illumina®

TruSeq™ Exome Enrichment Kit

Offering pre-enrichment sample pooling and the most comprehensive exome coverage for cost-effective, scalable exome sequencing studies.

Highlights

- **Most Comprehensive Exome Coverage:** Highly uniform coverage across 62 Mb of exomic sequence, including 5' UTR, 3' UTR, microRNA, and other non-coding RNA.
- **Most Cost-Effective Exome Sequencing:** Streamlined protocol for pre-enrichment pooling of up to six samples dramatically reduces hands-on time and cost.
- **Integrated Solution:** Optimized for use with the TruSeq DNA Sample Preparation Kit, providing a gel-free protocol that requires the lowest DNA input.
- **Simplest and Most Scalable Workflow:** Automation-friendly with master-mixed reagents and plate-based processing for up to 96 reactions.

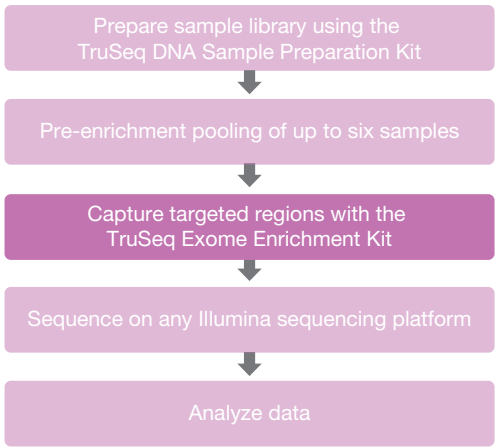
Introduction

Targeted resequencing allows researchers to concentrate their studies on a specific subset of the genome and more cost-effectively harness the power of next-generation sequencing (NGS) to discover disease causing variants. It is becoming evident that many rare complex disease causing variants lie within exonic, or coding regions, which comprise 1–2% of the genome[1]. Targeted exome sequencing enables researchers to take a closer look at these regions to discover causal variants for a range of complex human diseases.

Current approaches for targeting specific regions of the genome have severe limitations. PCR and Sanger sequencing are cost-prohibitive when working with many samples and regions. In addition, PCR demonstrates variable coverage between amplicons. Other targeted NGS solutions are available as well, but these tend to be time-consuming and offer less comprehensive content.

To overcome these challenges and make it easier and more affordable for researchers to take advantage of targeted exome sequencing, Illumina offers the TruSeq Exome Enrichment Kit, an in-solution sequence capture method for isolating exonic regions of interest in the human genome using hybrid selection. The TruSeq kit enables systematic detection of common and rare variants for high-throughput sequencing on any Illumina sequencer at a lower cost per sample. Designed and fully optimized for compatibility with the unique sample multiplexing capabilities of the Illumina TruSeq DNA Sample Preparation Kit, the TruSeq Exome Enrichment Kit provides a simple, scalable workflow as part of Illumina's integrated sample preparation sequencing solution (Figure 1).

Figure 1: Overview of Exome Sequence Capture as Part of Illumina Sequencing



The TruSeq Exome Enrichment Kit is an integral part of a fully supported, complete solution for targeted resequencing from Illumina.

