An RNA exome panel used to enrich transcript variants using cDNA libraries

Michael Bocek, Kristin Butcher, Yu Cai, Jean Challacombe, Derek Murphy, Esteban Toro

TWIST BIOSCIENCE

Abstract

Total RNA sequencing 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 the 7SK RNA or 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 bias towards known isoforms and allow for discovery of novel isoforms or 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 in RNA space 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.

Design strategy and content selection

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 CDS space in GenCode to categories of genes that were either protein-coding or with strong evidence for coding content in certain situations (see **Figure A**). 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.

We next had to decide on a tiling strategy. We considered three possibilities - first, tiling the probes against our content using the same strategy used for most DNA exomes (**Figure 1B**). This has the advantage of being conceptually simple and handling multiple isoforms with a minimum of redundancy, but it would be expected to selectively capture gDNA and pre-mRNA, as it contains exon-intron junctions. Second, we considered a straightforward tiling to the mature transcripts (**Figure 1C**) but found that this had significant probe redundancy and would likely select against novel isoforms or fusion transcripts, as it contains probes that span exon-exon boundaries. Finally, we placed probes such that every exon-exon boundary contained at least one non-spanning probe (**Figure 1D**). This reduced the number of distinct and redundant probes, avoided capturing intronic content, and avoided introducing additional bias towards content already represented in the design. We called this strategy the "exon-aware" design, and ultimately decided to move forward with this strategy for production.

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.

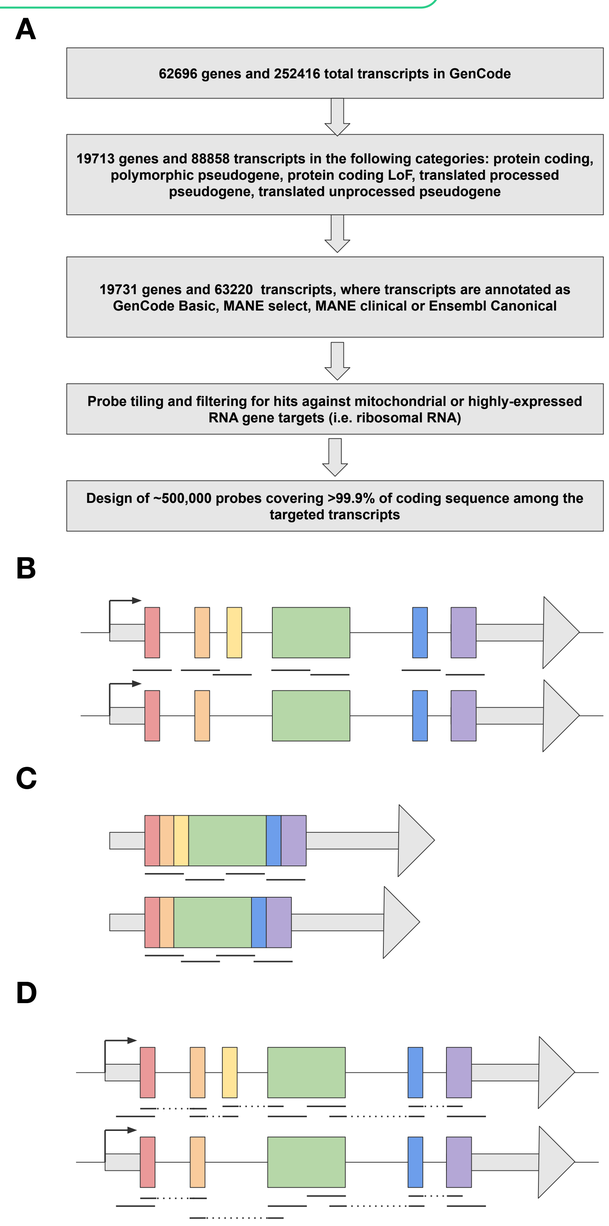

Figure 1: (A) Content curation process for the RNA exome. (B) Example of DNA-based tiling strategy, similar to what is adopted for most DNA-based exomes over two isoforms of an example gene. (C) Example of straightforward tiling of the transcript sequences with probes. (D) Example of Twist's exon-aware design strategy, which was ultimately adopted for the RNA exome design, over the two example transcripts.

Performance relative to uncaptured RNA-seq

Target capture is uniquely able to purify the subset of protein-coding genes. This design allows for improved efficiency without the need for a ribosomal depletion step. The Twist RNA Exome panel outperforms whole transcriptome sequencing (WTS) and 3' counting in having the least amount of intronic bases called and the most exonic content (expression profiling efficiency). More coding genes are detected with a lower 3' bias and percent duplication rate (**Figure 2A**).

