TECH NOTE

Efficient, Exon-Aware RNA Capture for Gene Expression and Novel Fusions

INTRODUCTION

Total RNA sequencing (RNA-seq) provides a relatively unbiased view of the transcriptional state of a population of cells. However, most total RNA-seq experiments must contend with a large number of reads that are not helpful for gene-expression analysis, including reads from highly abundant non-coding transcripts like ribosomal RNA, intronic reads from pre-mRNA, or contaminating genomic DNA. Target enrichment provides a way to focus sequencing on the informative parts of the genome, allowing for more sensitive detection of low-abundance transcripts, or for profiling only specific genes of interest.

Here we present capture sequencing experiments using Twist's new RNA Exome panel, which uses a novel design strategy to specifically target every protein-coding isoform in Gencode v41 Basic. Although the design natively targets the transcriptome, our design strategy also places probes to minimize design bias and allow for discovery of novel fusion genes. We evaluate panel performance in expression quantification, showing that relative transcript abundances are preserved after hybrid capture. This allows for accurate and reproducible quantification of transcripts that are present across many orders of magnitude. We show gains in sequencing efficiency from our targeted approach and demonstrate the ability to capture novel structural variants, such as RNA fusions common in cancers. Additionally, we discuss our bioinformatic approach to evaluating capture performance for RNA, and discuss specific challenges in the analysis of RNA-seq experiments. In summary, we provide evidence that the Twist Targeted Enrichment for Gene Expression solution is an effective way to efficiently profile gene expression and detect gene fusions.

RESULTS

DESIGN STRATEGY AND CONTENT

Our first step in generating the RNA exome was to decide on both a content curation strategy and a strategy for how we would design capture probes against a transcript. Content curation was performed using the GenCode gene definitions (v41 on hg38)—our aim was to focus our design on the coding regions of protein-coding genes. To this end, we pared down the total defined coding sequence (CDS) space in GenCode to categories of genes that were either protein-coding or with strong evidence for coding content in certain situations. In addition, we covered the 3' and 5' untranslated regions of some genes (such as those involved in recurrent fusions) to ensure that the panel had maximum sensitivity to these events. From these genes, we chose to tile a set of well-described transcript models, with the aim of natively covering the majority of isoforms that are of general interest to most researchers. Importantly, the content selected from these transcript models constitute the set of high-confidence exons within these genes.

To avoid capturing either contaminating genomic DNA, or pre-mRNA, we did not target flanking intronic sequences of genes. We thus decided on a tiling strategy that directly targeted the mature mRNA forms of transcripts. The naive approach to covering these transcripts would be to tile them with probes end-to-end **(Figure 1A)**. However, this has the drawback of biasing capture towards known isoforms, and biasing capture against fusion transcripts. Instead, we employed a new "exon-aware" design strategy that avoids placing probes across exon-exon boundaries **(Figure 1B)**. By doing this, we can ensure that novel isoforms or fusion transcripts can be detected efficiently, as the probes do not select for known exon-exon junctions.

After tiling the design using the exon-aware strategy above, we collapsed exact duplicate probes and removed probes with low-sequence complexity and/or homology towards non-coding RNAs that would reduce sequencing efficiency (i.e., mitochondrial and nuclear ribosomal RNAs and tRNAs). With this design finalized, we used Twist's DNA printing technology to synthesize our probes using our standard target enrichment panel process.

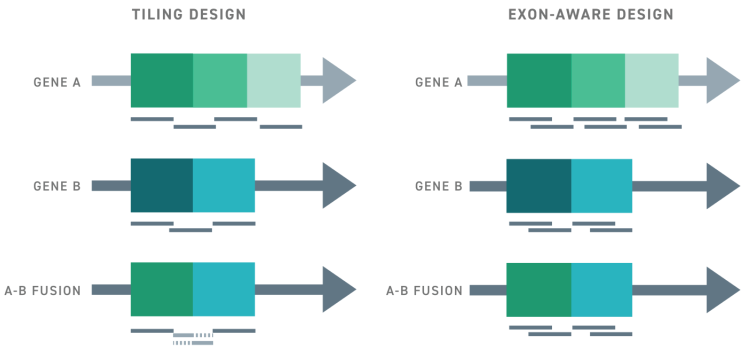

Figure 1. (A) Schematic of gene fusion detection using naive tiling across RNA transcripts. Dotted lines indicate mismatches (B) Schematic of gene fusion detection with the exon-aware tiling strategy.

