Fusion V5 (On-Target count)

Bioinformatics Development

Sandy

2025.03.24

Ref. issue:
- https://actg.atlassian.net/browse/ABIE-971

413 target exons:
- /mnt/RD_Develop/sandyteng/ACTFusionV5/code/fusionv4_annoloci2bed_test/targetexonbed/fusionv4.MANE.v0.95.GENCODE.r38.candidate.exons.transcript.bed

QC metrics overview

- Tools & fusion workflows

	STAR (arriba's workflow: STAR + arriba)	Fusion v4 (bwa-based)
Alignment analysis	STAR (to genome)	bwa-mem (to preferred transcriptome, MANE, GENCODE-r38)
(I) # of primary mapped reads	samtools flagstats (~81.7% from Twist NextSeq data)	samtools flagstats (~88.6% from Twist NextSeq data)
(II) % of on-target/probe-anchored reads	calculate_probe_reads.sh (in-house utility: samtools + bedtools) (~54% On-Target reads, Twist NextSeq data) **=> May over-estimate** **=> count the same read twice**	calculate_probe_reads.sh (in-house utility: samtools + bedtools) (~71.85% On-Target reads, Twist NextSeq data) **=> Remark: The reads are merged and went through isoform filtering. The value is calculated using merged single-end reads, while Arriba-STAR used paired-end reads.** **=> Convert probe region to preferred exon regions** **=> 413 preferred exons as target regions**
Read trimming	NA	trimadap
Counting (expression quantification)	HTseq ("htseq-count"), FeatureCounts ("featureCounts")	quantify_preferred_exons.v2.py 1. (transcript-level) via "htseq-count" => Need gtf file for preferred transcripts **=> Some arguments are not applicable for bwa (no 'NH' tag)** 2. (transcript-level) obtain alignments from *callingresult.txt file for each sample **=> Use "WILDTYPE" reads produced by the caller to quantify gene expression**
Fragmentation size	NA	fastp (insert size → peak, source file: *.fastp.merge.json) (129-153 bp insertion size, Twist NextSeq data)
Duplication rate	NA	fastp (duplication → rate, source file: *.fastp.merge.json) (29%-34% duplication rate, Twist NextSeq data)

ACT GENOMICS ™

On-Target %

- Tools
  - **samtools flagstats**
  - **calculate_probe_reads.sh**
- Example
  - **AANB02_202_AD02_AA-23-08153**

ACT GENOMICS ™

NextSeq, Twist 8 RNA data

Generated by "get_probe_reads.sh"

```
# AANB02_202_AH01_AA-23-08153_probe_report.txt
Total Primary Alignments: 5651046
Probe-Anchored Primary Alignments: 3812418
Percentage: 67.46%
```

Generated by "samtools flagstats <input.aligned.bam>" ("get_flagstats.sh")

```
# AANB02_202_AH01_AA-23-08153.flagstats.txt
10880521 + 0 in total (QC-passed reads + QC-failed reads)
5651046 + 0 primary
5019766 + 0 secondary
209709 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
9777437 + 0 mapped (89.86% : N/A)
4547962 + 0 primary mapped (80.48% : N/A)
5651046 + 0 paired in sequencing
2825523 + 0 read1
2825523 + 0 read2
4535110 + 0 properly paired (80.25% : N/A)
4547962 + 0 with itself and mate mapped
0 + 0 singletons (0.00% : N/A)
3164 + 0 with mate mapped to a different chr
1854 + 0 with mate mapped to a different chr (mapQ>=5)
```

- % Covered region anchored reads
- % On-Target reads = % of Primary mapped reads * % Covered region anchored reads

uuid	% of Primary mapped reads	% Covered region anchored reads	% On-Target reads
AANB02_202_AA02_AA-23-08153	83.64	66.64	55.74%
AANB02_202_AB02_AA-24-00324	81.09	67.42	54.67%
AANB02_202_AC02_AA-24-00324	82.33	64.01	52.70%
AANB02_202_AD02_AA-23-08153	80.89	68.06	55.05%
AANB02_202_AE02_AA-23-08153	83.37	66.53	55.47%
AANB02_202_AF02_AA-24-00324	79.86	66.53	53.13%
AANB02_202_AG02_AA-24-00324	81.7	63.54	51.91%
AANB02_202_AH01_AA-23-08153	80.48	67.46	54.29%