Simple, Scalable Workflow

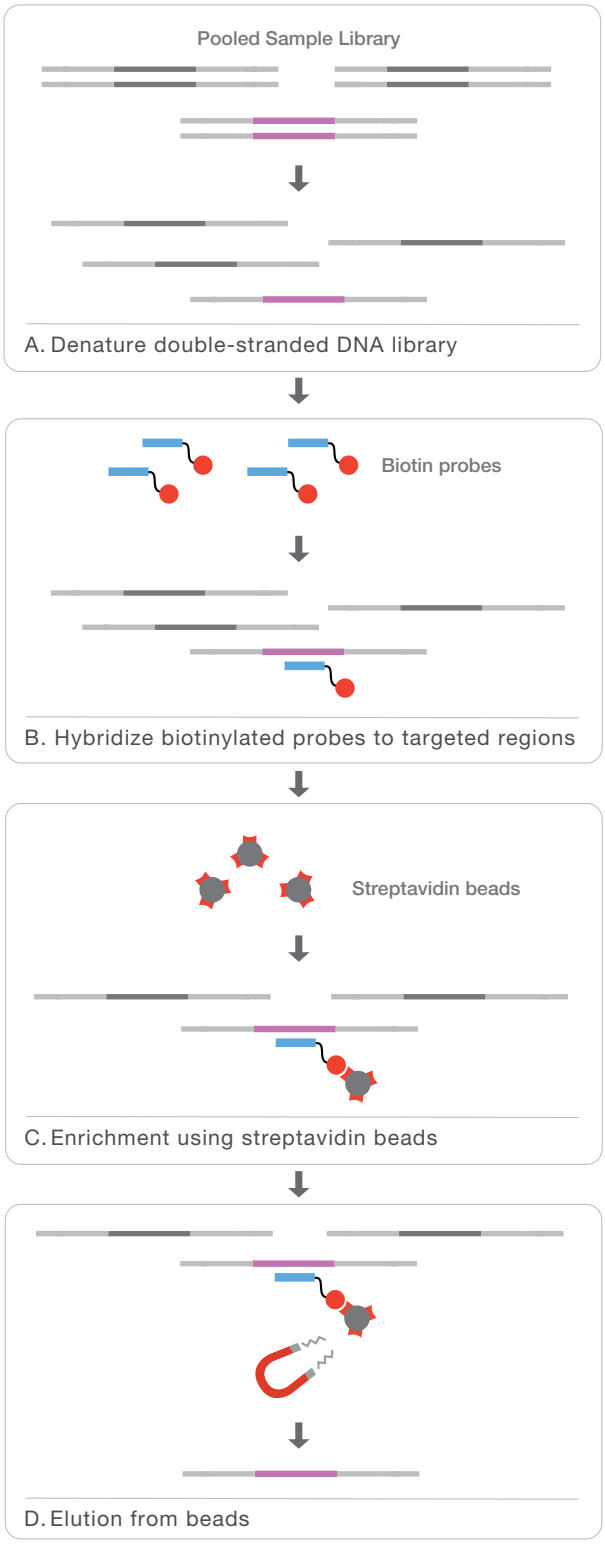
With a simple and scalable workflow, the exome enrichment method offers the most flexible and efficient solution for targeted resequencing. Master-mixed reagents are coupled with plate-based processing for up to 96 reactions and volumes are optimized for liquid handlers, making the process automation-friendly for even higher throughput.

Prior to exome enrichment, libraries are prepared using the TruSeq DNA Sample Preparation Kit, which provides robust 24-sample indexing of libraries. Multi-sample pooling of up to six samples in a single enrichment reaction dramatically reduces hands-on time compared to other available methods, making large high-throughput studies feasible and economical.

TruSeq DNA Sample Preparation Kits

Prior to exome enrichment, libraries are prepared using the TruSeq DNA Sample Preparation Kit. This kit provides master-mixed reagents, optimized index adapter design, a gel-free protocol, and a flexible workflow for preparing 24 multiplexed samples that can be pooled prior to sequencing, increasing the throughput and decreasing the cost of sequencing studies.

Figure 2: TruSeq Exome Enrichment Workflow



The TruSeq Exome Enrichment Kit provides a simple and streamlined method for isolating targeted regions of interest from samples prepared using the TruSeq DNA Sample Preparation Kit. There are two successive rounds of enrichment in the TruSeq Exome Enrichment workflow.

Exome Enrichment Workflow

Exome enrichment workflow begins with pooled indexed libraries of up to six samples prepared using the TruSeq DNA Sample Preparation Kit. These sample libraries are denatured into single-stranded DNA (Figure 2A) and then hybridized to biotin-labeled probes specific to the targeted region (Figure 2B). The pool is then enriched for the desired regions by adding streptavidin beads that will bind to the biotinylated probes (Figure 2C). Biotinylated DNA fragments bound to the streptavidin beads are magnetically pulled down from the solution (Figure 2D). The enriched DNA fragments are then eluted from the beads and hybridized for a second enrichment reaction. After amplification, a targeted library is ready for cluster generation and subsequent sequencing.

