A highly multiplexed target enrichment approach for sample identification and tracking using the NEBNext Direct Genotyping Solution

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

[1]New England Biolabs, Inc[1], Ipswich, MA; [2]Directed Genomics, Ipswich, MA

INTRODUCTION

Next-generation sequencing is increasingly being adopted for genetic screening and clinical diagnostics. To prevent false reporting of results, it is imperative that patient samples are tracked throughout sample processing and data analysis. A reliable method to track sample identity throughout a workflow is to monitor single nucleotide polymorphisms (SNPs) that are highly discriminatory across individuals. In order to incorporate a routine sample tracking method into diagnostic workflows, the method should be reliable, high-throughput, and cost-effective.

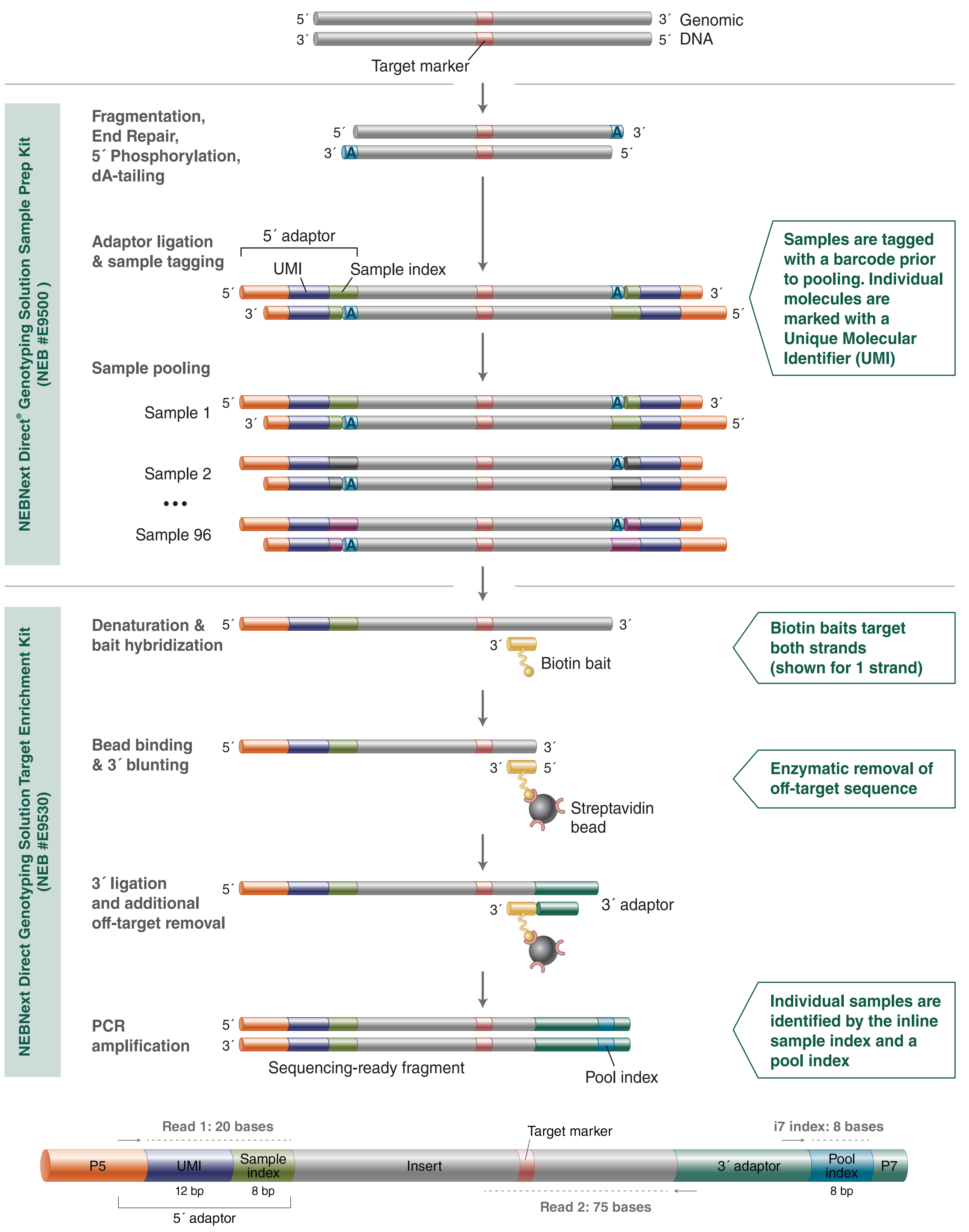
To address the need for high-throughput genotyping assays, we developed the NEBNext Direct® Genotyping Solution. This approach enables multiplexing of up to 96 samples in a single hybridization reaction that targets between 100 to 5000 SNPs. Here we demonstrate the power of this approach to distinguish 24 unique human samples from each other using a sample identification panel of highly discriminatory SNP targets. Using this approach, minimal sequencing reads were required per sample to obtain sufficient data for germline variant calling. With a one-day target enrichment and library preparation protocol and an approximately 12 hour sequencing strategy, we went from DNA samples to data in less than 24 hours. Our approach offers a convenient and reliable method to ensure that data integrity is maintained in a diagnostic workflow.