Coding sequences (CDS's) are generally the most informative part of a gene for detecting fusion events and are generally easier to uniquely assign reads to when compared with UTRs. As the RNA exome is primarily designed against CDS's, we obtain substantially more coding reads than other techniques (**Figure 2B**).

Since capture uses a limited quantity of probes, we wondered whether there might be a leveling effect where our capture probes become saturated. However, comparing gene counts in a WTS sample to our captured counts shows that enrichment is more or less even across the full 5 orders of magnitude of gene expression (**Figure 2C**).

Figure 2: (A) Comparison of sequencing metrics for enriched, whole transcript, and 3'-counting methods on identical reference samples. (B) Breakdown of signal from 3' counting, RNA exome, and WTS by genome compartment. (C) Correlation between RNA exome and WTS showing enrichment in raw counts per gene.

Capture of damaged/low-mass templates

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 the RNA exome is able to efficiently recover coding sequences from a library, we asked whether issues in FFPE tissue could be rescued by exome capture. Our results indicate that the RNA exome enriches equally efficiently in FFPE as in non-FFPE samples (**Figure 3A**), while reducing duplicate rates (**Figure 3B**), reducing incorrect strand percent (**Figure 3C**), and increasing the number of detected genes (**Figure 3D**) compared to WTS.

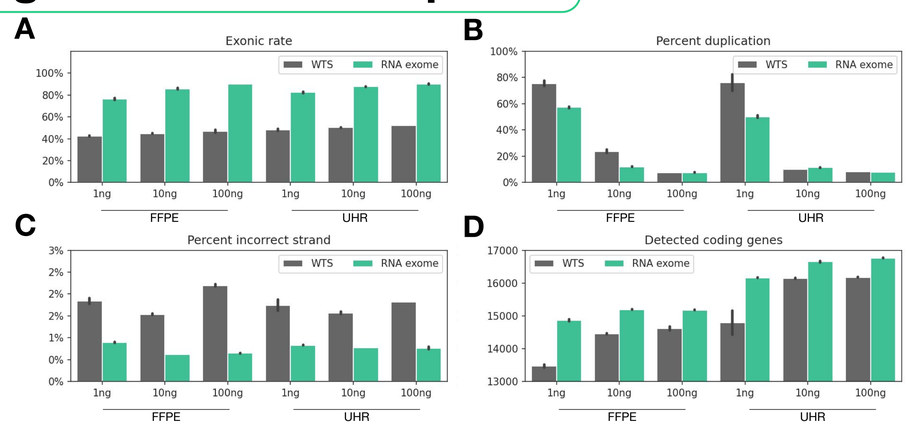

Figure 3: (A) Exonic rate (expression profiling efficiency) from FFPE and UHR RNA at mass inputs of 1ng, 10ng and 100ng. (B) Percent duplication as determined from UMI and mapping position from FFPE and UHR RNA at mass inputs of 1ng, 10ng and 100ng. (D) Percent of reads mapping to the incorrect strand from FFPE and UHR RNA at mass inputs of 1ng, 10ng and 100ng. (B) Number of detected protein-coding genes and defined by GenCode from FFPE and UHR RNA at mass inputs of 1ng, 10ng and 100ng. In all cases, error bars represent SEM.

Differential expression

One important application of RNA sequencing, particularly in oncology applications, is differential expression. Although capture does introduce some bias into gene expression estimates (**Figure 2C**), this bias is extremely consistent for the same genes between runs. We thus asked whether we could preserve differences in gene expression and recover similar estimates for WTS and RNA exome capture. To test this, we took 3 replicates of paired Tumor/Normal RNA reference samples through both WTS and RNA exome capture (**Figure 4A**).

We tested both high- (100ng) and low-input (10ng) conditions to evaluate whether limited material behaves differently in capture and WTS. Our results indicate that differential expression estimates are similar between the two experimental workflows (**Figure 4B**), but the increased read counts from capture provide better statistical power (**Figure 4C**), and identifies more genes that are significantly altered between the tumor and normal conditions (**Figure 4D**).