Superior Exome Coverage and More

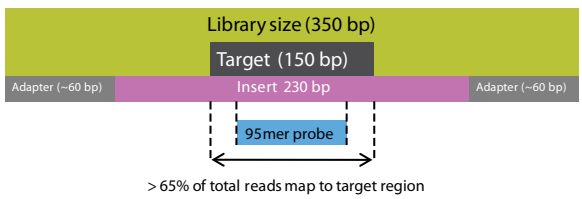
The TruSeq Exome Enrichment Kit features a highly optimized probe set that delivers comprehensive coverage of exomic sequence, starting from only 1 µg of DNA input (using TruSeq DNA Sample Preparation). The kit includes > 340,000 95mer probes, each constructed against the human NCBI37/hg19 reference genome. The probe set was designed to enrich > 200,000 exons, spanning 20,794 genes of interest (Table 1).

While the sum length of these probes is 32 Mb, the kit actually targets 62 Mb of the human genome (117.5 Mb if the 150 bp regions captured upstream and downstream of target are also considered). Each 95mer probe targets libraries of 300–400 bp (insert size of 180–280 bp), enriching 265–465 bases centered symmetrically around the midpoint of the probe (Figure 3)[2]. This means that, in addition to comprehensive coverage of the major exon databases (Table 2), the kit also provides broad coverage of non-coding DNA in exon-flanking regions (promoters and UTRs). Genome-wide association studies suggest that > 80% of disease-associated variants fall outside coding regions[3]. Analysis of these regions enables researchers to discover variants that effect gene function, at a more affordable price than whole-genome sequencing[4].

Table 1: Coverage Details

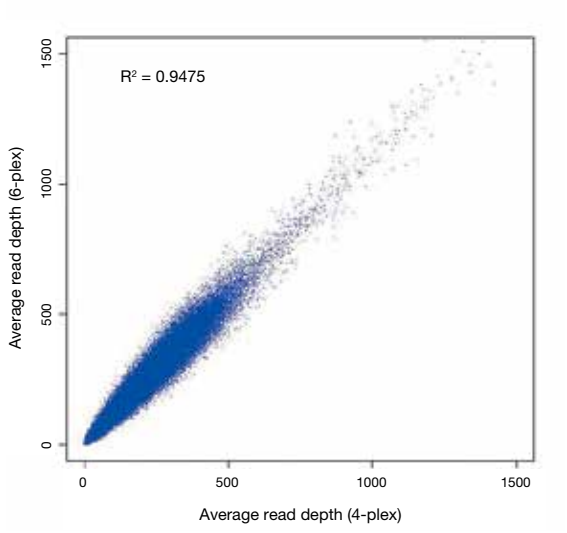
Target region size	62 Mb
Number of target genes	20,794
Number of target exons	201,121
Probe size	95-mer
Number of probes	340,427
Recommended library size	> 350 bp
Percent bases covered at 0.2x mean coverage	> 80%

Figure 3: Probe Footprint



With a 350 bp DNA library (mean insert size = 230 bp), the probe will enrich 365 bp (2 × Insert – Probe) centered around its midpoint.

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.

