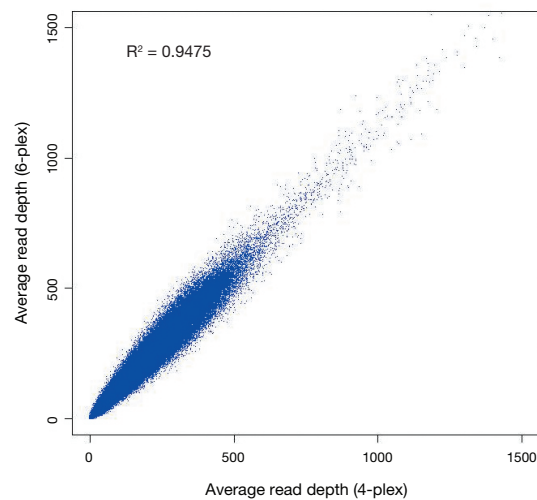


Figure 4: High Reproducibility



The same samples used in Figure 3 were analyzed for reproducibility. Results show a high level of concordance across replicates.

Highest Efficiency Protocol

For targeted resequencing, high enrichment efficiency and coverage uniformity ensure that all targeted regions are sequenced and minimize the required sequencing depth to accurately determine variants without bias. The TruSeq Exome Enrichment kit has been designed and optimized to deliver high enrichment rates and on-target specificity, while ensuring the highest coverage uniformity and reproducibility (Figures 4–6). Greater than 65% of reads that pass filter and map to

the reference genome will align to the targeted region, and > 75% will align within 150 bases of the targeted region. The kit is optimized for slightly larger libraries in order to effectively capture variance across library sizes. This not only increases the uniformity of coverage for smaller exons (< 150 bp), but also across long coding exons, UTRs, and non-coding RNA targets.

With the high-throughput processing power of Illumina sequencing systems, multiple exomes can be sequenced in a single run, reducing cost and minimizing hands-on time (Table 3)

Data Assessment

Sequence data generated from exome enrichment samples are analyzed using a script to generate two sets of statistics: post-alignment and post-CASAVA (Consensus Assessment of Sequence and Variation) analysis. Post-alignment analysis counts the number of reads that overlap any targeted region and defines whether a read falls within a target. Post-CASAVA analysis calculates the coverage at each base within a region. Data can be visualized to examine the on-target and off-target coverage in a sample using GenomeStudio® Data Analysis Software.

Enhanced Quality Controls

During the sample preparation process, artificial double-stranded DNA targets are incorporated into each of the three enzymatic steps: end repair, A-tailing, and ligation. To enrich for these sample preparation controls, there is a set of probes in the CTO (capture target oligos) pool that will specifically capture them. The control reagents can be used for a variety of library insert sizes ranging from 150–850 bp.

Control sequences appear in the final sequencing data as an indication that each of the enzymatic steps was successful. The built-in quality controls significantly assist in troubleshooting and are useful for identification of specific failure modes. Software for internal

Table 2: Databases Covered by the TruSeq Exome Enrichment Kit

Database	% Database Covered	Description	Web Address
CCDS coding exons (31.3 Mb; hg19)	97.2%	Core set of human protein coding regions that are consistently annotated and of high quality	http://www.ncbi.nlm.nih.gov/projects/CCDS/CcidsBrowse.cgi
RefSeq (regGene) coding exons (33.2 Mb; hg19)	96.4%	Known protein-coding genes taken from the NCBI RNA reference collection	http://www.ncbi.nlm.nih.gov/RefSeq/
RefSeq (regGene) exons plus (67.8 Mb; hg19)*	88.3%	Known protein-coding genes taken from the NCBI RNA reference collection along with non-coding DNA	http://www.ncbi.nlm.nih.gov/RefSeq/
Encode/Gencode coding exons (Encyclopedia of DNA Elements) (25.6 Mb; hg19)†	93.2%	Project to identify all functional elements in the human genome	http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=183763205&c=chr13&g=wgEncodeGencode
Predicted microRNA targets (9.0 Mb, hg19) ‡	77.6%	Includes predicted microRNA targets	http://www.microrna.org/microrna/getDownloads.do

* Includes coding exons, 5' UTR, 3' UTR, microRNA, and other non coding RNA.

† Manual V4

‡ mirbase 15 targets predicted by www.microrna.org.

