applied biosystems
by life technologies

Effective Strategies for Exon/Exon Junction Mapping and Fusion Transcript Detection

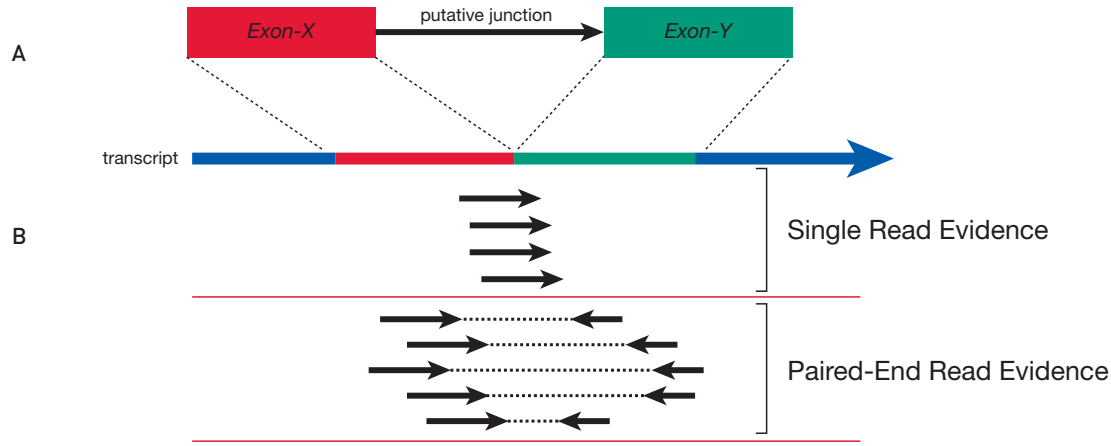

Figure 1. Mapping Exon/Exon Junctions Using Single Reads or Paired-End Reads. (A) Splicing can occur between two exons from the same gene, neighboring genes, or following fusion between two physically separated genes that are spliced to form a fusion transcript. (B) Exon/exon junctions can be mapped with either single reads or paired-end reads.

New RNA sequencing methods using high-throughput sequencing platforms such as the SOLiD™ System have revolutionized the study of RNA structure and advanced the discovery of both novel forms of known transcripts and novel transcripts arising from variations in chromosomal structure. Novel forms of known transcripts can be in the form of alternatively spliced exons of a gene or exons from other genes in the same region of the genome. Additionally, chromosomal rearrangements or translocations form novel fusion transcripts such as the BCR/ABL fusion in chronic myelogenous leukemia (CML). Detection of these novel events is driven by the number of uniquely mapped sequences generated, configuration of the sequencing reads, and properly tuned analysis software. This application note outlines how the integration of the SOLiD™ Total RNA-Seq Kit with the massively parallel sequencing output of the SOLiD™ System and SOLiD™ BioScope™ software 1.2.1 enables discovery of these events with extremely high validation rates.

Materials and Methods

In all experiments described, the sample used was 100 ng of poly(A)-selected RNA from Universal Human Reference RNA (Agilent). The SOLiD™ Total RNA-Seq Kit (P/N 4452437) was used to construct barcoded libraries following the standard protocol with the SOLiD™ RNA Barcoding Kits as specified in the appropriate protocol. Approximately 150 bp inserts were selected for the libraries. The resulting libraries were amplified onto beads by ePCR following standard protocols. Paired-end sequencing runs of 50 bp x 25 bp were conducted on a SOLiD™ 4 System, and the resulting sequences were mapped using SOLiD™ BioScope™ software 1.2.1.

In SOLiD™ BioScope™ 1.2.1, exon/exon junctions are called "present" when two unique single reads map to the junction; or, if paired-end reads are combined with single reads, one unique single read and one unique paired-end read must map to the junction. In order for a putative fusion transcript to be called 'present', two unique single reads must map to the junction; or if paired end reads were combined with single reads, two unique single reads as well as two unique paired-end reads must map to the junction sequence. This combination was found to be optimal for calling exon/exon junctions and fusion transcripts with a low false-positive rate in the samples used. However, you can adjust SOLiD™ BioScope™ software to be more or less stringent depending on your experimental needs.

