Highly Multiplexed, Targeted Sequencing for Genotyping Maize with the NEBNext Direct Approach

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

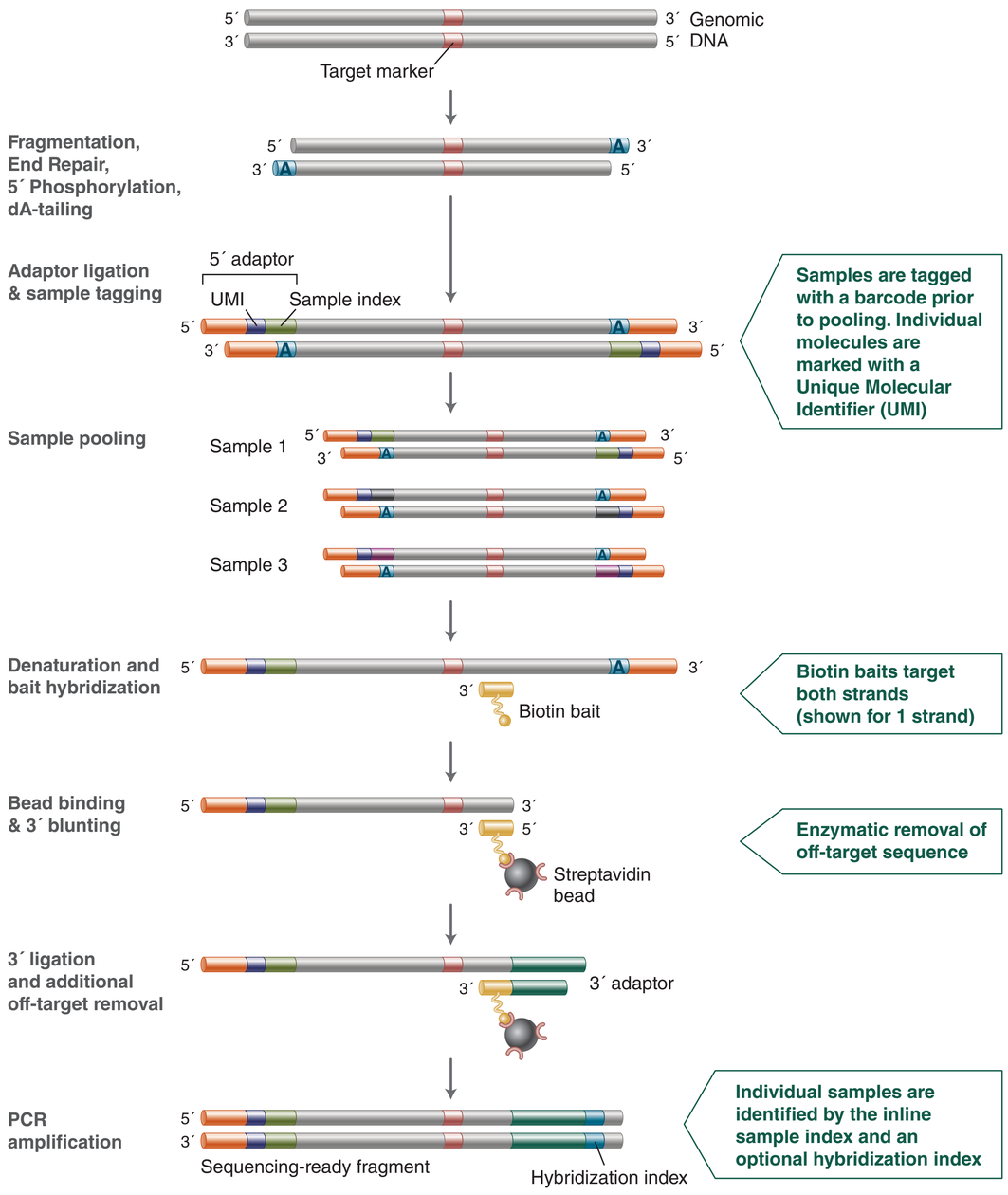
[1]Directed Genomics, Ipswich, MA; [2]New England Biolabs, Inc, Ipswich, MA; [3]New England Biolabs, GmbH, Frankfurt, Germany

Introduction

Targeted next-generation sequencing of molecular markers is a desirable approach to genotype crops for marker-assisted breeding. These methods offer several advantages over other genotyping approaches, including the ability to interrogate thousands of variant sites with a single assay while providing additional information on nearby sequences. However, NGS-based genotyping is typically more expensive than traditional genotyping methods, and for marker-assisted selection, many samples need to be screened to identify individuals to cross. Thus, it is important that the approach used to prepare samples for sequencing is high-throughput and that the cost per sample is low. Here we present the NEBNext Direct Genotyping Solution, a novel, capture-by-hybridization method that allows for processing of up to 9216 samples in a single 96-well plate.