COMPARISON OF EXOME CAPTURE TO WTS AND 3'-COUNTING

In addition to targeted sequencing, the common workflows for assessing gene expression are whole transcriptome sequencing (WTS), which uses random priming to select a relatively unbiased set of transcripts from ribosomal-depleted RNA, and 3'-counting, which uses an oligo-dT primer to isolate the 3'-ends of polyadenylated mRNA transcripts (primarily mRNAs). Broadly the benefit of performing WTS is that the user gets a relatively unbiased view of the transcriptome, at the expense of losing a substantial number of reads to introns and other relatively uninformative areas of the genome **(Figure 2A)**. Correspondingly, 3'-counting is more efficient at selecting exonic regions (CDS and UTR), but displays a strong bias towards the 3'-ends of transcripts **(Figure 2B)**. This would be expected to impact the ability to detect different isoforms, as only part of the transcript is profiled for longer genes.

To address these issues, we designed the RNA exome panel to profile the entire CDS of protein coding transcripts, which achieves a measured 3' bias and duplicate rate similar to what is observed with WTS **(Figure 2B)**. We also carefully excluded intronic and highly-expressed non-coding sequences from the design, which allows us to focus reads more efficiently into exons than either WTS or 3'-counting **(Figures 2A, 2B)**. Selecting for mature transcripts by hybrid capture had other advantages as well—the percentage of reads derived from the incorrect strand was reduced compared to either 3'-counting or WTS **(Figure 2B)**.

Since transcripts exist in concentrations that span roughly 6 orders of magnitude, we next asked whether RNA hybrid capture was equally efficient in enriching for both low- and highly-expressed transcripts. To do this, we correlated the counts from a WTS run to the counts obtained from an RNA exome capture in the same sample type **(Figure 2C)**. We found that enrichment was consistent across the entire range of expression, indicating that the capture system was not saturated even for highly expressed transcripts.

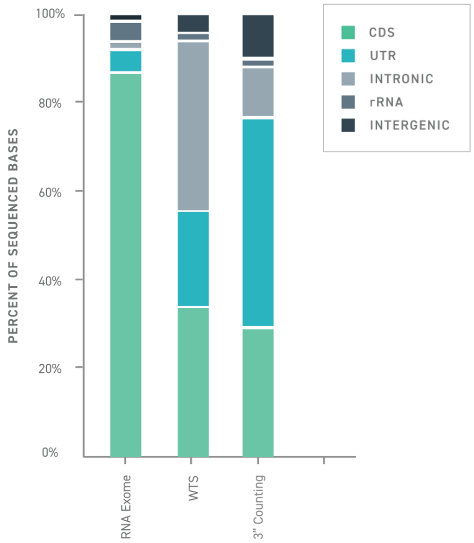

Figure 2A. Genomic distribution of reads in RNA exome capture, whole transcriptome sequencing (WTS), and 3'-counting.

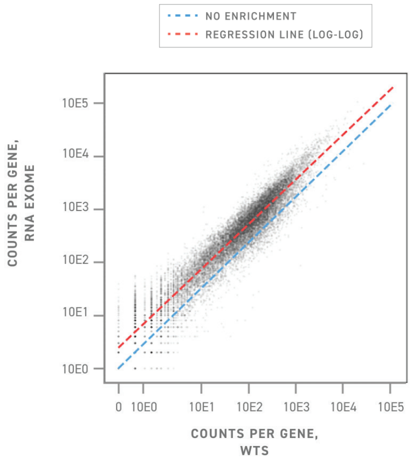

Figure 2C. Correlation of uncaptured counts (x-axis) to captured counts (y-axis) for protein-coding genes.