Ref. issue:
- https://actg.atlassian.net/browse/ABIE-971

Ref. directory
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_analysis.NextSeq/STAR-arriba/

Source files:
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_analysis.NextSeq/STAR-arriba/<uuid>.flagstats.txt
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_analysis.NextSeq/STAR-arriba/<uuid>_probe_report.txt

Scripts:
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_analysis.NextSeq/STAR-arriba/get_flagstats.sh
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_analysis.NextSeq/STAR-arriba/get_probe_reads.sh

ACT GENOMICS ™

On-target rate (bwa, preferred exons as target regions)

- % Covered region anchored reads
- % On-Target reads = % of Primary mapped reads * % Covered region anchored reads

uuid	% of Primary mapped reads	% Covered region anchored reads	% On-Target reads
AANB02_202_AA02_AA-23-08153	87.74	80.37	70.52%
AANB02_202_AB02_AA-24-00324	89.96	82.31	74.05%
AANB02_202_AC02_AA-24-00324	85.68	79.16	67.82%
AANB02_202_AD02_AA-23-08153	91.28	82.93	75.70%
AANB02_202_AE02_AA-23-08153	87.82	80.44	70.64%
AANB02_202_AF02_AA-24-00324	89.81	82.06	73.70%
AANB02_202_AG02_AA-24-00324	85.41	78.77	67.28%
AANB02_202_AH01_AA-23-08153	90.95	82.57	75.10%

91.28%*82.93%

Ref. issue:
- https://actg.atlassian.net/browse/ABIE-971

Ref. directory
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_analysis.NextSeq/bwa-fusionv4/ (=> % of Primary mapped reads)
- /mnt/RD_Develop/sandyteng/ACTFusionV5/code/fusionv4_calculate_probe_reads_test/ (=> % Covered region anchored reads)

