonic regions of NAM parental genomes with 7,128 SNP dense genes.

Introduction

Switchgrass (*Panicum virgatum* L.), a native North American C4 perennial grass species has been identified as a model species for bioenergy feedstock development for cellulosic ethanol production by US Department of Energy (US-DOE). Application of molecular breeding techniques can improve biomass production of switchgrass. The genome of lowland switchgrass cultivars grown in the Southern Plains is allotetraploid (2n = 4x = 36) with 18 linkage groups distributed into two highly homologous subgenomes and have the haploid genome of ~1.5Gb. NAM offers benefits of both biparental and association mapping to dissect complex traits. The objectives of this project are to develop a NAM population and construct a genetic map for this population, identify QTLs and molecular markers associated with biomass yield, feedstock quality and other agronomical important traits, and validate marker-QTL associations in breeding populations.

Materials and Methods

Fifteen diverse lowland switchgrass genotypes were crossed to a common parental genotype, AP13 and generated 15 pseudo F_1 families (Figure 1). Ten selected pseudo F_1 plants from each family were chain crossed. A total of 75-200 pseudo F_2 progenies from each family were randomly selected, which constituted the final NAM population of 2,000 progenies. They were evaluated along with parents and F_1s in the fields of Knoxville, TN and Ardmore, OK.

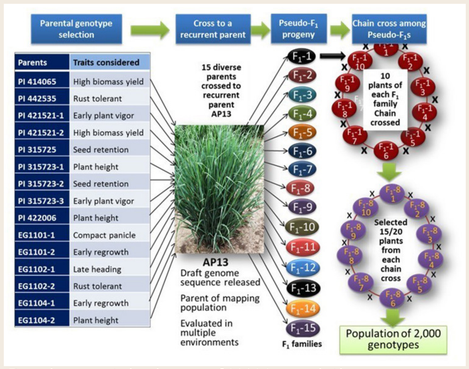

Figure 1. Developmental scheme of NAM population

Genomic shotgun sequencing of NAM parents were conducted at the USDOE Joint Genome Institute, Walnut Creek, CA as a part of the Community Sequencing Project. Parental sequences were mapped to switchgrass reference sequence of AP13 (*Panicum virgatum* v3.0: http://portal.nersc.gov/dna/plant/annotation/Pvirgatum/Pvirgatum_383_v3.1/) in Bowtie (v2.2.5). Variant calling was done using Samtools (v1.2) mpileup program. Variant filtering, file manipulations and color coded SNP frequency graphs in every 50 Kb interval were generated using custom Perl scripts.

Annotation of SNPs/InDels and assignment of variants in chromosomal locations were performed in Annovar software (http://www.openbioinformatics.org/annovar/; Wang et al. 2010) using Pvirgatum_383_v3.1.gene_exon.gff3 file (http://portal.nersc.gov/dna/plant/annotation/Pvirgatum/Pvirgatum_383_v3.1/).

Phylogenic trees created from the sequences of 2.01 million exonic SNP positions using K-Mers with Neighbor Joining (Bootstrapping) and Mahalanobis distance imputation procedure.

Results and Discussion

We sequenced 15 NAM parental genomic shotgun libraries that produced 37.99 to 60.07 Gb (258.97-400.45 Million Reads) of good quality sequence data with genome coverage of 25.33-40.05X (Table 1). Alignment of these sequences with reference genome sequence, AP13 (Panicum virgatum Version 3.0) showed that 97.28 to 99.00% of the sequence data were mapped with reference genome sequence.

Table 1: NAM parental genome sequence statistics

Genotype	Sequence Output				Mapped Sequence		
	Reads (Million)	Bases (Gb)	Coverage (X)	GC %	Reads (Million)	Bases (Gb)	% Bases mapped
NFGA02_06	258.97	38.85	25.90	44	254.63	38.20	98.32
NFGA04_01	328.73	49.31	32.87	49	323.36	48.50	98.37
NFGA09_02	360.13	54.02	36.01	44	356.29	53.44	98.94
NFGA09_05	318.51	47.78	31.85	43	315.33	47.30	99.00
NFGA15_11	400.45	60.07	40.05	46	394.73	59.21	98.57
NFGA16_02	278.09	41.71	27.81	43	272.87	40.93	98.12
NFGA16_05	321.21	48.20	32.13	44	312.95	46.94	97.40
NFGA16_12	289.09	43.36	28.91	45	285.18	42.78	98.65
NFGA32_06	333.97	50.10	33.40	45	324.88	48.73	97.28
NFGA34_08	329.25	49.39	33.93	47	322.58	48.39	97.98
NFGA34_10	335.17	50.28	33.52	44	330.02	49.50	98.46
NFGA36_05	376.11	37.99	25.33	45	371.18	37.48	98.67
NFGA36_09	373.18	55.98	37.32	44	367.60	55.14	98.50
NFGA37_03	326.17	48.93	32.62	44	319.50	47.93	97.96
NFGA37_05	327.60	49.13	32.73	43	324.14	48.62	98.96

We identified 28.20 million SNPs within NAM parental chromosomes. The polymorphic SNPs ranged from 6.43 million for parent, NFGA09_02 to 9.94 million for NFGA16_05 (Figure 2). A significant number of parental SNPs were monomorphic- 15.99 to 21.37 million. The ration of polymorphic to monomorphic SNP loci varies from 0.31 to 0.59.

