pbfusion: Detecting fusion genes and other transcriptional abnormalities from PacBio Iso-Seq data

Abstract # 9285

Roger Volden[1], Zev Kronenberg[1], Aaron Gillmor[2], Ted Verhey[2], Michael Monument[2], Donna Senger[2], Harsharan Dhillon[1], Jason Underwood[1], Khi Pin Chua[1], Elizabeth Tseng[1], Daniel Baker[1], Primo Baybayan[1], Michael A. Eberle[1], Jonas Korlach[1], Sorana Morrissy[2]

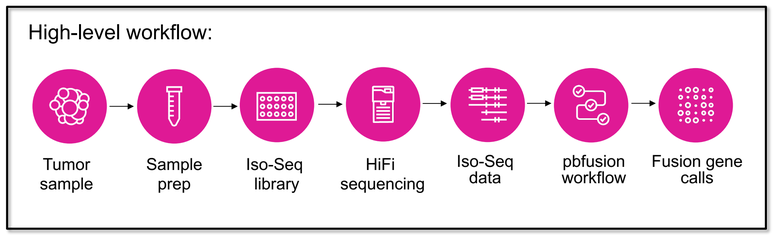
[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

[2]Charbonneau Cancer Institute, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

Abridged Abstract

Sarcomas are a broad group of soft tissue and bone cancers that can be difficult to treat leading to a high mortality rate. Sarcomas comprise two broad genomic classes: (1) simple karyotypes, where a single oncogenic structural variant (SV) clonally expands a subtype that is diagnostic and relevant to tumor burden tracking; and (2) complex karyotypes, with genomic instability resulting in heterogeneous cellular subtypes. Fusion genes have been discovered in one third of all sarcomas and are drivers of disease progression. PacBio full-length RNA sequencing (Iso-Seq method) detects fusion transcripts. However, to date, there are few dedicated long-read fusion detection tools, and none take advantage of the high read accuracy from Iso-Seq data. We present pbfusion, a fusion detection tool designed specifically for PacBio Iso-Seq data and apply it to twelve sarcoma samples. We demonstrate that pbfusion accurately identifies known and putative fusion genes (e.g.,TFE3-ASPSCR1). Simulation studies show that pbfusion is more sensitive than the alternative long-read fusion detection software while maintaining similar specificity.

Fusion gene finding using long-read sequencing



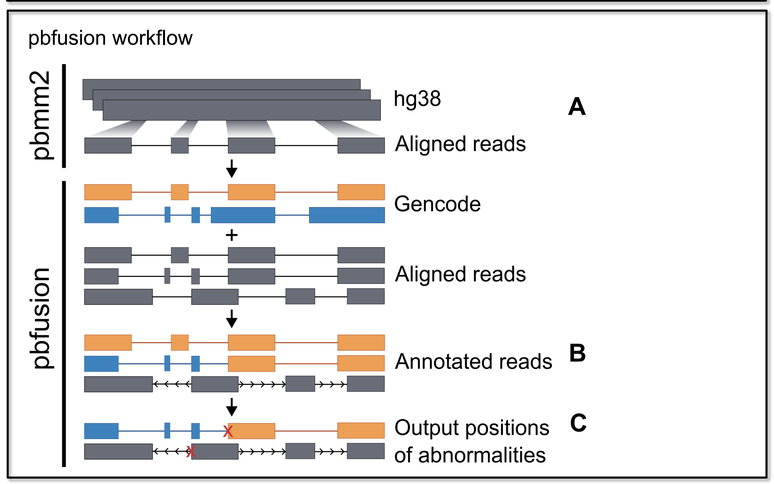


Figure 1. pbfusion workflow starting from Iso-Seq reads. (A) Iso-Seq reads are aligned to the reference genome using pbmm2. **(B)** pbfusion annotates alignment with reference transcripts, e.g., GENCODE GTF file. Transcripts with two different annotations, strand switches, and novel exons are clustered/chained by breakpoints. **(C)** The fusion gene variant-calls are output in a tabular BEDPE format containing support information that can be used for filtering false positive variants.