ACT GENOMICS ™

samtools flagstats

- % of Primary mapped reads
  - samtools flagstats aligned.bam

```
2327266 + 0 in total (QC-passed reads + QC-failed reads)
2121990 + 0 primary
0 + 0 secondary
205276 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
2142294 + 0 mapped (92.05% : N/A)
1937018 + 0 primary mapped (91.28% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

uuid	% of Primary mapped reads	% Covered region anchored reads	% On-Target reads
AANB02_202_AA02_AA-23-08153	87.74	80.37	70.52%
AANB02_202_AB02_AA-24-00324	89.96	82.31	74.05%
AANB02_202_AC02_AA-24-00324	85.68	79.16	67.82%
AANB02_202_AD02_AA-23-08153	91.28	82.93	75.70%
AANB02_202_AE02_AA-23-08153	87.82	80.44	70.64%
AANB02_202_AF02_AA-24-00324	89.81	82.06	73.70%
AANB02_202_AG02_AA-24-00324	85.41	78.77	67.28%
AANB02_202_AH01_AA-23-08153	90.95	82.57	75.10%

ACT GENOMICS ™

calculate_probe_reads.sh

- A tool for % Covered region anchored reads calculation (Probe covered reads percentage)
- Steps
  - Filter out secondary and supplementary alignments from the input BAM
  - Count total primary alignments
  - Extract probe-anchored primary alignments using bedtools intersect
  - Count primary alignments in probe-anchored BAM
  - Calculate probe-anchored read percentage

uuid	% of Primary mapped reads	% Covered region anchored reads	% On-Target reads
AANB02_202_AA02_AA-23-08153	87.74	80.37	70.52%
AANB02_202_AB02_AA-24-00324	89.96	82.31	74.05%
AANB02_202_AC02_AA-24-00324	85.68	79.16	67.82%
AANB02_202_AD02_AA-23-08153	91.28	82.93	75.70%
AANB02_202_AE02_AA-23-08153	87.82	80.44	70.64%
AANB02_202_AF02_AA-24-00324	89.81	82.06	73.70%
AANB02_202_AG02_AA-24-00324	85.41	78.77	67.28%
AANB02_202_AH01_AA-23-08153	90.95	82.57	75.10%

ACT GENOMICS ™

AANB02_202_AD02_AA-23-08153 (fusion v4 (bwa bam))

- AANB02_202_AD02_AA-23-08153_primary.bam => 2,121,990
- AANB02_202_AD02_AA-23-08153_probed.bam => 1,759,749

samtools view -b -F 0x900

Filter out secondary and supplementary alignments from the input BAM

2327266 + 0 in total (QC-passed reads + QC-failed reads)
2121990 + 0 primary
0 + 0 secondary
205276 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
2142294 + 0 mapped (92.05% : N/A)
1937018 + 0 primary mapped (91.28% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

Aligned.bam 1

2121990 + 0 in total (QC-passed reads + QC-failed reads)
2121990 + 0 primary
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
1937018 + 0 mapped (91.28% : N/A)
1937018 + 0 primary mapped (91.28% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

Primary.bam 2

bedtools intersect -a primary.bam -b probe.bed

1759749 + 0 in total (QC-passed reads + QC-failed reads)
1759749 + 0 primary
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
1759749 + 0 mapped (100.00% : N/A)
1759749 + 0 primary mapped (100.00% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

Probed.bam 3

4 **Probe_report.txt**

Total Primary Alignments: 2121990
Probe-Anchored Primary Alignments 1759749
Percentage: **82.93%**