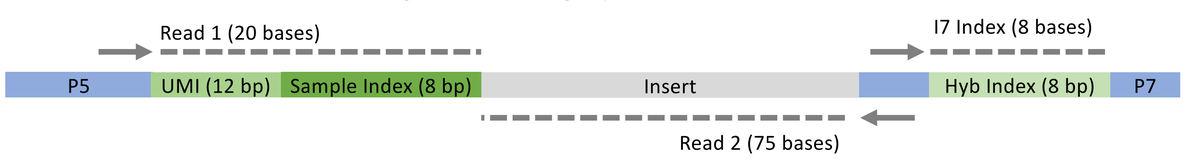
To demonstrate the capabilities of this approach, we applied the NEBNext Direct Genotyping Solution to genotype maize genomic DNA. We developed a panel of over 4600 legacy SNPs from the Panzea project and tested the ability of the panel to evenly enrich targets from 25 ng of maize DNA. Additionally, because the baits were individually synthesized, subsets of the panel could be rapidly generated to reduce sequencing costs when fewer targets were required. To demonstrate this ability, we selected a 100 marker subpool from the larger bait set and observed consistent coverage of the selected targets while maintaining the high specificity and uniformity of the panel. With this one day, highly multiplexed protocol, hundreds of samples could be processed in a high-throughput manner, making this approach ideal for genomic selection in maize.

Workflow



Methods

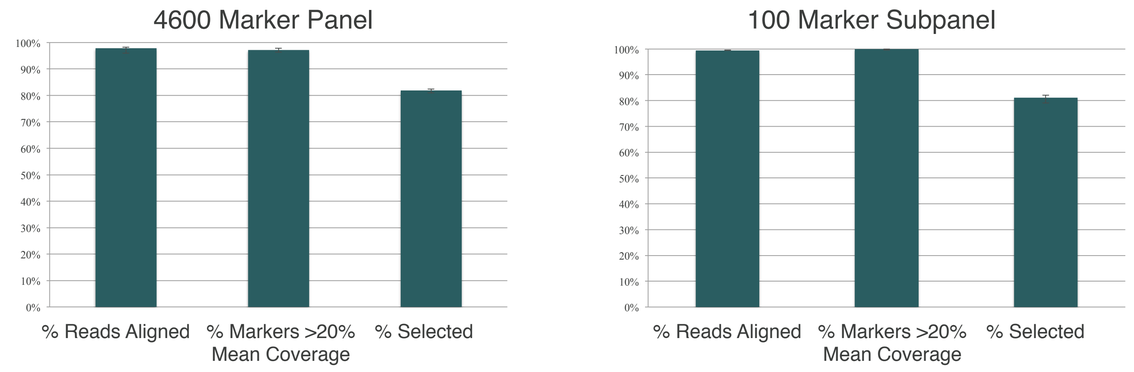
25 ng of 96 individual maize DNA samples were enzymatically fragmented and 5' tagged with an Illumina®-compatible P5 adaptor that incorporates both an inline sample index to tag each sample prior to pooling and an inline UMI to mark each unique DNA fragment within the samples, as shown in the workflow. The 96 samples were pooled and enriched in a single hybridization reaction with the 4600 genetic marker bait pool targeting legacy SNPs from the Panzea project. Following enrichment, Illumina libraries were prepared and PCR amplified. After purification and quantification, the 96-plex library was sequenced on a portion of a NextSeq® flowcell, 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 maize sequence. These same methods were then applied to enrich and sequence 96 maize samples with a 100 marker subset of the larger Panzea legacy SNP panel.



After sequencing, the reads were demultiplexed with a Picard-based workflow[1]. Sequencing reads were aligned to the B73 RefGen_v2 reference genome[2] using BWA-MEM[3] and PCR duplicates were identified using the UMIs[4].

