RNA-seq for biomarker identification using the xGen™ Broad-Range RNA Library Prep Kit and xGen Custom Hyb Panels

A capable workflow for transcriptome sequencing and targeted RNA-seq with FFPE samples

Abstract

Biomarker research provides potential targets that can indicate a normal biological process, a pathogenic process, or whether a drug is effective in its mode of action. An efficient workflow for biomarker research and discovery from formalin-fixed, paraffin-embedded (FFPE) samples is presented here. This workflow includes library preparation from isolated FFPE RNA using the xGen Broad-Range RNA Library Prep Kit, target enrichment using an xGen Custom Hyb Panel, followed by sequencing on an Illumina® platform. The workflow shows consistency across technical replicates and multiple mass inputs. The transcript-specific target enrichment capture panel design bypasses the need for rRNA depletion and produces deeper target coverage, thereby improving identification of low-expressed and rare fusion transcripts over whole transcriptome sequencing with FFPE RNA. Using highly characterized ALK-RET-ROS1 reference standards, this application note shows the ability to capture and call known gene fusions and presents a new hybridization capture panel design approach for identifying novel gene fusions to support oncology research at the biomarker discovery phase. By using a targeted RNA-seq approach, gene fusions were identified at higher rates than transcriptome libraries while using 10-fold less sequencing reads—drastically saving on sequencing costs.

Introduction

Since the discovery of RNA's role as an intermediary between genotype and phenotype, researchers have been investigating how changes in gene expression and DNA structural variations can lead to disease states. With the advancement of next generation sequencing (NGS) technologies, RNA-seq has been developed as a high-throughput technology to profile transcriptomes. RNA-seq allows for quantification of gene expression and identification of sequence alterations such as isoforms, alternative splicing events, and gene fusions across different tissues or conditions, which is valuable in understanding transcriptome dynamics during diseased and normal states [1].

One of the challenges with RNA-seq is the cost associated with the high coverage depth needed to identify low-expressed transcripts or rare gene fusions. By using a targeted sequencing approach that allows for deeper sequencing of areas of interest, reads from transcripts not relevant to a study's design are eliminated. In addition to the cost savings, the deeper sequencing with a targeted approach increases the ability to identify low-expressed and rare target RNAs, such as gene fusions.

Gene fusions are important biomarkers in oncology research as they are known to be cancer drivers. Structural rearrangements such as translocations, inversions and deletions can lead to the formation of gene fusions, while non-structural rearrangements such as transcript read-through of mRNA splicing can also lead to the formation of gene fusions [2]. Using target capture techniques provides the ability for deeper targeted sequencing and enables more sensitive identification of rare fusion events and novel fusion transcripts that would otherwise be difficult to identify by whole transcriptome sequencing [3].

FFPE tissue archives can be a valuable source of sample material for small—or large-scale RNA-seq studies—however, several factors can adversely affect FFPE sample quality and bring about challenges in creating high quality RNA-seq libraries and usable sequencing data. To assess RNA-seq library performance using FFPE samples, libraries were created from a highly characterized FFPE RNA sample containing droplet digital PCR (ddPCR)-confirmed EML4-ALK, CCDC6-RET and SLC34A2-ROS1 gene fusions, using two methods. The first workflow relied on ribosomal RNA depletion prior to generation of whole transcriptome libraries. The second workflow used hybridization capture of total RNA libraries to enrich for specific regions of the transcriptome using one of two xGen Custom Hybridization Capture Panels targeting a subset of non-small cell lung cancer (NSCLC)-related genes and gene fusions. Whole transcriptome and targeted sequencing results demonstrate a high-performing workflow with consistency between technical replicates and highly correlated gene expression information. While some gene fusions were not identified in the whole transcriptome libraries, all hybridization captured libraries allowed for the identification of all targeted gene fusions, at significantly higher rates than transcriptome libraries, and while using 10-fold less sequencing reads than transcriptome sequencing.

Targeted RNA-seq using the xGen Broad-Range RNA Library Prep Kit workflow

The xGen Broad-Range RNA Library Prep Kit can be used in combination with xGen Custom Hybridization Capture Panels to provide targeted RNA-seq data. This workflow can be used to identify known and novel fusions in a poor-quality sample such as degraded RNA from FFPE samples. The workflow is shown in Figure 1 and 2.

