TECH NOTE

Sensitive Capture of Full-Length Transcript Information with Targeted RNA-Seq

SMARTer Target RNA Capture for Illumina

Target-specific and sensitive RNA capture >>

Detection of rare gene fusion events at low sequencing depth >>

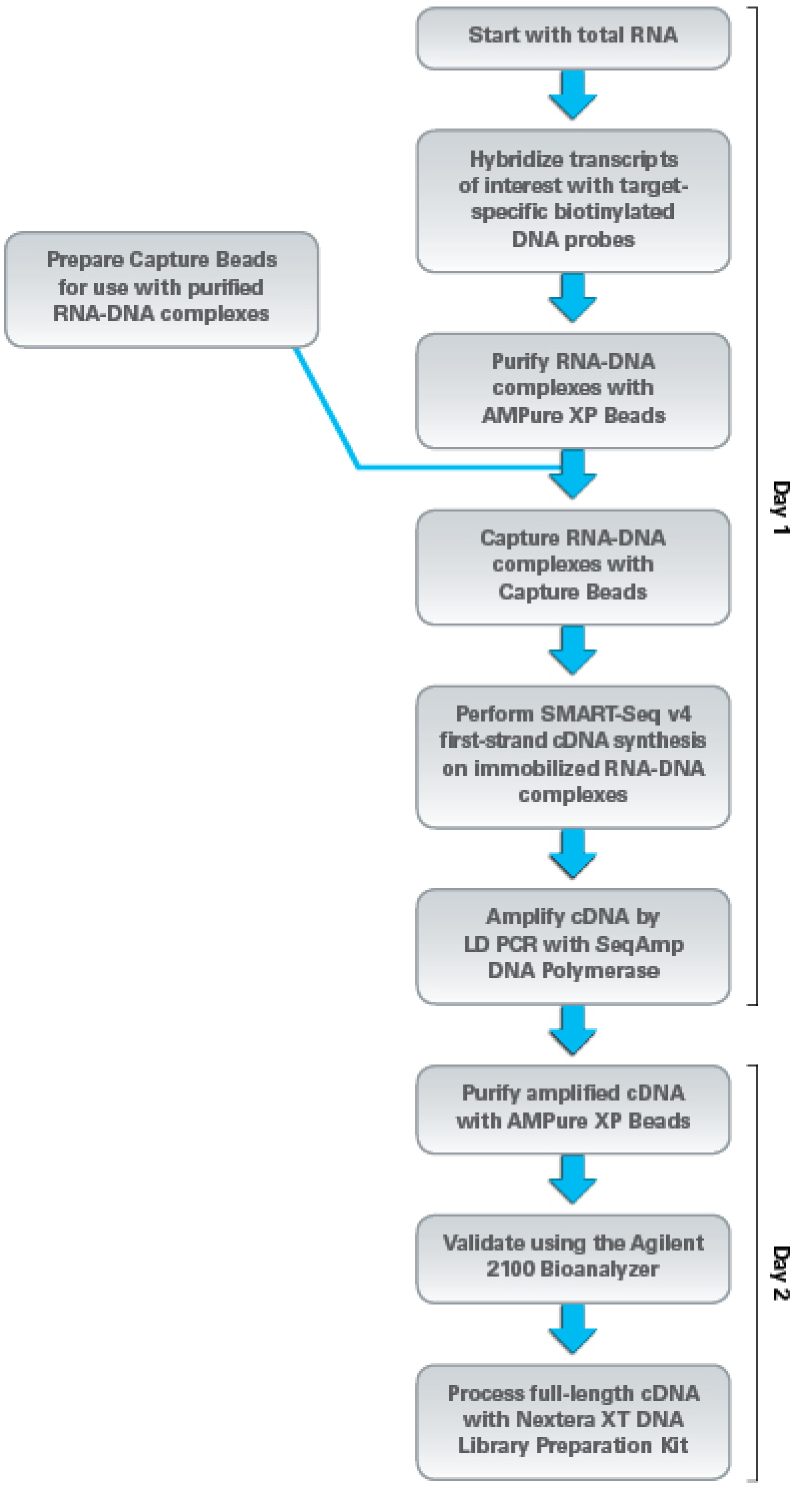
Consistent enrichment and coverage across input levels >>