METHODS

Pooling of 24 HapMap DNA samples
Using the NEBNext Direct Genotyping Solution Sample Prep Kit, 25 ng of 24 different HapMap DNA samples from the Coriell Institute for Medical Research 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. After index tagging, the samples were pooled together for hybridization.

Target enrichment and sequencing of pooled samples
Using the NEBNext Direct Genotyping Solution Target Enrichment Kit, the pool of 24 samples described above was hybridized to baits targeting 262 sample identification markers, which included several sex chromosome markers as well as the 24 SNPs identified by Pengelly et al[1] to have high discriminatory power across individuals. Following library prep and PCR amplification, the samples were sequenced on an Illumina Miseq® 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 sequence. After sequencing, the reads were demultiplexed with a Picard-based workflow[2]. Sequencing reads were aligned to the b38 (hs38DH) genome using BWA-MEM[3] and PCR duplicates were identified using the UMIs[4].

WORKFLOW



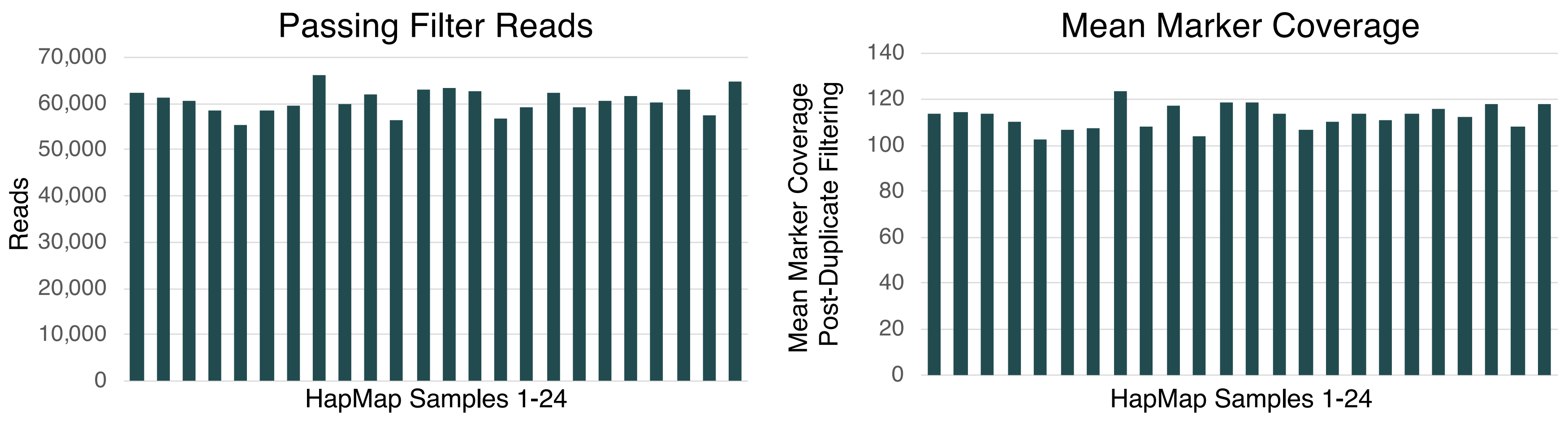
Samples are tagged with a barcode prior to pooling. Individual molecules are marked with a Unique Molecular Identifier (UMI)

Biotin baits target both strands (shown for 1 strand)

