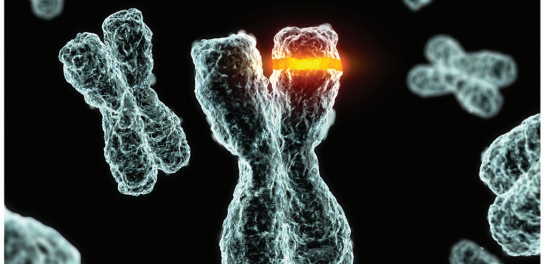
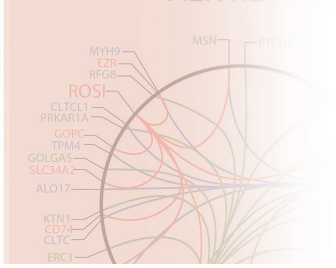




Technical Note

Fusion Detection in Archer™ Analysis Software

PN-MKT-0008

Description

This Technical Note contains a description of the fusion detection algorithm in Archer Analysis. The fusion detection algorithm in Archer Analysis relies on the specificity of the gene specific primers (GSPs) used in the amplification steps in the Anchored Multiplex PCR (AMP™) process. After the libraries are created, the samples are sequenced and the resulting FASTQ files are analyzed by the Archer Analysis software application.

This Technical Note describes the process in version 3.1, released Jan 2015.

Key Points

Gene fusion detection

Detect gene fusions in total nucleic acid samples with confidence

Increased accuracy

The Archer Analysis software makes optimal use of Archer's Anchored Mulitplex PCR technology to detect gene fusions

Visualization

Visualize gene fusion candidates in the JBrowse Genome Browser directly from the Archer Analysis application webpage

Summary of the Fusion Detection Algorithm

1. Adapter trimming of the reads
2. De-duplication of the reads using the random 8-mer molecular barcode
3. Mapping of reads to control regions
4. Mapping of reads to target regions
5. Mapping of reads to the human genome
6. Reads spanning two separate genes are considered fusion candidates
7. Binning of reads with the same breakpoint and initial consensus creation
8. Mapping of reads back to the initial consensus sequence
9. Annotations of fusion partners with information from RefSeq database
10. Applying filters to fusion candidates to remove false positive

Some of the steps are described in more detail below.

2. De-Duplication of the reads using the molecular barcode

The partially-functional Y-adapter used in the library creation contains a random 8-mer sequence (also known as a molecular barcode or MBC) and the ligation of these adapters to the fragments in the library allows for detection and removal of PCR duplicates which is important for many aspects of fusion and mutation detection.

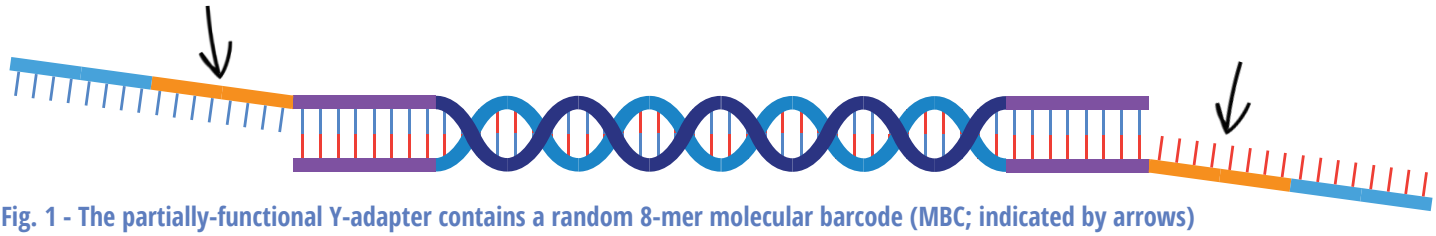


Fig. 1 - The partially-functional Y-adapter contains a random 8-mer molecular barcode (MBC; indicated by arrows)

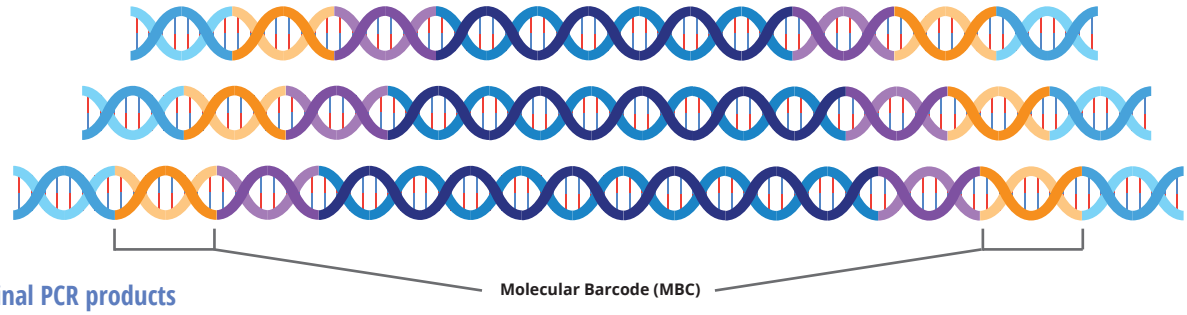


Fig. 2 - The final PCR products

Reads with the same molecular barcode are consolidated into a single consensus read, since they are considered PCR duplicates. Reads with different molecular barcodes have a very high probability to be derived from different original fragments. Fusion candidates identified from multiple original fragments provide higher confidence that the fusion is a real event and not a random error event or sequencing artifact.