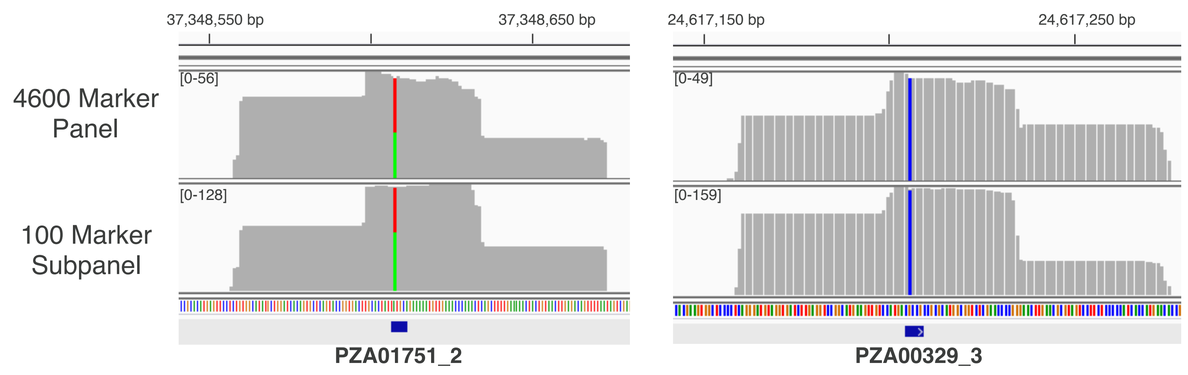
[1]http://broadinstitute.github.io/picard
[2]Andorf CM et al (2015). MaizeGDB update: new tools, data, and interface for the maize model organism database. Nucleic Acids Res D1: 44, D1195-D1201.
[3]Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]
[4]Fulcrum Genomics, https://github.com/fulcrumgenomics/fgbio

Panel Performance



The 96 libraries enriched with the full 4600 marker panel or a subpool of baits targeting 100 markers demonstrated a consistently high percent of reads aligning to the genome, a high percent of targets obtaining reads greater than 0.2X of the mean target coverage, and high specificity for targeted markers. Bar graph values represent averages from the 96 pooled samples. Error bars represent the range of observed values across the 96 samples.

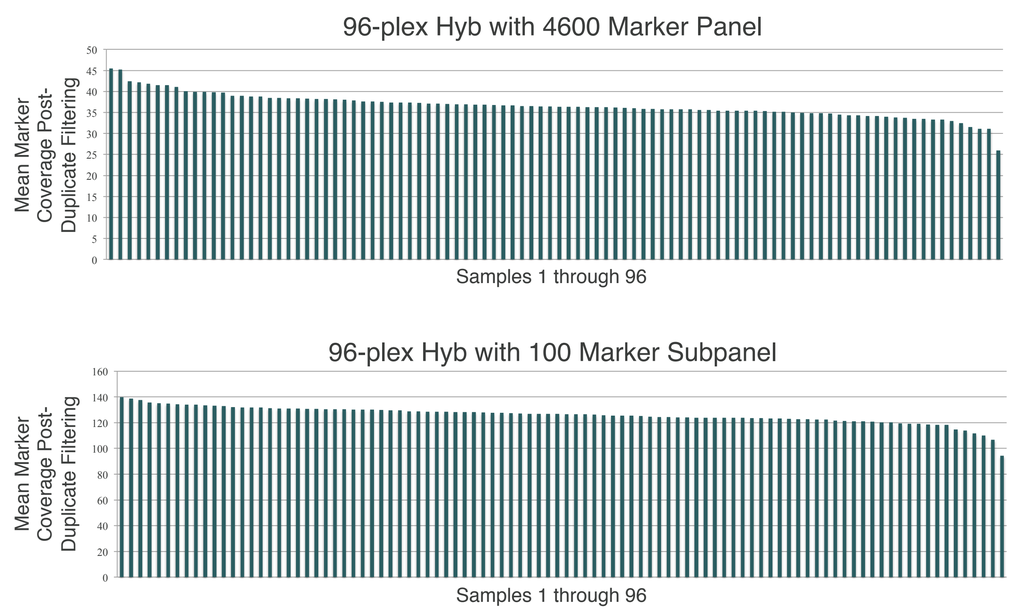
Target Coverage Maintained Across Panels



Two examples of coverage of targeted markers, as visualized in the Integrative Genome Browser (IGV)[5], demonstrating consistent coverage of the same targets in the context of different panel sizes and content. Examples shown are from single samples within a 96-plex hybridization using either the 4600 marker panel or the 100 marker subpanel. Reads were generated from targeting both strands of the genomic DNA and are deduplicated using UMIs.

[5]Robinson JT et al (2011) Integrative genomics viewer. Nat Biotech 29:24-26, and Thorvaldsdottir H et al (2013) Integrative Genomics Viewer (IGV): high-performance data visualization and exploration. Briefings in Bioinformatics. 14:178-192

Uniform Coverage Across 96 Multiplexed Samples



Uniform distribution of mean target coverage across all 96 pooled samples demonstrates the comparable performance of each sample during hybridization-based capture and library preparation with both the large 4600 marker panel and the 100 marker subpanel. The coverage values represent unique reads covering the selected targets after removing PCR duplicates using the UMIs.

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