A SMARTer approach to gene fusion identification

Rachel N. Fish*, Michael Young, Sally Zhang, and Andrew A. Farmer

Takara Bio USA, Inc., Mountain View, CA 94043, USA *Corresponding author

Abstract

Chromosomal rearrangements bring together coding sequences or regulatory elements of genes that are normally separated. Many of the resulting fusion genes have been identified as key drivers of tumor growth in cancer, and the chimeric protein products may serve as specific therapeutic targets. Traditional approaches for identifying whether a gene may be involved in a fusion can be challenging to perform or difficult to interpret. Furthermore, methods such as reverse transcription PCR (RT-PCR) rely on prior knowledge of both genes involved in the fusion, while techniques like fluorescent *in situ* hybridization (FISH) and immunohistochemistry (IHC) are unable to identify the fusion partner of the gene of interest. While whole transcriptome RNA sequencing (RNA-seq) has proven to be a powerful tool in discovering new gene fusions, several challenges remain, stemming from both the complexity inherent in such large-scale sequencing and the variable quality of samples and the RNA isolated from them. The large dynamic range of the transcriptome often means that a few highly abundant transcripts account for the majority of sequencing reads while less-abundant transcripts account for only a small percentage of sequencing reads. These rare transcripts therefore require high sequencing depth to be reliably detected. Additionally, clinical samples may have RNA of unknown quality, necessitating methods that can amplify both high-quality and severely degraded RNA, such as that from FFPE (formaldehyde fixed paraffin embedded) tissue. Targeted RNA-seq 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.

We leveraged SMARTer® 5′-RACE technology for direct amplification of the 5′ ends of transcripts of interest to develop a target-specific protocol with high sensitivity and low background starting from total RNA of any quality. The resulting gene-specific priming method demonstrates considerable enrichment from 10 ng to 1 µg of a variety of degraded total RNA inputs. The protocol enables the detection of gene fusions and other structural variation in expressed RNAs without knowledge of features 5′ of the junction being investigated. For example, a lung tissue FFPE sample [from a non-small cell lung cancer (NSCLC) patient] determined to be ALK fusion positive only by FISH was shown by our assay to have a rare retention of intron 19 of *ALK*. Furthermore, we identified all targeted gene fusions in an RNA fusion control mixture at an input of 5 ng in only 200,000 sequencing reads. Altogether, our data indicate that our SMARTer approach is a sensitive tool for identification of structural variation in transcripts of interest from degraded total RNA.

1 Fusions in cancer are complex

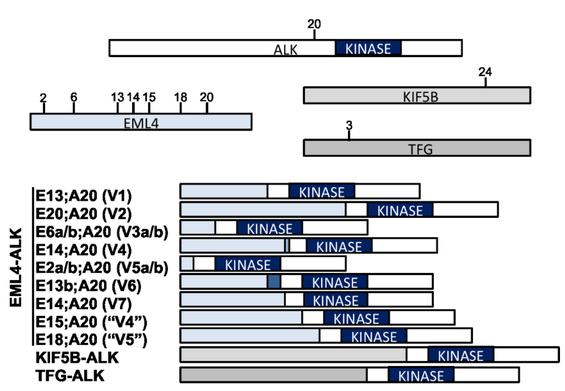

Figure 1. Schematic of *ALK* gene fusions found in lung cancer. The *ALK* gene has a common known breakpoint at exon 20. The upstream partners of *ALK* in NSCLC can vary, as shown by the figure. In addition, different portions of the upstream partner may be involved in the fusion to the *ALK* gene. Image: Lovly et al. (2014)

