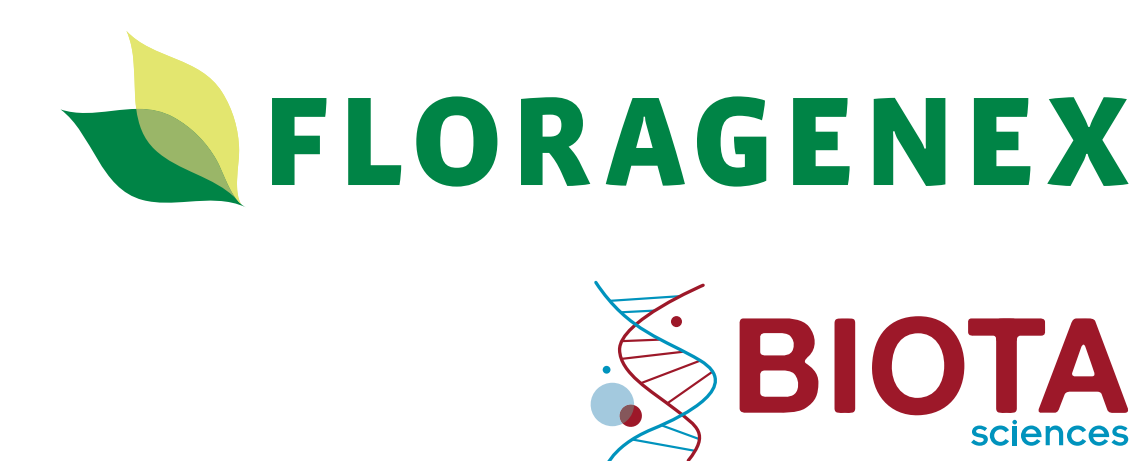


RAD LongRead: a SNP Discovery and *de novo* Sequence Assembly Strategy



Tressa S. Atwood, Jenna M. Gribbin, Jason Q. Boone, Rick W. Nipper, Nathan J. Lillegard and Eric A. Johnson

Floragenex, Inc. 1900 Millrace Drive, Eugene, Oregon, 97403 info@floragenex.com

Abstract

Accurate SNP discovery and *de novo* sequence assembly in complex plant genomes remains challenging despite ubiquitous second-generation sequencing technologies. Such platforms are encumbered with increased error rates and short read lengths which often do not provide sufficient information content to discriminate between highly similar genetic loci originating from paralogs or in duplicated, polyploid genomes. Longer DNA sequences provide enhanced resolving power for genome alignments, enable efficiencies in SNP detection and can uncover powerful haplotype information. Here we describe the RAD (Restriction-site Associated DNA) LongRead sequencing strategy as an efficient method to create *de novo* pyrosequence-length DNA contigs from paired-end Illumina/Solexa data. We report LongRead scans in the elite maize (*Zea mays*) inbreds B73 and Mo17 produced contigs ranging in size from 100 to 600 bp (N50: 375 bp), with extremely low sequence error rates (~0.05%). Preliminary analysis of sequence data indicates 92% of LongRead contigs anchor to single positions in the maize genome and identify SNP and InDels concordant with known polymorphisms.

