Twist pan-cancer synthetic RNA fusion control for assay development

Patrick Cherry, Jason Corwin, Yu Cai, Kit Fuhrman, Jean Challacombe, Derek Murphy, Esteban Toro

Poster#: 247

TWIST BIOSCIENCE

1. Abstract

Fusion genes are the result of genomic structural variants that arise when the coding region of two previously independent genes become joined together and can be a major driving cause of cancer development. Many gene fusion events have been classified as clinically-relevant and are targets for both diagnostic applications, such as TMPRSS2-ERG in prostate cancer, and therapeutic applications, such as CD74-ROS1 and EML4-ALK in non-small cell lung cancer. Given the potential clinical benefits of early detection, there is significant interest in developing highly sensitive and accurate diagnostic assays to detect cancer using RNA transcripts of these gene fusions as biomarkers. A significant barrier to the development of these diagnostic assays is the lack of a reliable and renewable gene fusion positive control for use in assay development and analytical validation.

Here, we describe the development and performance of a new fusion reference material: the Twist Pan-cancer RNA Fusion Control, a highly-multiplexed positive control designed to be spiked into a selected RNA background or as a stand-alone positive control. The Twist Pan-cancer RNA Fusion Control provides a wide selection of 80+ common and curated cancer targets for analytical validation and assay development. The synthetic RNAs are designed centered around the known and documented fusion sites with 750 nt of RNA 5′- of and 3′- of the fusion junction and is suitable for both short read sequencing and qPCR experiments. Pooled synthetic fusion RNAs are quantified, pooled and normalized to similar molar ratios, analyzed for uniformity by NGS, and ddPCR'd to establish a highly precise pool concentration. Finally, in fusion detection NGS experiments, we observe over a 90% recall rate of fusion events, compared to a 100% recall rate during QC, highlighting the contribution of panel and bioinformatic approach to detection and the importance of controls.

The Pan-cancer RNA Fusion Control is designed to serve as a spike-in or stand-alone positive control and specifically pairs well with a clinically-relevant target enrichment (TE) panel within the Twist Targeted Enrichment for Gene Expression solution. This control sample positive for RNA fusions can be applied in both qPCR and NGS workflows to validate panel/probe sets, establish limits of detection, and monitor ongoing assay performance. The Twist Pan-cancer RNA Fusion Control provides a valuable one-stop solution to assay developers seeking reference materials for a wide array of cancer-associated fusion transcripts.

For Research Use Only. Not for use in diagnostic procedures.

2. Introduction & Motivation

Fusion genes are the result of genomic structural variants that arise when the coding region of two previously independent genes become joined together, such as by a chromosomal rearrangement or duplication event. Due to the large intronic spans in many coding regions of the human genome, many different DNA breakpoints can lead to similar or the same cancer-associated mRNA or protein fusion.

Fusion genes can be a major driver of cancer development. While broad genomic instability that can lead to gene fusions is a hallmark consequence on the path of oncogenesis, gene fusions can also be founding causal drivers of cancer. Many gene fusion events have been classified as clinically-relevant and are targets for diagnostic applications, such as DNAJB1-PRKACA in liver fibrolamellar carcinoma, BCR-ABL1 in myelogenous leukemia, and TMPRSS2-ERG in prostate adenocarcinoma. Additionally, many clinically-relevant gene fusion events are targets for diagnostic and/or therapeutic applications, such as CD74-ROS1 and EML4-ALK in non-small cell lung cancer.

Given the potential clinical benefits of early detection, there is significant interest in developing highly sensitive and accurate diagnostic assays to detect cancer using these gene fusions as biomarkers. Detecting these fusions at the DNA level is difficult for two reasons: (1) the exact break points are often not known or (2) the exact break points are known to vary. These issues mean that the sequencing space (such as for a target enrichment capture panel) is too large to be practically deployed in high-throughput clinical diagnostics. RNA-seq is a more efficient sequencing method for discovering fusion events in new samples.