1,759,749 (primary reads)
/2,121,990 (probe anchored reads)

Remark:
3 reports are generated via “samtools flagstats”
- Aligned.bam
- Primary.bam
- Probed.bam

1 report is generated via “calculate_probe_reads.sh”

ACT GENOMICS ™

% Covered region anchored reads calulation workflows

- fusion v4
- arriba

ACT GENOMICS ™

% Covered region anchored reads calculation (fusionv4)

- Preferred exons to bed regions conversion (fusionv4_annoloci2bed.py)
  - Input files: preferred.genome.exons.annotation, preferred.transcriptome.exons.annotation
  - Output files: preferred.genome.exons.annotation.bed, preferred.transcriptome.exons.annotation.bed
- Bed coordinates sorting (sort-bed)
  - Input files: preferred.transcriptome.exons.annotation.bed, probe.bed
  - Output files: sorted.preferred.transcriptome.exons.annotation.bed, sorted.probe.bed
- Bed files intersection (bedtools intersect)
  - Input files: sorted.preferred.transcriptome.exons.annotation.bed, sorted.probe.bed
  - Output files: candidate.exons.bed
- Extract target exon list (uniq, awk)
  - Input file: candidate.exons.bed
  - Output file: target.exons.namelist.txt
- Extract transcript loci bed (grep -wf)
  - Input files: preferred.transcriptome.exons.annotation.bed, target.exons.namelist.txt
  - Output files: candidate.exons.transcript.bed
- % Covered region anchored reads calculation (calculate_probe_reads.sh: samtools + bedtools)
  - Input files / string: aligned.**fusionv4**.bam, **candidate.exons.transcript.bed**, sample.id (uuid string)
  - Output files: sample.id_primary.bam (&.bai), sample.id_probed.bam (& .bai), sample.id_probe_report.txt

ACT GENOMICS ™

% Covered region anchored reads calculation (arriba)

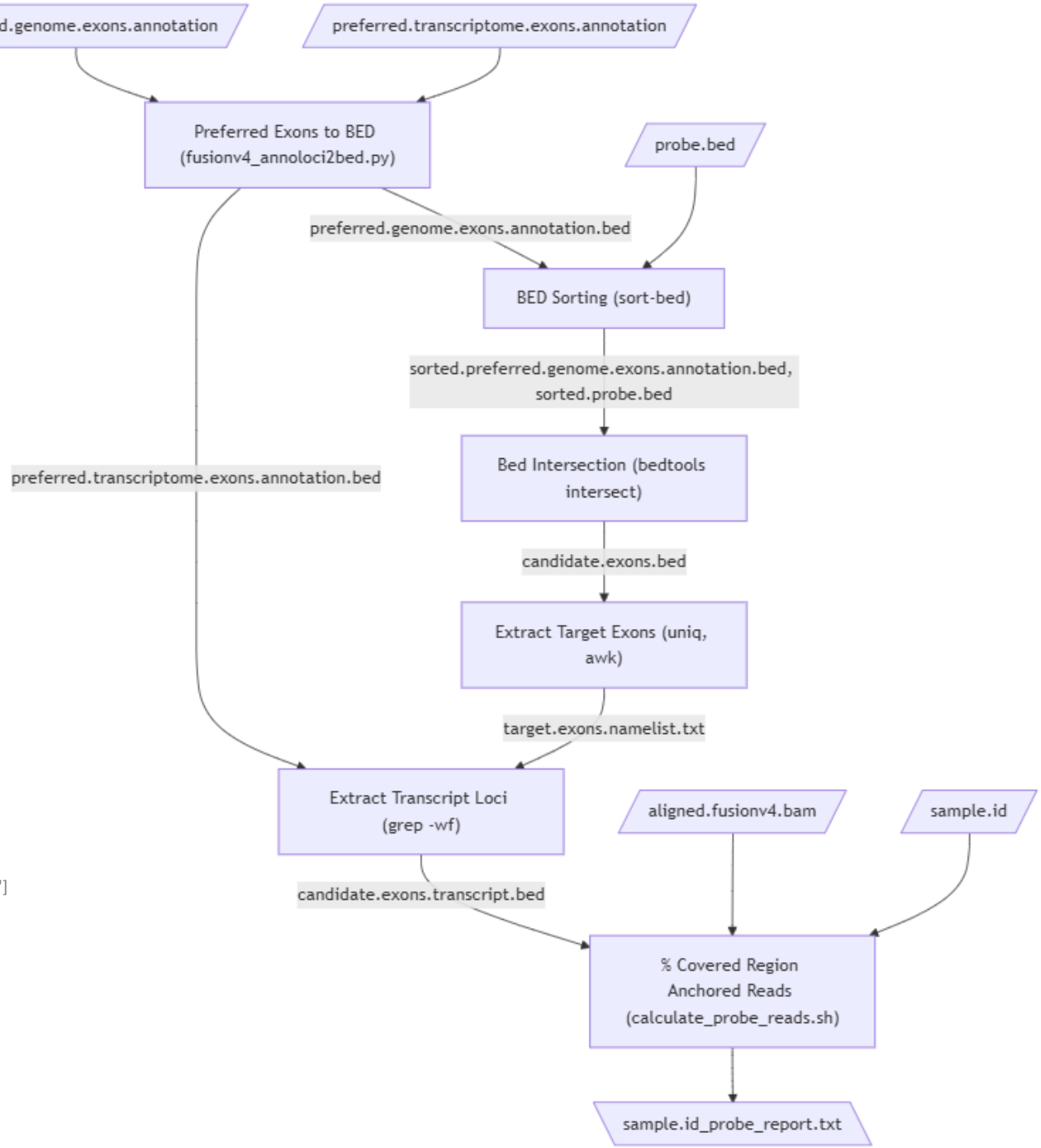
- Preferred exons to bed regions conversion (fusionv4_annoloci2bed.py)
  - Input files: preferred.genome.exons.annotation, preferred.transcriptome.exons.annotation
  - Output files: preferred.genome.exons.annotation.bed, preferred.transcriptome.exons.annotation.bed
- Bed coordinates sorting (sort-bed)
  - Input files: preferred.transcriptome.exons.annotation.bed, probe.bed
  - Output files: sorted.preferred.transcriptome.exons.annotation.bed, sorted.probe.bed
- Bed files intersection (bedtools intersect)
  - Input files: sorted.preferred.transcriptome.exons.annotation.bed, sorted.probe.bed
  - Output files: candidate.exons.bed
- Extract target exon list (uniq, awk)
  - Input file: candidate.exons.bed
  - Output file: target.exons.namelist.txt
- Extract transcript loci bed (grep -wf)
  - Input files: preferred.transcriptome.exons.annotation.bed, target.exons.namelist.txt
  - Output files: candidate.exons.transcript.bed
- % Covered region anchored reads calculation (calculate_probe_reads.sh: samtools + bedtools)
  - Input files / string: aligned.**arriba**.bam, **probe.bed**, sample.id (uuid string)
  - Output files: sample.id_primary.bam (&.bai), sample.id_probed.bam (& .bai), sample.id_probe_report.txt

ACT GENOMICS ™

Workflow

- Fusion v4 (full)