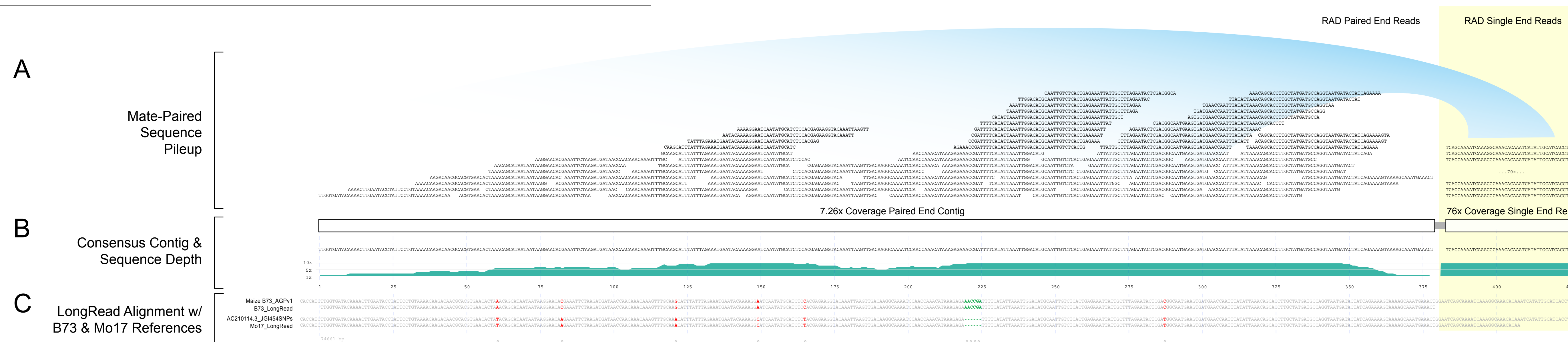


Figure 1. Assembled maize B73 and Mo17 LongRead contig alignment to reference genome(s). This illustration displays how a single LongRead contig is constructed from mate-paired Illumina/Solexa sequence. A) Paired end data from a clonal set of RAD single end reads (shown at right) is depicted as a pileup. There were 76 paired end reads (2x45 bp) incorporated into this assembly. In B) sequence coverage for every nucleotide in the LongRead contig is shown on the teal scale. Average coverage over the contig was 7.26x and ranged between 1x and 18x. Approximately 85% of the contig is covered by 3 or more reads. C) Alignment of the assembled B73 contig to the AGPv1 reference genome shows 100% identity between the two sequences. A homologous LongRead contig from the Mo17 cultivar is shown, along with a sequence annotated with available polymorphisms between B73 and Mo17 in the area of interest. All seven SNPs and Insertion/Deletions (Indels) in this region were detected by LongRead.

1 Introduction

Discovering genetic variation in species without an available reference genome often requires the development and assembly of large islands of DNA sequence surrounding the polymorphism of interest. A common example of this strategy in plant genomics is *de novo* EST/transcriptome sequencing, which identifies both genic sequence and sequence variation in parallel.

Here we present a novel approach for SNP development in unsequenced genomes. Based on the Restriction site Associated DNA (RAD) system, the innovative modification, called LongRead, is designed to increase the length and quality of sequence reads. As in classic RAD markers, LongRead interrogates tracts of DNA sequence flanking restriction enzyme digestion loci in the target genome. However, unlike traditional RAD markers, which are restricted to between 30 - 50bp in length, LongRead sequences can span hundreds of basepairs.

2 Approach

To test the performance of RAD LongRead in a well-studied plant genome, we selected two elite maize (*Zea mays* ssp *mays*) inbred lines; B73 and Missouri 17 (Mo17) for sequencing and technical benchmarking of the system. The availability of genomic resources for B73 and Mo17, allow us to examine the fidelity and accuracy of LongRead contigs compared to known standards.

The RAD LongRead protocol is shown below in Figure 2. First, DNA is digested with a restriction enzyme, followed by an adapter ligation step, then sonicated. Sheared RAD fragments are size-selected and a final adapter is ligated. The two adapters direct the sequencing of DNA adjacent to restriction enzyme cleavage sites and the randomized paired end (1,2). The overlapping RAD sequences from the sheared end are then computationally reassembled into 100 - 500bp contigs.

