Fusion V5 (Sample QC)

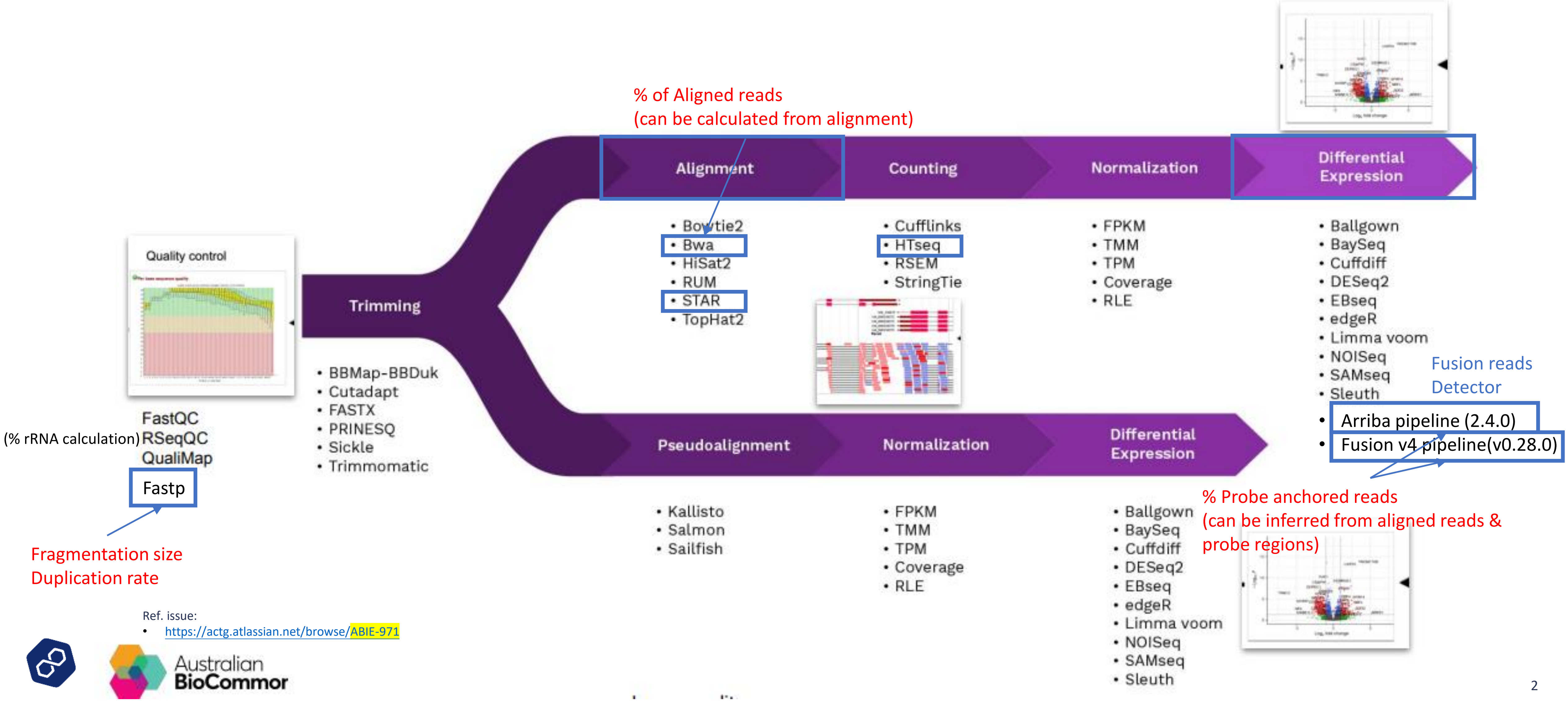
Bioinformatics Development

Sandy

2025.03.24

Available tools for RNA-Seq

Ref:
- https://zenodo.org/records/8323208
  [Getting started with RNAseq: Transforming raw reads into biological insights]



Trimming:
- BBMap-BBDuk
- Cutadapt
- FASTX
- PRINESQ
- Sickle
- Trimmomatic

Alignment:
- Bowtie2
- Bwa
- HiSat2
- RUM
- STAR
- TopHat2

Counting:
- Cufflinks
- HTseq
- RSEM
- StringTie

Normalization:
- FPKM
- TMM
- TPM
- Coverage
- RLE

Differential Expression:
- Ballgown
- BaySeq
- Cuffdiff
- DESeq2
- EBseq
- edgeR
- Limma voom
- NOISeq
- SAMseq
- Sleuth

