Fusion V5 (Gene quantification)

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)	calculate_probe_reads.sh (in-house utility: samtools + bedtools) **To-do**
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 ™

Counting (expression quantification)

- Quantification scenarios
  - Htseq (+ arriba.STAR.bam)
  - FeatureCounts (+ arriba.STAR.bam)
  - quantify_preferred_exons.v2.py (in-house script) (+ fusionv4.bwa.bam)

- Analysis workflow
  - Gene count quantification (via htseq, featurecounts, quantify_preferred_exons.v2.py)
  - Target gene count extraction (only compare the 220 target genes defined in twist.covered.bed (via grep -wf))

- Result summary
  - Gene count obtained from htseq and featurecounts are similar (correlation 99.9%)
  - Gene count quantified form fusion v4 and arriba workflows are similar (correlation 99.3%)

Target gene:
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Twist_kit/Covered_Regions_RNA_Fusions_4X_TE-98493102_GRCh38.gene.list.txt (Twist)

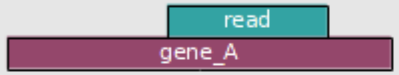
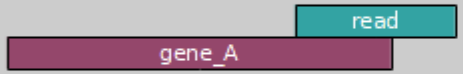
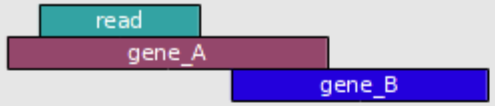
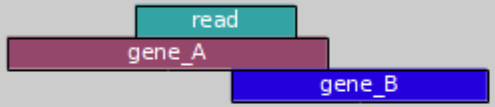
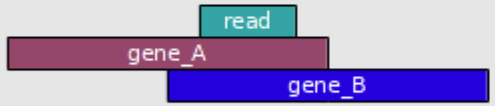
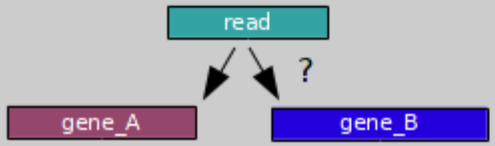
ACT GENOMICS ™

HTSeq-count

- Default options for feature count (gene count)

- **-t exon**
  (default feature type => 3rd column in GTF file)
- **-i gene_id**
  (default id attribute => feature ID)
- **-m union**
  (default read overlapping handling)
- **--nonunique none**
  (default mode for reads aligned to more than one feature in the "-m" option)

Ref. link
- Htseq-count docs

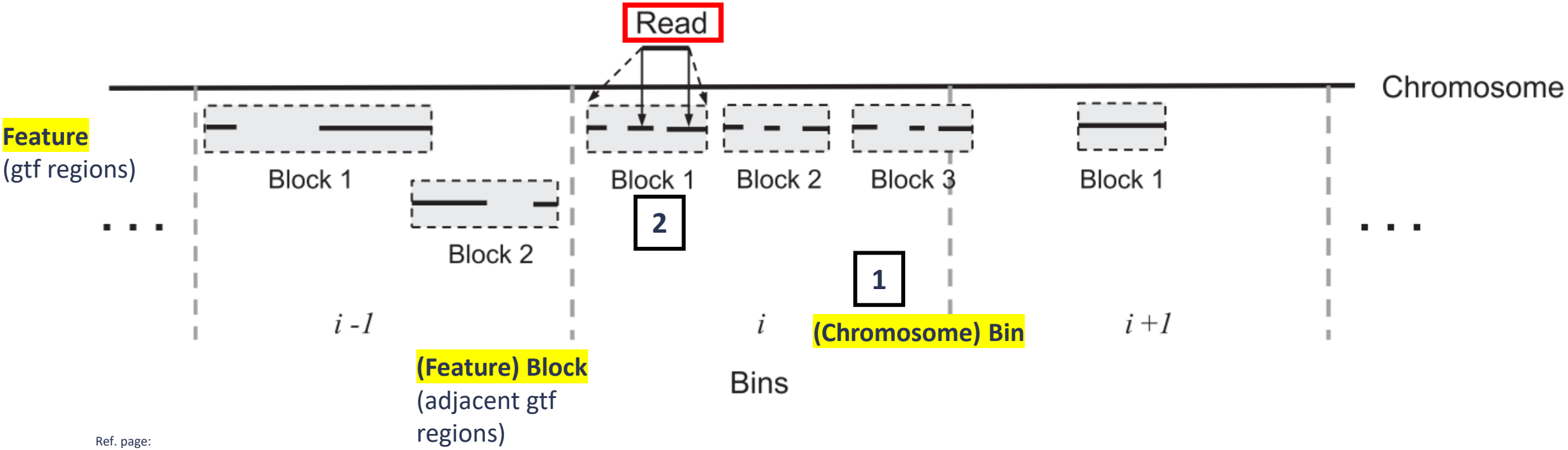
	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