Results

The SOLiD™ Total RNA-Seq Kit workflow is similar to that of the Whole Transcriptome Analysis kit it replaces. However, in order to obtain optimal libraries for paired-end sequencing, it is necessary to select excised cDNA molecules of 150–200 nt in size. The insert size must be longer than the combined length of both the forward (50 bp single-read) and reverse (25 bp paired-end) reads in order to assure optimal mapping of both reads. Libraries with shorter cDNA molecules are satisfactory for single 50 bp sequencing.

Fifty base single reads are satisfactory for mapping to the genome and exons. However, in order to have the highest confidence for mapping a single read to both exons of an exon/exon junction requires that the junction to be close to the middle of the read. Reads meeting this criteria will represent only a small percentage of the total reads. A schematic representation of how single reads can be used to map exon/exon junctions is shown in Figure 1B. The advantage of single reads is the direct evidence that two regions that are physically distant on a genomic map are now joined. On the other hand, paired-end reads have the advantage of spanning larger genomic regions due to the fact that the two reads are separated by a known distance and there are a relatively large number of reads that can span a putative exon/exon junction, thus increasing the chances of detecting exon/exon junctions. However, the presence of an exon/exon junction is inferred and not directly detected because the mapping results obtained from the two reads are inconsistent with the expected mapping from the reference sequence. This is also illustrated in Figure 1B.

Detecting Exon/Exon Junctions

The ability of single reads and paired-end reads to detect exon/exon junctions was directly studied by research scientists at Life Technologies. Multiple RNA-Seq libraries were constructed from UHR poly(A) RNA, barcoded using RNA barcode kits, and sequenced using 50 bp forward reads and 25 bp reverse reads (50 bp x 25 bp). The reads were mapped, and exon/exon junctions were called using SOLiD™ BioScope™ software 1.2.1 for individual libraries as well as reads pooled from two of the samples. Putative exon/exon junctions found in the NCBI database are referred to as "known". The number of RefSeq exon/exon junction detected for each data set was graphed. The results are shown in Figure 2.

Figure 2 shows approximately 80,000 known junctions (50% of the total known exon/exon junctions) detected by 16–17 M uniquely mapping single 50 bp reads. The number of junctions detected goes up to 100,000 (61% of total) when single reads from the two libraries are pooled (Figure 2, bar labeled "S1&S2/SR"). Additionally, there are a number of putative novel exon/exon junctions detected in the samples as shown by the dark blue regions.

When the same number of 50 bp x 25 bp paired-end reads are analyzed in SOLiD™ BioScope™ software, more known exon/exon junctions are detected as shown in the S1/SR+PE bar; and S2/SR+PE bar, where approximately 100,000 (60% of total) known junctions are detected. In the case of the pooled paired-end sequences (S1+S2/SR+PE), approximately 119,000 (73%) of known exon/exon junctions are called. Additionally, a larger number of putative novel junctions are called in this sample.

Therefore, while single reads can detect exon/exon junctions, the same number of paired-end reads detects approximately another 20,000 junctions. Additionally, many more putative novel exon junctions are identified when paired-end reads are used for detection.