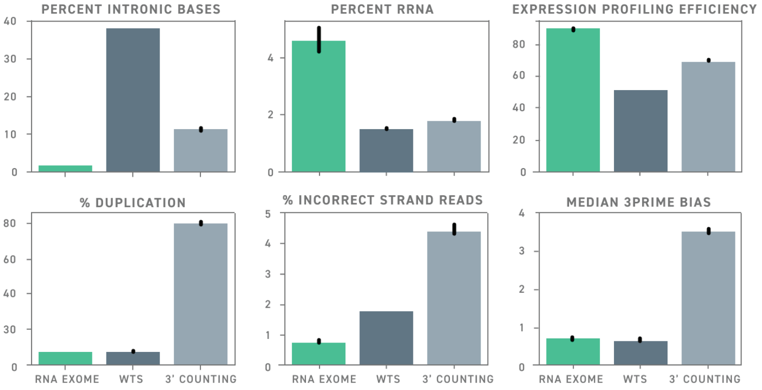

Figure 2B. Comparison of sequencing metrics between RNA exome capture, WTS and 3'-counting.

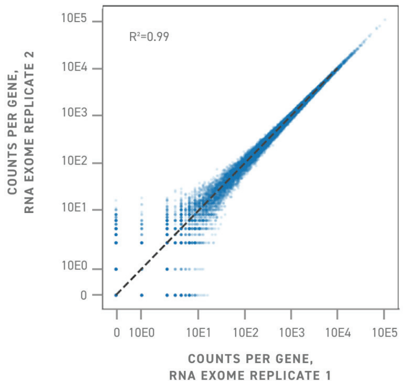

Figure 2D. Correlation of two technical replicate captures performed with the RNA exome.

PERFORMANCE OF THE TWIST RNA EXOME ON LOW-MASS INPUTS AND FFPE SAMPLES

Formalin-fixed paraffin-embedded (FFPE) tissue is tissue that has been preserved for histology. Although this process damages nucleic acids, FFPE tissue is nonetheless often used for RNA-seq because the samples are readily available as clinical specimens. As previous applications of RNA capture in the literature have focused on FFPE samples (Jang et al 2021, Pennock et al 2019, Vahrenkamp et al 2019), we evaluated the performance of the RNA exome on FFPE samples at three different mass inputs (1 ng, 10 ng and 100 ng). Since the RNA exome selects efficiently for coding content, we also examined 5 levels of read sampling between 10M and 30M reads for both WTS and RNA exome to establish an approximate equivalence between the reads required to detect a particular number of genes in each workflow. We looked both at coding genes as detected by alignment and standard feature counting **(Figure 3A)**, as well as the number of detected isoforms using a k-mer based approach **(Figure 3B)**. In both cases, a cutoff of 5 supporting reads was used to define detection.

Our results show that the RNA exome dramatically improves the number of detected coding genes and transcripts at all mass levels. We find at high mass inputs that we detect similar numbers of coding genes with 15M sampled reads compared to a WTS sample with 30M reads. The results were particularly striking for 1 ng of FFPE input, where the TE sample detected comparable numbers of coding genes as higher input quantities in the TE sample, while even 30M reads in the WTS workflow was unable to detect approximately 1,000 low-expressed genes that were measurable with TE **(Figure 3A)**. The patterns for the number of detected transcripts were similar, with a measurable increase at all levels of read sampling that was more striking for low mass inputs. For 1 ng of FFPE, we were able to detect a comparable number of transcripts with 10M reads with the RNA exome as were detected with 30M reads using WTS **(Figure 3B)**.