ACT GENOMICS ™

FeatureCounts

Steps:
- Overlap of reads with features
- Multiple overlaps
- Chromosome hashing
- Genome bins and feature blocks

- FeatureCounts (algorithm)



Ref. page:
- featureCounts: an efficient general purpose program for assigning sequence reads to genomic features

Ref. file:
- /mnt/RD_Develop/sandyteng/ACTFusionV5/code/plot_exon_distribution_test/ENST00000397752.8.exon_distribution.pdf

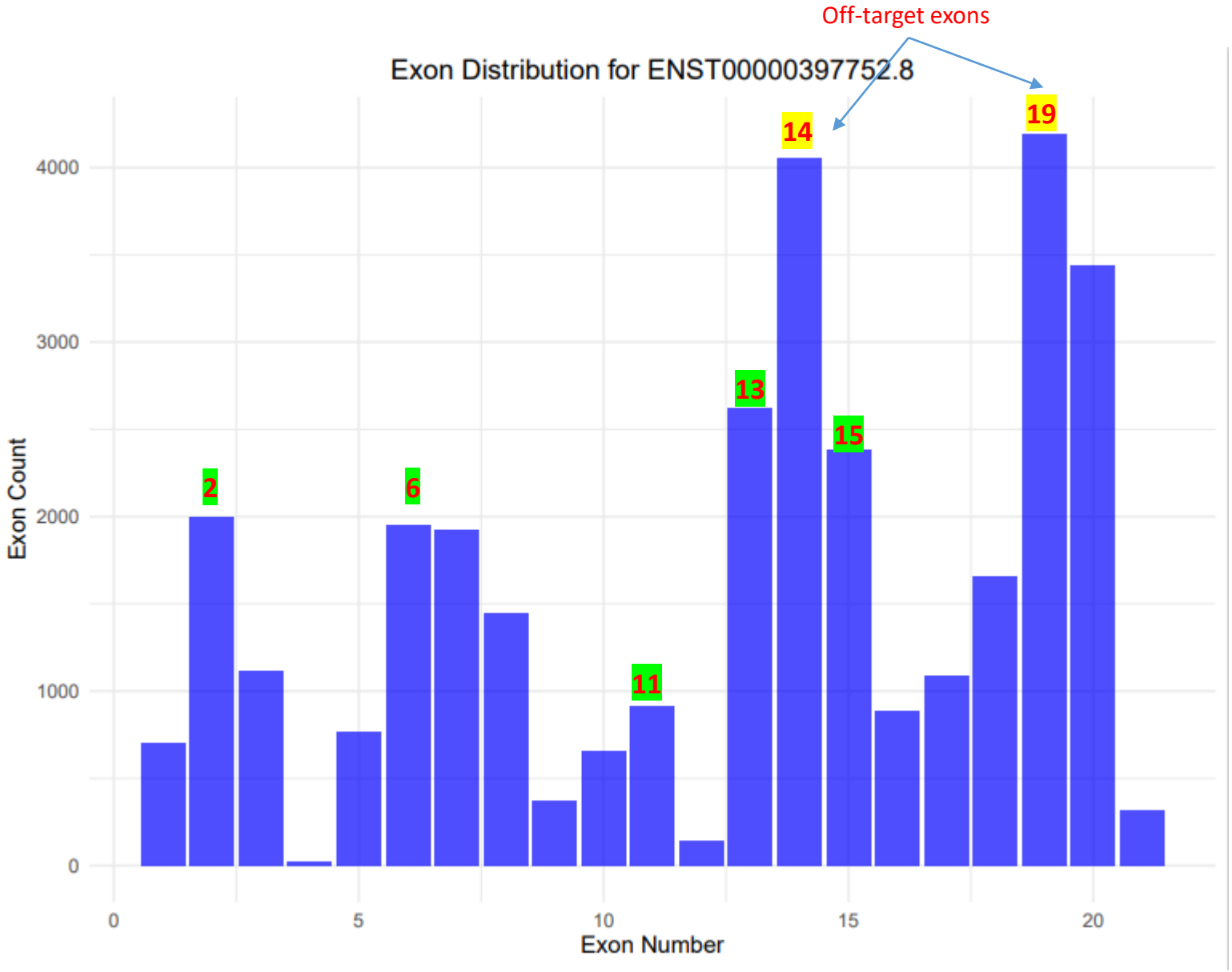
Script:
- /mnt/RD_Develop/sandyteng/ACTFusionV5/code/plot_exon_distribution_test/get_ENST00000397752.8.MET.exon.plot.sh