Enriched targets maintain relative expression levels >>

Introduction

While recent advancements in next-generation sequencing technology have greatly improved research in whole transcriptome RNA-seq, several challenges still remain, stemming from the complexity inherent in such large-scale sequencing. The large dynamic range of the transcriptome often means that a few highly abundant transcripts account for the majority of sequencing reads while only representing a small percentage of RNAs. In contrast, less-abundant transcripts (representing a majority of RNAs) account for only a small percentage of sequencing reads (Jiang et al., 2011). Targeted RNA-seq aims to overcome this problem by improving sequence coverage of transcripts of interest that may be present in low amounts, thus saving costs and simplifying analysis. This enrichment enables the capture of information about transcripts that would otherwise be missed or would require a much greater number of sequencing reads to be detected, including chimeric gene fusions, transcript isoforms, and splice variants (reviewed in Byron et al., 2016).

SMARTer Target RNA Capture for Illumina provides a streamlined, sensitive method for enriching for transcripts of interest from total RNA prior to first-strand cDNA synthesis. Target-specific probes are hybridized to transcripts of interest, the resulting RNA-DNA hybrids are captured with streptavidin-coated Capture Beads, and SMART-Seq v4 technology is used for first-strand cDNA synthesis. The Capture Beads have been specifically selected for their low adsorption of protein and nucleic acids, and lack of interference with downstream reactions. SMART-Seq v4 cDNA synthesis enables high sensitivity, low background, and ensures the production of full-length cDNAs.