Enzymatic removal of off-target sequence

Individual samples are identified by the inline sample index and a pool index

PERFORMANCE ACROSS 24 SAMPLES



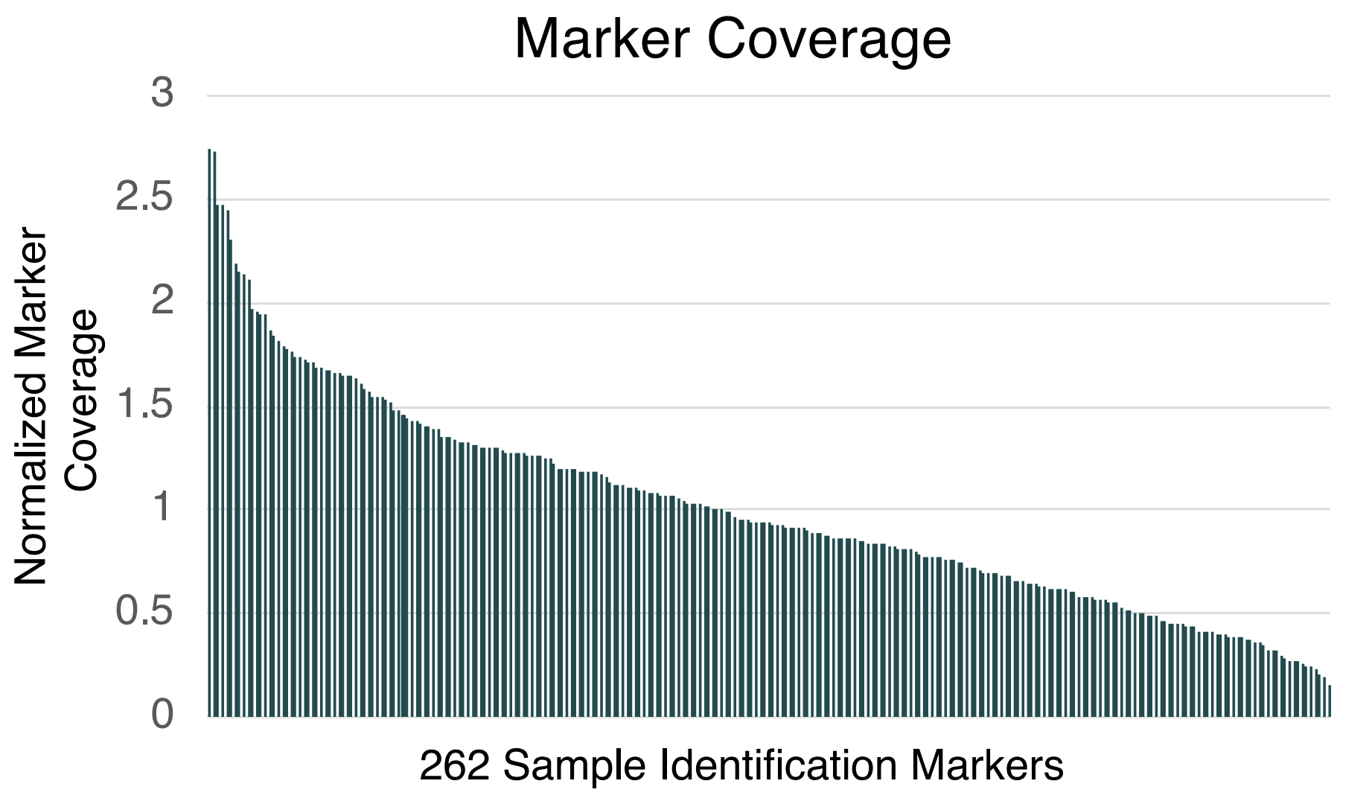
Uniform distribution of passing filter reads and mean marker coverage across the 24 HapMap DNA samples demonstrates comparable performance of each sample during hybridization-based capture and library preparation.

PANEL PERFORMANCE

Panel Metrics

# of Markers	262
Input Amount	25 ng
PF reads	50,964
%Aligned	99.3%
%Inserts On Markers	85.1%
Mean Marker Coverage	100
%Bases < 20% of Mean	2.0%

Values reported represent averages across the 24 pooled HapMap samples after downsampling the sequencing reads to a depth of 100X marker coverage post-duplicate filtering.