```
graph TD;
  %% Initial Inputs
  I1[/preferred.genome.exons.annotation/]
  I2[/preferred.transcriptome.exons.annotation/]
  I3[/probe.bed/]
  I4[/aligned.fusionv4.bam/]
  I6[/sample.id/]

  %% FusionV4 Workflow
  I1 --> A1["Preferred Exons to BED (fusionv4_annoloci2bed.py)"]
  I2 --> A1
  A1 -->|preferred.genome.exons.annotation.bed| B1["BED Sorting (sort-bed)"]
  A1 -->|preferred.transcriptome.exons.annotation.bed| E1

  I3 --> B1
  B1 -->|sorted.preferred.genome.exons.annotation.bed, sorted.probe.bed| C1["Bed Intersection (bedtools
intersect)"]

  C1 -->|candidate.exons.bed| D1["Extract Target Exons (uniq, awk)"]

  D1 -->|target.exons.namelist.txt| E1["Extract Transcript Loci (grep -wf)"]

  E1 -->|candidate.exons.transcript.bed| F1["% Covered Region Anchored Reads (calculate_probe_reads.sh)"]

  I4 --> F1
  I6 --> F1
  F1 --> O1[\sample.id_probe_report.txt\]
```

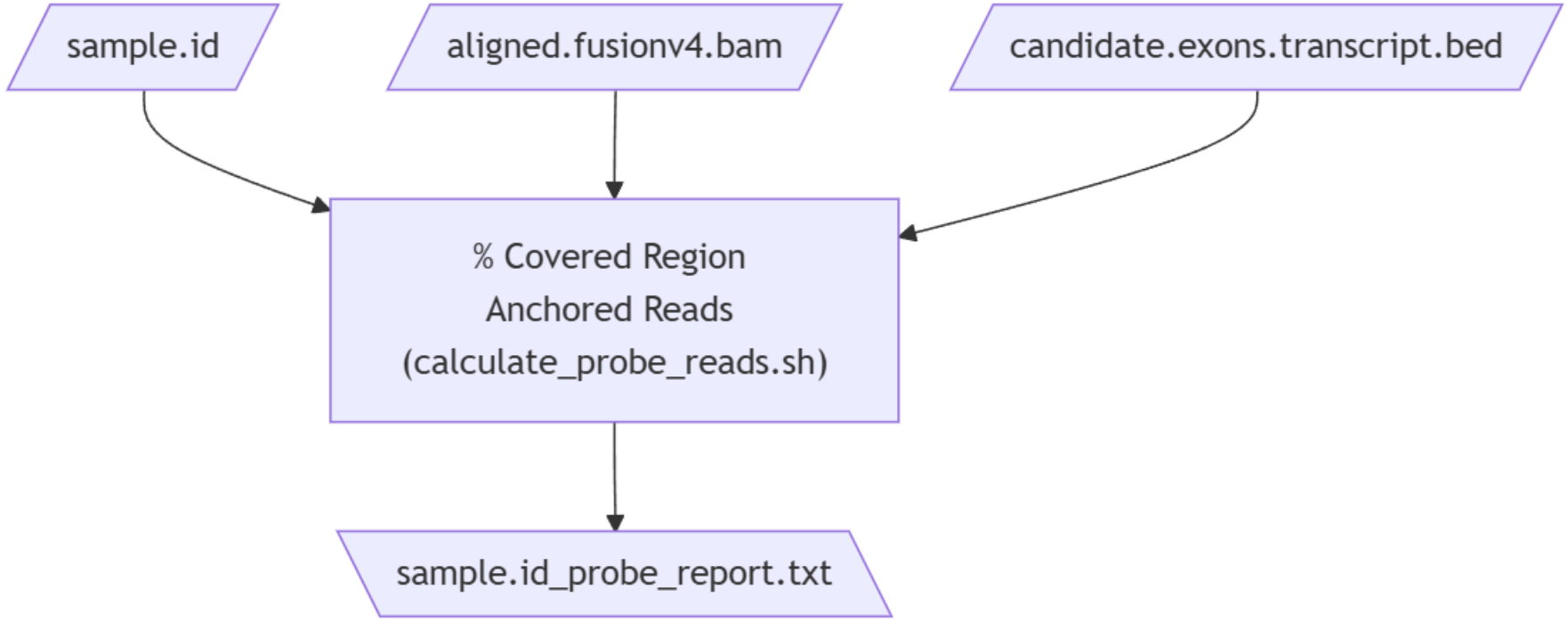


ACT GENOMICS ™

Workflow

- Fusion v4