A significant barrier to the development of these diagnostic assays is the lack of a reliable and renewable sample source harboring gene fusions of interest for use as a positive control in assay development and analytical validation. An ideal control sample would be well documented for all fusions in the design, be positive for all intended fusions, and negative for any unintended fusions. Also, a scalable solution for high-throughput applications would be ideal.

The challenges that this control material seeks to solve are:

Firstly, currently available RNA fusion positive controls do not include enough unique fusions to test modern high-throughput assays. **Secondly**, currently available RNA fusion positive controls may not precisely document the RNA fusions present in the pool (*e.g.*, with genomic coordinates). And **thirdly**, currently available RNA fusion positive controls frequently contain contaminating fusion transcripts that are not documented in the product description, potentially leading to false positives.

3. Design

We bioinformatically designed and synthetically produced all the RNA fusions within the Twist Pan-cancer Fusion Control. First, the fusion targets to be included were selected by a combination of a literature search curation and collating fusion abundance observations from databases. Next, the fusion constructs were prioritized for inclusion based on their clinical relevance, actionability in diagnostics, or in treatment availability/clinical trial availability. The bioinformatic design unambiguously informs the documentation of the fusions present in the product. Documentation of each fusion contains the left and right HGNC (HUGO Gene Nomenclature Committee), last/first exon number, breakpoints in hg19 or hg38 genomic coordinates associated with it, and additional information where available (*e.g.*, ENSEMBL and Refseq ids).

We designed the synthetic fusion RNA products to be uncapped and non-polyadenylated 1,500 nucleotide RNA molecules composed of 750 nt on each side of the fusion breakpoint (where fusion transcript is available, which was in most cases), including UTRs.

RNA Fusion design method:

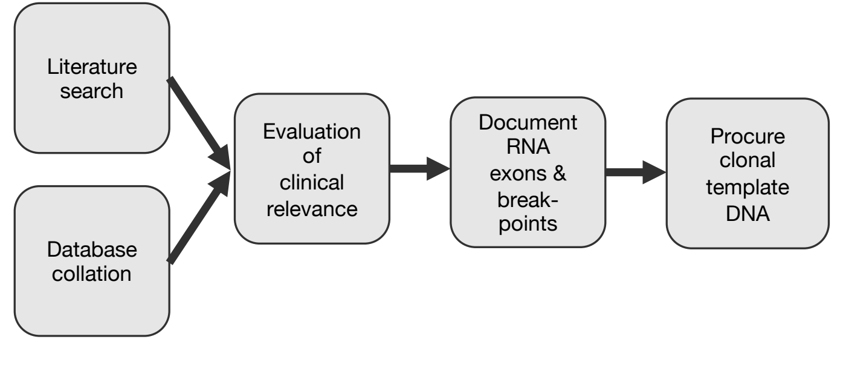

Literature search
Database collation
Evaluation of clinical relevance
Document RNA exons & break-points
Procure clonal template DNA

Figure 1: Schematic describing the process of configuring the fusion content of the control.