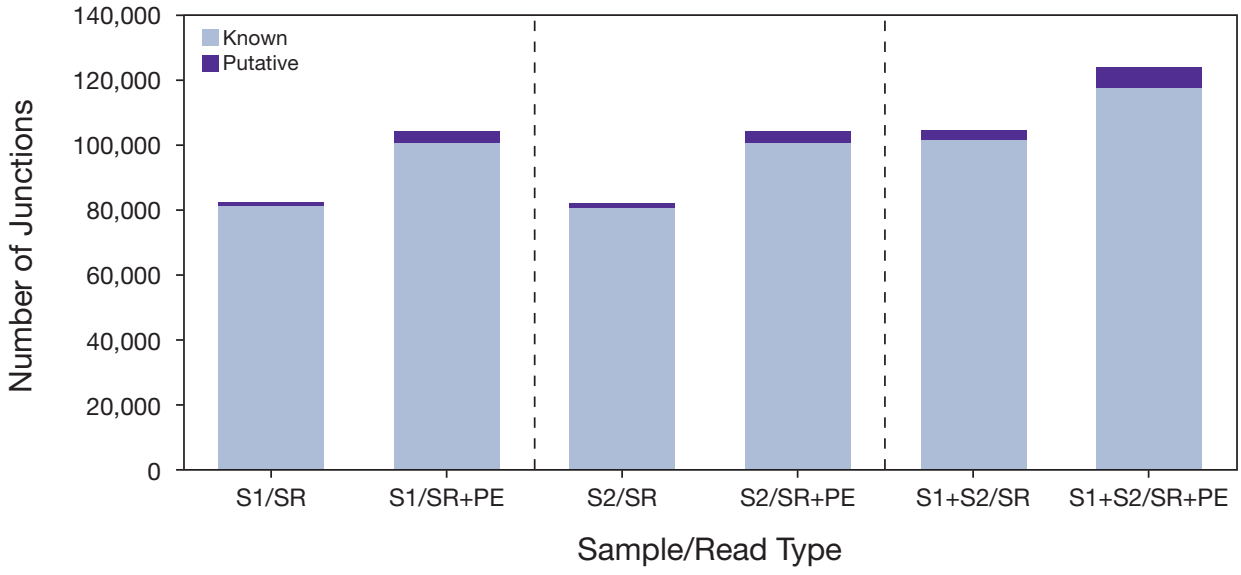

Figure 2. Detection of Known Junctions and Putative Junctions. Each library yielded approximately 16 (S1) or 17 million (S2) uniquely mappable reads, and the pooled sample (S1&S2) represents 33 million uniquely mappable reads. The y-axis represents the number of exon/exon junctions detected. The total number of known exon/exon junctions are approximately 163,000, RefSeq (hg18).

Using Paired-End Reads to Detect Fusion Transcripts

Detection of fusion transcripts is an extension of the rationale for detecting exon/exon junctions. Fusion transcripts arise by chromosomal rearrangements where transcripts on different chromosomes are produced as one transcript, and the exons from each transcript are now spliced together to form a novel exon/exon junction. This was first shown to be the case in the BCR/ABL translocation found in CML patients [1]. The fusion protein has been shown to be an effective target for imatinib mesylate therapy. Since this work, fusion transcripts and fusion proteins have been shown to exist in a number of cancers [2]. The UHR RNA samples are ideal for demonstrating the ability to detect fusion transcripts, since a number of these transcripts have been previously detected and a small number have been subsequently validated [2].

The same samples used for the data in Figure 2 were analyzed with SOLiD™ BioScope™ software 1.2.1; but this time, putative fusion transcripts were identified using either single reads (SR) or a combination of paired-end and single reads (PE+SR). The result of this analysis is shown in Figure 3.

Again, it is clear that single reads are capable of detecting putative fusion transcripts—58 or 67—in the two individual sample data sets. However, when putative fusion transcripts were identified using paired-end reads in combination with single reads, only 9 or 10 were detected in the same data sets. In the pooled data, 169 putative fusions were called by single reads, but 16 were called using paired-end and single reads. The fusions called using a combination of the two read types are always a subset of the fusions called using single reads. This suggests that the combination of reads does not find new putative fusions that were not previously identified by single reads only. In other words, single reads have a low false negative rate. False positives can be eliminated by subsequent validation, which requires additional time and expense. The ideal case is where all true positives are detected, but there is a low false positive rate.

Because it comprises RNA pooled from 10 different tumor lines (one of them, CML, is known to contain at least the BCR/ABL gene fusion [1]), UHR is an excellent sample for detection of fusion transcripts. The study of this RNA has shown the presence of at least one additional validated fusion transcript (GAS6/RASA3). Therefore, the effectiveness of our software can be demonstrated using these known fusion transcripts. The 16 putative fusion transcripts identified in Figure 3 were then selected for validation using custom TaqMan® assays designed to the sequences surrounding the junctions.