```
graph TD;
  %% Initial Inputs
  I3[/candidate.exons.transcript.bed/]
  I5[/aligned.fusionv4.bam/]
  I6[/sample.id/]
%% FusionV4 Workflow
  I6 --> F1["% Covered Region Anchored Reads (calculate_probe_reads.sh)"]
  I5 --> F1
  I3 --> F1
  F1 --> O1[\sample.id_probe_report.txt\]
```

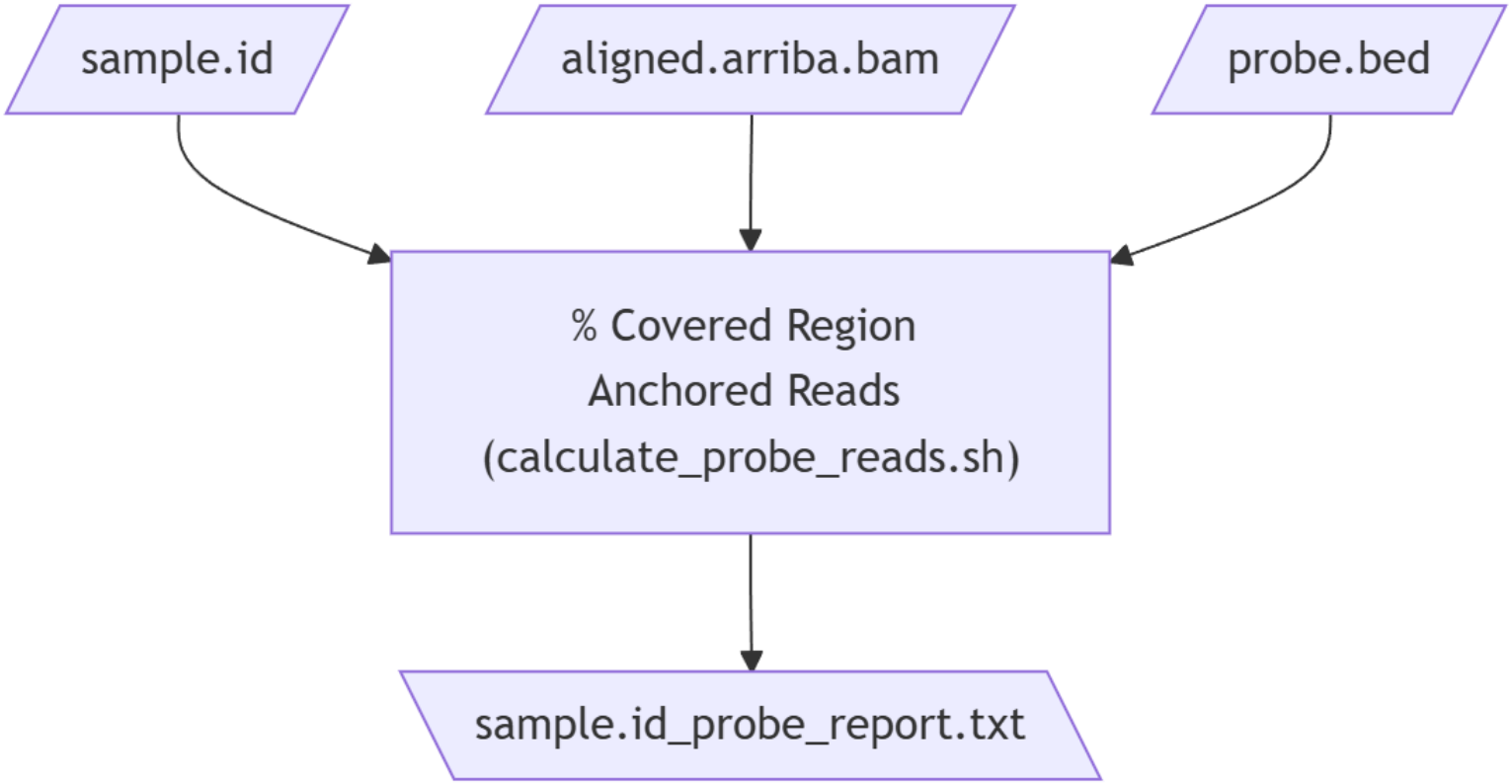


ACT GENOMICS ™

Workflow

- Arriba

```
graph TD;
  %% Initial Inputs
  I3[/probe.bed/]
  I5[/aligned.arriba.bam/]
  I6[/sample.id/]
%% Arriba Workflow
  I6 --> F1["% Covered Region Anchored Reads (calculate_probe_reads.sh)"]
  I5 --> F1
  I3 --> F1
  F1 --> O1[\sample.id_probe_report.txt\]
```



ACT GENOMICS ™

Make Personalized Medicine Accessible to All

ACT GENOMICS ™