3, 4 and 5. Mapping of the reads

Archer Analysis software makes optimal use of the fact that the Archer AMP technology is a targeted sequencing technology. To increase mapping specificity of the reads to the correct location on the human genome, the Archer Analysis software first maps the reads to the targeted locations on the genome before mapping the remainder of the reads to the human genome. The first reads that are mapped are to the control targets defined for this assay. Secondly, reads that do not map to any of the control targets are mapped to the target regions of the assay. Lastly, any remaining reads are mapped directly to the human genome.

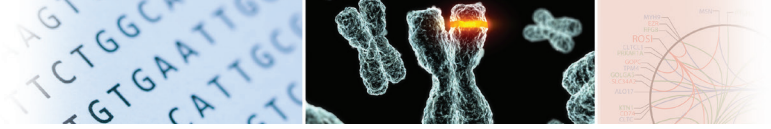
6. Gene fusion detection

The software requires a single read spanning two separate genes to be considered a fusion candidate. At least 23 bp need to be mapped on either side of the apparent breakpoint to be a valid fusion candidate read.

NOTE: Reads from a paired-end read library where each read maps completely to a single gene but each to a *separate gene* is NOT considered a fusion candidate due to the high false positive rate of these “chimeric” reads.

7 and 8. Binning of reads and consensus creation

Each fusion candidate read that spans the same apparent breakpoint between two genes is grouped together and an initial consensus sequence is constructed by concatenating the two (or more) reference sequence fragments that are spanned by the supporting reads. The original fusion candidate reads are mapped back to this initial consensus sequence to construct the final consensus sequence (including any InDels or mutations).



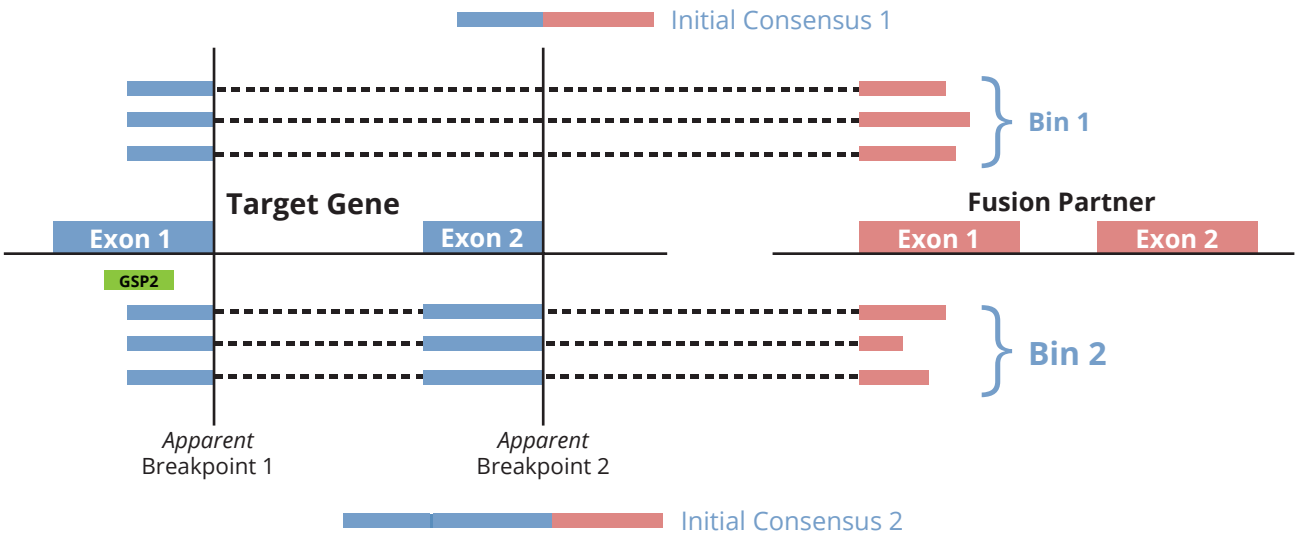


Fig. 3 - Reads with the same apparent breakpoint are binned together to create an initial consensus sequence

9. Annotations of fusion partners

The final consensus sequence is used to annotate the two (or more*) fusion partners. The consensus sequence is compared to the human genome with BLAST and the annotations from the RefSeq database are used to annotate the final fusion partners.

10. Applying filters to remove false positives

To separate the potential false positive fusion calls from the true positive fusion calls, a set of filters is applied to the fusion partners. If any of the following findings are *true*, the fusion is considered a potential false positive and categorized as a fusion candidate candidate with "weak evidence".