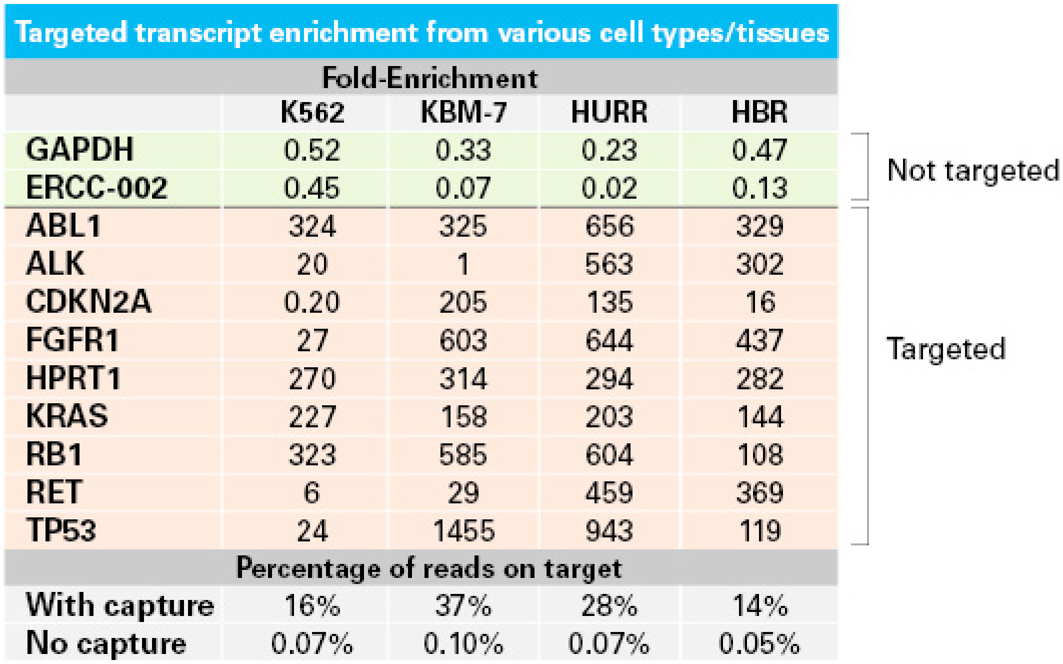

Start with total RNA
Hybridize transcripts of interest with target-specific biotinylated DNA probes
Prepare Capture Beads for use with purified RNA-DNA complexes
Purify RNA-DNA complexes with AMPure XP Beads
Capture RNA-DNA complexes with Capture Beads
Perform SMART-Seq v4 first-strand cDNA synthesis on immobilized RNA-DNA complexes
Amplify cDNA by LD PCR with SeqAmp DNA Polymerase
Purify amplified cDNA with AMPure XP Beads
Validate using the Agilent 2100 Bioanalyzer
Process full-length cDNA with Nextera XT DNA Library Preparation Kit
Day 1
Day 2

Protocol overview. The protocol through validation is completed over two days, with just 2.5–3 hours of hands-on time and without the need for additional rRNA removal methods or kits.

Results

Target specific and sensitive RNA capture

In order to assess the ability to enrich for targeted genes, SMART-Seq v4 cDNA synthesis was performed on samples both with and without first carrying out the capture part of the protocol. In this manner, nine genes were targeted in four RNA samples, using 100 ng each of starting material. In general, targeted genes from each sample were enriched, often at the expense of highly expressed transcripts (e.g., GAPDH, ERCC-002) that were not targeted. By capturing genes of interest, 14–37% of sequencing reads were on-target compared with less than 0.1% with a no-capture protocol. This increase in the percentage of reads on target reduces the number of sequencing reads and overall costs required to obtain desired sequencing data, as compared to whole-transcriptome RNA-seq. The targeted enrichment also showed significantly better results for on-target reads when capture was part of the protocol. We have observed that genes with very low expression in certain RNA samples can result in fewer on-target reads for those transcripts being identified in those samples (e.g., ALK in RNA from KBM-7 cells and CDKN2A in RNA from K562 cells).



Targeted transcript enrichment from various cell types/tissues

	Fold-Enrichment				
	K562	KBM-7	HURR	HBR	
GAPDH	0.52	0.33	0.23	0.47	Not targeted
ERCC-002	0.45	0.07	0.02	0.13	
ABL1	324	325	656	329	Targeted
ALK	20	1	563	302	
CDKN2A	0.20	205	135	16	
FGFR1	27	603	644	437	
HPRT1	270	314	294	282	
KRAS	227	158	203	144	
RB1	323	585	604	108	
RET	6	29	459	369	
TP53	24	1455	943	119	
	Percentage of reads on target				
With capture	16%	37%	28%	14%	
No capture	0.07%	0.10%	0.07%	0.05%	

Fold enrichment of targeted transcripts from various cell types and tissues. cDNA synthesis was performed from four different RNA samples, with and without transcript capture, but with the same input and number of PCR cycles. Two non-targeted but highly expressed transcripts were included in the measurements: GAPDH and ERCC-002. GAPDH is a highly expressed housekeeping gene and ERCC-002 is present at many copies in the RNA (~18 million copies in ERCC Mix 1 and ~36 million copies in ERCC Mix 2). Fold enrichment was calculated as the total number of sequencing reads that mapped to the targeted region after capture, divided by the number of sequencing reads that mapped to the targeted region without capture. The percentage of on-target reads was calculated as the total number of sequencing reads that mapped to the targeted genes, divided by the number of uniquely mapped reads.

Identification of rare gene fusions at low sequencing depth

The highly sensitive targeted-enrichment protocol enables the detection of rare gene fusion events at low sequencing depth while targeting for only one of the gene fusion partners. For example, K562 cells are derived from chronic myelogenous leukemia and harbor a BCR-ABL1 fusion in about 30% of ABL1 transcripts (data not shown). In a sequencing library made from human brain RNA spiked with K562 RNA at 1%, it is expected that the BCR-ABL fusion would be present in less than 1 in 200 reads that map to the fusion locus. With the power of SMART technology for amplifying full-length transcript information and by using probes targeting only ABL1, this fusion was evident using a modest sequencing library of fewer than 2 million total mapped reads (data not shown).

