High-Throughput Screening of 2300 Genetic Markers in *S. lycopersicum* using the NEBNext Direct Multiplexed Genotyping Approach

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

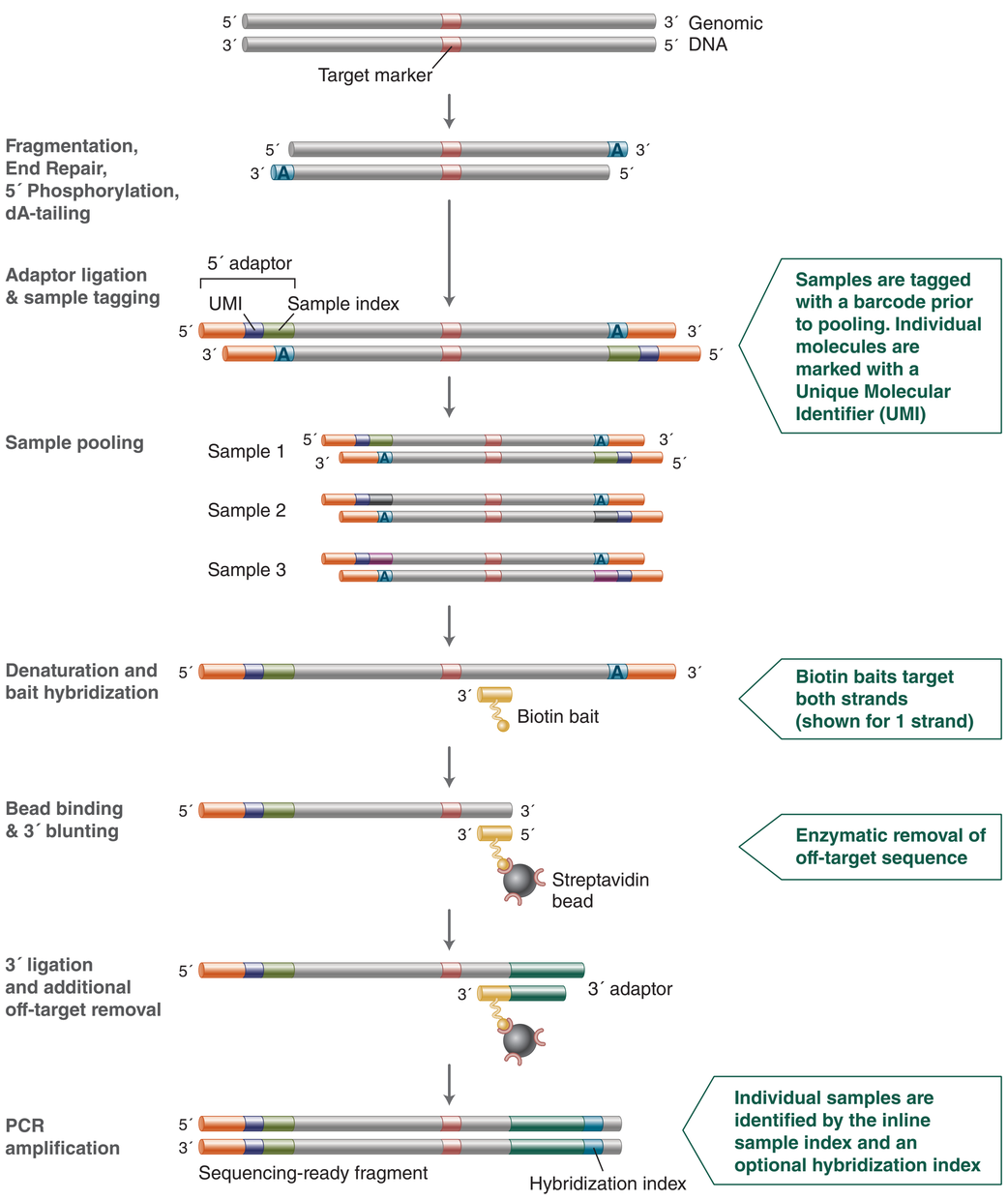
[1]New England Biolabs, GmbH, Frankfurt, Germany; [2]Directed Genomics, Ipswich, MA; [3]New England Biolabs, Inc, Ipswich, MA

Introduction

Targeted DNA sequencing is rapidly being adopted for the molecular screening of markers during selective crop breeding. For these applications, the need for cost-effective and high-throughput technologies to process large numbers of samples is imperative. Here we describe a novel capture-by-hybridization method for targeted genotyping of crops. This simple workflow allows processing of up to 9216 samples in a single 96-well plate in one day and is easily automated.