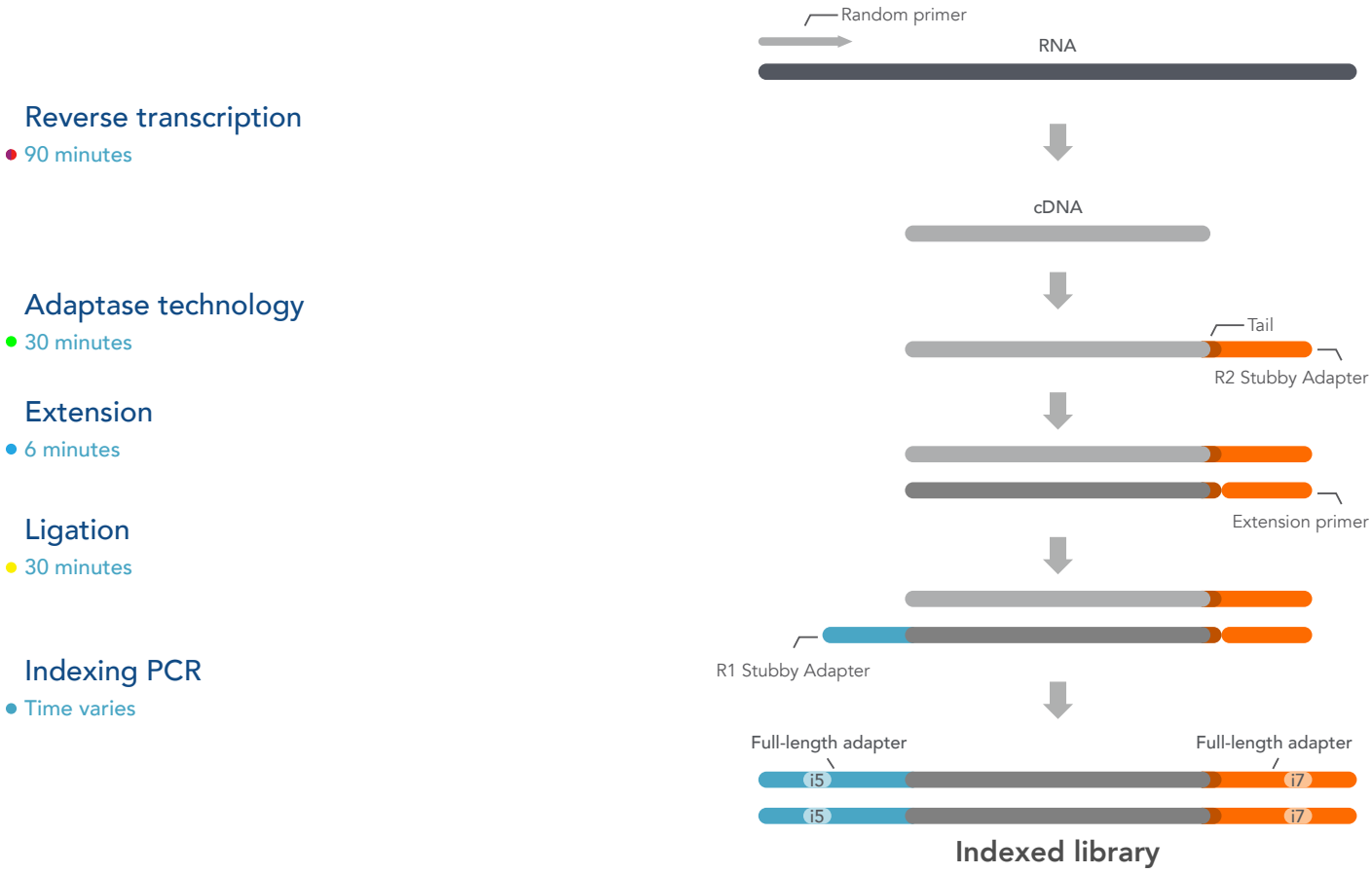

Figure 1. Overview of the xGen Broad-Range RNA Library Prep Kit workflow. After RNA fragmentation, the reverse transcriptase step uses random primers to generate the first-strand cDNA. Next, Adaptase technology simultaneously performs tailing and ligation to incorporate the R2 Stubby Adapter to the 3' ends of the cDNA molecules. The extension step produces a dsDNA duplex, while ligation adds the R1 Stubby Adapter to the 3' ends of the primer-extended cDNA molecules. Finally, indexing PCR increases library yield, incorporates single or dual indexes, and results in full-length adapters at the ends of each molecule. In addition, bead-cleanup steps are needed after extension, ligation, and final indexing PCR steps.

