Application of a Novel, Targeted Sequencing-Based Genotyping Approach for Cost Effective Marker Assessment in *O. sativa*.

Andrew Barry[1], Amy B. Emerman[2], Bjoern Textor[3] Kruti M. Patel[2], Sarah K. Bowman[2], Scott M. Adams[2], Brendan S. Desmond[2], Jonathon S. Dunn[2], Susan E. Corbett[2], Charles D. Elfe[2], Evan Mauceli[2], and Cynthia L. Hendrickson[2]

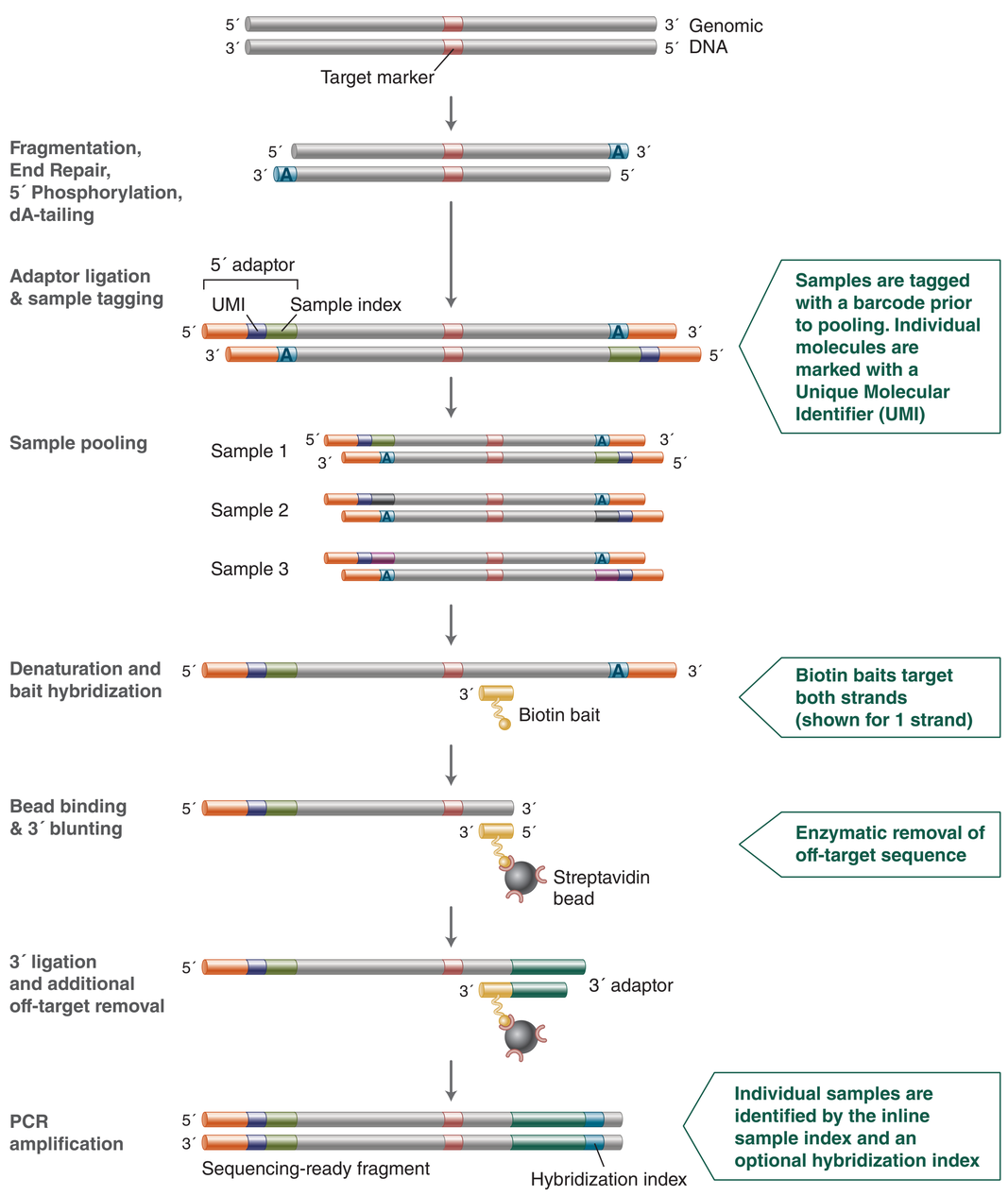
[1]New England Biolabs, Inc., Ipswich, MA; [2]Directed Genomics, Ipswich, MA; [3]New England Biolabs, GmbH, Frankfurt, Germany

Introduction

Decreases in sequencing costs have increased the availability of public SNP databases while necessitating development of targeted genotyping assays for use in marker assisted genomic selection for a variety of crop species. The NEBNext Direct Genotyping Solution is a novel, hybridization-based target enrichment approach that has been optimized for use in genotyping applications to increase the number of assays that can be performed in a single reaction, while providing sequencing coverage depth suitable for SNP identification. The approach enables high-levels of multiplexing of both isolates and markers, allowing enrichment of hundreds of thousands of SNP targets in a single hybridization reaction, and the protocol is easily completed in a single day.