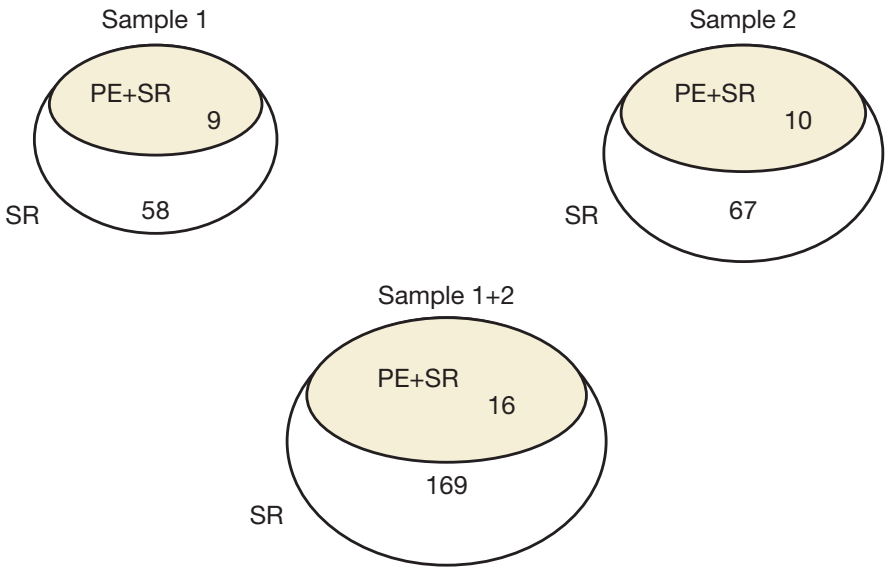

Figure 3. Fusion Transcripts Detection. Single reads (SR) or paired-end + single reads (PE+SR) were analyzed for possible fusion transcripts using SOLiD™ BioScope™ software 1.2.1 with default parameters in each data set.

Table 1 shows the results of the validation of the 16 putative fusion transcripts called in UHR RNA.

Two of the putative fusion transcripts identified in UHR by SOLiD™ BioScope™ software 1.2.1 were the BCR/ABL and GAS6/RASA3 fusions previously shown to be present by another study [2] and validated here by TaqMan® assays. There were four putative fusion transcripts for which custom TaqMan® assays could not be designed. There are several explanations for this, including failure to find satisfactory primer/probe sequences in the region immediately flanking the junction or multiple possible priming sites in the genome, among others. Ten of the twelve custom assays designed showed significantly lower C_t scores relative to the No Template Control samples run at the same time, strongly suggesting they are true fusion transcripts. Eighty-three percent of the putative fusion transcripts identified by SOLiD™ BioScope™ software 1.2.1 that had TaqMan® assays were validated. This is an excellent validation rate for fusion transcripts. Even if we include putative fusions for which no custom assays could be designed, 62% of all putative transcripts were validated using TaqMan® assays. Each of these values is at least 10-fold better than if putative fusions were called by single reads alone.

Conclusions

Accurate and efficient detection of exon/exon junctions—regardless of whether they are alternative splicing events or fusion transcripts resulting from structural variation—is an increasingly important field of study. The data presented here demonstrate the ability of the SOLiD™ Total RNA Seq-Kit plus the SOLiD™ 4 System (generating 50 bp x 25 bp paired-end reads) and SOLiD™ BioScope™ software 1.2.1 to detect known and putative novel exon/exon junctions in a reference RNA. Known and putative fusion transcripts are detected with an 83% validation rate using TaqMan® assays.

Table 1. Validation of Putative Fusion Transcripts.

METRIC	RESULT	BCR/ABL	GAS6/RASA3
Putative Fusion Detected by Single Reads	169	+	+
Putative Fusions Detected by Single Reads + Paired-End Reads	16	+	+
Successful Assays Designed	12	+	+
Fusions Validated With TaqMan® Assay	10	+	+
% Fusions Validated With TaqMan® Assays	83% (10 of 12)	--	--
% Overall Validation of all Putative Fusions	62% (10 of 16)	--	--
% Overall Validation of Putative Fusions from Single Reads	6% (10 of 169)	--	--

References

1. Muller AJ, Young JC, Pendergast AM et al. (1991) BCR first exon sequences specifically activate the BCR/ABL tyrosine kinase oncogene of Philadelphia chromosome-positive human leukemias. *Mol Cell Biol* 11(4):1785–1792.
2. Maher CA, Palanisamy N, Brenner JC et al. (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A* 106(30):12353–12358.

Headquarters
5791 Van Allen Way | Carlsbad, CA 92008 USA | Phone +1 760 603 7200 | Toll Free in North America 800 955 6288
www.lifetechnologies.com

life technologies™