Add blocking oligos

Hybridize targets to capture probes

Incubate with magnetic streptavidin beads

Isolate targets with magnet

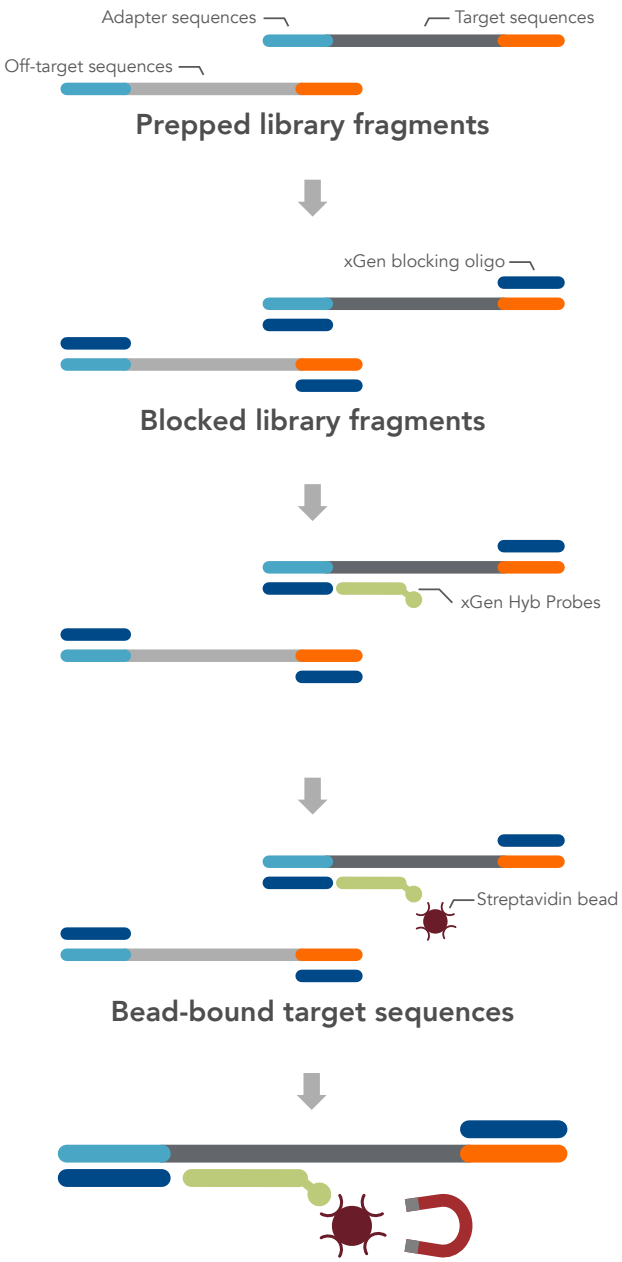

Figure 2. Overview of xGen Hybridization Capture. First, xGen Universal Blockers are mixed with prepared library fragments to prevent adapter-to-adapter hybridization. Blocked library fragments are then annealed to the 5′ biotinylated oligonucleotide probes from an xGen Predesigned Hyb Panel or an xGen Custom Hyb Panel. The probe and fragment duplexes are then separated from the unbound fragments by streptavidin-coated magnetic bead purification. The resulting library is highly enriched for fragments of interest.

Methods

Whole transcriptome libraries from FFPE samples

The ALK-RET-ROS1 targeted FFPE RNA Fusion Reference Standard is a highly characterized control material used to assess fusion identification for a variety of assays, including RNA-seq. FFPE RNA (Horizon HD784) was extracted using the Qiagen RNeasy™ FFPE Kit with deparaffinization solution according to the manufacturer's protocol. RIN scores of 3.0–3.3 and DV_{200} scores (>70%) were determined using an Agilent 2100 bioanalyzer. 10 ng or 50 ng of extracted total RNA was rRNA depleted in triplicate using the Lexogen RiboCop™ rRNA Depletion Kit HMR V2.

Transcriptome libraries were generated from all rRNA-depleted RNA (n = 6) following the **xGen Broad-Range RNA Library Prep Kit protocol** and using modifications outlined in Appendix C: Adjustments for xGen Hybridization Capture Panels. Libraries were sequenced using a NextSeq™ 550 (Illumina) using 2 x 150 paired-end reads and subsampled to 40 million reads per sample. For analysis, 10 bases were trimmed from the beginning of each read due to the low complexity tail added during the Adaptase step (for more information, see the technical note **Tail trimming for better data**).

Software packages

Alignment and filtering were performed using STAR (v2.6.1b), general target enrichment metrics used Picard (v2.18.9), transcript counts were calculated with Kallisto (v0.46.2) and STARFusion (v1.9.1) was used for fusion identification.

Targeted RNA libraries from FFPE samples

RNA was extracted as described above from ALK-RET-ROS1 FFPE RNA standard (Horizon HD784) and libraries were generated from 10 ng and 50 ng total RNA inputs following the **xGen Broad-Range RNA Library Prep Kit protocol** and using modifications outlined in Appendix C: Adjustments for xGen Hybridization Capture Panels.

For target enrichment, two fusion-identifying capture panel design approaches—a Known Fusion Design and a Discovery Fusion Design capture panel—were evaluated, one targeting known gene fusions, and one designed to discover unknown fusions in an implicated region of interest (**Figure 3**). The ALK-RET-ROS1 known fusion probes were designed to identify three digital PCR-confirmed fusions found within the RNA reference sample, CCDC6_RET, EML4_ALK, and SLC34A2_ROS1. The known fusion transcript probe design consists of three 120 nt hyb capture probes per gene fusion: the first probe is 120 bases upstream of the 5′ gene breakpoint, the second probe is 120 bases downstream of the 3′ gene breakpoint, and the final probe contains 60 bases flanking the known breakpoint for the 5′ and 3′ genes, with the breakpoint centered in the probe (**Figure 3A**). For the Discovery Fusion Design panel approach, one gene partner, *ROS1*, from the confirmed fusion SLC34A2_ROS1 was used as proof of concept. Here, the xGen Custom Hyb Capture panel contains a set of 24X-tiled probes with 5 nucleotide offsets designed to cover each of three exons within *ROS1* (**Figure 3B**). Though confirmed in our sample, the Discovery Fusion capture probe design does not rely on knowledge of the exact breakpoint, rather only that *ROS1* is implicated as the target gene of interest; therefore, the panel design is ultimately a strategy for gene fusion discovery.