pbfusion features

- Works with bulk and single-cell Iso-Seq data (with or without MAS-Seq)
- Human and machine-readable output (BEDPE)
- Single binary tool that is easy to install and use

pbfusion detects known and putative fusion genes in sarcoma samples

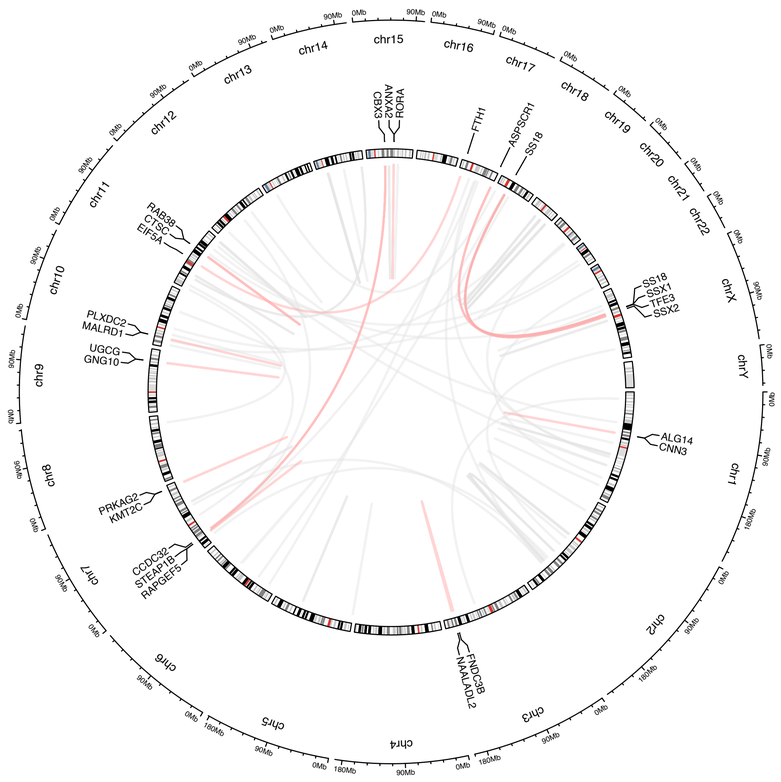


Figure 2. Chromosomal ideogram of fusion genes detected by pbfusion in sarcoma Iso-Seq data. Red lines denote known fusions while grey lines are putative fusions. All fusions across the twelve sarcoma samples are condensed into this plot.

Number of samples	Fusions (filtered)	Fusions known	Putative fusions
12	122	23	99

Table 1. Raw fusion calls were filtered for genes not related to ribosomal and immunoglobulin genes, no more than three gene partners, at least five supporting reads, and breakpoints that are at least 100 kb apart.

Visualizing pbfusion breakpoints and measures of accuracy

Visualizing bp resolved fusion genes with pbfusion read annotation

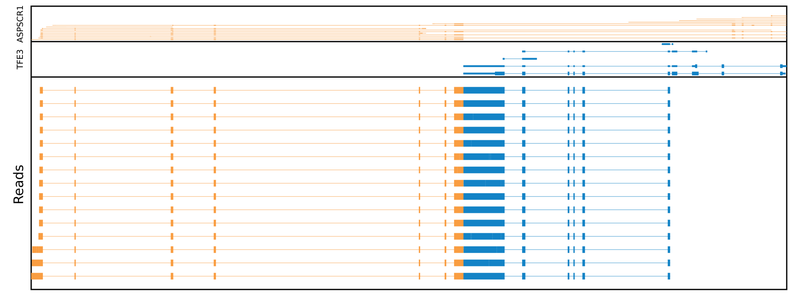


Figure 3. Visualization of TFE3-ASPSCR1 fusion gene detected by pbfusion. The top two panels contain the annotated transcripts for each gene, with ASPSCR1 in orange and TFE3 in blue. Bottom panel shows a subsampling of the reads that align to both loci, colored by gene annotation. This fusion gene is a known mutation for ASPS samples and was successfully detected by pbfusion.

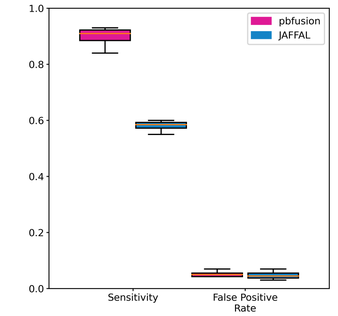


Figure 4. Sensitivity and false positive rate for pbfusion and JAFFAL on simulated data. Fusion reads were simulated by randomly selecting 100 pairs of protein coding genes and splicing them together at exonic boundaries at 100% accuracy. The remainder of the dataset are random transcripts taken from GENCODE. pbfusion has a mean sensitivity of 90% vs. JAFFAL's 60%. Both tools share a similar false positive rate where most false positives are from mapping errors.

Conclusions

- **pbfusion discovers 23 known and 99 novel fusion gene partners in sarcoma samples (Figure 2).**
- **pbfusion provides useful functions including single-cell annotations and plotting scripts (Figure 3).**
- **pbfusion leverages the accuracy of HiFi Iso-Seq data to make accurate fusion calls (Figure 4).**

Availability

Software and documentation can be found at:

https://github.com/PacificBiosciences/pbfusion

https://www.pacb.com/cancer-research