SNP Discovery Using Restriction-Site Associated DNA (RAD) LongRead Sequencing in *Brassica napus*, a Polyploid Species

Shunxue Tang[1], Michelle Wiggins[1], Rick Nipper[2], Jenna Gribbin[2], Eric Johnson[2], Nathan Lillegard[2], Thomas Greene[1], Steve Thompson[1], Siva Kumpatla[1]

[1] Dow AgroSciences LLC, 9330 Zionsville Road, Indianapolis, IN 46268, USA

[2] Floragenex, Inc. 1900 Millrace Drive, Eugene, OR 97403, USA

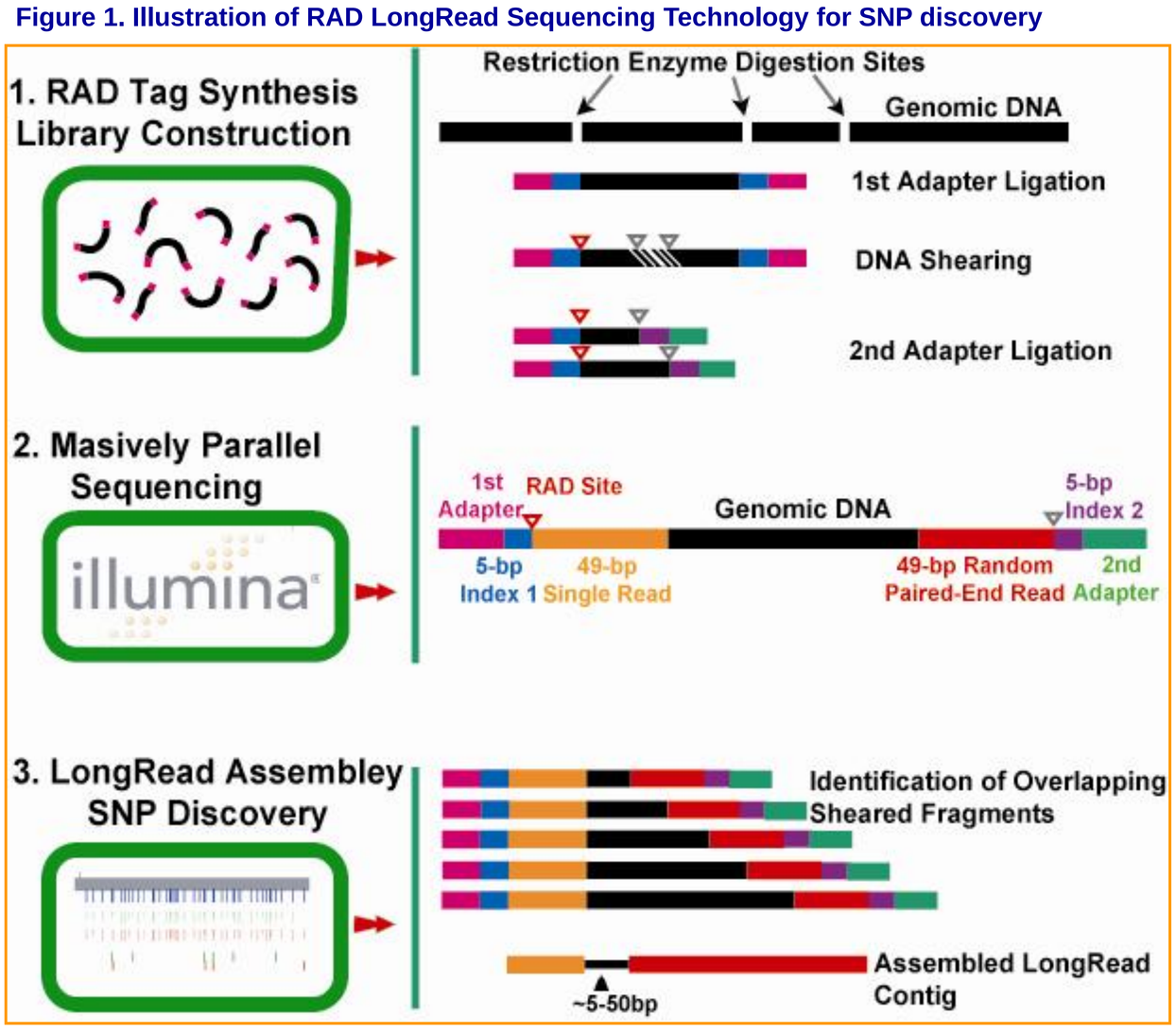
Introduction

Whole-genome association studies and fine mapping of complex traits in crops like oilseed rape (*Brassica napus* L.) require high density of single nucleotide polymorphism (SNP) markers uniformly spread over the entire genome. To design high-throughput, multiplexed SNP genotyping assays such as Illumina GoldenGate or Infinium, researchers must first discover a large numbers of SNPs by extensively sequencing the genomic fragments of multiple lines.