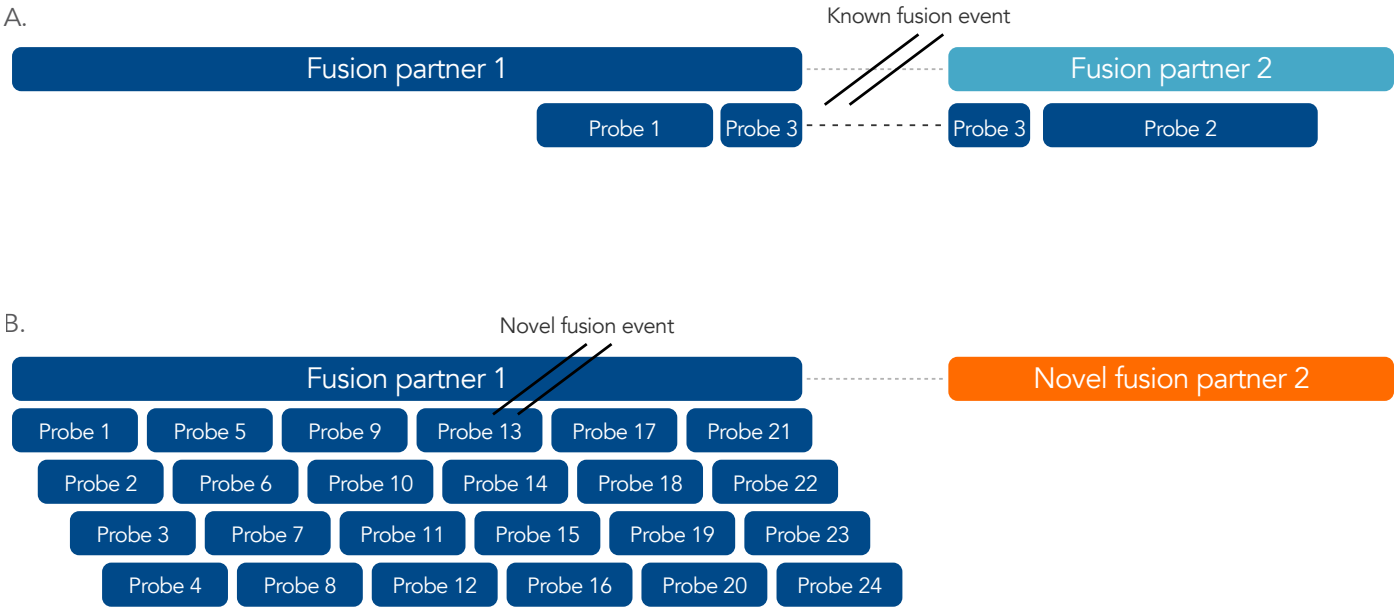

Figure 3. Schematic of known gene fusion vs. discovery gene fusion probe design strategies in the xGen Custom Hyb Panels. (**A**) The xGen Custom Hyb Capture panel based on known gene fusion breakpoints uses a design containing three probes per known gene fusion; one probe to each known partner, and a third probe spanning the junction point. (**B**) The xGen Custom Hyb Capture panel design for gene fusion discovery has 24X probe tiling of the exon(s) of interest to ensure that probes will capture the fusion event without any prior knowledge of the fusion junction.

For hybridization capture, total RNA libraries were individually hybridized overnight with one of the two xGen Custom Hyb Capture panels spiked into a backbone non-small cell lung cancer (NSCLC) Hyb Capture panel following the xGen Hyb Capture protocol. The Known Fusion Hyb Capture panel was tested with two total RNA inputs (10 ng and 50 ng), while the Discovery Fusion Hyb Capture was tested using 50 ng total RNA input libraries. Hyb captured libraries were sequenced on a NextSeq™ (Illumina) using 2 x 150 paired-end reads subsampling to 4 million reads per library. Analysis was performed using hg38 as the human reference genome. For analysis, 10 bases were trimmed from the beginning of each read due to the low complexity tail added during the Adaptase step (for more information, see the technical note Tail trimming for better data).

Results

Whole transcriptome libraries from FFPE samples

First, the quality and consistency of whole transcriptome RNA-seq libraries generated from FFPE RNA was assessed. The data demonstrates that rRNA-depleted FFPE transcriptome libraries have consistently high mapping efficiency and high percent-strandedness (Figure 4A and 4B). The number of transcripts identified between library inputs is also similar (Figure 4C). For whole transcriptome sequencing of degraded samples like FFPE, rRNA-depletion is commonly used in place of poly(A) enrichment as poly(A)-selection would only provide 3′ coverage of the fragmented transcripts. Without removal of the rRNA, up to 90% of the reads would likely come from ribosomal RNA, resulting in very low coverage of other genomic features. rRNA depletion removes the ribosomal RNA without skewing the remaining transcripts to only cover the 3′ ends. The sequencing results show only 2–4% of transcriptome sequencing reads aligned to rRNA after upstream rRNA depletion (Figure 5). Furthermore, ~20% of the bases sequenced were coding regions, while ~36% were intronic regions. The exonic rate for FFPE samples is low due to the bias toward intronic reads that is introduced during formalin fixation. The bias toward intronic reads can result in a need for deeper sequencing to achieve the desired coverage for mature transcript identification [4].

As shown in Figure 6, a very high Pearson's correlation for gene expression level was seen, demonstrating consistency among both technical replicates and between varying mass inputs when comparing FFPE transcriptome libraries generated from 10 ng and 50 ng total RNA input into ribo-depletion, despite the low RIN scores of 3.0–3.3 and DV_{200} scores (>70%) and low mass input quantities. All metrics show very little variability within and between inputs, indicating consistency in library preparation across technical replicate and sample input amounts.