A fusion candidate is considered to be a potential false positive if any one of the following is true:

1. Fusion partner is NOT a known fusion partner of the target gene
2. Fusion partner is similar to GSP2 for the target gene
3. Fusion partners are NOT exon-exon fusions

Weak evidence fusions are still reported, but are binned into a separate category on the Results page.

Results

Two sequencing libraries were created from Ambion® Human Lung RNA and from a mix of the RNA from 3 cell lines (H2228, LC-2/ad and HCC78) using the Archer FusionPlex™ ALK/RET/ROS1 panel, sequenced on an Illumina MiSeq® and analyzed with Archer Analysis version 3.0.

* The final consensus can combine pieces of two or *more* fusion partners. Since each exon of a gene is annotated separately, multiple isoforms of fusion partners are considered separate fusion partners.

The result of the Analysis is shown in Figure 4. The TriplePOS sample shows the expected fusions (SLC34A2-ROS1, CCDC6-RET and EML4-ALK, the latter shown twice since two isoforms were detected) while the negative control shows no signs of any gene fusions.

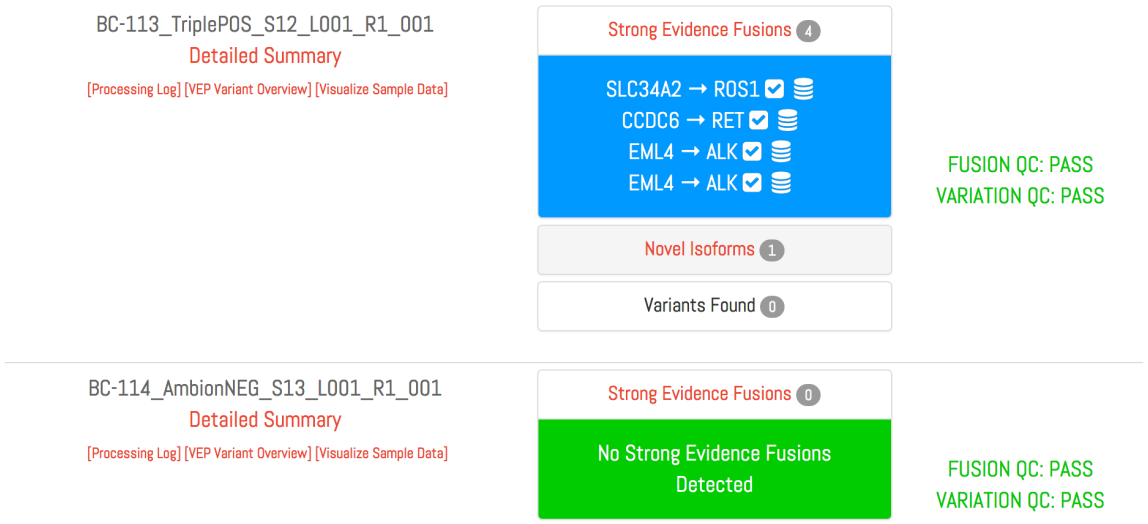


Fig. 4 - Gene fusion analysis results in the column marked "Strong Evidence Fusions". The third column shows the Quality Control filter status. BC-113 is the Triple Positive cell line mixture, BC-114 is the Ambion Lung negative control sample

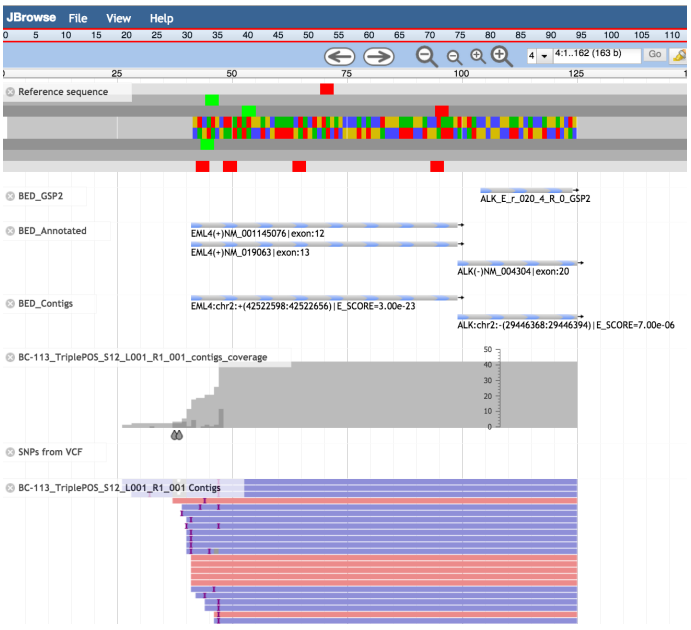


Fig. 5 - Visualization of the reads supporting the gene fusion in the JBrowse genome browser

To verify the results of the fusion detection algorithm, the de-duplicated reads mapped to the initial consensus sequence were visualized directly in the Archer Analysis software using the JBrowse genome browser (Genome Res. 2009. 19: 1630-1638). Figure 5 shows that there is a number of reads that span the breakpoint of the ALK-EML4 fusion candidate.

ArcherDX
2477 55th Street, Suite 202
Boulder, CO 80301
(877) 771-1093

For more information visit www.archerdx.com