Exon number vs Count

- MET

gene count
gene_id count
MET **18002** => identical to # of WILDTYPE MET reads

exon count
=>
(-i) *callingresult.txt ("ENST00000397752.8" => "MET")
167 WILDTYPE MET:15,16
280 WILDTYPE MET:15,16,17
53 WILDTYPE MET:16
372 WILDTYPE MET:16,17
12 WILDTYPE MET:16,17,18
⇒ (-o) exon.count ("ENST00000397752.8" exon 16)
⇒ ENST00000397752.8 16 884 (167+280+53+372+12 = 884)



Ref. file:
- /mnt/RD_Develop/sandyteng/ACTFusionV5/code/plot_exon_distribution_test/ENST00000397752.8.exon_distribution.pdf

Script:
- /mnt/RD_Develop/sandyteng/ACTFusionV5/code/plot_exon_distribution_test/get_ENST00000397752.8.MET.exon.plot.sh

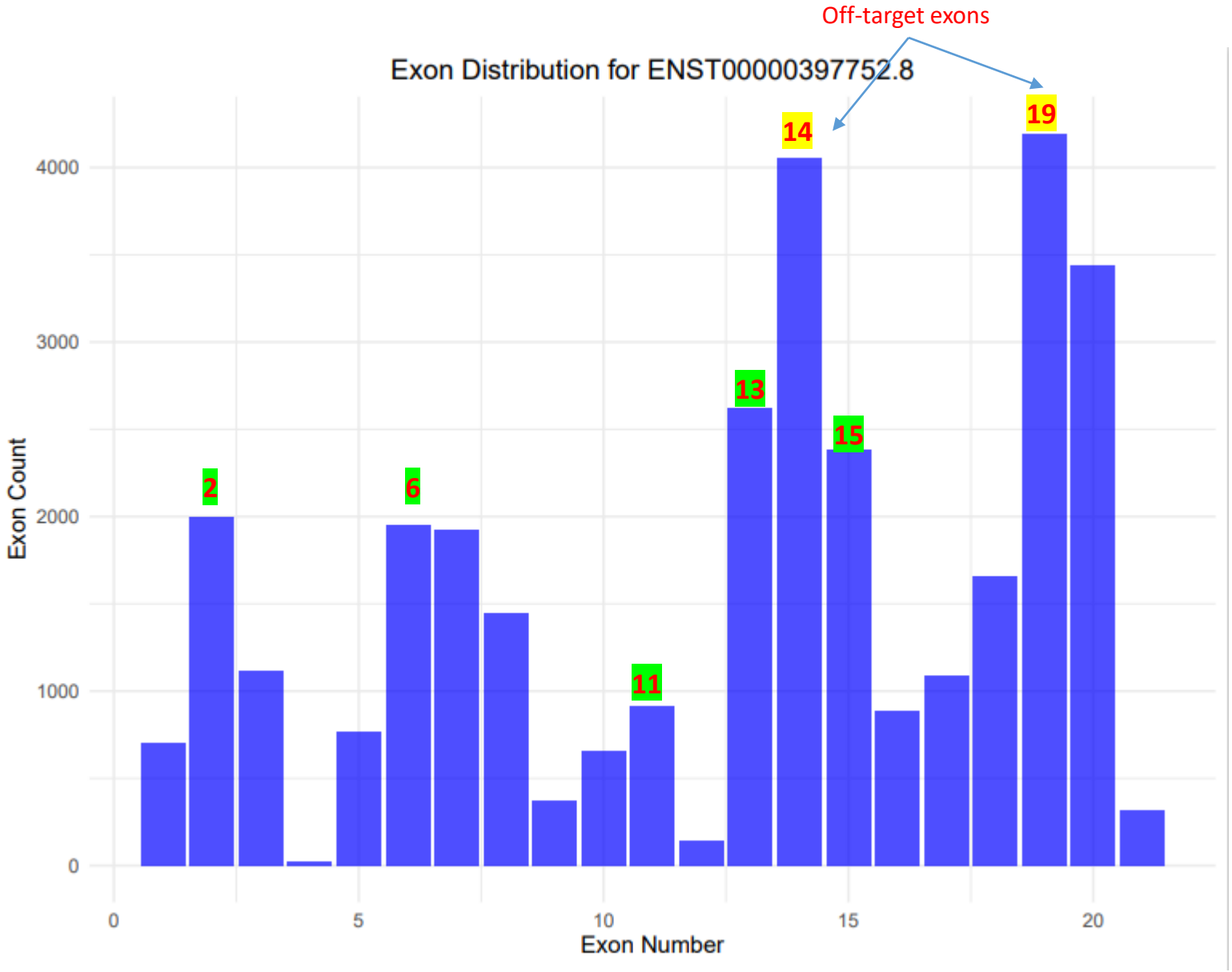
Exon number vs Count

- MET

chr7	116699071	116700284	MET-exon-fusionv4-2
chr7	116755355	116755515	MET-exon-fusionv4-6
chr7	116763050	116763268	MET-exon-fusionv4-11
chr7	116771498	116771654	MET-exon-fusionv4-13
chr7	116774881	116775111	MET-exon-fusionv4-15

gene_id count
MET 18002
=> identical to
# of WILDTYPE MET reads

ENST00000397752.8	1	707
ENST00000397752.8	2	2000
ENST00000397752.8	3	1117
ENST00000397752.8	4	27
ENST00000397752.8	5	763
ENST00000397752.8	6	1952
ENST00000397752.8	7	1928
ENST00000397752.8	8	1443
ENST00000397752.8	9	373
ENST00000397752.8	10	653