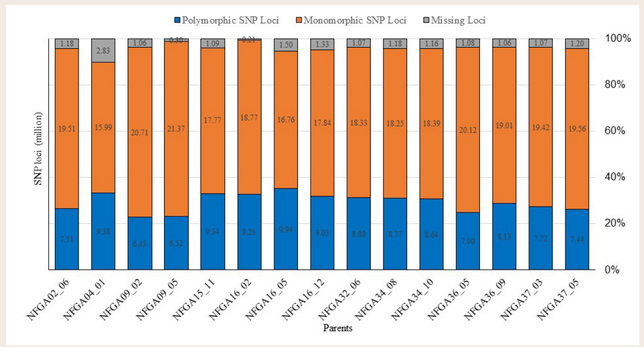

Figure 2. Polymorphic, monomorphic and missing SNP loci in parental genomes of the NAM parents.

We recorded 27.78 million bi-allelic SNP loci in the chromosomes of NAM parents with the criteria of presence one polymorphic SNP loci in at least one of the 15 parents. The number of SNP loci reduced to almost half when we applied filtering criteria of SNPs present in at least 3 parents (Figure 3). Only 0.69 million SNPs possessed by all of the 15 parents.

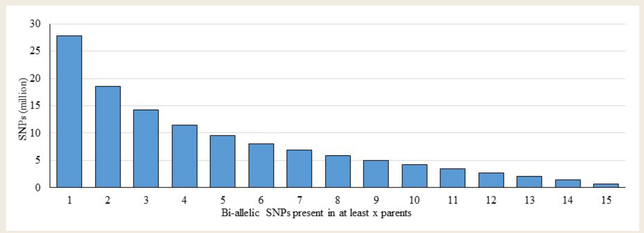

Figure 3. Bi-allelic SNP filtering

Each of the 18 switchgrass chromosomes produced 1.13 to 2.20 million SNPs (Figure 4a). Normalization of SNP numbers with chromosome size revealed that presence of 1 SNP every 48 and 64 bp genome sequences for chromosome III-N and V-K, respectively (Figure 4b).

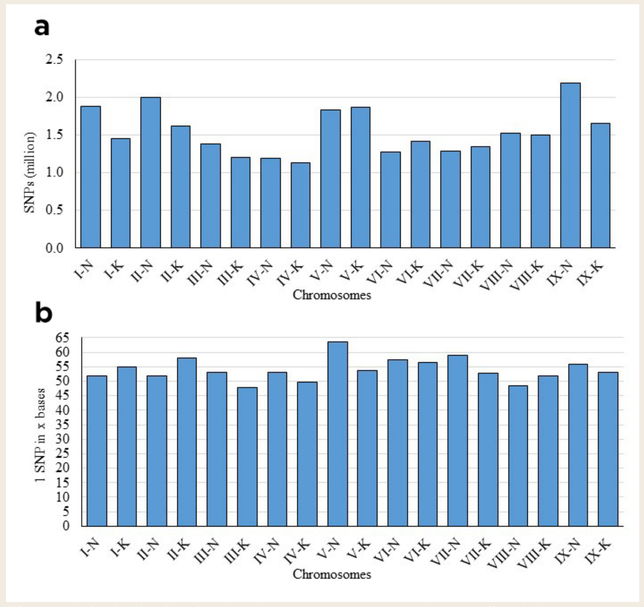

Figure 4. SNP distribution in chromosomes (a) and normalized (by chromosome size) SNPs in chromosomes (b).

Plotting of SNP frequencies in every 50 Kb interval across each of the chromosomes identified as many as 3,764 SNPs in 50 Kb chromosomal regions which indicated the presence of 1 SNP in every 13 bp in some of the chromosomal regions (Figure 5). We identified 2.01, 3.47 and 19.86 million variants in exonic, intronic and intergenic regions, respectively, in the NAM parental chromosomes (Table 2). The proportion of intronic to exonic SNP set was 1.73. Among the exonic SNPs, we observed 1.09 million non-synonymous SNPs with non-synonymous SNPs over-numbered by 1.67X as compared to the synonymous SNPs.