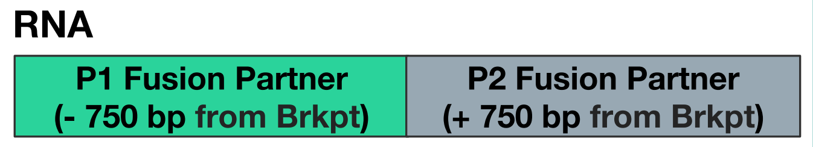
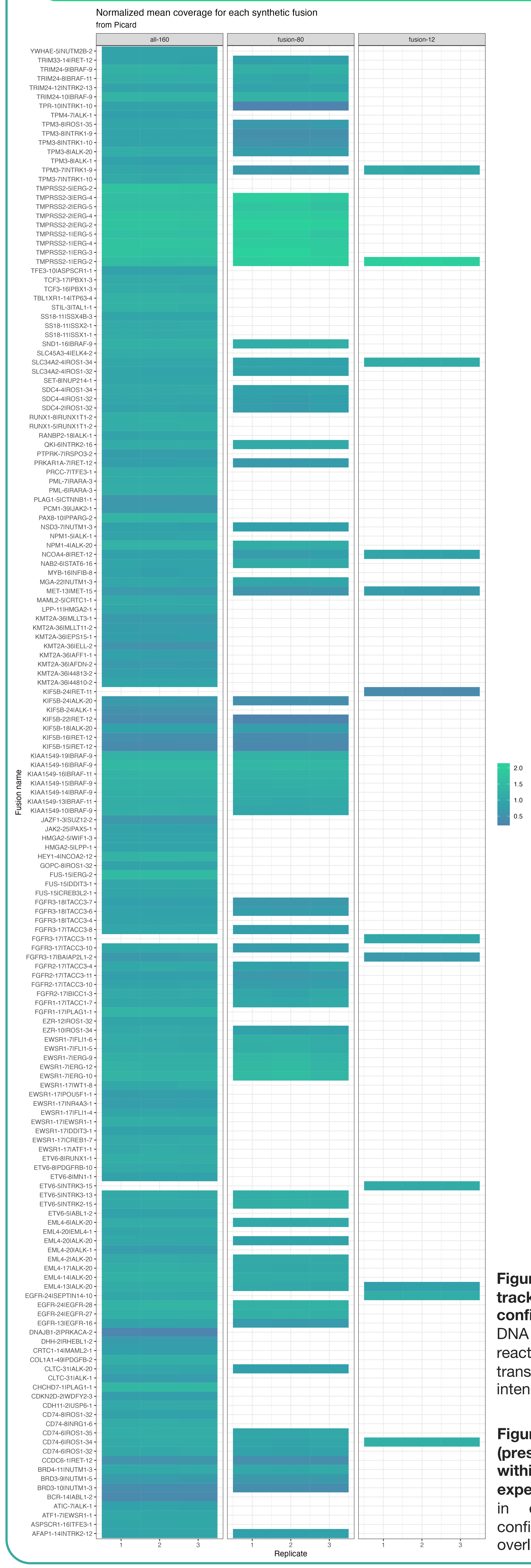

RNA
P1 Fusion Partner (- 750 bp from Brkpt)
P2 Fusion Partner (+ 750 bp from Brkpt)

Fig 2: Schematic of synthetic fusion RNA structure depicting a P1 (partner 1) fusion exon on the left, including 750 nucleotides of sequence 3′- of the breakpoint, and a P2 (partner 2).

Financial Disclosures: All authors are employees and shareholders of Twist Bioscience