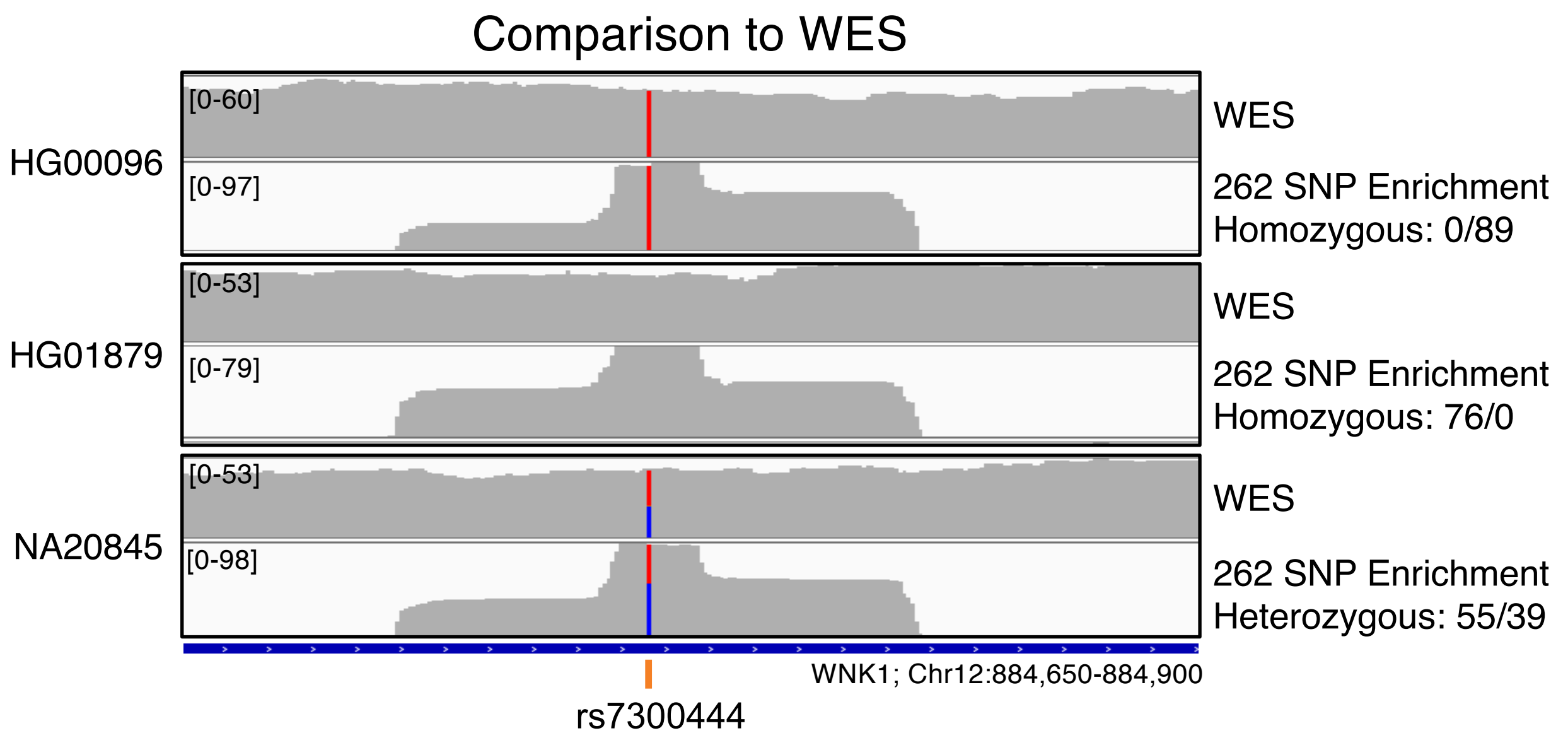


Normalized marker coverage from the HapMap sample HG01051. When sequenced to a depth of 100X, the lowest covered marker had 12X coverage.