Since FFPE RNA tends to be highly fragmented, we asked whether target enrichment might be able to select for a subset of less degraded sequences. We sequenced an FFPE sample using both whole transcriptome sequencing and the RNA exome, and plotted the inferred size distribution of fragments based on alignment directly to RNA transcripts. The size distribution showed a clear upward shift for the RNA exome sample **(Figure 3C)** indicating that the RNA exome does indeed select for more intact material.

Finally, we wanted to get a sense of how robust RNA exome capture was to FFPE material of differing quality. We thus extracted RNA from 5 commercially available FFPE standards, and subjected these samples to both whole-transcriptome sequencing and capture with the RNA exome. We find that the RNA exome is able to significantly increase the number of detected genes for all tested samples, irrespective of their level of degradation **(Figure 3D)**.

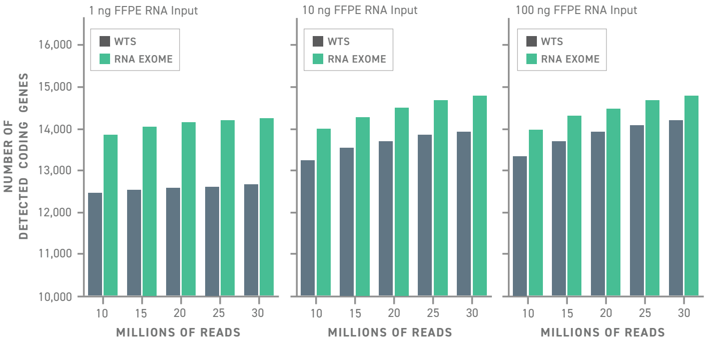

Figure 3A. Number of protein coding genes detected (y-axis) with different levels of downsampling (x-axis) in 3 different mass inputs of FFPE material.

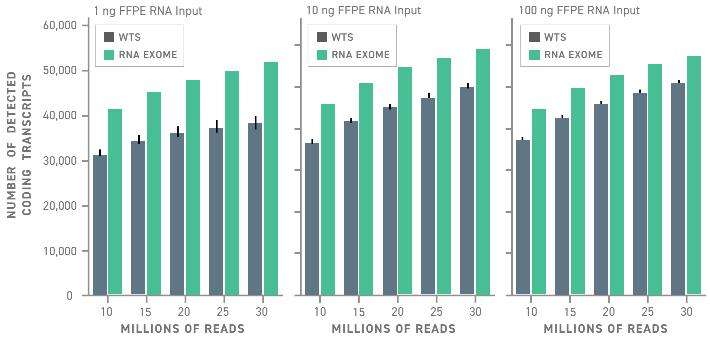

Figure 3B. Number of protein coding transcripts (including different isoforms of the same gene) detected (y-axis) with different levels of downsampling (x-axis) in 3 different mass inputs of FFPE material.

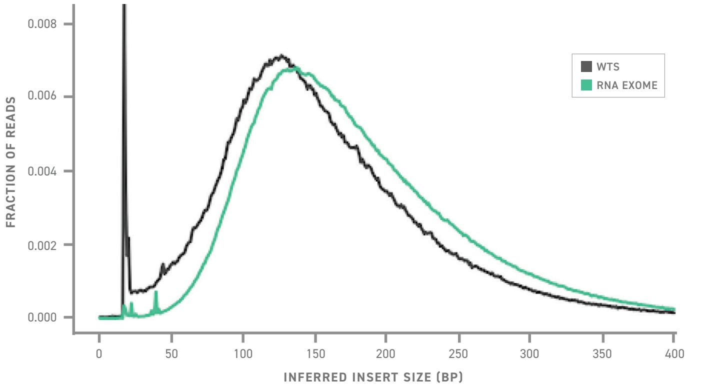

Figure 3C. Size distributions of captured (RNA exome) and uncaptured (WTS) reads from an FFPE sample.