2 SMARTer RNA Fusion Kit amplifies all sequences upstream of a location of interest

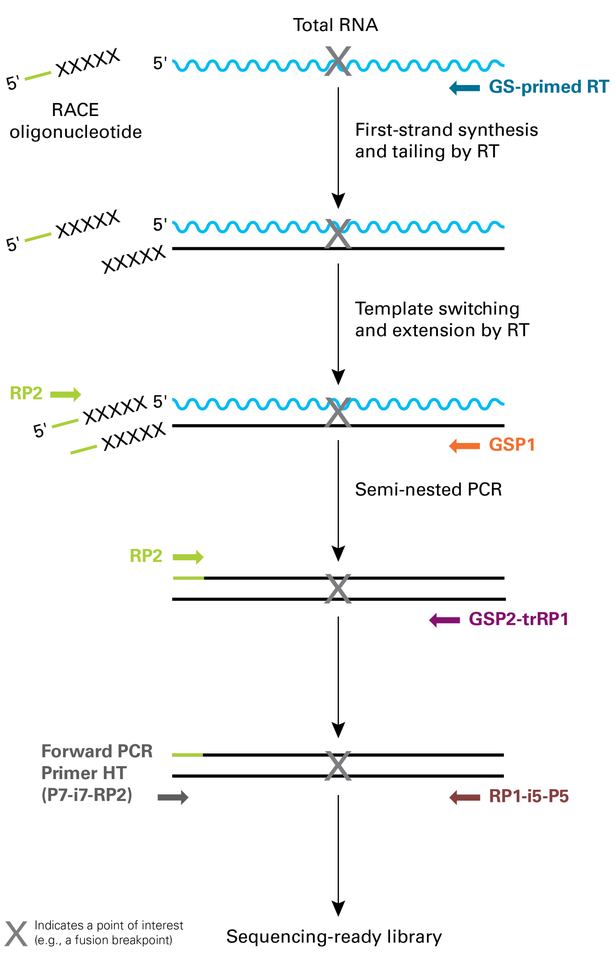

Figure 2. Schematic of SMARTer fusion RNA chemistry. The SMARTer fusion RNA protocol uses a 5′-RACE-based approach to amplify specific targets with minimal bias to detect fusion genes. No ligation is used to add adapter sequences, and no prior knowledge of features 5′ of the junction being investigated is required. Total RNA of any quality can be used as input, and no rRNA removal is required. A gene-specific RT generates the first-strand cDNA. Non-templated nucleotides are added when the RT reaction reaches the other end. A RACE oligonucleotide hybridizes to the non-templated nucleotides and incorporates adapter sequences. This oligo provides a new template for the RT to use, thereby allowing it to extend to the 5′ end of the fragment. This is followed by two rounds of nested gene-specific PCR and a final PCR reaction that adds Illumina® indexes.