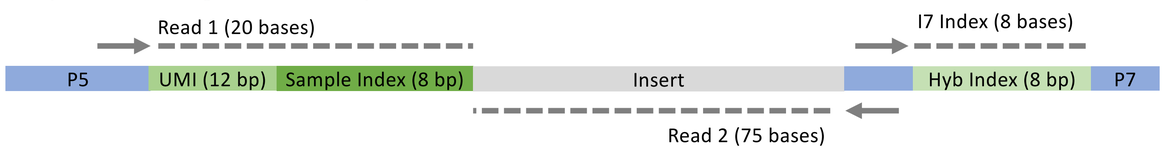
The NEBNext Direct Genotyping Solution can target 100 to 5000 markers from up to 96 samples within a single hybridization. Here we developed a panel targeting 2300 SNPs in the tomato crop, *Solanum lycopersicum*. Baits were placed within 75 nucleotides of the targeted SNPs, allowing for an efficient sequencing run of 75 bases of target sequencing, 8 bases of sample barcode, 8 bases of hybridization barcode, and 12 bases of a unique molecular identifier (UMI) for filtering PCR duplicates. After an initial screening of the panel, the bait concentrations were adjusted by performance to ensure uniform coverage of the targets. The optimized panel resulted in greater than 90% of the sequencing reads mapping to targeted regions and highly uniform coverage. As a result, this approach reduced the cost and increased the throughput of crop sequencing while generating robust data to reliably genotype multiple varieties of *S. lycopersicum*.

Workflow



Methods

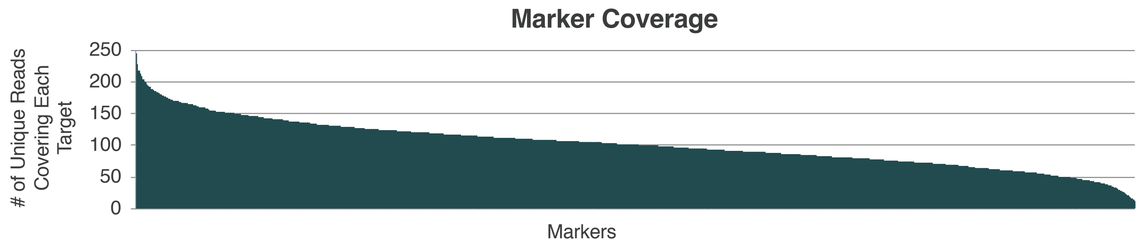
25 ng of 96 individual tomato DNA samples were enzymatically fragmented and 5' tagged with an Illumina®-compatible P5 adaptor that incorporates both an inline sample index to tag each sample prior to pooling and an inline UMI to mark each unique DNA fragment within the samples, as shown in the workflow. The 96 samples were pooled and enriched using the 2300 genetic marker bait pool targeting common markers from the Solanaceae Coordinated Agricultural Project (SolCAP) 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 MiSeq® 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 tomato sequence.



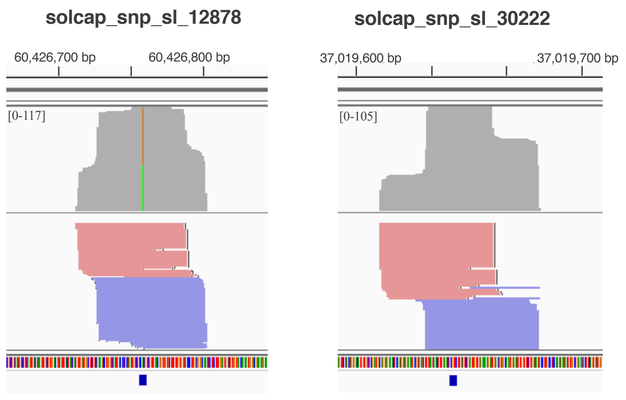
After sequencing, the reads were demultiplexed with a Picard-based workflow[1]. Sequencing reads were aligned to the SL2.40 reference genome[2] using BWA-MEM[3] and PCR duplicates were identified using the UMIs[4].