Fusion reads Detector:
- Arriba pipeline (2.4.0)
- Fusion v4 pipeline(v0.28.0)

Pseudoalignment:
- Kallisto
- Salmon
- Sailfish

Normalization:
- FPKM
- TMM
- TPM
- Coverage
- RLE

Differential Expression:
- Ballgown
- BaySeq
- Cuffdiff
- DESeq2
- EBseq
- edgeR
- Limma voom
- NOISeq
- SAMseq
- Sleuth

% of Aligned reads
(can be calculated from alignment)

% Probe anchored reads
(can be inferred from aligned reads & probe regions)

Fragmentation size
Duplication rate

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

Interval count approaches

Sorted.bam for
- HTSeq (sort by name if -r name is specified)
- FeatureCounts (sort by coordinates)

- Tools for RNAseq gene count analysis
  - **HTSeq** (not sensitive to duplication FLAG => lack of duplicate handling)
    - ~~htseq-count -f bam -r name -s no -t exon -i gene_id aligned_reads.bam ref.gtf > genes_htseq.count~~
    - htseq-count -f bam -r name -s no -t exon -i gene_id --minaqual 10 aligned_reads.bam ref.gtf > genes_htseq.count
  - **FeatureCounts** (duplication FLAG sensitive)
    - ~~featureCounts -T 8 -p --countReadPairs -s 0 -t exon -g gene_id -Q 10 -a ref.gtf -o genes_featureCounts.count aligned_reads.bam~~
    - featureCounts -T 8 -p --countReadPairs -s 0 -t exon -g gene_id -Q 10 --primary -a ref.gtf -o genes_featureCounts.count aligned_reads.bam

- Tools for DNA probe coverage analysis
  - bedtools coverage (not FLAG sensitive) [Yu-Feng's issue: ABIE-976]
    - bedtools coverage -a probe.bed -b <aligned.bam (processed.bam)> -d
      # -d report the depth at each position in each feature (defined in -a)
  - samtools depth (duplication FLAG sensitive) [Yu-Feng's issue: ABIE-976]
    - samtools depth -b probe.bed <aligned.bam (processed.bam)>