Figs. 4 & 5: QC: RNA-seq of neat pools



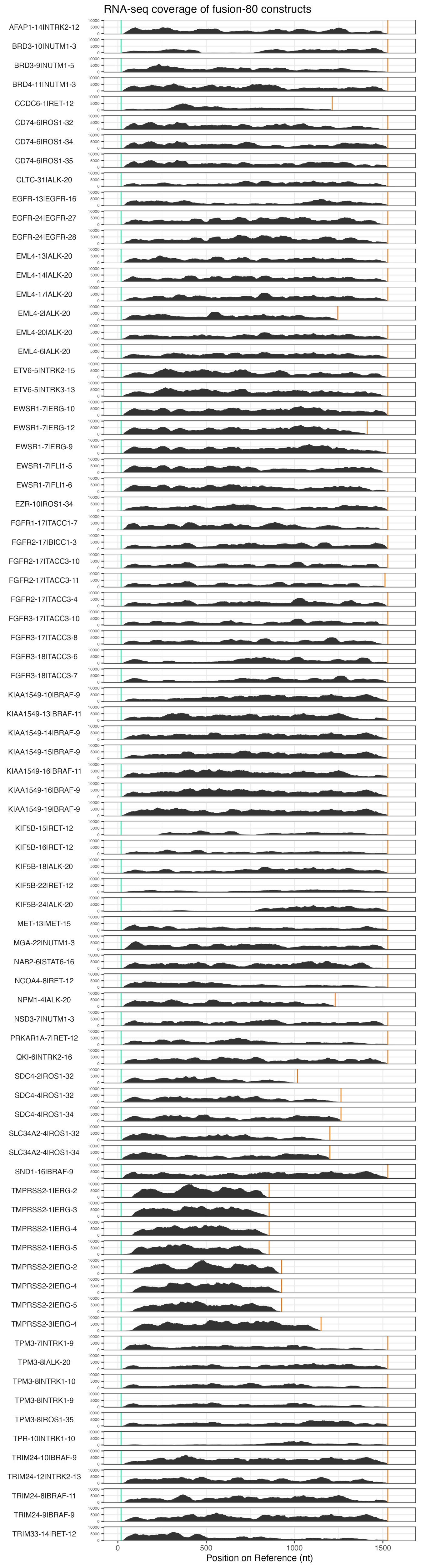


Figure 4: Above: Coverage plots of the reference tracks of one replicate of the fusion-80 configuration. Coordinates start and end with the DNA template input of the RNA transcription reaction. Green bars mark the intended transcription start site; orange bars indicate the intended transcription termination site.

Figure 3: Left: Heatmap of fusion representation (presence/absence and normalized coverage) within each sample in a neat pool QC experiment. All expected fusions spiked-in (100%) in each sample are detected, but each configuration is different and may have non-overlapping fusions in the design.

4: Quality Control

The RNA Fusion controls are extensively quality controlled during their production, both in-line (during production) and as final products. First, all template DNA that serves as input to the RNA production is analyzed by NGS to ensure the correct sequence prior to RNA synthesis. Next, post-purification RNA yields are assessed for the individual fusion RNAs. This concentration measurement also allows for equimolar normalization to be coordinated. Next, once the pool has been composed, RNA-seq libraries are prepared and sequenced; coverage, quantitative representation of the fusion sites, and lack of coverage of extraneous sequence are evaluated.

RNA-seq library prep followed by 2 x 75 paired-end sequencing on an Illumina MiSeq and bwa alignments of the data to the reference templates from which the transcripts were generated indicates that all fusions intended to be present in the control are detected (**Fig.3**), and no cross-contamination between configurations was detected. Coverage plots (**Fig. 4**) show that intended transcribed regions are present, whereas regions outside the intended regions are not present. **Table 1** below shows recall rates.

Sample	n fusions detected	n fusions in config	recall rate (%)
all-160	160	160	100
fusion-80	80	80	100
fusion-12	12	12	100