ENST00000397752.8	11	918
ENST00000397752.8	12	145
ENST00000397752.8	13	2624
ENST00000397752.8	14	4055
ENST00000397752.8	15	2385
ENST00000397752.8	16	884
ENST00000397752.8	17	1088
ENST00000397752.8	18	1655
ENST00000397752.8	19	4196
ENST00000397752.8	20	3443
ENST00000397752.8	21	314



Gene quantification (FusionV4)

Ref. issue
https://actg.atlassian.net/browse/ABIE-987
https://actg.atlassian.net/browse/ABIE-988

- FusionV4 processes
  - R1.fq.gz, R2.fq.gz (input files) -> mergefastq ("mergefastq") -> trimadap -> fastp -> bwaisoform -> **bwase** -> **fusioncalling** -> fuscall2QC
- **bwase** (“bwa”) -> **fusioncalling** (“ACTGfuscall.py”) -> quantifygene (“quantify_preferred_exons.v2.py”)
  - bwase
    - Input files: preferred.transcriptome.fasta, preferred.transcriptome.fasta.indices
    - Output files: aligned.fusionv4.bam, aligned.fusionv4.bam.bai
  - Fusioncalling
    - Input files:
      aligned.fusionv4.bam, preferred.transcriptome.exons.annotation,
      protein.fasta, protein.fasta.meta, qc.thresholds.config
    - Output files:
      callingresult.txt, gspcallingresult.txt, protein_seq.meta.txt, callingform.txt
  - quantifygene
    - Input file: callingresult.txt
    - Output files: gene.count, exon.count

ACT GENOMICS ™

Gene quantification (Arriba)