SAMPLE IDENTIFICATION ACROSS 24 SAMPLES

	rs2229546	rs1410592	rs497692	rs10203363	rs2819561	rs4688963	rs309557	rs2942	rs17548783	rs4735258	rs1381532	rs10883099	rs4617548	rs7300444	rs9532292	rs2297995	rs4577050	rs2070203	rs1037256	rs9962023	rs2228611	rs10373	rs4148973	rs4675
HG00096	■	⬜	▨	⬜	▨	▨	▨	⬜	▨	⬜	▨	⬜	⬜	■	⬜	⬜	▨	▨	■	■	⬜	■	▨	▨
HG00268	■	■	▨	▨	■	▨	▨	■	🟥	▨	⬜	⬜	▨	▨	■	■	⬜	▨	▨	▨	⬜	⬜	⬜	▨
HG00419	⬜	▨	▨	⬜	▨	▨	⬜	▨	■	▨	▨	⬜	■	■	⬜	■	■	⬜	▨	▨	▨	⬜	⬜	⬜
HG00759	⬜	⬜	▨	▨	▨	▨	⬜	▨	▨	■	■	■	■	■	⬜	■	▨	⬜	■	■	■	⬜	▨	⬜
HG01051	■	▨	▨	▨	⬜	▨	▨	■	▨	▨	▨	▨	■	■	▨	▨	■	▨	▨	■	■	🟥	▨	▨
HG01112	■	▨	⬜	■	■	▨	⬜	▨	▨	▨	■	■	▨	▨	■	▨	▨	▨	▨	▨	⬜	▨	▨	▨
HG01500	⬜	■	⬜	▨	■	⬜	■	■	■	▨	⬜	⬜	▨	▨	⬜	■	▨	■	▨	■	▨	▨	▨	⬜
HG01565	■	▨	■	⬜	■	▨	▨	▨	■	▨	▨	■	■	▨	⬜	▨	▨	▨	▨	▨	■	⬜	⬜	▨
HG01583	■	▨	■	▨	■	⬜	⬜	■	▨	⬜	▨	■	⬜	▨	▨	■	■	▨	▨	■	▨	■	▨	⬜
HG01595	⬜	⬜	■	■	▨	▨	⬜	▨	▨	■	▨	▨	⬜	▨	⬜	▨	▨	■	⬜	▨	⬜	▨	⬜	■
HG01879	▨	▨	▨	▨	■	▨	■	■	▨	⬜	■	■	■	⬜	■	■	▨	▨	■	⬜	⬜	⬜	⬜	■
HG02568	⬜	▨	▨	▨	■	▨	▨	⬜	▨	⬜	▨	▨	⬜	⬜	■	▨	⬜	▨	⬜	■	⬜	▨	■	▨
HG02922	▨	■	⬜	■	■	▨	▨	■	▨	⬜	▨	■	■	⬜	⬜	■	■	⬜	▨	▨	⬜	⬜	▨	⬜
HG03006	▨	■	■	▨	▨	⬜	▨	■	■	▨	■	▨	▨	▨	▨	▨	■	▨	⬜	■	⬜	▨	■	■
HG03052	▨	■	⬜	▨	■	▨	⬜	▨	▨	▨	▨	⬜	⬜	⬜	▨	■	■	⬜	▨	▨	⬜	▨	▨	▨
HG03642	▨	■	⬜	■	■	■	⬜	⬜	▨	▨	▨	■	⬜	⬜	⬜	■	▨	⬜	▨	■	⬜	■	▨	■
HG03742	■	⬜	■	▨	■	■	■	▨	▨	▨	⬜	▨	▨	▨	▨	■	■	⬜	■	⬜	⬜	▨	⬜	⬜
NA18525	⬜	▨	▨	⬜	▨	■	■	■	■	■	■	■	⬜	▨	⬜	▨	▨	⬜	▨	■	■	⬜	▨	▨
NA18939	⬜	▨	▨	▨	■	■	▨	■	⬜	■	▨	■	⬜	▨	⬜	▨	▨	▨	■	⬜	▨	🟥	⬜	■
NA19017	■	⬜	⬜	⬜	▨	⬜	■	⬜	■	■	▨	■	▨	⬜	▨	⬜	■	▨	■	■	■	▨	■	■
NA19625	■	⬜	⬜	▨	■	⬜	▨	⬜	■	■	▨	⬜	⬜	■	⬜	▨	▨	▨	■	▨	▨	⬜	⬜	⬜
NA19648	■	■	■	■	⬜	⬜	■	■	▨	■	■	■	⬜	■	▨	⬜	■	▨	■	■	⬜	▨	■	⬜
NA20502	▨	⬜	▨	⬜	▨	■	⬜	▨	⬜	▨	⬜	▨	▨	▨	▨	▨	■	▨	■	■	▨	■	▨	▨
NA20845	▨	■	■	▨	■	▨	▨	▨	⬜	▨	⬜	⬜	▨	▨	▨	■	▨	▨	⬜	■	■	■	▨	▨

Sequencing data from the 24 HapMap samples assayed with the 262 SNP sample identification panel was downsampled to 100X marker coverage post-duplicate filtering (average of 50,964 PF reads per sample). Germline variants were called using the HaplotypeCaller from GATK[5]. Variant calls for the 24 discriminatory SNPs identified by Pengelly et al[1] are shown above for each of the 24 HapMap samples. White indicates homozygous for the reference allele, light blue indicates heterozygous for the reference/alternate alleles, dark blue indicates homozygous for the alternate allele, and red are genotypes that were incorrectly called by the HaplotypeCaller.



Coverage of the SNP rs7300444 from whole exome sequencing (WES) or the 262 SNP sample identification panel as visualized in the Integrative Genome Browser (IGV)[6]. Three HapMap samples with different genotypes are shown.

References:
[1] Pengelly, RJ et al (2013) A SNP profiling panel for sample tracking in whole-exome sequencing studies. Genome Medicine. 5(9):89
[2] http://broadinstitute.github.io/picard
[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
[5] McKenna A et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297-303, and https://software.broadinstitute.org/gatk/best-practices
[6] 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

NEBNextDirect.com

For Research Use Only