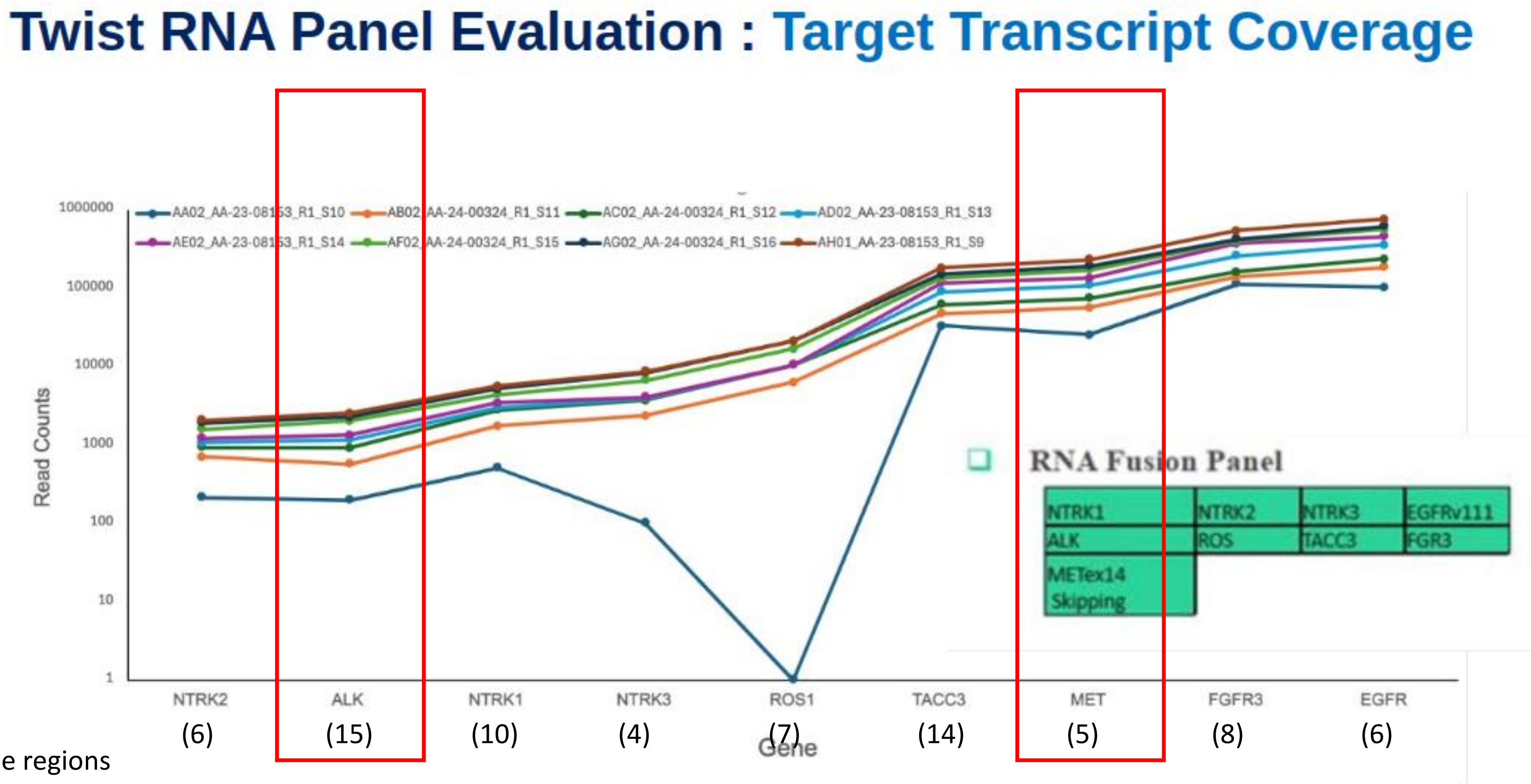
Ref. issues:
- https://actg.atlassian.net/browse/ABIE-976

ACT GENOMICS ™

Probe regions vs Target gene expression

Ref. files:
- 9 target gene expression generated by Chien-Hung
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Twist_kit/Covered_Regions_RNA_Fusions_4X_TE-98493102_GRCh38.bed

- Probe regions vs Gene expression (Twist test data)



Twist RNA Panel Evaluation : Target Transcript Coverage

□ RNA Fusion Panel

NTRK1	NTRK2	NTRK3	EGFRv111
ALK	ROS	TACC3	FGR3
METex14 Skipping			

	NTRK2	ALK	NTRK1	NTRK3	ROS1	TACC3	MET	FGFR3	EGFR
# of Probe regions	(6)	(15)	(10)	(4)	(7)	(14)	(5)	(8)	(6)