Ref. issue
https://actg.atlassian.net/browse/ABIE-987
https://actg.atlassian.net/browse/ABIE-988

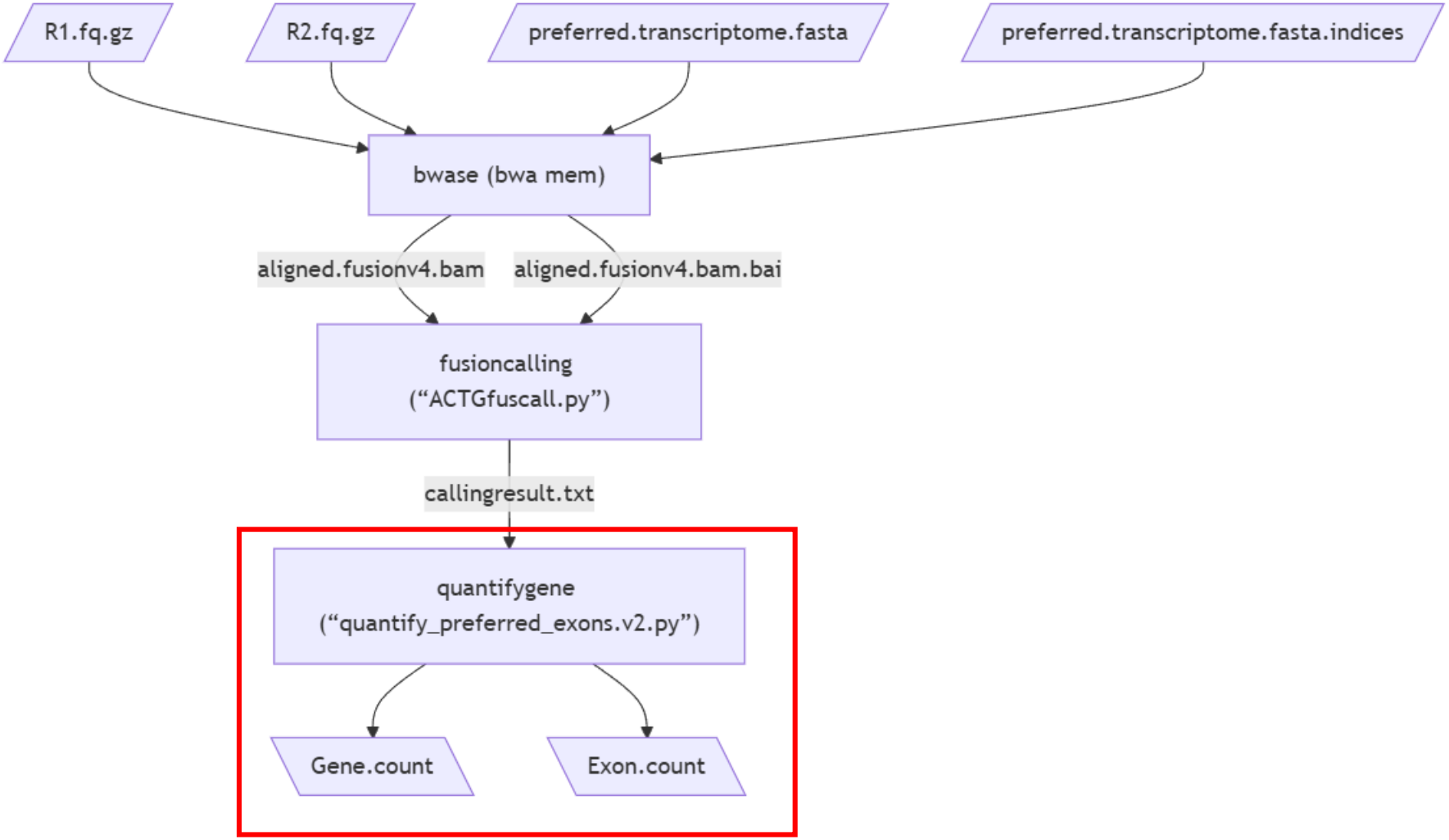
- Arriba processes
  - R1.fq.gz, R2.fq.gz (input files) -> **STAR** -> Arriba
- **STAR** (“STAR”) -> quantifygene (“HTSeq” / “featureCounts”)
  - STAR
    - Input files: GRCh38.fa, GRCh38.fa.indices, RefSeq_hg38.gtf
    - Output files: aligned.arriba.bam
  - quantifygene (htseq)
    - Input files: aligned.arriba.bam, RefSeq_hg38.gtf
    - Ouptut files: genes_htseq.count (gene.count)
  - quantifygene (featureCounts)
    - Input files: aligned.arriba.bam, RefSeq_hg38.gtf
    - Ouptut files: genes_featureCounts.count (gene.count), genes_featureCounts.count.summary (gene.count.summary)