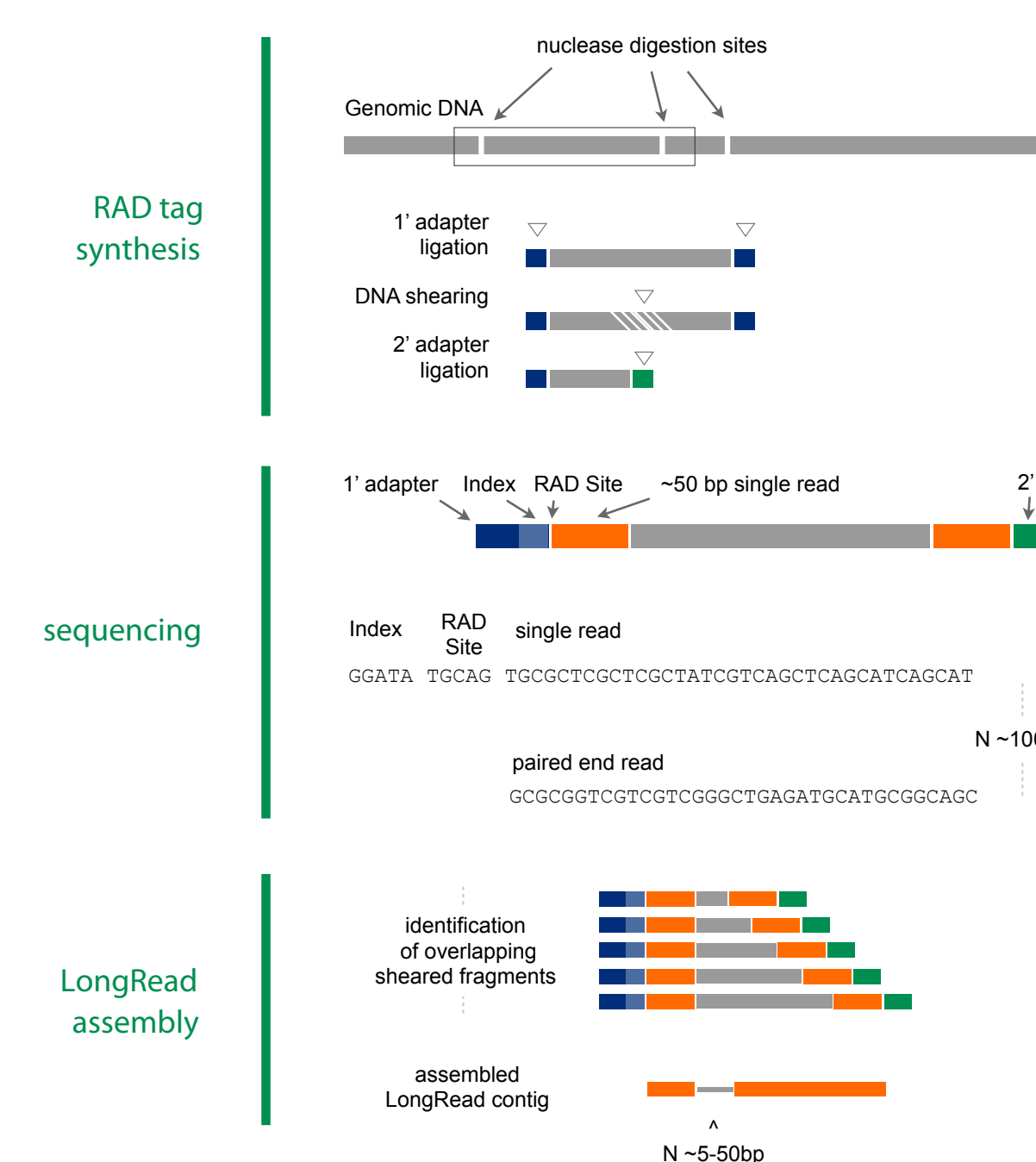


Figure 2. Illustration of the RAD LongRead protocol.

3 Methods

Germplasm, DNA Isolation and Library Preparation

B73 and Mo17 seeds (accessions PI 550473 and PI 558532) were obtained from the USDA / ISU NCRPIS stock center and germinated in potting soil for 10 days. Young leaf tissue from was snap frozen under liquid nitrogen, pulverized and DNA extracted using a modified Qiagen PureGene Genra protocol. High quality genomic DNA from each line was then processed into an Illumina-GAI compatible RAD library using the enzyme SbfI based on the methods of Baird, et al 2008 (1,2).