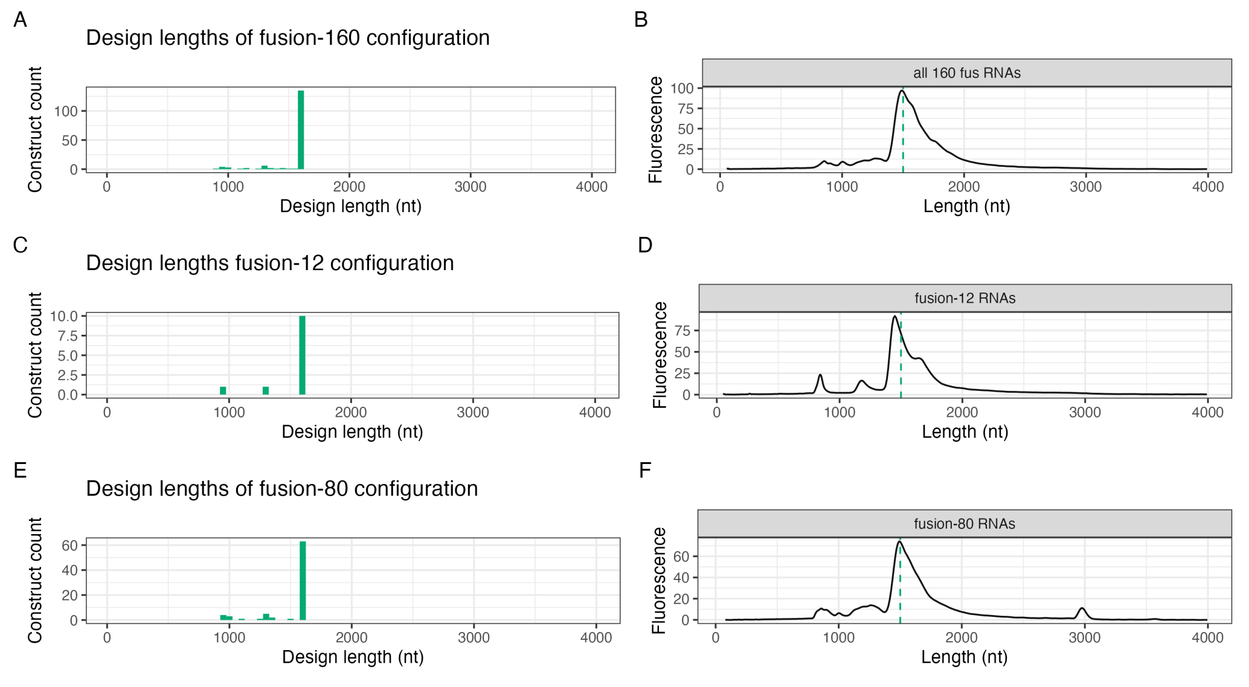


Figure 5: RNA size distributions of the pools. In pairs by rows, histograms (A, C, and E) on the left depicting the distribution of designed sizes of RNAs within each configuration. On the right, Bioanalyzer electropherograms (B, D, and F, respectively) show the measured length distributions of the indicated pool. In the electropherograms, a dashed green line marks the 1500 nucleotide size, the length of most RNA constructs.

5. Performance 1: Dilutions

In order to empirically determine and verify a reasonable dilution scheme for the neat pool of fusion RNAs into a background RNA sample, we carried out a dilution series in Universal Human Reference RNA (UHR) (Invitrogen, QS0639) in multiple configurations (fusion-12, fusion-80, and fusion -160). We used an internal RNA-seq library prep kit (with 1000 ng input total RNA) and target enrichment with an internally-developed Twist Target Enrichment (TE) panel specific for RNA fusions. Samples were sequenced on a NextSeq 550 high output kit.

We used STAR-Fusion to detect and quantify fusions (**Fig. 6**) without *a priori* information of fusion presence/absence or configuration and compared the results to the known construct designs. Additionally, we compared STAR-Fusion results with a direct string search of the exact sequence ±15bp of the breakpoint for each fusion sequence within the raw FASTQ files. STAR-Fusion (Haas 2019) is fusion identifying software that uses the STAR aligner's output of chimeric and discordant reads (and some filtering parameters) to identify fusion RNAs in a complex sample.

Quantitatively, the mean fusion fragments per million (FFPM) quantities trend down ten-fold with every ten-fold serial dilution; however, the all-160 panel, which has significant fusion content non-overlap with the TE panel used, shows more below-trend points (and more dropout), perhaps indicating inefficient capture of this subset the fusion designs. Recall is also steady for the fusion-12 configuration and reduced slightly for the fusion-80 (the configuration to which the TE panel was specifically tailored); recall rate decreases significantly for the fusion-160 configuration over the dilution series, due to the lower absolute abundance of each fusion RNA in the mixture, and likely the configuration interacting the the TE panel used in capture (which did not include probes specifically designed for all fusions in the control). Recall rates were also significantly impacted by limitations of STAR-Fusion to detect exon-skipping fusions independent of the spike-in concentration.