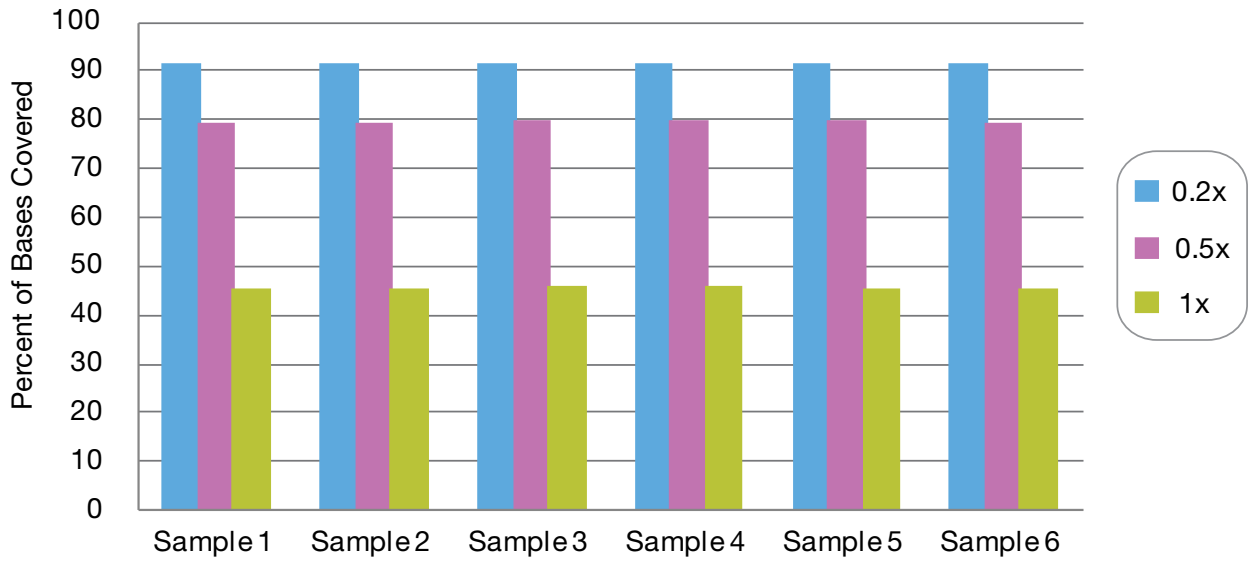
Control sequences appear in the final sequencing data as an indication that each of the enzymatic steps was successful. The built-in quality controls significantly assist in troubleshooting and are useful for identification of specific failure modes. Software for internal controls is supported by RTA [version 1.10 (HiSeq Systems) and version 1.9 (Genome Analyzer)] to recognize the sequences and to isolate the sequences from sample data.

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/CcdsBrowse.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

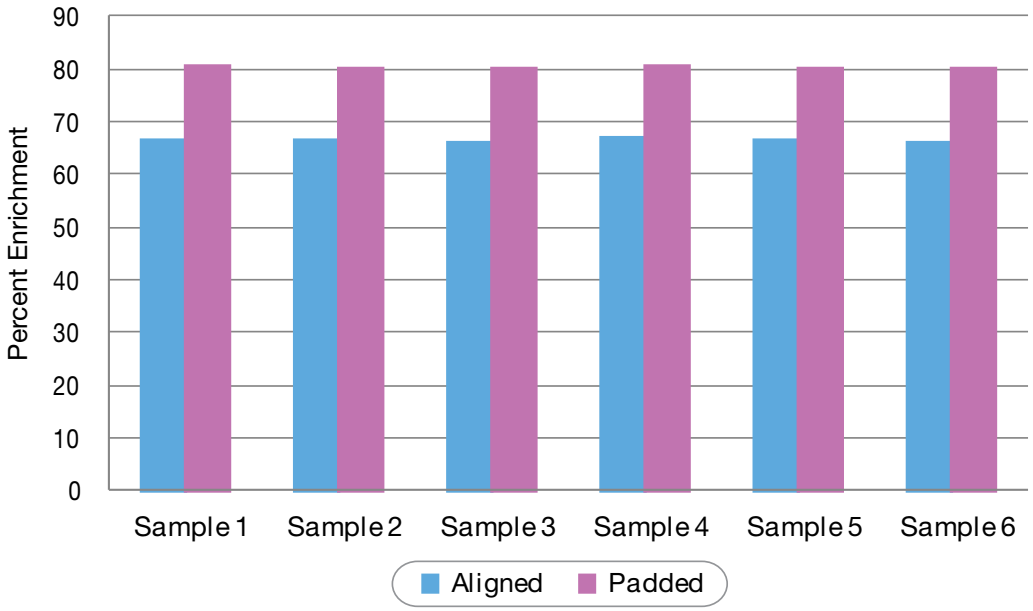
* Includes coding exons, 5' UTR, 3' UTR, microRNA, and other non coding RNA.
† Manual V4
‡ mirbase 15 targets predicted by www.microrna.org.

Figure 5: Highest Coverage Uniformity



Coverage uniformity is given for six samples with respect to the percentage of bases covered across the 62 Mb target region at varying mean normalized read depths. The six samples were prepared and pooled into a single reaction using the gel-free TruSeq DNA Sample Preparation Kit protocol, and then simultaneously enriched using the TruSeq Exome Enrichment Kit. The pooled samples were sequenced across five lanes of a HiSeq™ flow cell, generating mean read depths of 58 –72x (varying for each sample). Although there was some expected minor sample-to-sample variability, this data demonstrates that sample pooling does not effect coverage uniformity across the entire 62 Mb target region. Over 90% of bases were covered at 0.2x mean coverage, which means that at an average read depth of 64, over 90% of bases were covered at 12.8x depth (64 × 0.2x = 12.8x).