Consistent enrichment and coverage across input levels

SMARTer Target RNA Capture for Illumina has been optimized across a range of input levels (10 ng–1 µg), as shown below. Human Brain Reference RNA (HBRR; Thermo Fisher Scientific) or Human Universal Reference RNA (HURR;

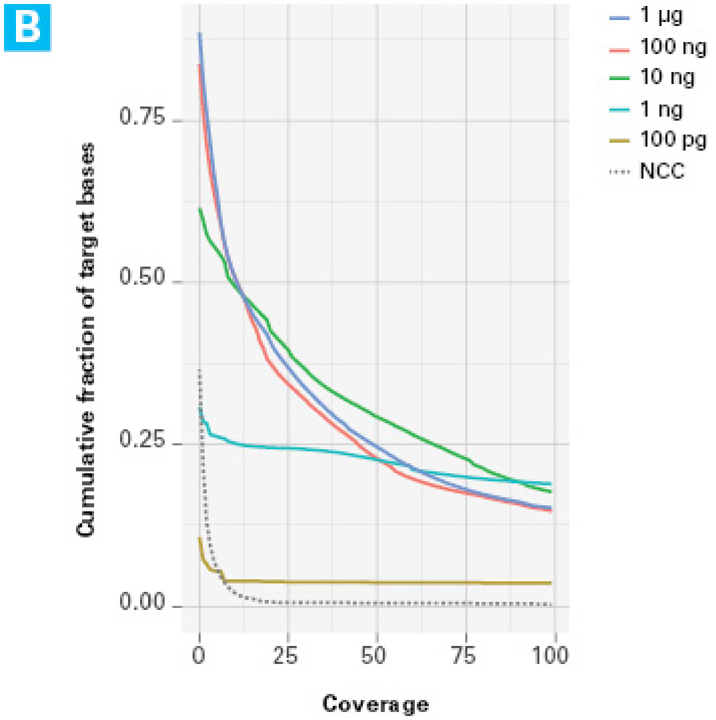
Agilent) total RNA ranging from 100 pg to 1 µg was used as input for this kit, with 1 pmol of probes against 60 genes. The kit displayed strong capabilities in the percent of on-target reads, fold-enrichment, and the number of targeted genes identified.

The on-target percentage for HBRR and HURR samples was ~25% and ~45%, respectively. This variation is due to inherent differences between the two RNAs, with many reads mapping against mitochondrial sequences in HBRR, but not in HURR. The higher percentage of mitochondrial HBRR sequences is also observed in the no-capture control (NCC), indicating that it is a characteristic of HBRR in general. The number of target genes identified (out of 60) represents the number of targeted genes that had a read greater than one.

Enrichment led to a 90-fold increase over the no-capture control and was fairly similar for the two RNA types. 80% of the target genes were captured independent of sequencing depth. While there were less than one million reads for each sample, the data do not indicate a compromise in sensitivity, as a higher sequencing depth would not necessarily result in the identification of additional transcripts. For example, the 10-ng HBRR and HURR samples were run on a NextSeq with 150 x 150 paired-end reads, with ~100 million reads per sample (data not shown).The same libraries created from the 10-ng samples were also run on a MiSeq instrument, where the increase in sequencing depth improved the fraction of bases covered at a given depth, as expected, but only increased the number of target genes identified by one (HBRR) or two (HURR) transcripts. It is important to note that attaining 100% of on-target reads is nearly impossible regardless of sequencing depth, because the target region was set as the union of all possible transcript isoforms and not all isoforms are expressed for this RNA. Additionally, we expect that identification of some transcripts may be stochastic due to low expression level, low total RNA input, and/or insufficient sequencing depth.