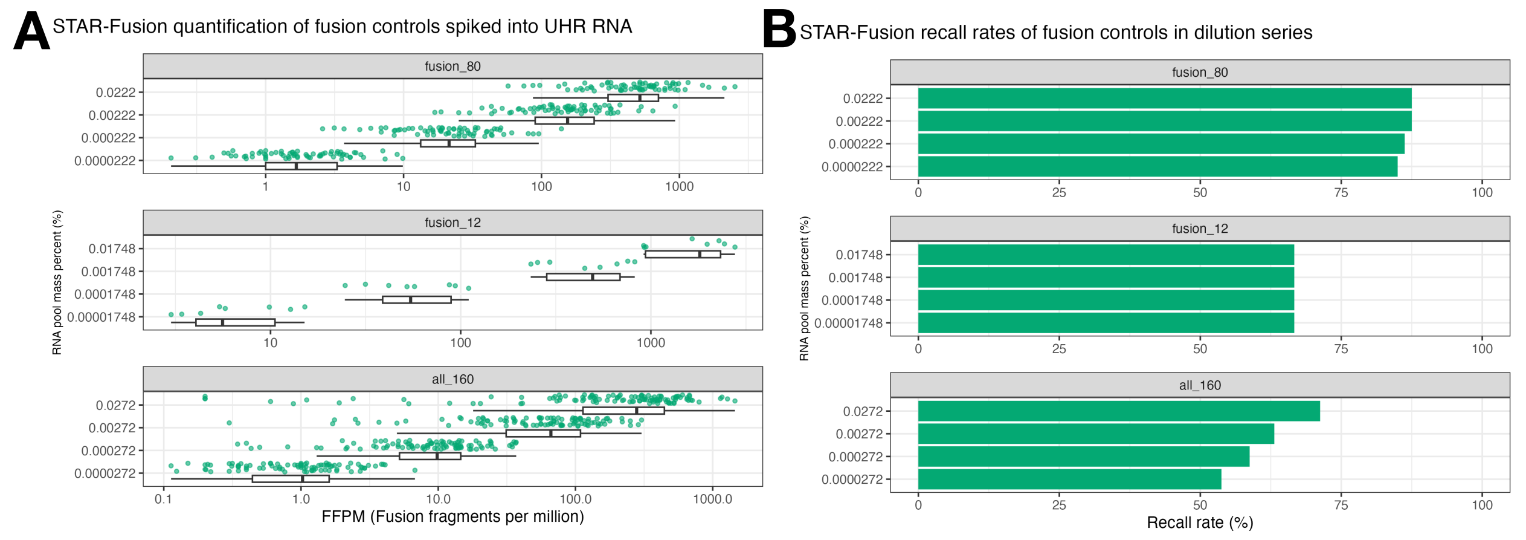


Figure 6: **A**: Categorical point plot showing the STAR-Fusion-observed FFPM (fusion fragments per million [total fragments] [a measure of abundance]) versus dilution level, measured in Fusion pool mass percent (spiked into UHR RNA). Each point represents one unique fusion. **B**: Bar plot depicting the recall rate observed from STAR-Fusion calls on the RNA-seq data.

6. Performance 2: Fusion Calling

In order to empirically evaluate and verify the standards in realistic use conditions, we carried out library preps on UHR RNA with and without the fusion-80 configuration (0.0022% by mass) of positive control spiked-in. We used an internal RNA-seq library prep kit (with 100 ng input total RNA) and TE with an internally-developed Twist TE panel specific for RNA fusions. Samples were sequenced on a NextSeq 550 high output kit.

STAR-Fusion called 72 out of the 80 fusions present in the sample as present (**Fig. 7A**), representing a 90% recall rate. Many of the fusions that are not detected are "exon skipping" fusion events, which can be missed by STAR-Fusion as splice variants, not RNA fusions.

When using string searching (**Fig. 7B**), 79 out of 80 fusions can be detected, a 98.8% recall rate. This is slightly higher than the STAR-Fusion recall rate, which illustrates the complications of bioinformatic pipeline development, the importance of prior knowledge of the exact fusion sequence expected, and the effect of bioinformatic analysis on sensitivity of fusion detection. Furthermore, even with string searching, a third-party 6-fusion standard was only able to identify three of the six fusions using the same bioinformatic methods.