Figure 6: High Target Specificity



Percent Enrichment is defined as the number of reads mapping to the targeted regions out of the total reads produced in a sequencing run (on a per-sample basis). The six samples shown here were prepared and pooled into a single reaction using the gel-free TruSeq DNA Sample Preparation Kit protocol, and then simultaneously enriched using the TruSeq Exome Enrichment Kit. A 65% enrichment (blue bars) was achieved when considering only the reads that mapped to bases in the regions targeted by the capture probes. An increase to 80% enrichment (purple bars) is observed when the assessed reads are expanded to include those that map to regions +/- 150 bp flanking the probe-targeted region. This occurs because each probe is designed to capture more sequence than just the absolute targeted region.

Table 3: Exomes Sequenced Per Run

Instrument	Output per run	Approximate number of exomes sequenced per lane/run					
	(2 × 100 bp reads)	50× coverage per lane	50× coverage per run	75× coverage per lane	75× coverage per run	100× coverage per lane	100× coverage per run
HiScan™SQ	150 Gb	3.5	28	2.4	18	1.8	14
Genome Analyzer$_{IIx}$	64 Gb	1.5	12	1.0	8	0.8	6
HiSeq 1000	300 Gb	7.1	56	4.7	37	3.5	28
HiSeq 2000	600 Gb	7.1	112	4.7	74	3.5	56

The approximate number of exomes per lane/run, at varying coverage levels, is given for Illumina sequencers. The output required per exome takes into account the exome size, enrichment efficiency, and depth of coverage desired. Assuming ~65% enrichment, which includes target enrichment and reads passing filters, the total amount of available post-enrichment data can be determined. For example, 3.1 Gb of data is required for 50x coverage of a 62 Mb exome. On the HiSeq 2000 with version 3 reagents, this would allow for approximately 112 exomes per run, or about 7 exomes per lane (of the 8-lane flow cell).

Demo Kit Available

For researchers interested in trying TruSeq Exome Enrichment on their own samples, Illumina offers a TruSeq Exome Enrichment Demo Kit that includes sufficient materials to prepare 24 total samples (4 reactions/kit). For further details, contact your local sales representative or Illumina Customer Service (orders@illumina.com).

Summary

By harnessing the power of Illumina sequencing, the TruSeq Exome Enrichment Kit enables highly efficient resequencing studies that deliver comprehensive coverage of exomic regions across the human genome. An integrated workflow for pre-enrichment sample pooling, combined with the high-throughput power of Illumina sequencers, delivers the most streamlined, cost-effective method for analyzing multiple exomes in a single sequencing run.

Learn more, visit:
www.illumina.com/applications/sequencing/targeted_resequencing.ilmn.

Ordering Information

Kit	Reactions/ Kit	Samples/ Kit (6-plex)	Catalog No.
TruSeq Exome Enrichment Kit	8	48	FC-121-1008
	24	144	FC-121-1024
	48	288	FC-121-1048
	96	576	FC-121-1096
	192	1152	FC-121-1192
	480	2880	FC-121-1480
	960	5760	FC-121-1960
TruSeq DNA Sample Preparation Kit, Set A (12 indexes)		48	FC-121-2001
TruSeq DNA Sample Preparation Kit, Set B (12 indexes)		48	FC-121-2002

References

1. Shaheen R, Faqeih E, Sunker A, Morsy H, Al-Sheddi T, et al. (2011) Recessive Mutations in DOCK6, Encoding the Guanidine Nucleotide Exchange Factor DOCK6, Lead to Abnormal Actin Cytoskeleton Organization and Adams-Oliver Syndrome. The American Journal of Human Genetics 89: 328-333.
2. Optimizing Coverage for Targeted Resequencing Technical Note. www.illumina.com/documents/products/technotes/technote_optimizing_coverage_for_targeted_resequencing.pdf
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex disease. Nature 4618: 747-753.
4. Bainbridge, MN, Wang M, Wu YQ, Newsham I, Muzny DM, et al. (2011) Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biology 12(7):R68.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

illumina®