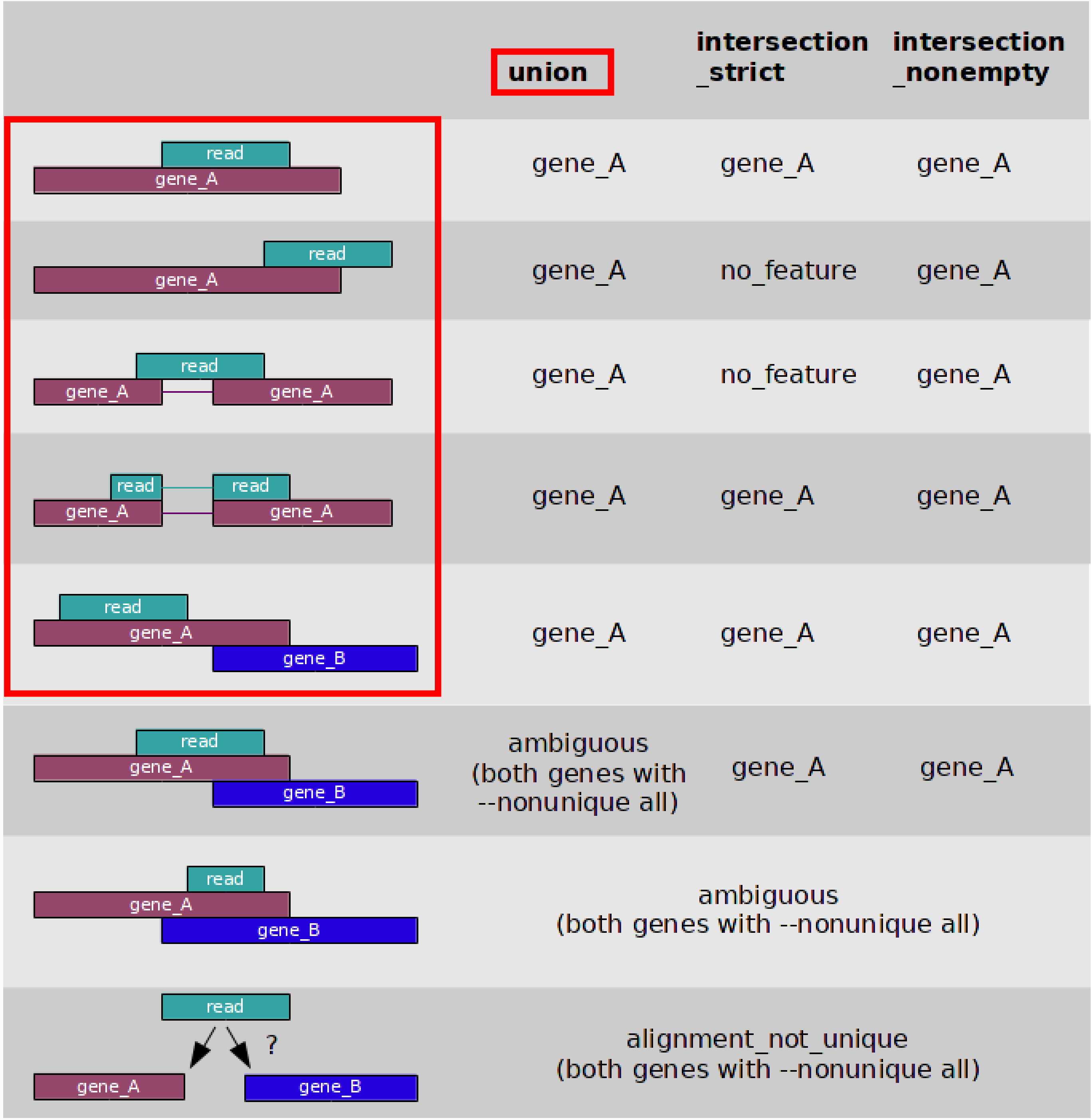
HTSeq-count

- 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

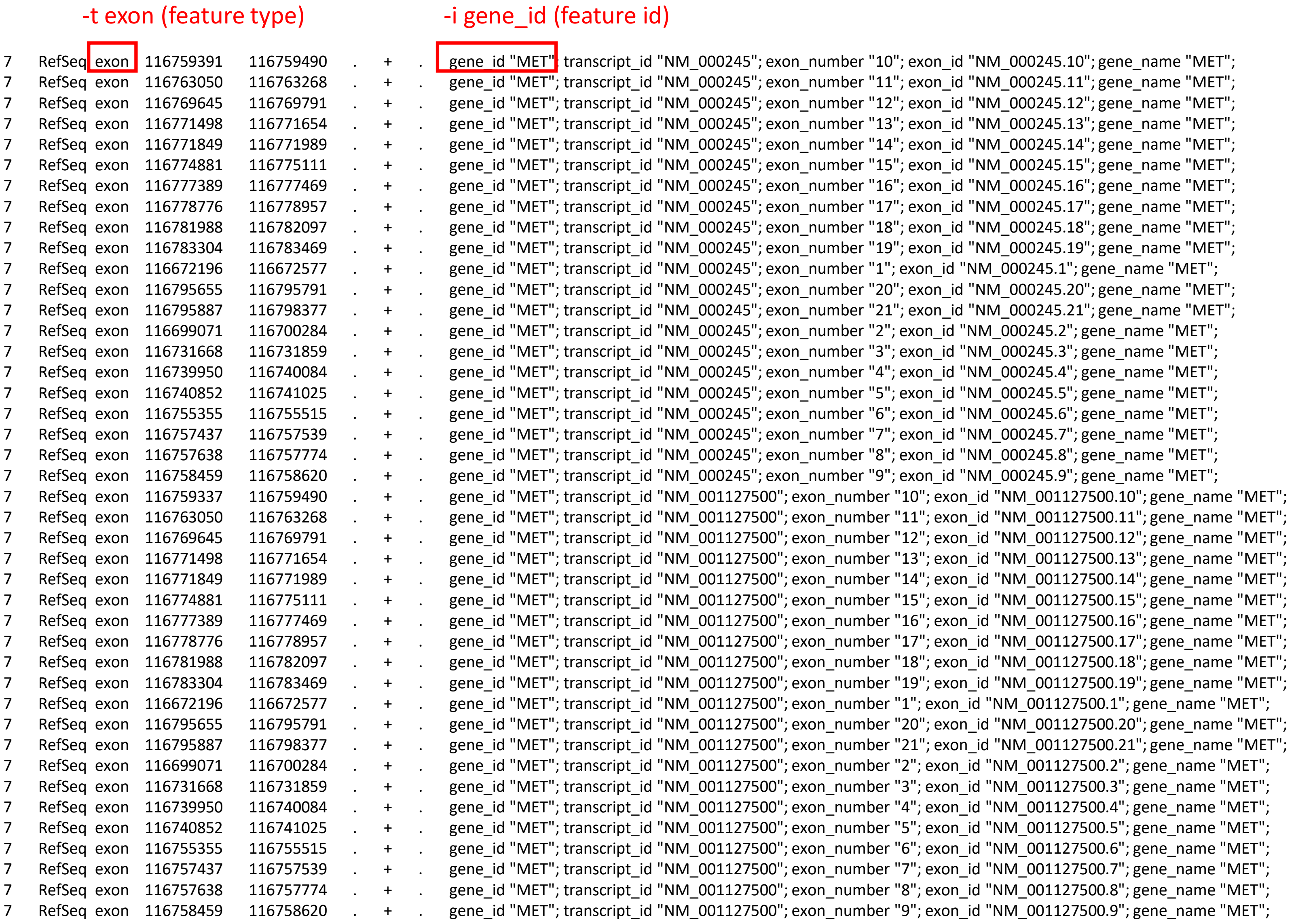
- Htseq-count docs



	union	intersection_strict	intersection_nonempty
read / gene_A	gene_A	gene_A	gene_A
read / gene_A	gene_A	no_feature	gene_A
read / gene_A gene_A	gene_A	no_feature	gene_A
read read / gene_A gene_A	gene_A	gene_A	gene_A
read / gene_A / gene_B	gene_A	gene_A	gene_A
read / gene_A / gene_B	ambiguous (both genes with --nonunique all)	gene_A	gene_A
read / gene_A / gene_B	ambiguous (both genes with --nonunique all)		
read ? / gene_A gene_B	alignment_not_unique (both genes with --nonunique all)		

Arriba’s GTF file

- MET



-t exon (feature type)

-i gene_id (feature id)