A

Input	On-target reads	Fold-change	Target genes identified
HBRR			
100 pg	17%	47	32
1 ng	31%	91	33
10 ng	28%	89	51
100 ng	22%	67	58
1 µg	24%	78	58
HURR			
100 pg	43%	91	29
1 ng	46%	107	46
10 ng	41%	93	55
100 ng	51%	120	60
500 ng	1%	33	59

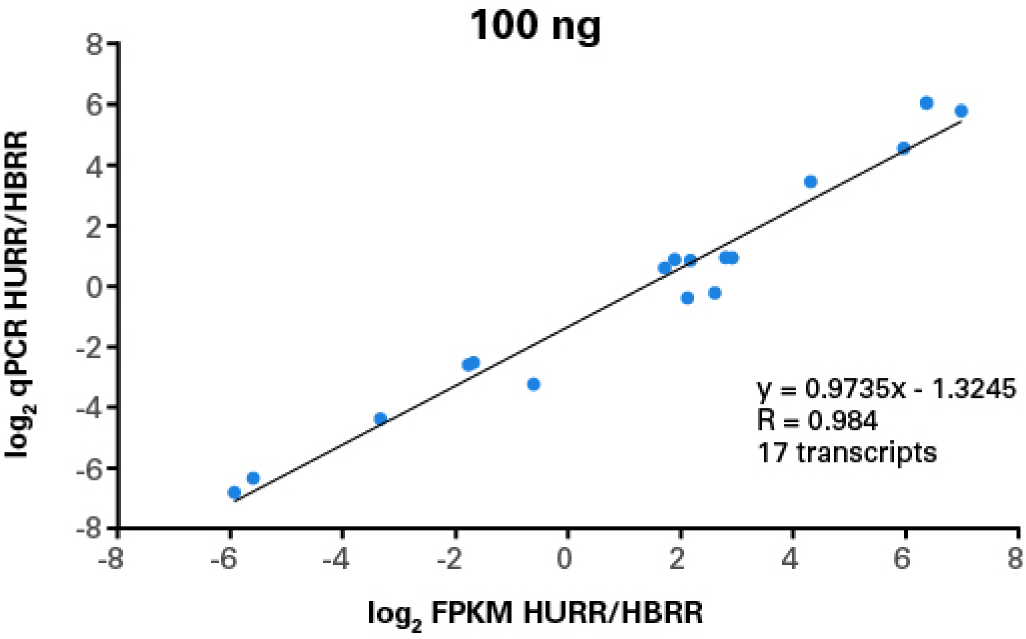


Capture across input levels for HBRR and HURR total RNA with ~0.9 M reads per sample. Panel A. The percentage of on-target reads was calculated as the total number of sequencing reads against the targeted genes, divided by the number of uniquely mapped reads as calculated by STAR. Fold-enrichment was calculated as the total number of sequencing reads against the target obtained after capture, divided by the number of sequencing reads obtained without capture. The 500-ng input for HURR displayed difficulties due to contaminants present in the RNA buffer that inhibited the hybridization reaction. **Panel B. On-target cumulative coverage plot for a titration of HBRR samples with ~1 M reads per sample.** The lines show the cumulative fraction of the target bases along the y-axis with a read coverage of at least x. With lower input, the probability of capturing any specific gene decreases. The total number of target genes is 60, comprising 1,200 exons and a target region of ~0.46 Mb.

Enriched targets maintain relative expression levels

Although the main goal of targeted transcript capture is to allow identification of transcripts with greater sensitivity, there has been some question as to whether it can maintain relative expression levels in a single sample or maintain differential expression between two samples. The Microarray Quality Control (MAQC) studies used TaqMan qPCR data to evaluate differential expression of ~1,000 genes between HURR and HBRR on microarray platforms. To evaluate this kit, the NGS expression data (expressed as FPKM) from the experiment above was compared against the TaqMan data. 17 of the 60 genes in the probe set also appear in the MAQC data.

There was a strong relationship between the orthogonal technologies, suggesting that the targeted capture method is maintaining similar relationships for differential expression ratios for these two RNA samples. Too few MAQC genes were captured for the 1-ng and 100-pg inputs so analysis was not possible.



MAQC analysis of targeted capture data. The y-axis shows the $\log_2$ of the ratio of the TaqMan qPCR Ct values and the x-axis shows the $\log_2$ of the ratio of FPKM values obtained with SMARTer Target RNA Capture for Illumina. Similar relationships for differential expression ratios were seen with 10 ng or higher (up to 1 µg) of input RNA.