3 High on-target performance and identification of known fusions

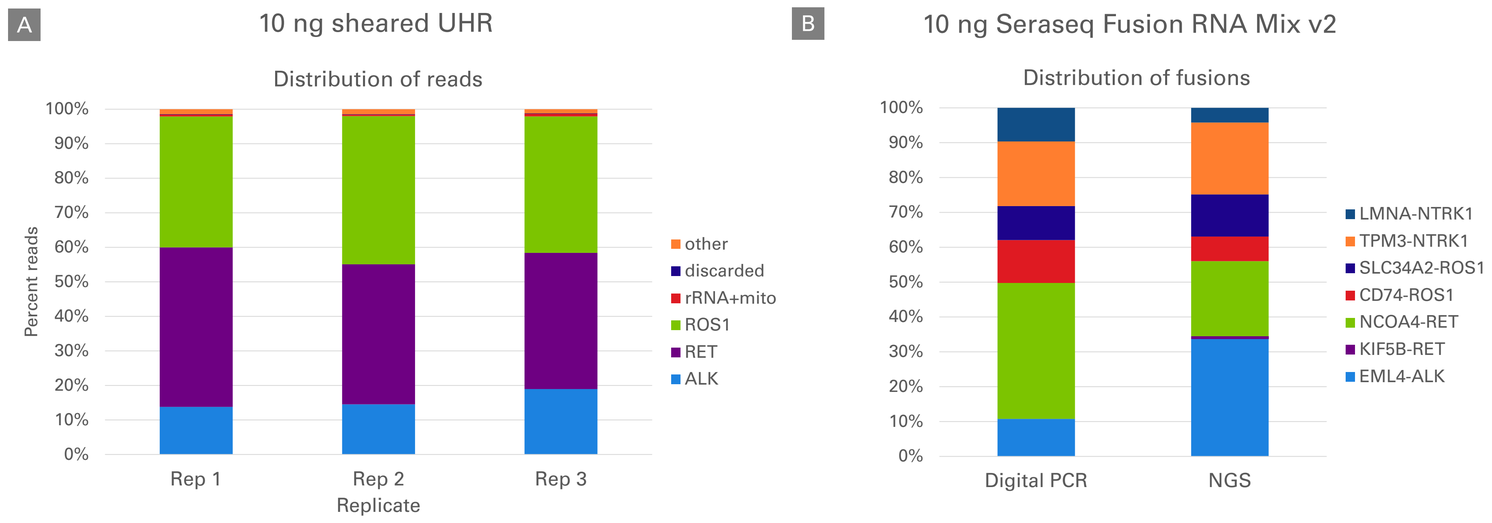

Figure 3. Enrichment of fusion genes. 10 ng RNA was used as input into the SMARTer RNA Fusion Kit's protocol. Gene-specific RT and two rounds of gene-specific nested PCR were all multiplexed with primers for the indicated genes, and for common breakpoints in each of these genes. Following the final PCR, the libraries were sequenced on a NextSeq® 500 or MiniSeq™ instrument at 2 x 75 PE. Sequencing data was down-sampled to 200,000 reads and analyzed using CLC Bio v10 software. **Panel A.** Chemically sheared (2.5 minutes at 94°C in Fragmentation Buffer) Universal Human RNA (Takara Bio qPCR Human Reference Total RNA, Cat. # 636690) was used as input RNA. Sequencing data shows a high on-target percentage and a relatively consistent distribution of reads. Very few reads are discarded or mapped to other sequences. **Panel B.** 10 ng of Seraseq Fusion RNA Mix v2 was used as input RNA. This is a control RNA developed by SeraCare to include synthetic transcripts for 15 clinically significant fusions prevalent across different types of solid tumors in a wild-type background. This includes *ALK*, *RET*, *ROS1*, and *NTRK1* fusions. It is designed to mimic a lightly fixed sample, so the RNA is slightly fragmented (RIN 6.6). SeraCare used digital PCR to quantify the amount of each of the fusion transcripts in a wild-type background. The distribution of only the genes targeted in this experiment are shown. The overall on-target rate was 91%. (i.e., 91% of the 200,000 reads mapped to the targeted gene or a fusion with that gene). The distribution is similar to that determined by digital PCR.

4 Targeted genes are enriched and rearrangements identified in FFPE samples from NSCLC patients

A

% Reads to *ALK*	FFPE1 *ALK+*	FFPE2 *ALK-*	FFPE3 *ALK+*
Targeted (SMARTer fusion RNA protocol)	96%	97%	64%
Non-targeted (random-primed WTA protocol)	0.012%	0.015%	0.0019%