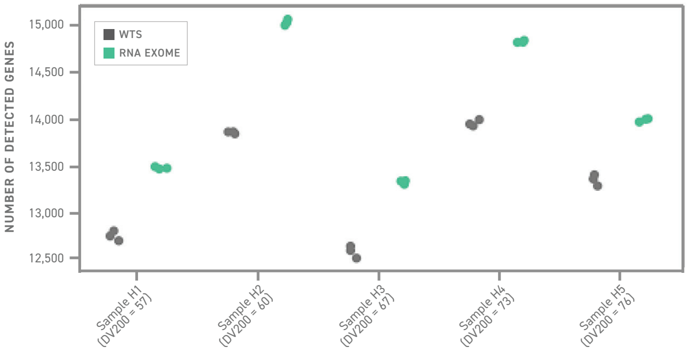

Figure 3D. Comparison of captured (RNA exome) and uncaptured (WTS) counts for a variety of FFPE samples with different integrity.

DIFFERENTIAL EXPRESSION ANALYSIS WITH THE TWIST RNA EXOME

RNA-seq is often used to gain biological insight into how specific stimuli (such as drug treatments or gene knockouts) can alter a cell's transcriptional landscape. Since the RNA exome is able to efficiently enrich transcripts across the full range of expression **(Figure 2C)** with a consistent degree of enrichment **(Figure 2D)**, we asked whether the counts from the RNA exome could be used directly to assess changes in expression between two conditions.

As a model for these changes, we made use of a commercially available pair of matched breast tumors and normal breast tissue samples. We sequenced each sample in triplicate at two mass conditions (10 ng and 100 ng) using both whole-transcriptome sequencing and capture with the Twist RNA exome. After quantifying counts per gene and performing differential expression analysis, we compared the fold-changes detected in WTS to those detected with capture. We found good agreement between these estimates **(Figure 4A)**, indicating that counts from capture can be directly used to estimate fold-changes in gene expression between conditions. Since RNA capture increases the total counts of genes **(Figure 2C)** we asked whether this would lead to better statistical power for differential calls. As expected, we observed an overall upward shift in the volcano plot for the captured sample indicating that most calls are overall more statistically significant between the conditions **(Figure 4B)**. Making pairwise comparisons between the FDR-adjusted p-values for individual genes, we find that the majority of genes are detected with increased power **(Figure 4C)**. Thus, the Twist RNA exome both preserves the existing relationships between differentially expressed genes, and by increasing the fraction of useful reads, allows for greater statistical power in determining differences.

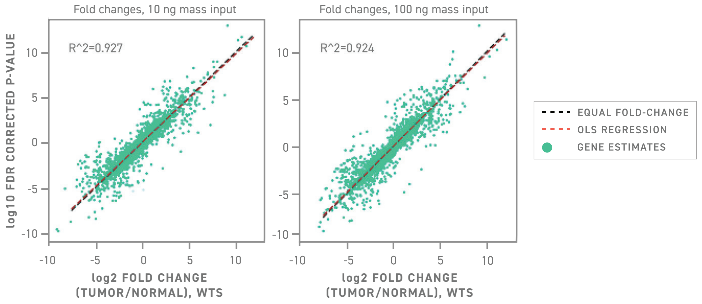

Figure 4A. Correlation between uncaptured (WTS) and captured (RNA exome) log2 fold-change estimates between a matched Tumor-Normal pair at two different mass inputs.

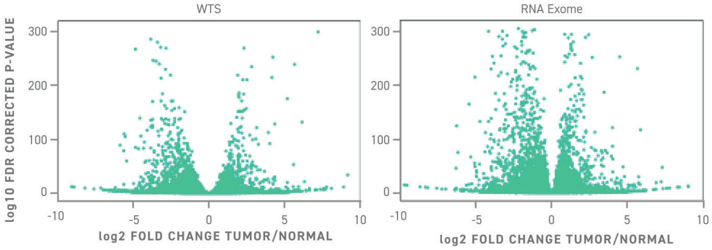

Figure 4B. Volcano plots illustrating the estimated log2 fold changes and log10 p-values for uncaptured (WTS) and captured (RNA exome) samples.