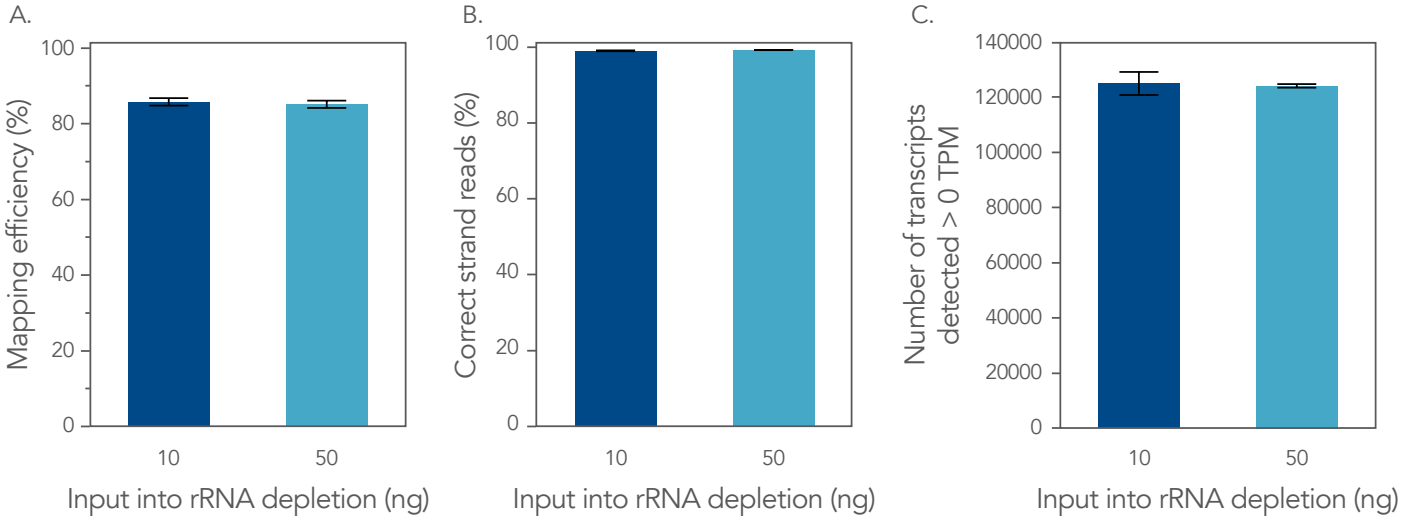

Figure 4. RNA-seq metrics for transcriptome libraries. (**A**) 10 ng and 50 ng of total FFPE RNA was rRNA-depleted and used to create xGen Broad-Range RNA libraries which show similar high mean mapping efficiencies, (**B**) High percent strandedness, (**C**) and number of transcripts identified. Sequencing was performed on a NextSeq™ (Illumina) using 2 x 150 paired-end reads and the data subsampled to 40 million reads/sample. Values in the charts represent the mean for three technical replicates per condition.

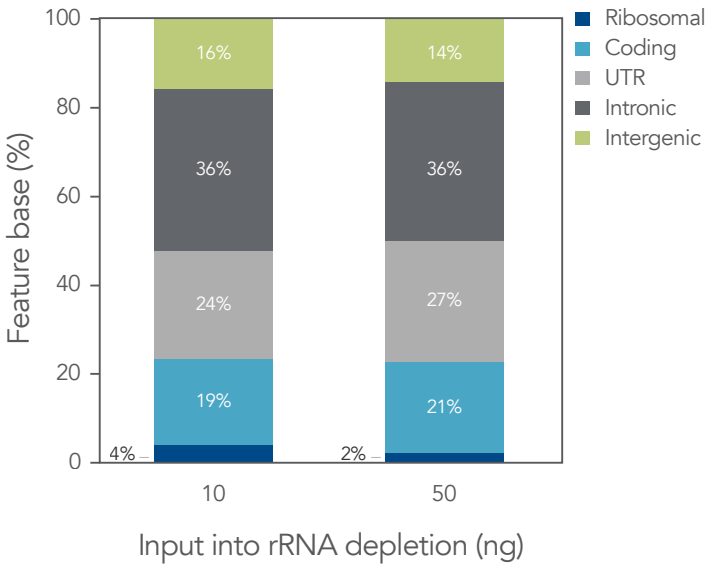

Figure 5. Featured bases chart for transcriptome libraries. 10 ng and 50 ng of total FFPE RNA was rRNA-depleted and used to create xGen Broad-Range RNA libraries. Transcriptome libraries show similar percentages ribosomal, coding, UTR, intronic and intergenic bases in final libraries across both input amounts. Sequencing was performed on a NextSeq™ (Illumina) using 2 x 150 paired-end reads and the data subsampled to 40 million reads/sample. Values in the chart represent the mean percentage of feature bases for three technical replicates per condition.