Sequencing and LongRead Contig Assembly

RAD libraries were sequenced on an Illumina Genome Analyzer IIx using 2 x 54 bp paired-end chemistry. Approximately 1M reads were obtained for each accession.

Accession	Number of Reads
B73	1,212,238
Mo17	912,293

To assemble RAD LongRead contigs, several filtering and processing steps were used. First, any raw sequences with >5 poor Illumina quality scores (Q10 or lower) were discarded. Reads passing filters were then grouped together based on Illumina single end data. A minimum of 60 redundant single end reads (60x depth) were required for each locus. The cognate paired end sequences were isolated and used for LongRead contig construction using a modified version of Velvet (3). Both B73 and Mo17 LongRead contig builds were completed independently without the aid of the reference genome. After initial assembly, an additional round of processing removed fragmented contigs with at least one gap in the paired-end assembly.

4 Results

Evaluation of LongRead Contigs

Table 1 provides general assembly information from the B73 and Mo17 LongRead builds. Contigs assembled from both cultivars displayed similar contig lengths (Figure 3) and sequence coverage. The increased number of contigs seen in B73 is likely due to the difference in the number reads obtained between the samples.

Table 1. LongRead Contig Statistics

	B73	Mo17
Number of Contigs	2,583	1,884
N50 Contig Length (bp)	375	362
Average Contig Coverage	6.86x	6.47x
<i>de novo</i> Sequence Generated (kb)	860.1	606.6