B

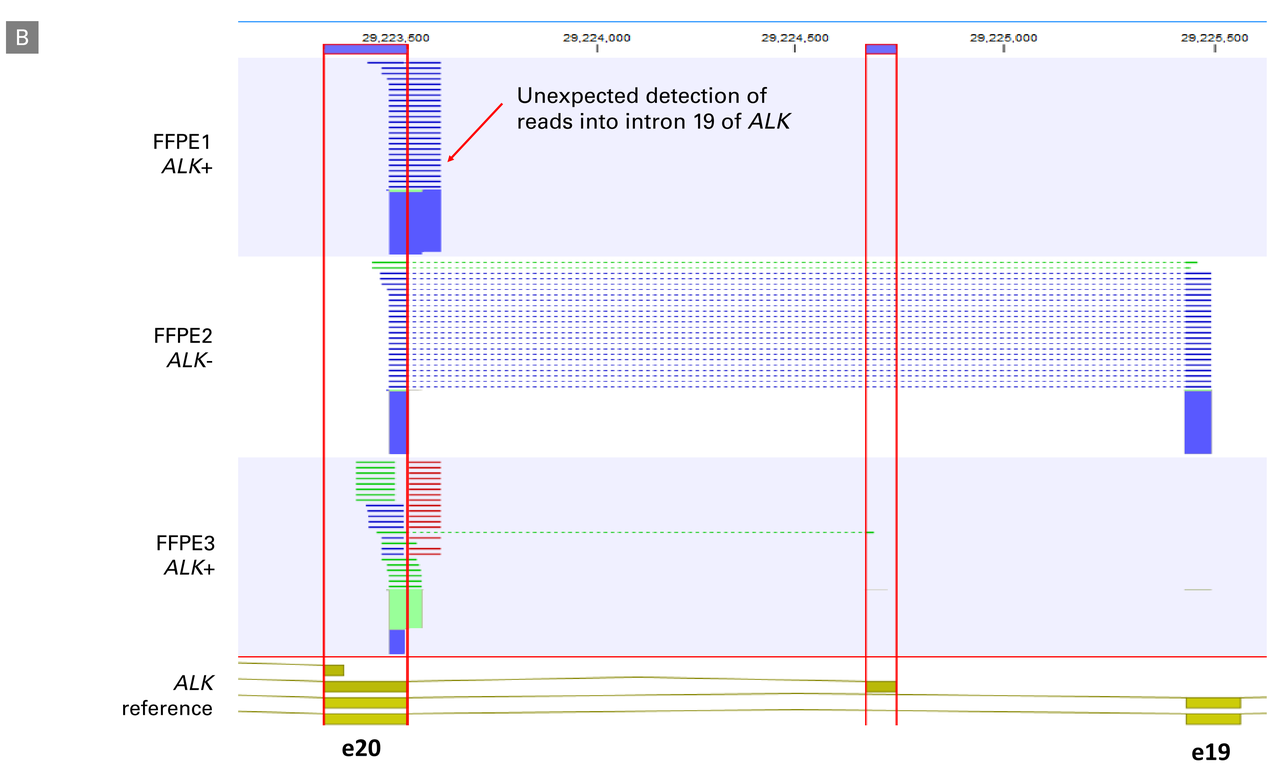

Figure 4. Detection of fusion genes. FFPE samples were obtained from lung tissue from three different patients. All three were diagnosed with non-small cell lung cancer (NSCLC). Two had an ALK positive status and one was ALK negative. The ALK status was determined by FISH. RNA extracted from the FFPE tissue had DV200 of 24–68. NGS libraries were prepared following the SMARTer RNA fusion protocol targeting *ALK* or using a random-primed kit for whole transcriptome amplification (WTA). The resulting libraries were sequenced and analyzed as described in Figure 3. **Panel A.** Targeting resulted in a significant increase in the number of reads going to the gene of interest (64% to 97% vs. 0.0019% to 0.015%). **Panel B.** Alignment of sequencing reads to the *ALK* gene are shown (hg38 reference sequence). For the *ALK*- sample (FFPE2), there was no detected fusion, and our reads extended to exon 19 as expected. When there was an expected translocation, the reads started in exon 20 but were not shown in exon 19. Surprisingly, reads extending into intron 19 of *ALK* were detected in the two *ALK*+ samples. This was unexpected, as one would expect to see exon-exon reads only. *ALK* is 3′ to 5′ on the reference sequence. The SMARTer fusion RNA primers are in exon 20 and extend 5′ towards exon 19.

5 Identification of a rare subtype of an *ALK* fusion event

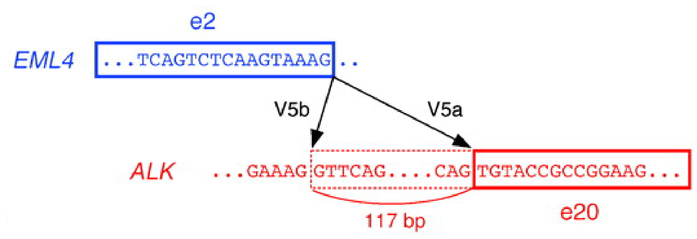

Figure 5. Confirmation of *ALK* intron 19. Sanger sequencing of the final PCR product was used to confirm the presence of *ALK* intron 19 upstream of *ALK* exon 20 in the sample. BLASTing the sequence showed that there is a known *EML4-ALK* fusion that includes a portion of intron 19 of *ALK*. Our FFPE samples were presumably too fragmented to include the fusion partner sequence. This fusion gene has been previously reported by Takeuchi et al. (2008). It is formed as the result of a small inversion within the short arm of chromosome 2 that joined exon 2 of *EML4* to a position within intron 19 of *ALK*, 117 base pairs upstream of exon 20. The EML4-ALK protein thus contains the amino-terminal half of EML4 and the intracellular catalytic domain of ALK. As concluded by Takeuchi et al., replacement of the extracellular and transmembrane domains of ALK with this region of EML4 results in constitutive dimerization of the kinase domain of ALK and a consequent increase in its catalytic activity.

Conclusions

Library generation without adapter ligation or known sequences

- SMART® technology in combination with a 5′-RACE-based approach allows for unbiased capture of all structural variation upstream of a location of interest

Discovery of fusions in cancer research samples

- Rare fusion events identified in degraded RNA and RNA purified from FFPE tissue without ribosomal RNA removal

Confirmation of fusions in a control fusion sample

- This approach identifies all targeted fusions in a reference material for fusion RNA

References

Lovly, C., L. Horn, W. Pao. ALK in Non-Small Cell Lung Cancer (NSCLC). *My Cancer Genome* https://www.mycancergenome.org/content/disease/lung-cancer/alk/ (2014).

Takeuchi, K. *et al*. Multiplex reverse transcription-PCR screening for EML4-ALK fusion transcripts. *Clin Cancer Res.* **14**(20), 6618-24 (2008).

Takara Bio USA, Inc.
United States/Canada: +1.800.662.2566 • Asia Pacific: +1.650.919.7300 • Europe: +33.(0)1.3904.6880 • Japan: +81.(0)77.565.6999

Scan to download your copy of this poster and view additional posters, or visit takarabio.com/NextGenDx-SMARTer-fusion

800.662.2566
Visit us at www.takarabio.com

Clontech TaKaRa cellartis