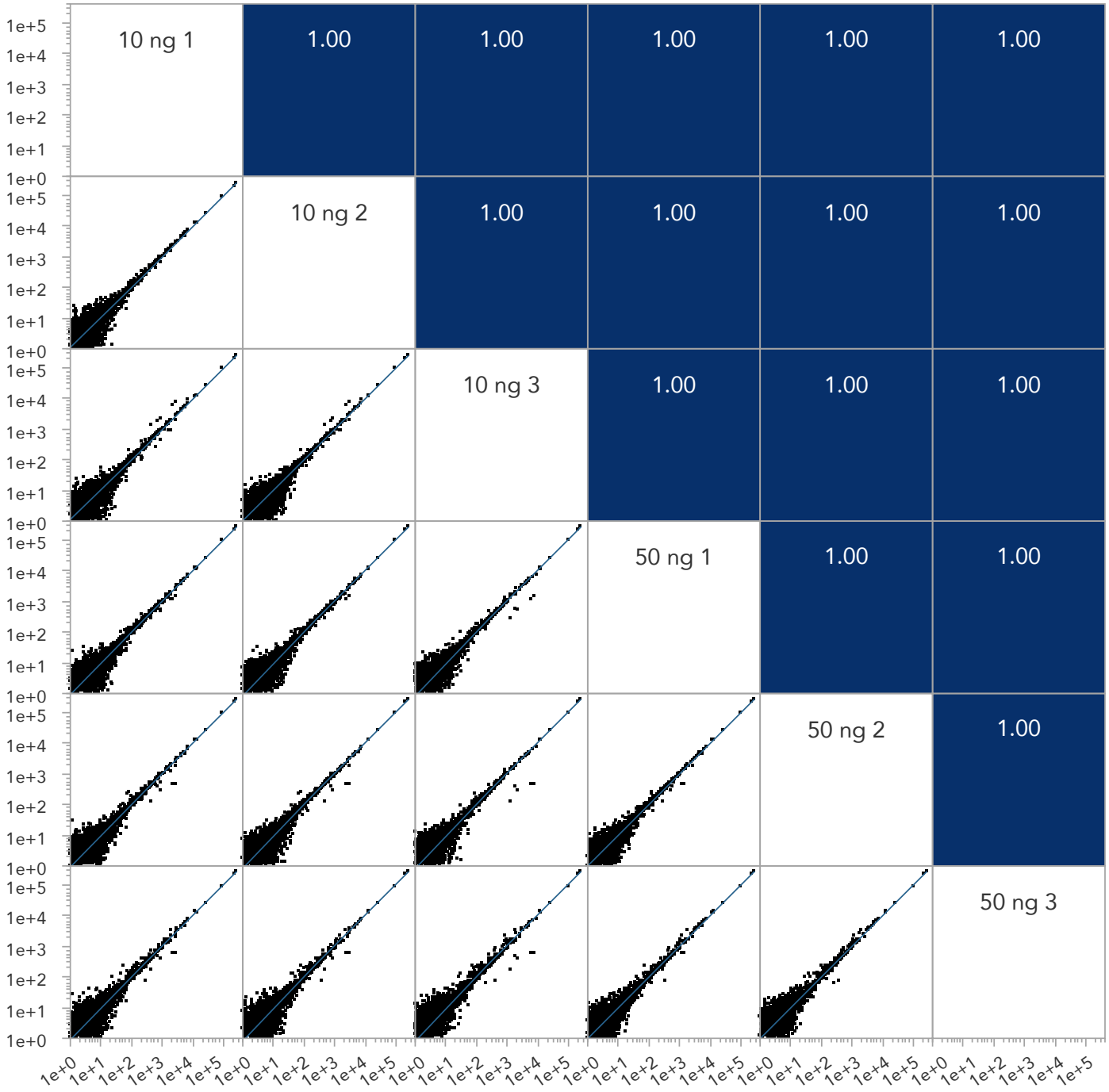

Figure 6. Expression correlation using transcripts per million (TPM) values for transcriptome libraries. Gene expression information for xGen Broad-Range RNA libraries generated from ribo-depleted FFPE RNA. 10 ng and 50 ng indicate the total RNA input into ribosomal depletion. A high Pearson's Correlation between technical replicates (r = 1.00) and high correlation between inputs (r = 1.00) is seen. Sequencing performed on a NextSeq™ (Illumina) using 2 x 150 paired-end reads and the data subsampled to 40 million reads/sample.

Targeted RNA libraries from FFPE samples

To assess the ability to identify known gene fusions, an xGen Custom Hyb Panel targeting the coding regions of five genes associated with non-small cell lung cancer (NSCLC) was used as a base capture panel to compare fusion-identifying probe designs. Since the hybridization capture reaction is target specific, it was not necessary to perform rRNA-depletion for these experiments.

Hybridization capture results

For hybridization capture, the two xGen Custom Hyb Capture panels (Known Fusion Design and Discovery Fusion Design) were separately spiked into the NSCLC Hyb Capture base panel. Results show mapping rates of over 85% (Figure 7A), greater than 71% on-target rate (Figure 7B), and ≥99.8% stranded reads (Figure 7C) across all replicates regardless of input or panel design. The consistency between technical replicates is noted by small standard deviations between the triplicate samples.

Regarding the necessity for rRNA depletion, greater than 66% of the sequenced bases aligned with the coding region and showed only ~14% rRNA bases (Figure 8), suggesting hybridization capture is a suitable strategy for enriching FFPE RNA without the need for upstream ribosomal RNA depletion. This is a notable observation, as it will save researchers both time and costs by eliminating the need for upstream depletion modules. As seen in Figure 5, FFPE ribo-depleted transcriptome samples tend to be skewed towards intronic reads. Target enrichment removes the skew towards intronic regions from >36% of the sequenced bases to <3% of the sequenced bases, and increased coding bases from ~20% to ~70% of sequenced bases. By focusing sequencing reads on coding regions and minimizing reads that are not of interest, researchers can save money on sequencing costs and enable increased sample throughput.