[1]http://broadinstitute.github.io/picard
[2]https://solgenomics.net/help/index.pl
[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 2300 Markers



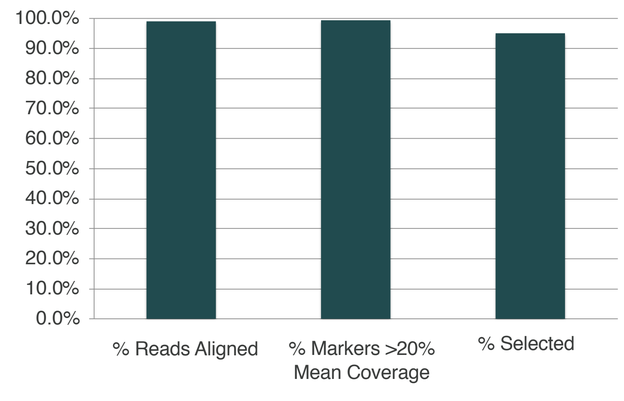
Unique read depth of the targeted 2300 markers from a single sample within the 96-plex enrichment, after removing PCR duplicates using the UMIs, demonstrating even enrichment of targets.



Two examples of the coverage of targeted markers within a single sample from the 96-plex enrichment as visualized in the Integrative Genome Browser (IGV)[5]. Reads shown are deduplicated using UMIs. Baits target both strands of the input DNA, as indicated by the red and blue aligning reads.

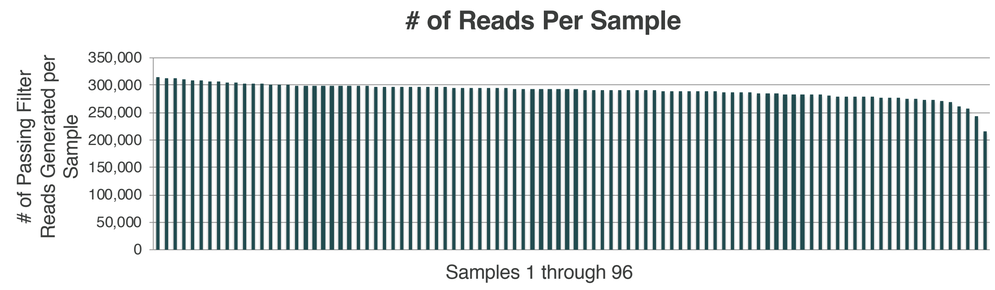
[5]Robinson JT et al (2011) Integrative genomics viewer. Nat Biotech 29:24-26, and Thorvaldsdottir H et al (2013) Integrative Genomics Viewer (IGV): high-performance data visualization and exploration. Briefings in Bioinformatics. 14:178-192

Consistent Panel Performance Across Samples

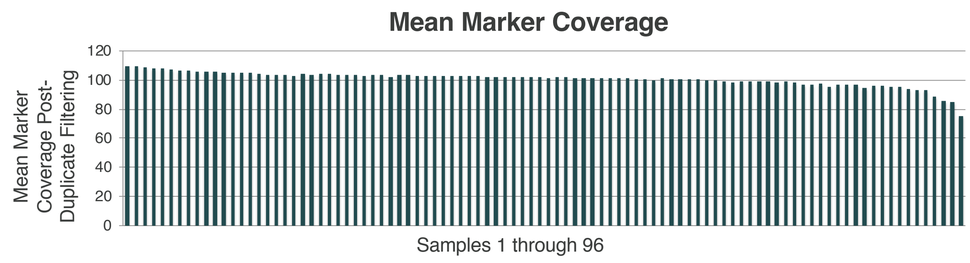


The 96 libraries enriched with the 2300 marker tomato panel demonstrated a high percent of reads aligning to the genome, uniform target coverage, and high specificity for targeted markers. Bar graph values represent averages across the 96 pooled samples, with each individual sample varying from the averages by less than 0.3%

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.



After removing the PCR duplicates, the 96 pooled samples had an average of 100 unique reads covering the targeted markers in a single MiSeq 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