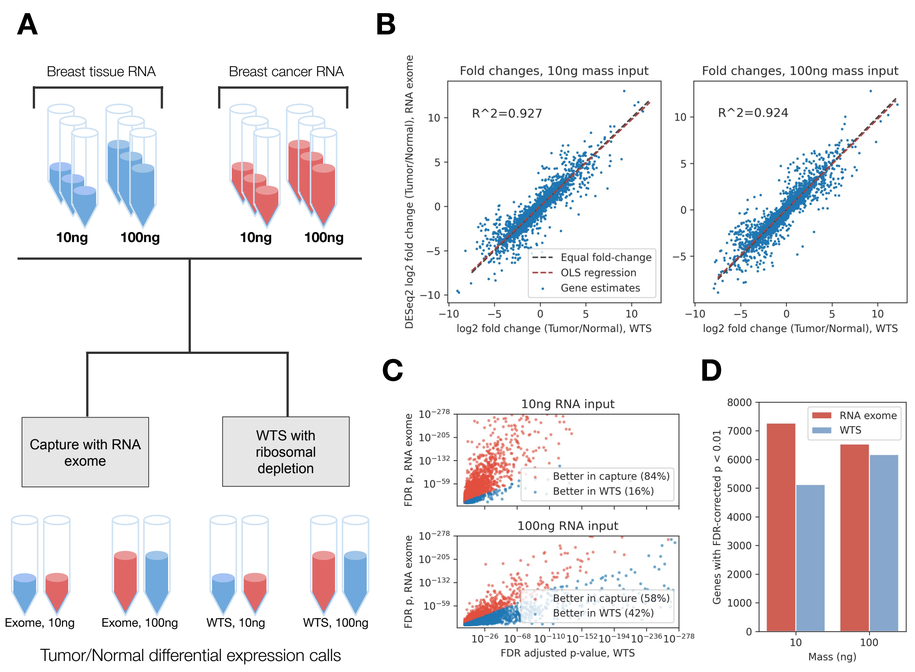

Figure 4: (A) Summary of differential expression experiment design. (B) Correlation of tumor/normal fold-change estimated from WTS (x-axis) to tumor/normal fold-change estimated from RNA exome capture (y-axis). (C) Comparison of false discovery rate (FDR) adjusted p-values from differential expression experiment in WTS and RNA exome capture comparing significance in each experiment. (D) Number of genes with FDR-corrected p-value <0.01 in RNA exome and WTS experiments at both mass conditions

Fusion RNA detection

In addition to gene quantification, an important application of RNA-seq is to discover certain classes of structural variants (such as gene fusions) that are difficult to discover in DNA space. One potential challenge with RNA capture is that it might introduce bias towards transcripts in the design space and cause these fusion transcripts to be underrepresented.

To determine whether our RNA capture is able to detect novel fusions, we sequenced material containing two fusions common in solid tumors (EML4-ALK and SLC34A2-ROS1). After mapping reads to the consensus sequences of the fusion variants, we looked for reads spanning the breakpoints (**Figures 5A, 5B**). We additionally quantified the fusion and normal transcripts, and compared their ratios (**Figure 5C**), showing that capture preserves detection of fusions across a range of mass conditions.

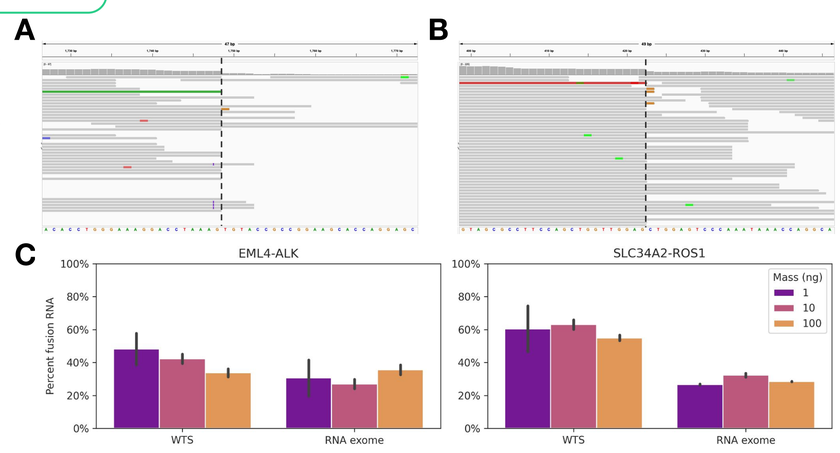

Figure 5: (A) Genome browser view of reads aligned to an EML4-ALK fusion transcript present in a cell-line derived standard - dotted black line represents the gene breakpoint. (B) Same as in (A), but for an SLC43A-ROS1 fusion also present in the cell line. (C) Ratio of fusion/normal transcripts from samples in both WTS and RNA-exome capture.

Materials and methods

To test the Twist RNA Exome panel, 1ng, 10ng, or 100ng 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 and gene quantification was performed using FeatureCounts with GenCode v41 gene annotations. Metrics were calculated using Picard CollectRnaSeqMetrics. 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 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.

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.

Patro, R., Duggal, G., Love, M. *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14, 417–419 (2017). https://doi.org/10.1038/nmeth.4197

Conflict of interest statement:
All authors are employees and shareholders of Twist Bioscience