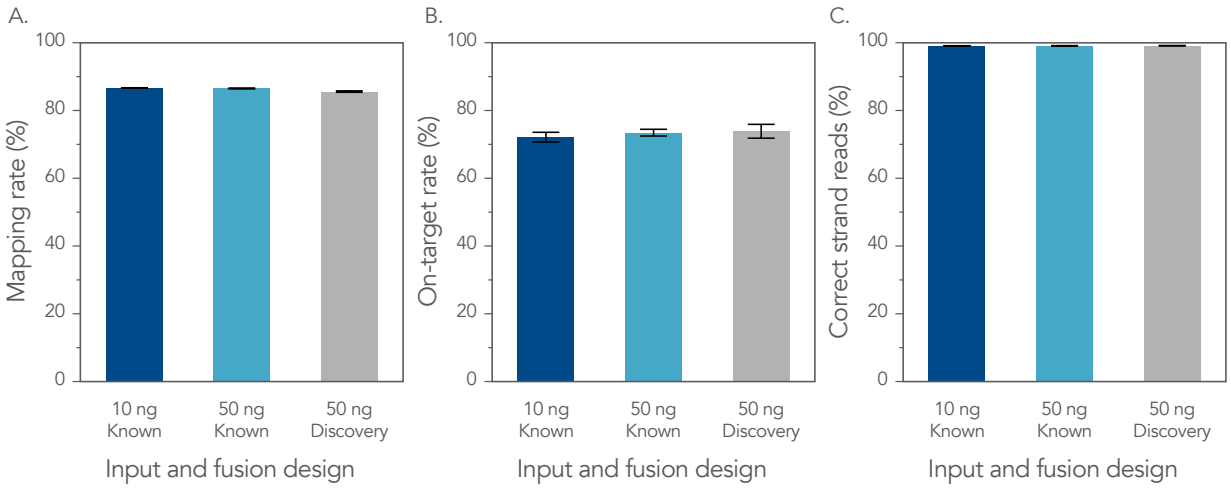

Figure 7. RNA-seq metrics for targeted enriched libraries. xGen Broad-Range RNA libraries from 10 ng and 50 ng of total FFPE RNA were captured with a Custom Hyb Capture Panel designed to capture known gene fusions. A subset of libraries generated from 50 ng were enriched with a Custom Hyb Capture Panel designed for gene fusion discovery. All libraries show (**A**) high mapping rate, (**B**) high on-target, and (**C**) high percent strandedness for libraries generated from both input amounts and when using either fusion Hyb Capture panel designs. Sequencing was performed on a NextSeq™ (Illumina) using 2 x 150 paired-end reads and the data subsampled to 4 million reads/sample. Each bar represents the mean of three technical replicates.

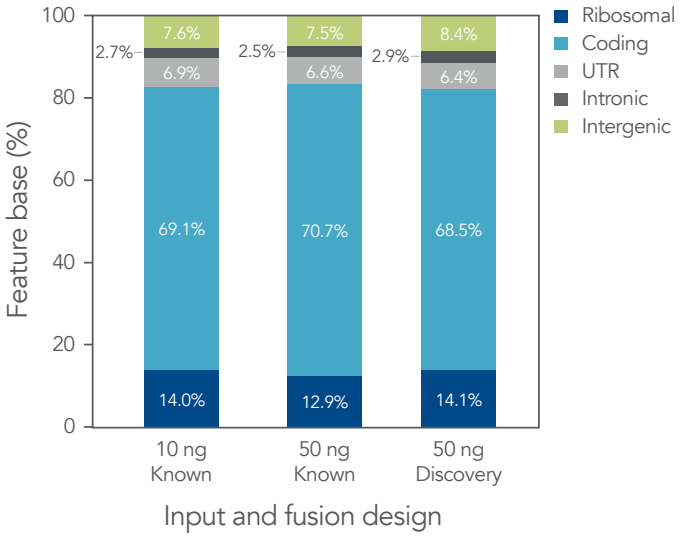

Figure 8. Featured bases chart for hyb captured libraries without upstream ribosomal RNA depletion. Values within the chart represent the mean for the three replicates of hybridization captured xGen Broad-Range RNA libraries. Sequencing performed on a NextSeq™ (Illumina) using 2 x 150 paired-end reads and the data subsampled to 4 million reads/sample.

Fusion identification results

When assessing libraries enriched with the Known Fusion Design capture panel, all expected fusions were identified in libraries prepared with both 10 ng and 50 ng of total FFPE RNA, indicating this fusion panel design strategy can capture and correctly identify fusion events across a range of RNA input amounts. The results also show a higher junction and spanning read count for samples with higher mass input which is expected (Table 1).

When comparing whole transcriptome and hybridization captured libraries, all three known fusions were identified in each of the libraries. Spanning reads were consistently identified in all hyb captured libraries but inconsistently identified in the transcriptome libraries. Additionally, with 10X fewer reads, the captured libraries found substantially more junction and spanning reads for each of the fusion events (Table 1). The higher number of reads identifying the fusion gene increases the confidence that the fusion event is not an artifact, but an actual fusion found in the sample. Hybridization capture target space is significantly smaller than the entire transcriptome which enables an increased sequencing coverage and therefore an increased ability to identify gene fusions. Furthermore, many more libraries can be sequenced on a single flow cell, which is a cost-effective approach to oncology research studies.