ACT GENOMICS ™

Workflow

- FusionV4

```
graph TD;
  %% Initial Inputs
  I1[/R1.fq.gz/]
  I2[/R2.fq.gz/]
  I7[/preferred.transcriptome.fasta/]
  I8[/preferred.transcriptome.fasta.indices/]
%% FusionV4 Workflow
  I1 --> A1
  I2 --> A1
  I7 --> A1["bwase (bwa mem)"]
  I8 --> A1
  A1 -->|aligned.fusionv4.bam| B1
  A1 -->|aligned.fusionv4.bam.bai| B1["fusioncalling (“ACTGfuscall.py”)
  B1 -->|callingresult.txt| C1["quantifygene (“quantify_preferred_exons
  C1 --> O1[\Gene.count\]
  C1 --> O2[\Exon.count\]
```

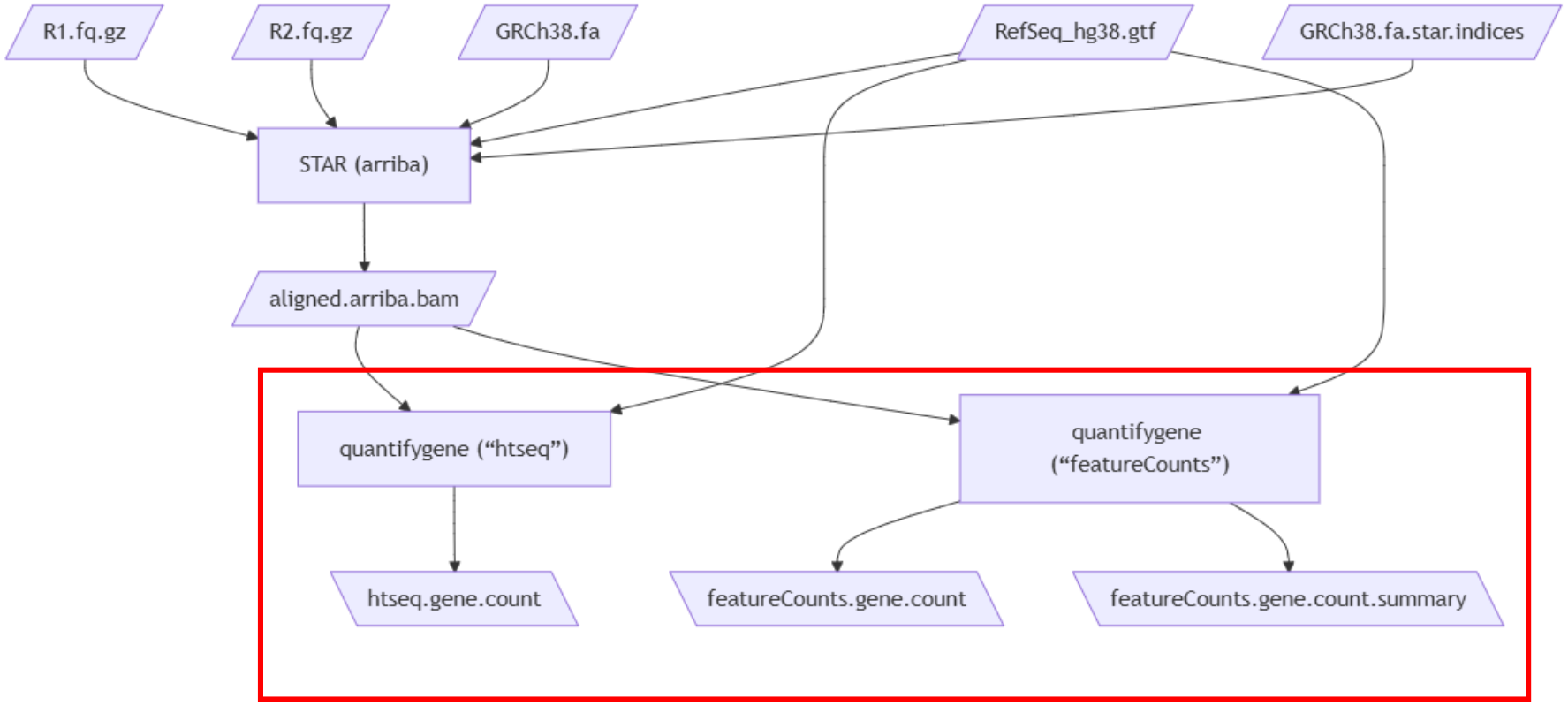


ACT GENOMICS ™

Workflow

- Arriba

```
graph TD;
  %% Initial Inputs
  I1[/R1.fq.gz/]
  I2[/R2.fq.gz/]
  I6[/GRCh38.fa/]
  I7[/GRCh38.fa.star.indices/]
  I8[/RefSeq_hg38.gtf/]
%% FusionV4 Workflow
  I1 --> A1
  I2 --> A1
  I6 --> A1
  I7 --> A1["STAR (arriba)"]
  I8 --> A1
  I8 --> B1
  I8 --> C1
  A1 --> I9[/aligned.arriba.bam/]
  I9 --> B1["quantifygene (“htseq”)"]
  I9 --> C1["quantifygene (“featureCounts”)"]
  %%A1 -->|aligned.arriba.bam| B1["quantifygene (“htseq”)"]
  B1 --> O1[\htseq.gene.count\]
  %%A1 -->|aligned.arriba.bam| C1["quantifygene (“featureCounts”)"]
  C1 --> O2[\featureCounts.gene.count\]
  C1 --> O3[\featureCounts.gene.count.summary\]
```