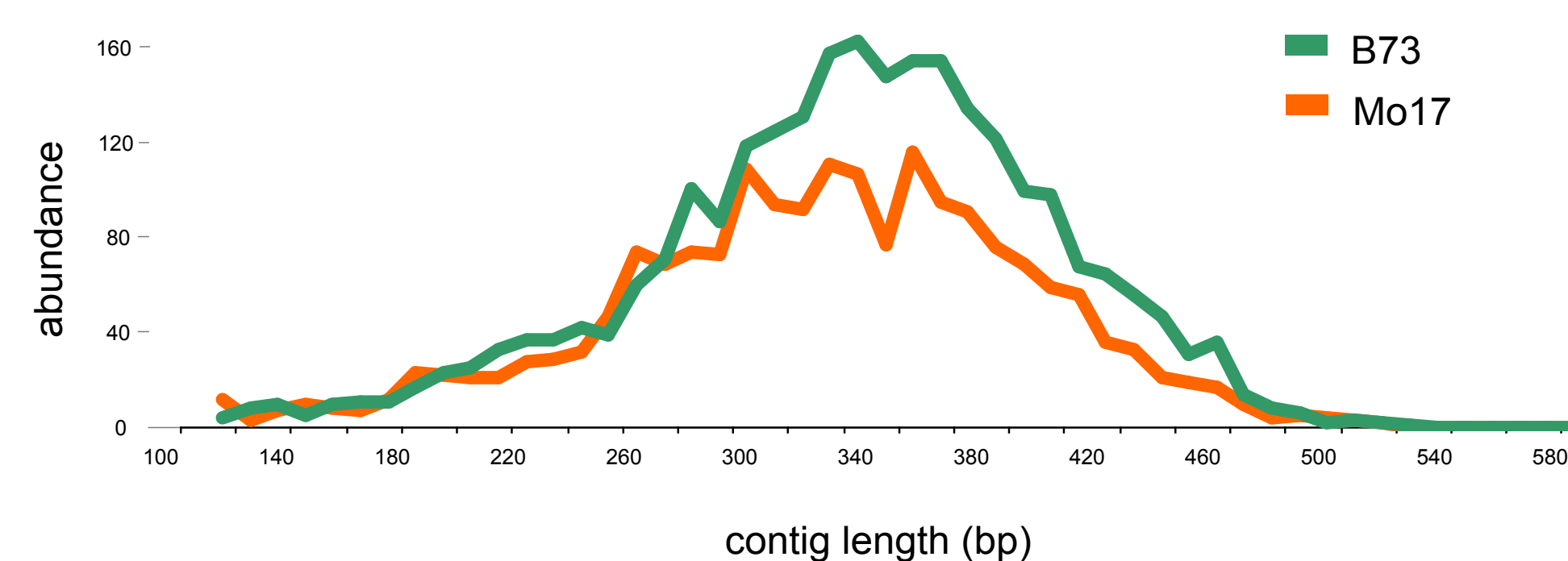


Figure 3. Histogram of RAD LongRead contig lengths for B73 and Mo17. Contig lengths for both maize accessions are noted in orange and green lines above. LongRead contig lengths display a Poisson distribution, consistent with DNA fragmentation through random shearing. Both accessions share a peak maxima at approximately 345 bp.

LongRead Assembly Quality

To determine the reliability and accuracy of RAD LongRead contigs, we aligned all 2,583 B73 contig assemblies to the *Zea mays* B73 reference genome (AGP v1.0) with SSAHA2 using Sanger read-length stringency parameters (4,5,6). A representative LongRead contig uniquely aligning to linkage group 9 is shown in Figure 1 above. A summary of statistics from the comprehensive genome-wide analysis is shown below in Table 2.

Table 2. LongRead Whole Genome Alignment

Number of B73 LongRead Contigs	2,583
Number of Uniquely Anchoring Contigs (UACs)	2,396 (92.7%)
Number of UACs w/ 100% Identical Sequence Alignment to B73 AGPv1	2,207 (92.1%)
Overall Nucleotide Identity between B73 LongRead contigs & B73 AGPv1	99.95%

We identified a large number of B73 LongRead contigs (92.7%) that successfully anchored to single loci on the maize physical sequence suggesting LongRead sequences provide sufficient information content for mapping in a complex plant genome. Examination of the alignment files indicates that the overall nucleotide identity between B73 LongRead contigs and the AGPv1 genome exceeds 99.9%, consistent with a high-quality LongRead assembly.

SNP and InDel Detection

Over 1.2M SNPs and InDels identified between B73 and Mo17 have been made publicly available as part of ongoing genome sequencing projects (5). To determine if SNPs identified from RAD LongRead contigs matched known B73 x Mo17 polymorphisms, we analyzed a small set of contigs. Figure 1C, above, displays an typical alignment between the RAD contigs, the B73 genome and shotgun 454 sequence from the Mo17 cultivar. We observe a high level of concordance between polymorphisms identified through LongRead and established genetic variation in B73 versus Mo17.

5 Conclusions

Our findings suggest LongRead is an efficient and accurate tool for SNP detection and *de novo* sequence development. We envision future applications will include Genome Survey Sequencing, SNP and InDel discovery, haplotype analysis in polyploid genomes and *de novo* genome assembly.

6 References

- Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* 3(10): e3376 doi:10.1371/journal.pone.0003376
- Faculty of 1000 Biology: evaluations for Baird NA et al. *PLoS ONE* 2008 3(10): e3376 <http://www.f1000biology.com/article/doi/10.1371/journal.pone.0003376>
- Zerbino, DR and Birney E. 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*. 18(8):821-829
- Schnable, et al. 2009. The B73 Maize Genome: Complexity, Diversity and Dynamics. *Science*. Vol. 326. no. 5956, pp. 1112 - 1115. DOI: 10.1126/science.1178534
- <http://www.phytozome.net/maize.php>
- <http://www.maizesequence.org/>
- Produced from Genome Sequencing Center at WUSTL
- Ning, Z, Cox, AJ and Mullikin, JC. 2001. SSAHA: a fast search method for large DNA databases. *Genome Research* 11: 10: 1725-9

7 Acknowledgements

The authors wish to thank the USDA ISU North Central Regional Plant Introduction Station for providing germplasm for this project. The database of 1.2M B73 x Mo17 Single Feature Polymorphisms was obtained from the Phytozome4.1 FTP server, released as part of the DOE-JGI Mo17 sequencing effort.