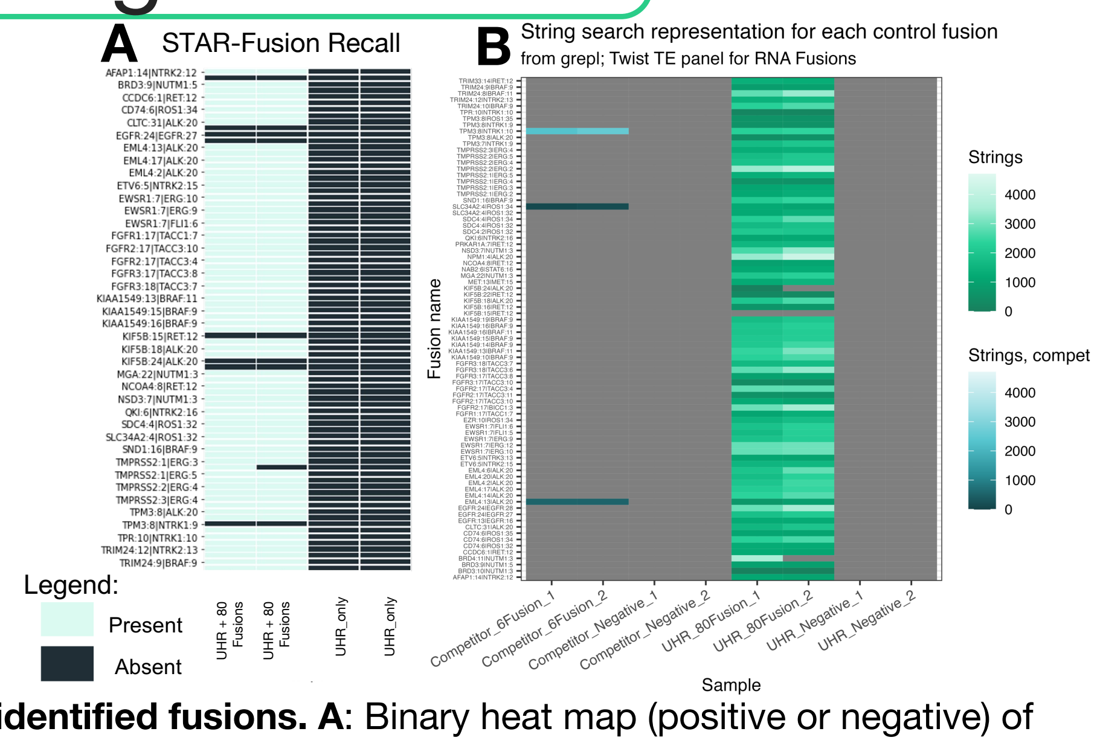


Figure 7: Heat map of identified fusions. A: Binary heat map (positive or negative) of STAR-Fusion-identified fusions in the fusion-80 configuration and in the negative control UHR RNA background material. Dark blue is not detected, and light teal is positively detected. **B**: Heap map (linear color scale) showing string-search-identified and -quantified. (Grey is not detected.)

7. Conclusions

- •The Twist RNA Fusion Controls support wet-lab and bioinformatic process testing to ensure robust RNA fusion detection.
- •QC experiments indicate the production process is making RNAs at the intended lengths, all of which are present in the final pools.
- •The configurations are expansive enough for high-throughput multiplexed evaluations of NGS assay sensitivity and can be diluted as desired.
- •The superior documentation of the fusion configurations, including precise breakpoint information, allows for straightforward verification of TE panel compatibility and verification of recall in high-throughput NGS experiments.
- •The difference in abundance and recall rates in the fusion-calling application compared to the QC experiment show the control is useful in revealing analytical gaps in an assay.