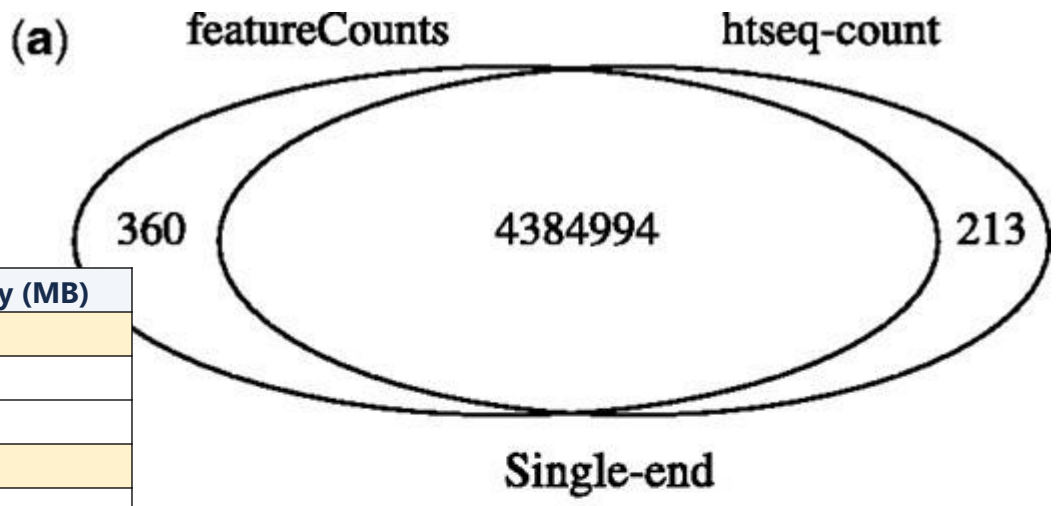


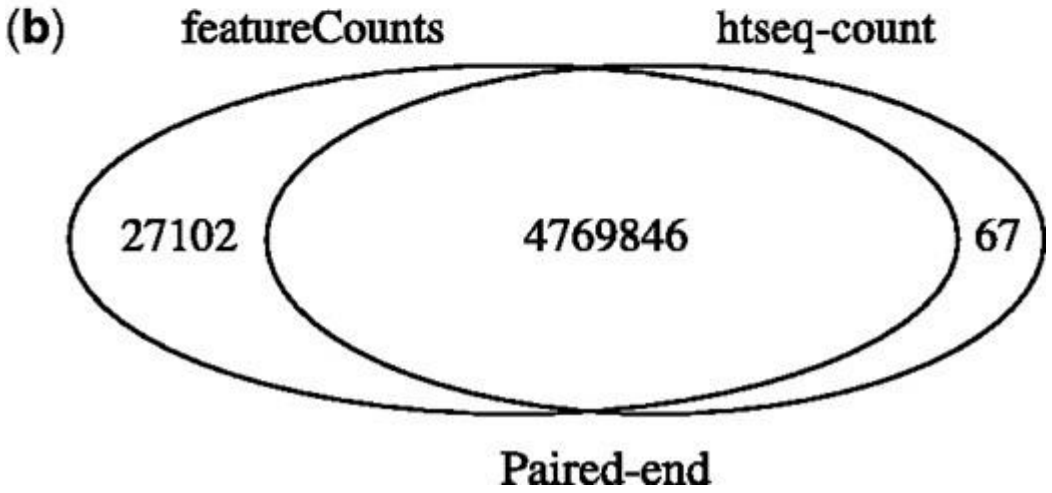
ACT GENOMICS ™

Gene count comparison (paper)

- HTSeq/FeatureCounts

Method	Number of fragments	Time (min)	Memory (MB)
featureCounts	5 392 155	0.9	4
CountOverlaps (whole genome at once)	5 392 155	24.4	7000
CountOverlaps (by chromosome)	5 392 155	36.6	783
htseq-count (union)	4 978 050	36	31
htseq-count (intersection-nonempty)	4 993 644	35.7	31
coverageBED	5 366 902	4.4	41





Ref:
[featureCounts: an efficient general purpose program for assigning sequence reads to genomic features]

ACT GENOMICS ™

Gene count comparison (AANB02_202_AD02_AA-23-08153)

- HTSeq/FeatureCounts
- FeatureCounts/quantify_preferred_exons.v2.py

Pearson's r (B, C) (featureCounts, htseq)	0.999996602
n	244
t-value (r*SQRT(n-2))/(SQRT(1-r^2))	5967.755077
p-value TDIST(x, deg_freedom, tails)	0
Pearson's r (C, D) (featureCounts, quantify_preferred_exons.v2.py)	0.993358852
n	244
t-value	134.307269
p-value	2.8426E-229

Ref. issue
https://actg.atlassian.net/browse/ABIE-987
https://actg.atlassian.net/browse/ABIE-988

ACT GENOMICS ™

Ref:
- HTSeq/FeatureCounts
- bedtools coverage/samtools depth
- quantify_preferred_exons.v2.py

Tool overview

	HTSeq/FeatureCounts	bedtools coverage /samtools depth	quantify_preferred_exons.v3.py
Quantification level	Gene Level (predefined intervals within gtf => gene id recognition)	Base Level (bedtools coverage -d /samtools depth) Interval Level (bedtools coverage)	Gene Level + Exon Level
Limitations	Some arguments are not applicable for bwa (no"NH"tag) Count gene using the predefined gtf (merged the same gene_id) => limit to predefined gene intervals (may not encompass MANE 1.4 transcripts)	samtools depth is preferred for SAM FLAG sensitivity (duplication removal) ABIE-976: "bedtools coverage" vs. "samtools depth" Done bedtools coverage fails to identify read fragment => extra care is required for result interpretation see details for https://github.com/ACTGenomics/panel_gene_coverageConnect your Github account	Only quantify exons defined in the preferred transcripts (MANE 0.95 + GENCODE-r38) => one may change the preferred transcripts to MANE 1.4 Rely on fusion v4 calling result => only work for fusion v4 pipeline

Ref. issue
https://actg.atlassian.net/browse/ABIE-987
https://actg.atlassian.net/browse/ABIE-988

ACT GENOMICS ™

Make Personalized Medicine Accessible to All

ACT GENOMICS ™