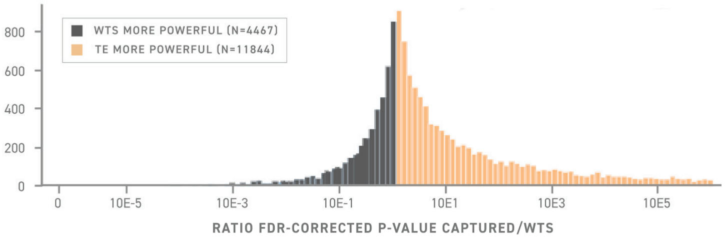

Figure 4C. Histogram showing distribution of ratios of FDR-corrected p-values for captured (RNA exome) and uncaptured (WTS) samples. Bins colored in orange show more significant p-values in RNA exome, while bins colored blue show more significant p-values in WTS.

DETECTION OF FUSION TRANSCRIPTS WITH THE TWIST RNA EXOME

Because RNA-seq detects mature transcripts, it is uniquely able to detect novel structural variants with functional impact on the cell. These events include fusion genes, a common driving mechanism in cancer where a novel exon-exon junction is formed between two distinct genes. Since the Twist RNA exome improves detection of transcripts that are used in the design space, we additionally asked whether the Twist RNA exome could efficiently discover these novel transcripts that were not included in the design space.

To look into this, we made use of a cell-line-derived FFPE standard material that contained well-defined fusion events. We built an index from transcripts in GenCode v41 combined with the sequences of the expected fusions, and then classified 10M sampled reads from either a WTS experiment or an RNA exome capture against this set of transcripts. We find that the RNA exome significantly enriches for number of detection events for both fusion transcripts **(Figure 4A)**. As these events may occur at a low frequency in a tumor sample, these data establish RNA capture as an efficient method for improving the detection rate of these important events. To ensure that the RNA exome was truly detecting intact reads spanning the fusion junction, we similarly examined reads aligned to the sequence of the fusion transcript from these samples, finding that a large number of reads crossed the expected junction site in transcript space **(Figure 5A, 5B, and 5C)**. These results demonstrate that the improvement in targeted reads does not come at the expense of the ability to profile novel exon-exon junctions, and in fact that the increased number of targeted reads allows for increased sensitivity in detecting these important events.

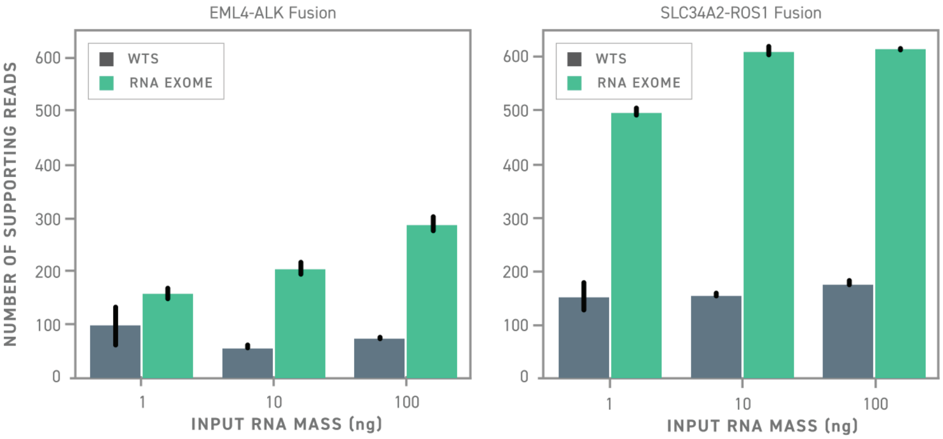

Figure 5A. Number of reads detecting 2 fusion transcripts (EML4-ALK and SLC34A2-ROS1) for different input masses in captured (RNA exome) and uncaptured (WTS) samples.