A large number of SNPs have been discovered through the joint efforts of *Brassica* SNP Discovery Consortium via *in silico* methods using existing *B. napus* expressed sequence tags (ESTs) or 454 sequencing of *B. napus* transcripts and gene families (Chatel et al. 2008)). Because of the uneven distribution of *Brassica* genes in genomes, SNP markers from non-genic regions are needed to fill the gaps in the genetic maps. However, the development of SNP markers in oilseed rape, an allotetraploid without reference genome sequences, is very challenging. Particularly, polyploidy greatly increases the sequencing depth for SNP discovery since more sequences from each line are needed to distinguish the allelic variants (intragenomic SNPs) from the non-allelic (intergenomic) variants at SNP positions in contigs. Thus, the reduction of genome complexity for sequencing is very important for the SNP discovery in oilseed rape.

The Floragenex Restriction-Site Associated DNA (RAD) Technology, which combines the genome complexity reduction technology with high throughput next-generation sequencing, provides a rapid, effective and cost-efficient approach for SNP discovery in model and non-model organisms (Baird et al. 2008). Here, we describe the SNP discovery through the sequencing of the RAD fragments from two canola inbred lines using the Illumina Genome Analyzer IIx platform.

Figure 1. Illustration of RAD LongRead Sequencing Technology for SNP discovery



Materials and Methods

DNA samples of two canola inbred lines, DAS01 and DAS02, were digested with restriction enzyme *Pst* I, followed by adapter ligation and DNA fragment sonication. Sheared RAD fragments were size-selected and a 2nd adapter was ligated. The two adapters directed the sequencing of DNA adjacent to *Pst* I cleavage sites and the randomized paired-ends. The RAD libraries were sequenced on an Illumina Genome Analyzer IIx using 2 x 54 bp paired-end chemistry (Figure 1).

de novo assembly of RAD LongRead contigs, included several filtering and processing steps. First, any raw reads with >5 nucleotides of the quality scores Q10 or lower were discarded. Reads passing filters were then grouped together based on Illumina single end data. A minimum of 20 redundant single end reads (20x depth) were required for each locus. The overlapping RAD sequences from the randomized paired-ends were assembled into 100-500bp contig using a modified version of Velvet (Zerbino and Birney 2008). A SNP discovery pipeline based on sequencing depth in each canola line at SNP positions was used to distinguish allelic variants (simple SNPs and hemi-SNPs) from non-allelic (paralog) variants (false SNPs, nucleotide polymorphisms between paralogs from A and C genomes) (Figures 1 and 2).

Results and Conclusions

•A total of approximately 20 million raw reads were obtained from both inbred lines, ~10M reads from DAS01 passed the quality filter, and were used for initial *de novo* assembly. A total of 52,048 contigs were obtained, ranging in sizes from 201 to 631 bp with an N50 length of 280 bp, and with an average contig sequencing depth of 9.0x.

•All reads of high quality from both inbred lines were aligned to assembled contigs for SNP discovery. A total of 108,551 putative polymorphisms were identified in 27,155 contigs.

•A total of 1,094 simple SNPs, 1,634 hemi SNPs, and 6,644 paralog variants from 8,389 contigs passed the sequencing depth test (Figure 2). Accounting for the Illumina reads from both inbred lines, the average sequencing depth at the SNP positions was 21.0x.

•More than 2,000 candidate SNPs discovered from this project will be used for assay development towards validation on a panel of oilseed rape inbred lines followed by mapping in different populations.

•This study demonstrated that RAD LongRead sequencing technology was a cost-efficient method for high-throughput SNP discovery in a polyploid crop like canola.

References

Chatel G, Guigou G, Kalogerakis M, Dayton L, Guo Y, Cheung WY (2008) *Brassica* SNP Discovery And Validation. Plant & Animal Genome Conference XVI, San Diego, CA, USA. Jan 12-16, 2008.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA and Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3: e3376.

Zerbino DR and Birney E (2008) Velvet: Algorithms for *de novo* short read assembly using de Bruijin graphs. Genome Research 18:821-829.

Figure 2. Assembled LongRead contig alignments indicated the (A) simple SNPs, (B) hemi-SNPs and (C) paralog variants (false SNPs) identified between two canola inbred lines DAS01 and DAS02. The sequence reads of the randomized paired-ends from DAS01 and DAS02 were highlighted in blue and red, respectively; the contig sequences assembled from the short paired-end reads were highlighted in black; the black rectangles indicated the locations of the SNPs or paralog variants in the contigs.