Conclusions

SMARTer Target RNA Capture for Illumina is a targeted RNA-seq kit that leverages the sensitivity of SMART-Seq v4 technology to generate high-quality, full-length cDNA from enriched transcripts. The high level of sensitivity enables the detection of rare structural events such as gene fusions that might otherwise be missed in whole transcriptome RNA-seq, even at low sequencing depths. With lower sequencing depth required, analysis time and experimental costs are reduced. Performance is consistent across input levels and different RNA types, and relative expression levels are maintained post-enrichment.

Methods

Highly specific and sensitive RNA capture

Fold enrichment for nine genes from four different RNA samples was measured with and without targeted capture (but same RNA input and same number of PCR cycles). Each RNA sample input was 100 ng. ERCC Mix 1 was spiked into K562 RNA (BioChain Inc., Cat. No. R1255820-50) and HURR (Agilent, Cat. No. 740000), and ERCC Mix 2 was spiked into KBM-7 RNA (purified from cells) and Human Brain Total RNA (Cat. No. 636530). The genes of interest were targeted with 1 pmol of probe against 9 genes (~160 exons, ~55 kb total target region).

Sequencing libraries were run on an Illumina MiSeq (75 x 75 paired-end reads), and sequencing reads were aligned to the human reference GRCh38 release 23 (Gencode) with the TopHat or STAR aligner. The number of read counts were determined for each targeted gene after downsampling to the same number of reads (~2 M). The fold enrichment was calculated as the total number of sequencing reads against the target obtained after capture, divided by the number of sequencing reads obtained without capture. The percentage of on-target reads was calculated as the total number of sequencing reads against the targeted genes, divided by the number of uniquely mapped reads.

For the BCR-ABL1 fusion, K562 RNA was spiked into Human Brain Total RNA at 1% for 10 ng of combined RNA. 1 pmol probes against only ABL1 were used. Sequencing libraries were run on an Illumina MiSeq (75 x 75 paired-end reads), and sequencing reads were aligned to the human reference with the TopHat aligner and then downsampled to the same number of reads (~2 M). The number of reads to the fusion was determined with a custom script that directly queried reads for evidence of a translocation.

Consistent enrichment and coverage across input levels

HBRR (Thermo Fisher Scientific, Cat. No. AM6050) or HURR total RNA was used as input into the SMARTer Target RNA Capture for Illumina protocol, with amounts ranging from 100 pg to 1 µg. 1 pmol of probes against 60 genes were used (1,200 exons, ~0.46 Mb). Sequencing libraries were run on a MiSeq (75 x 75 paired-end reads), and sequencing reads were aligned with STAR to the GRCh38 release 23 with Gencode annotations. The number of read counts and FPKMs were determined for each targeted gene after downsampling to the same number of reads. The percentage of on-target reads was calculated as the total number of sequencing reads against the targeted genes, divided by the number of uniquely mapped reads as calculated by STAR. The fold-enrichment was calculated as the total number of sequencing reads against the target obtained after capture, divided by the number of sequencing reads obtained without capture. The target region was set as the union of all possible transcript isoforms and not all isoforms are expressed for this RNA. The duplicate rate ranged from 23–24% within the target region for the HBRR samples. Duplicate rates for given genes are related to the expression of those genes, was not correlated with input, and removal of the duplicates did not alter the results. For MAQC analysis, the $\log_2$ ratio of FPKMs from HURR/HBRR was plotted against the $\log_2$ of the ratio of HURR/HBRR derived from qPCR TaqMan probes.

References:

Byron, S. A., *et al.,* (2016) *Nat. Rev. Genet.* **17**(5):257–271.
Jiang, L., *et al.,* (2011) *Genome Res.* **21**(9):1543–1551.

Visit the product page >>

View web page >>

http://www.clontech.com/US/Products/cDNA_Synthesis_and_Library_Construction/NGS_Learning_Resources/Technical_Notes/Targeted_RNA-Seq