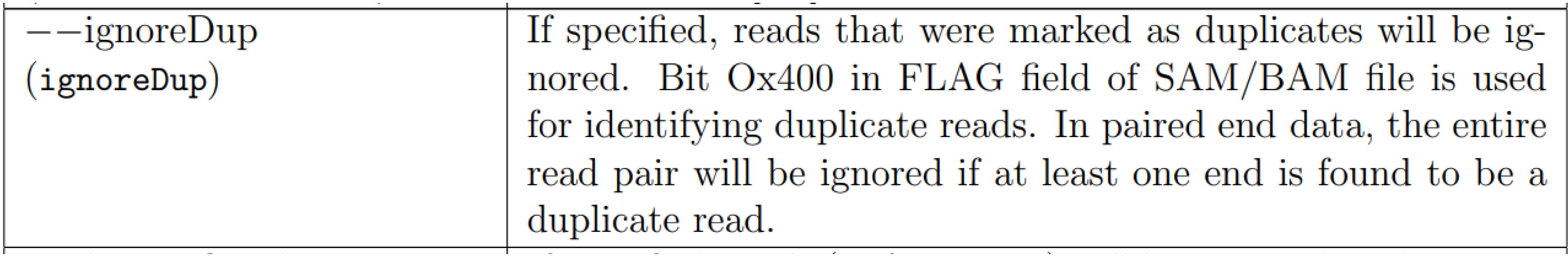
```
7	RefSeq	exon	116759391	116759490	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "10"; exon_id "NM_000245.10"; gene_name "MET";
7	RefSeq	exon	116763050	116763268	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "11"; exon_id "NM_000245.11"; gene_name "MET";
7	RefSeq	exon	116769645	116769791	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "12"; exon_id "NM_000245.12"; gene_name "MET";
7	RefSeq	exon	116771498	116771654	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "13"; exon_id "NM_000245.13"; gene_name "MET";
7	RefSeq	exon	116771849	116771989	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "14"; exon_id "NM_000245.14"; gene_name "MET";
7	RefSeq	exon	116774881	116775111	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "15"; exon_id "NM_000245.15"; gene_name "MET";
7	RefSeq	exon	116777389	116777469	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "16"; exon_id "NM_000245.16"; gene_name "MET";
7	RefSeq	exon	116778776	116778957	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "17"; exon_id "NM_000245.17"; gene_name "MET";
7	RefSeq	exon	116781988	116782097	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "18"; exon_id "NM_000245.18"; gene_name "MET";
7	RefSeq	exon	116783304	116783469	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "19"; exon_id "NM_000245.19"; gene_name "MET";
7	RefSeq	exon	116672196	116672577	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "1"; exon_id "NM_000245.1"; gene_name "MET";
7	RefSeq	exon	116795655	116795791	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "20"; exon_id "NM_000245.20"; gene_name "MET";
7	RefSeq	exon	116795887	116798377	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "21"; exon_id "NM_000245.21"; gene_name "MET";
7	RefSeq	exon	116699071	116700284	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "2"; exon_id "NM_000245.2"; gene_name "MET";
7	RefSeq	exon	116731668	116731859	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "3"; exon_id "NM_000245.3"; gene_name "MET";
7	RefSeq	exon	116739950	116740084	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "4"; exon_id "NM_000245.4"; gene_name "MET";
7	RefSeq	exon	116740852	116741025	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "5"; exon_id "NM_000245.5"; gene_name "MET";
7	RefSeq	exon	116755355	116755515	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "6"; exon_id "NM_000245.6"; gene_name "MET";
7	RefSeq	exon	116757437	116757539	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "7"; exon_id "NM_000245.7"; gene_name "MET";
7	RefSeq	exon	116757638	116757774	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "8"; exon_id "NM_000245.8"; gene_name "MET";
7	RefSeq	exon	116758459	116758620	.	+	.	gene_id "MET"; transcript_id "NM_000245"; exon_number "9"; exon_id "NM_000245.9"; gene_name "MET";
7	RefSeq	exon	116759337	116759490	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "10"; exon_id "NM_001127500.10"; gene_name "MET";
7	RefSeq	exon	116763050	116763268	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "11"; exon_id "NM_001127500.11"; gene_name "MET";
7	RefSeq	exon	116769645	116769791	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "12"; exon_id "NM_001127500.12"; gene_name "MET";
7	RefSeq	exon	116771498	116771654	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "13"; exon_id "NM_001127500.13"; gene_name "MET";
7	RefSeq	exon	116771849	116771989	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "14"; exon_id "NM_001127500.14"; gene_name "MET";
7	RefSeq	exon	116774881	116775111	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "15"; exon_id "NM_001127500.15"; gene_name "MET";
7	RefSeq	exon	116777389	116777469	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "16"; exon_id "NM_001127500.16"; gene_name "MET";
7	RefSeq	exon	116778776	116778957	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "17"; exon_id "NM_001127500.17"; gene_name "MET";
7	RefSeq	exon	116781988	116782097	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "18"; exon_id "NM_001127500.18"; gene_name "MET";
7	RefSeq	exon	116783304	116783469	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "19"; exon_id "NM_001127500.19"; gene_name "MET";
7	RefSeq	exon	116672196	116672577	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "1"; exon_id "NM_001127500.1"; gene_name "MET";
7	RefSeq	exon	116795655	116795791	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "20"; exon_id "NM_001127500.20"; gene_name "MET";
7	RefSeq	exon	116795887	116798377	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "21"; exon_id "NM_001127500.21"; gene_name "MET";
7	RefSeq	exon	116699071	116700284	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "2"; exon_id "NM_001127500.2"; gene_name "MET";
7	RefSeq	exon	116731668	116731859	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "3"; exon_id "NM_001127500.3"; gene_name "MET";
7	RefSeq	exon	116739950	116740084	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "4"; exon_id "NM_001127500.4"; gene_name "MET";
7	RefSeq	exon	116740852	116741025	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "5"; exon_id "NM_001127500.5"; gene_name "MET";
7	RefSeq	exon	116755355	116755515	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "6"; exon_id "NM_001127500.6"; gene_name "MET";
7	RefSeq	exon	116757437	116757539	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "7"; exon_id "NM_001127500.7"; gene_name "MET";
7	RefSeq	exon	116757638	116757774	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "8"; exon_id "NM_001127500.8"; gene_name "MET";
7	RefSeq	exon	116758459	116758620	.	+	.	gene_id "MET"; transcript_id "NM_001127500"; exon_number "9"; exon_id "NM_001127500.9"; gene_name "MET";
```