	50 ng transcriptome–40M reads						50 ng Known Fusion Panel–4M reads						10 ng Known Fusion Panel–4M reads					
	Library 1		Library 2		Library 3		Library 1		Library 2		Library 3		Library 1		Library 2		Library 3	
Fusion Name	Junction	Spanning	Junction	Spanning	Junction	Spanning	Junction	Spanning	Junction	Spanning	Junction	Spanning	Junction	Spanning	Junction	Spanning	Junction	Spanning
SLC34A2-ROS1	16	3	19	5	16	0	761	249	732	277	851	266	282	112	259	91	257	100
CCDC6-RET	6	0	6	0	4	0	251	59	262	87	251	68	82	6	83	21	99	31
EML4-ALK	2	2	9	1	6	1	179	57	203	80	206	51	63	13	55	21	55	26

Junction Read = single read overlaps fusion
Spanning Read = paired-end read which maps to both sides of fusion

Table 1. Fusion identification. A comparison of gene fusion identification between transcriptome and hyb captured xGen Broad-Range RNA libraries enriched using a Hyb Capture Panel designed to identify known gene fusions shows correct identification of all known gene fusions in all hyb captured libraries, while some gene fusions are missed or found in significantly less abundance in whole transcriptome libraries. The number of gene fusions identified in hyb captured libraries is several fold higher than what is found in transcriptome libraries, while using 10X less sequencing reads. 2–3X more fusions were identified in libraries prepared with the higher input amount. Sequencing was performed on a NextSeq™ (Illumina) using 2 x 150 paired-end reads, with transcriptome libraries subsampled to 40 million reads per sample and hyb captured libraries subsampled to 4 million reads.

The Custom Hyb Panel designed to for gene fusion discovery utilized 24X tiled probes spaced 5 nucleotides apart for the targeted gene exon(s). As a proof of concept, this design was tested on three exons of *ROS1*, therefore, one would expect to identify the SLC34A2-ROS1 fusion which is located in that exonic region. The sequencing results (Table 2) show correct identification of this gene fusion. When looking at libraries generated equivalent library input amounts, the sequencing read counts for junction reads and spanning reads were comparable to the number of junction reads seen using the Known Fusion Design in data Table 1, therefore, the capture panel using the Discovery Fusion Design enables gene fusion identification with hybrid capture technology, regardless that the design does not take into account the break-point of the fusion and did not target SLC34A2. This design strategy can enhance the ability to discover new gene fusions that can potentially become important biomarkers for oncology.

	50 ng Discovery Fusion Panel–4M reads					
	Library 1		Library 2		Library 3	
Fusion Name	Junction	Spanning	Junction	Spanning	Junction	Spanning
SLC34A2-ROS1	704	298	664	299	615	287

Junction Read = single read overlaps fusion
Spanning Read = paired-end read which maps to both sides of fusion

Table 2. Fusion discovery. xGen Broad-Range RNA libraries prepared using 50 ng total FFPE RNA (*n* = 3) were captured using a hyb panel design for gene fusion discovery. The fusion discovery Hyb Capture Panel was designed to interrogate three exons of *ROS1* as a region of interest in a highly tiled manner and does not include any probes specific to the SLC34A2 target or fusion target. Results show the correct identification of the SLC34A2-ROS1 fusion when targeting only one gene fusion partner indicating this panel design strategy can be used for gene fusion discovery.

Conclusion

RNA-seq continues to play an important role in understanding how structural variants affect disease state. Here the data have shown how using the xGen Broad-Range RNA Library Prep Kit in conjunction with xGen Custom Hyb Panels results in sequencing results that enable fusion identification of both known and novel fusions. The results demonstrate that hybridization capture for target enrichment of RNA-seq libraries can replace an upstream ribosomal RNA depletion step as the percentage of rRNA bases was far lower than is typically seen in total RNA libraries [5]. By eliminating rRNA depletion, the workflow time and cost is reduced. Furthermore, the consistency between technical replicates and between varying input amounts using a FFPE RNA sample highlight the consistency of the workflow.

Both the Known Fusion Design and the Discovery Fusion Design capture panels were used to correctly identify ddPCR-confirmed gene fusions with 10x fewer reads than transcriptome sequencing. Targeted sequencing results in substantial cost savings as higher target coverage can be obtained from 10X less sequencing compared to whole transcriptome sequencing. By continuing to identify novel fusions and gaining better resolution of fusion gene isoforms, researchers can gain a better understanding of fusion gene biology and identify potentials for further research.

References

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009; 10(1):57-63. doi.org/10.1038/nrg2484
2. Latysheva NS, Babu MM. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res.* 2016;44(10):4487-4503. doi:10.1093/nar/gkw282
3. Heyer, E.E., Deveson, I.W., Wooi, D. et al. Diagnosis of fusion genes using targeted RNA sequencing. *Nat Commun.* 2019;10(1):1388. doi.org/10.1038/s41467-019-09374-9
4. Lin, X., Qiu, L., Song, X. et al. A comparative analysis of RNA sequencing methods with ribosome RNA depletion for degraded and low-input total RNA from formalin-fixed and paraffin-embedded samples. *BMC Genomics.* 2019;20(1):831. doi.org/10.1186/s12864-019-6166-3
5. O'Neil, D., Glowatz, H. & Schlumpberger, M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr Protoc Mol Biol.* 2013;Chapter 4. doi.org/10.1002/0471142727.mb0419s103

RNA-seq for biomarker identification using the xGen™ Broad-Range RNA Library Kit and xGen Custom Hyb Panels

Technical support: applicationsupport@idtdna.com

Doc ID: RUO21-0603_001 06/22