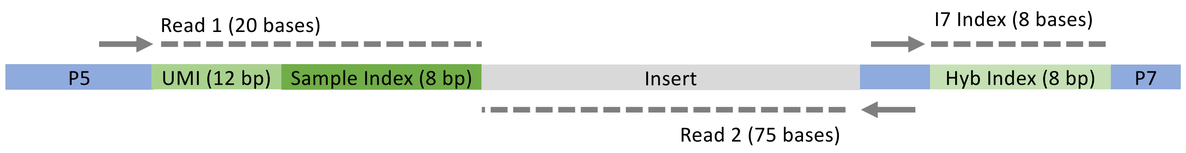
We developed a panel covering the 1,996 single nucleotide polymorphisms previously identified as markers for polymorphism detection in *O. Sativa*. Here, we demonstrate the application of this panel to cost-effectively enrich defined SNP markers in a highly specific and uniform manner prior to next-generation sequencing.

Workflow



Methods

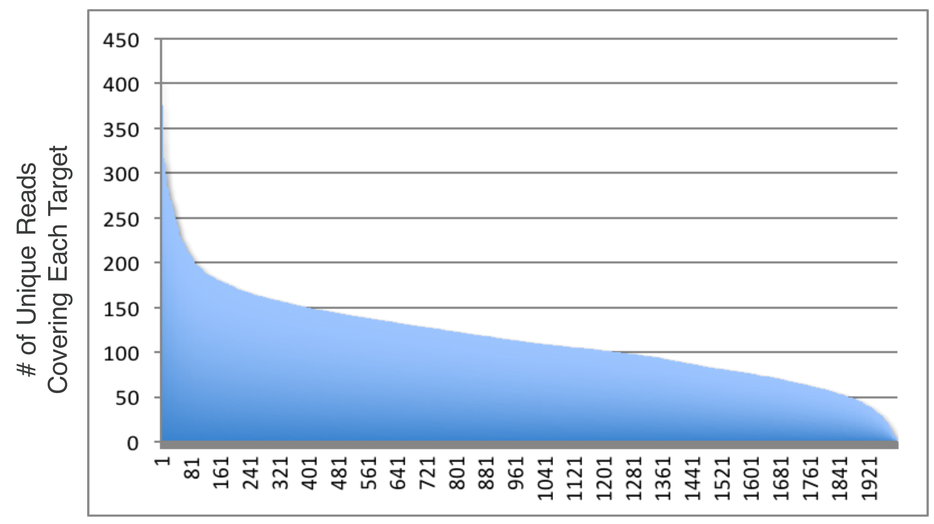
25 ng of 96 individual rice DNA samples were enzymatically fragmented and 5' tagged with an Illumina®-compatible P5 adaptor that incorporates both an inline sample index to tag each sample prior to pooling and an inline UMI to mark each unique DNA fragment within the samples, as shown in the workflow. The 96 samples were pooled and enriched using the 1,996 genetic marker bait pool targeting common markers from the Generation Challenge Programme to interrogate polymorphic sites within and between the main Oryza groups in a single hybridization reaction, followed by library preparation and 16 cycles of PCR amplification. After purification and quantification, the 96-plex library was sequenced in a single NextSeq® run as shown in the diagram below, where Read 1 captures the inline UMI and sample barcode, the i7 read (Index 1) captures a second index added to all samples in the same hybridization-based enrichment, and Read 2 captures the target rice sequence.



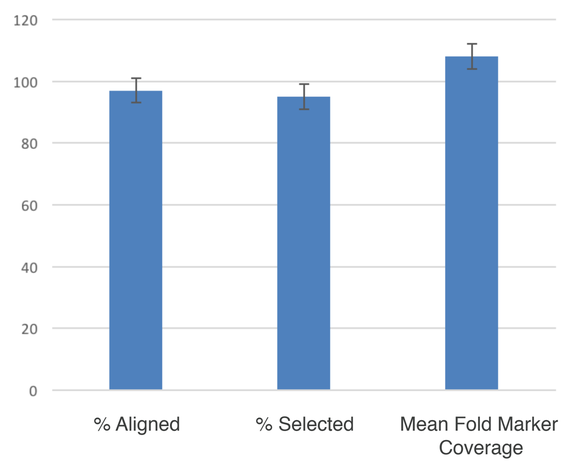
After sequencing, the reads were demultiplexed with a Picard-based workflow[1]. Sequencing reads were aligned to the Osativa_193 reference genome[2] using BWA-MEM[3] and PCR duplicates were identified using the UMIs[4].