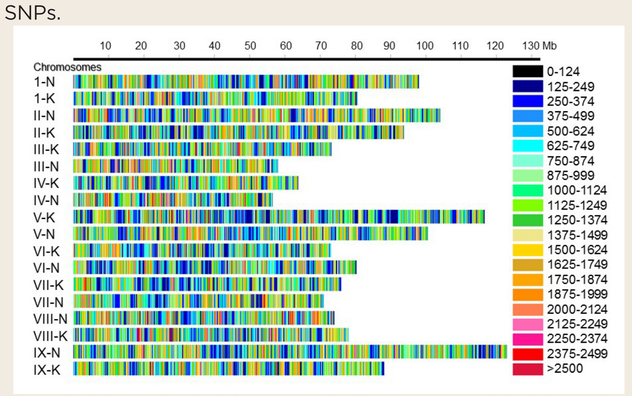

Figure 5. Color coded SNP frequency distribution in every 50 Kb interval across 18 chromosomes.

Table 2. Distribution of variant types

Variant Types		Variants (Chromosomes)	
		SNPs	InDels
Exonic		2,012,884	
	Synonymous	653,590	
	Non-synonymous	1,093,317	
	Stop Gain	30,616	4185
	Stop Loss	4,940	531
	Frameshift Deletion		87,676
	Frameshift Insertion		138,099
Intronic		3,469,802	
Intergenic		19,863,732	
5'UTR		599,560	
3'UTR		893,414	

Annotation and Gene Ontology (GO) analysis revealed that 39.83 % of the nonsynonymous SNPs belongs to 23.22% of the switchgrass v3.0 annotated genes (31,068 of 133,775 total transcripts) that have GO defined. Among the GO defined SNP set, 154,976 SNPs (within 9,934 genes), 21,599 SNPs (within 1,774 genes), and 7,860 (within 796 genes) belongs to molecular function, biological process and cellular component categories, respectively (Figure 6).

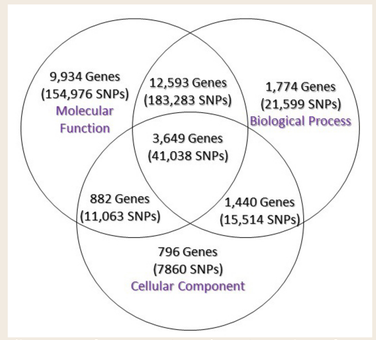

Figure 6. Venn diagram of gene ontogoly categories of non-synonymous SNPs

We estimated nonsynonymous SNP frequencies for each of the annotated 62,203 genes. The range of SNPs per gene varies from 1-344 with an average of 13.30±14.03. We observed 7,128 (11.46%) high-SNP density genes with the criteria of SNPs present at a frequency of at least 27 (mean SNPs+1SD) (Table 3).

Table 3. Statistics of SNP frequency per gene

Range	1-344
Mean with SD	13.30±14.03
Genes with SNP frequency of 1-12	38,582
Genes with SNP frequency of 13-26	16,493
Genes with SNP frequency of ≥ 27 (Mean + 1 SD) per Gene	7,128

Phylogenic tree created from the sequences of 2.01 million exonic SNP positions indicated that 15 parental genotypes and AP13 clustered in 5 distinct groups (Figure 7) with NFGA02_06 showed distantly related standalone cluster. AP13 clustered with parents, NFGA04_01, NFGA09_02 and NFGA37_05. Four members of NFGA15 and NFGA16 clustered together. Similarly, NFGA36 as well as NFGA34 clustered with their own members.

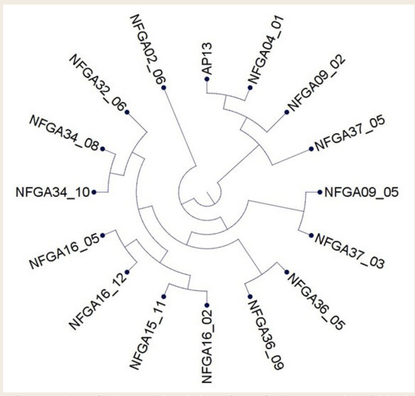

Figure 7. Phylogenic relationship (circular cladogram) of NAM parental genomes

Summary

- Up to 99.0 % of the parental sequences mapped to AP13 v2.0/3.0 reference genome.
- Cataloged a total of 27.78 million bi-allelic SNPs within NAM parental genome, among them there were 6.43-9.38 million polymorphic SNPs.
- On an average 1 SNP identified in every 48 to 64 bp of the genome with a maximum of 1 SNP in every 13 bp in SNP dense regions.
- Identified 2.01 million SNPs in exonic region with intronic to exonic SNPs ration of 1.72.
- Identified 1.09 million nonsynonymous SNPs in the exonic regions with 7,128 SNP dense genes.
- NAM parents clustered in 5 distinct groups with AP13 clustered with parents, NFGA04_01, NFGA09_02 and NFGA37_05 and NFGA02_06 positioned in standalone cluster.

Acknowledgements

The authors acknowledge Josh Barbour, The Samuel Roberts Noble Foundation, Ardmore, OK, for collecting of the field data. We also acknowledge Cheryl Dalid and Santosh Nayak, University of Tennessee, Knoxville, TN for maintaining and collecting data from NAM population planted at Noxville location.
The project was funded by U.S. DOE (Project funding #DE-SC0008781) and sequencing was done at the U.S. DOE Joint Genome Institute, Walnut Creek, California.