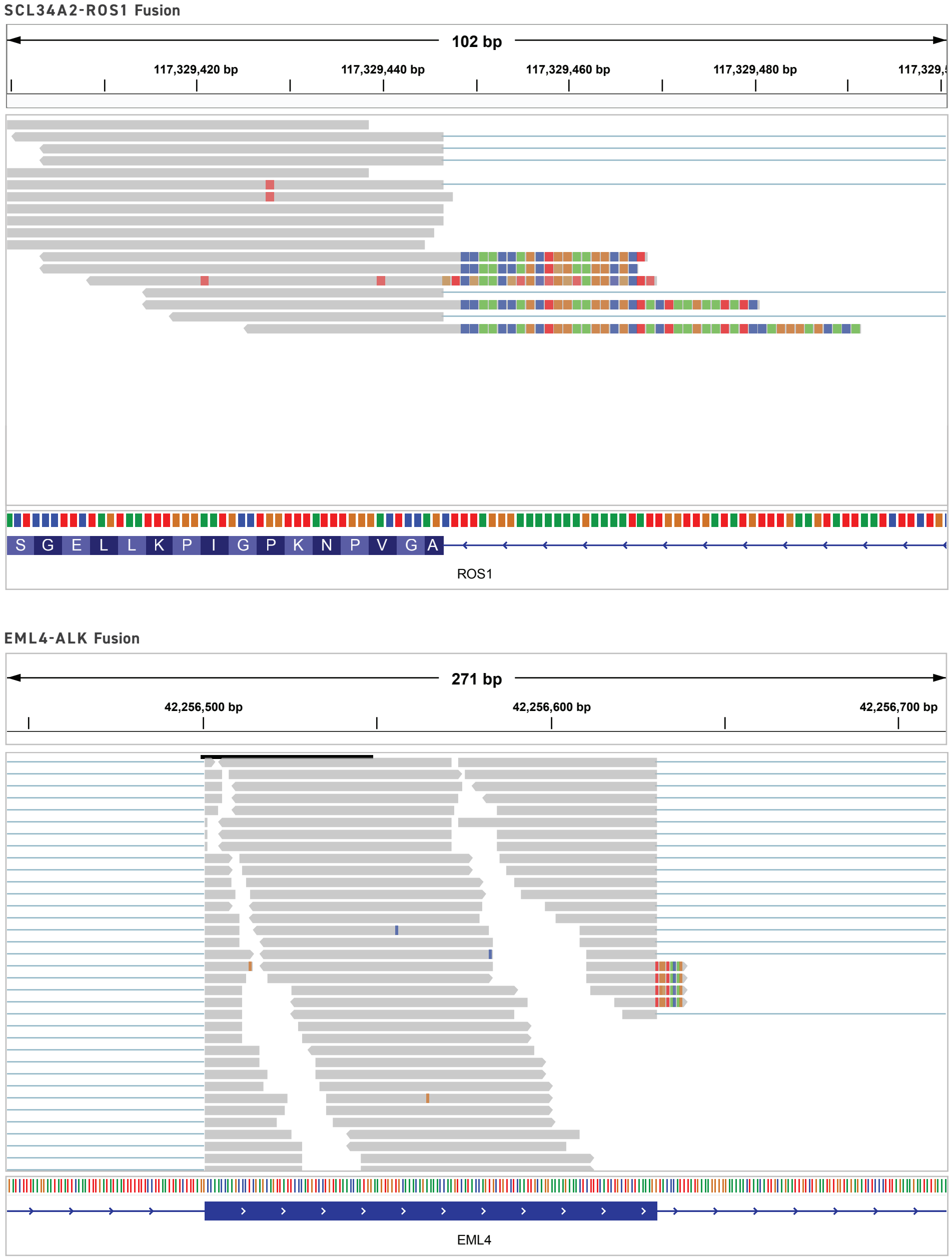

Figure 5B and C. Genome browser view showing fusion detection at ROS1 (top) and EML4 (bottom) breakpoints. Colored blocks at the end of reads indicate bases supporting a fusion partner instead of aligning to the reference sequence.