[1]http://broadinstitute.github.io/picard
[2]http://www.plantgdb.org/XGDB/phplib/download.php?GDB=Os
[3]Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]
[4]Fulcrum Genomics, https://github.com/fulcrumgenomics/fgbio

Uniform Coverage Across 1,996 Markers

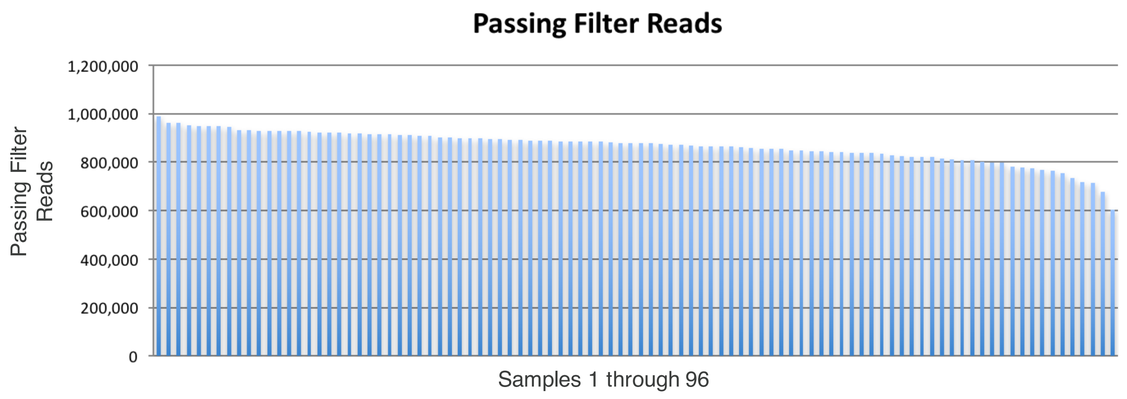


Unique read depth of the targeted 1,996 markers from a single sample within the 96-plex enrichment, after removing PCR duplicates using the UMIs, demonstrating uniform enrichment across SNP marker targets.

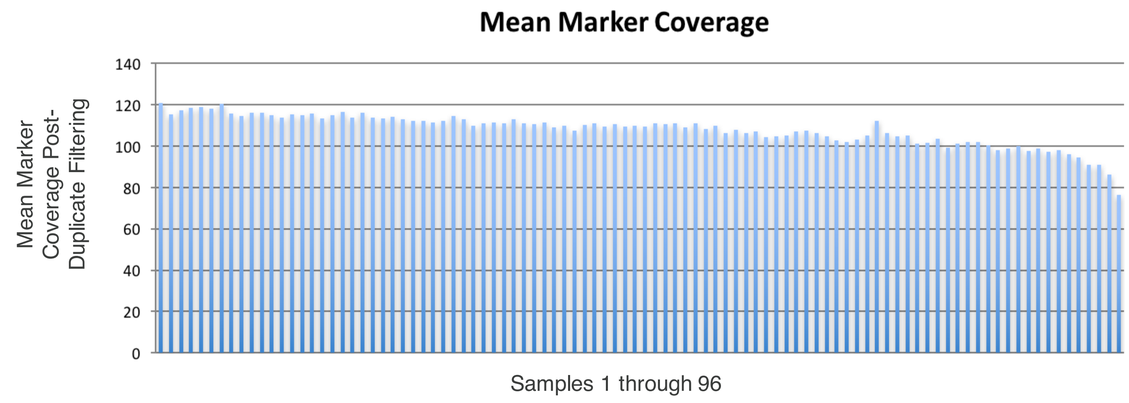


The 96 libraries enriched with the 1,996 marker rice panel demonstrated a high percentage of reads aligning to the *O. Sativa* genome, high specificity for targeted markers, and consistent coverage across marker loci. Bar graph values represent averages across the 96 pooled samples, with error bars indicating standard error across data derived from each of the 96 samples assayed.

Uniform Coverage Across 96 Multiplexed Samples



Uniform distribution of sequencing reads across all 96 pooled samples demonstrates the comparable performance of each sample during hybridization-based capture and library preparation. The number of reads reported represent the number of 75 base sample reads generated.



Following removal of PCR duplicates, the 96 pooled samples had an average of 107 unique reads covering the targeted markers in a portion of the NextSeq® run, with very little variation in the mean coverage across the samples.

Advantages

- Robust, user-friendly protocol to generate Illumina-compatible, target-enriched libraries within one day
- Multiplexes samples upfront to reduce cost and increase throughput
- Scalable from 100-5000 markers or more
- Processes up to 9216 samples in a single 96-well plate
- Flexible multiplexing: Same protocol can pool 4 to 96 samples into a single hybridization
- Maximized on-target bases by enzymatic removal of off-target sequences
- Column purification of DNA samples is not required for most plants
- Safe workflow stopping points throughput the protocol
- Automation-friendly