SUMMARY

Our results indicate that RNA capture with the Twist RNA exome is a powerful approach for profiling transcription. We show that the RNA exome is more efficient at profiling mature transcripts than either 3'-counting or whole transcriptome sequencing **(Figure 2)**. We quantify the expected savings on reads from this increased efficiency and further demonstrate the particular utility of RNA exome capture on highly degraded or low input samples **(Figure 3)**. Finally we show that exome capture can increase the statistical power of differential expression calls without introducing bias **(Figure 4)**, and that the RNA exome demonstrates increased sensitivity to fusion transcripts **(Figure 5).**

MATERIALS AND METHODS

To test the Twist RNA Exome panel, 1 ng, 10 ng, or 100 ng of Universal Human Reference RNA (Agilent P/N 740000) or FFPE RNA Fusion Reference Standards (Horizon Discovery P/N HD784) was added to the Twist RNA-seq Library Preparation Kit. Prior to making libraries, FFPE material was extracted using the Qiagen RNeasy® FFPE Kit. Target enrichment was performed using 500ng of library and the Twist Target Enrichment Standard Hybridization v2 Protocol with a 16-hour hybridization reaction time. Sequencing was performed with the Illumina NextSeq platform and 76 bp paired-end reads.

Analysis was performed by sampling FASTQ files to a fixed number of reads (10M pairs/20M reads unless otherwise specified). Alignment was performed against hg38 using STAR (Dobin et al 2013) and gene quantification was performed using FeatureCounts (Liao et al 2014) with GenCode v41 gene annotations. Metrics were calculated using Picard CollectRnaSeqMetrics. Differential expression was assessed with DESeq2 (Love et al 2014). Data processing and visualization were performed with Pandas and Seaborn using custom Python scripts. Genome browser visualization was performed with IGV. Fusion transcript quantification was performed using Salmon (Patro et al 2017) with an index built from the GenCode v41 transcript sequences concatenated to the fusion transcript sequences.

REFERENCES

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.

Jang JS, Holicky E, Lau J, McDonough S, Mutawe M, Koster MJ, Warrington KJ, Cuninngham JM. Application of the 3' mRNA-Seq using unique molecular identifiers in highly degraded RNA derived from formalin-fixed, paraffin-embedded tissue. BMC Genomics. 2021 Oct 24;22(1):759. doi: 10.1186/s12864-021-08068-1. PMID: 34689749; PMCID: PMC8543821

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656. Epub 2013 Nov 13. PMID: 24227677.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PMID: 25516281; PMCID: PMC4302049.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017 Apr;14(4):417-419. doi: 10.1038/nmeth.4197. Epub 2017 Mar 6. PMID: 28263959; PMCID: PMC5600148.

Pennock ND, Jindal S, Horton W, Sun D, Narasimhan J, Carbone L, Fei SS, Searles R, Harrington CA, Burchard J, Weinmann S, Schedin P, Xia Z. RNA-seq from archival FFPE breast cancer samples: molecular pathway fidelity and novel discovery. BMC Med Genomics. 2019 Dec 19;12(1):195. doi: 10.1186/s12920-019-0643-z. PMID: 31856832; PMCID: PMC6924022.

Vahrenkamp JM, Szczotka K, Dodson MK, Jarboe EA, Soisson AP, Gertz J. FFPEcap-seq: a method for sequencing capped RNAs in formalin-fixed paraffin-embedded samples. Genome Res. 2019 Nov;29(11):1826-1835. doi: 10.1101/gr.249656.119. Epub 2019 Oct 24. PMID: 31649055; PMCID: PMC6836741.

LEARN MORE

TWISTBIOSCIENCE.COM/NGS
SALES@TWISTBIOSCIENCE.COM