Ref. file:

- /mnt/RD_Develop/sandyteng/FusionCaptureTools/refdb_arriba/RefSeq_hg38.gtf

FeatureCounts

- Ignore duplicates (−−ignoreDup)



−−ignoreDup (ignoreDup)	If specified, reads that were marked as duplicates will be ignored. Bit Ox400 in FLAG field of SAM/BAM file is used for identifying duplicate reads. In paired end data, the entire read pair will be ignored if at least one end is found to be a duplicate read.

Appendix: FeatureCounts (bitwise FLAG/tag sensitive arguments)

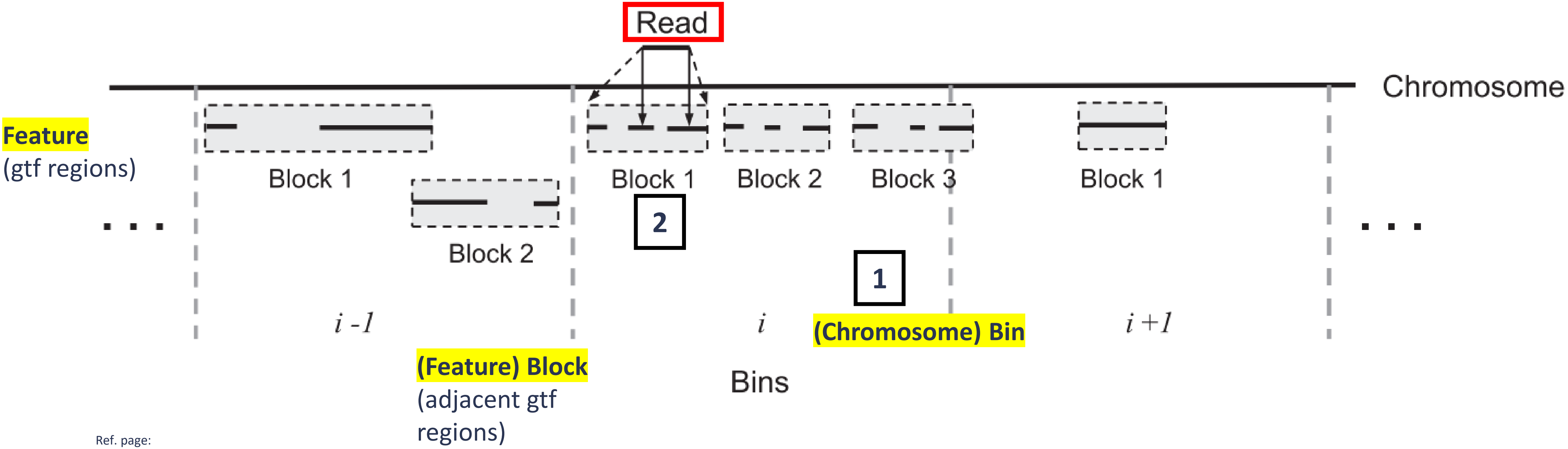
- FLAG sensitive feature count calculation
- FLAGs
    - −−ignoreDup (Bit Ox400 in FLAG field)
    - −−primary (Primary and secondary alignments are identified using bit 0x100 in the Flag field)
    - −−fraction (fractional count for ‘NH’ tag)
    - -M (countMultiMappingReads) (‘NH’ tag)
    - -B < int > (nBestLocations) (Specify the maximal number of equally-best mapping locations to be reported for a read. 1 by default.) (‘NH’ tag)

FeatureCounts

Steps:

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

- FeatureCounts (algorithm)



Ref. page:

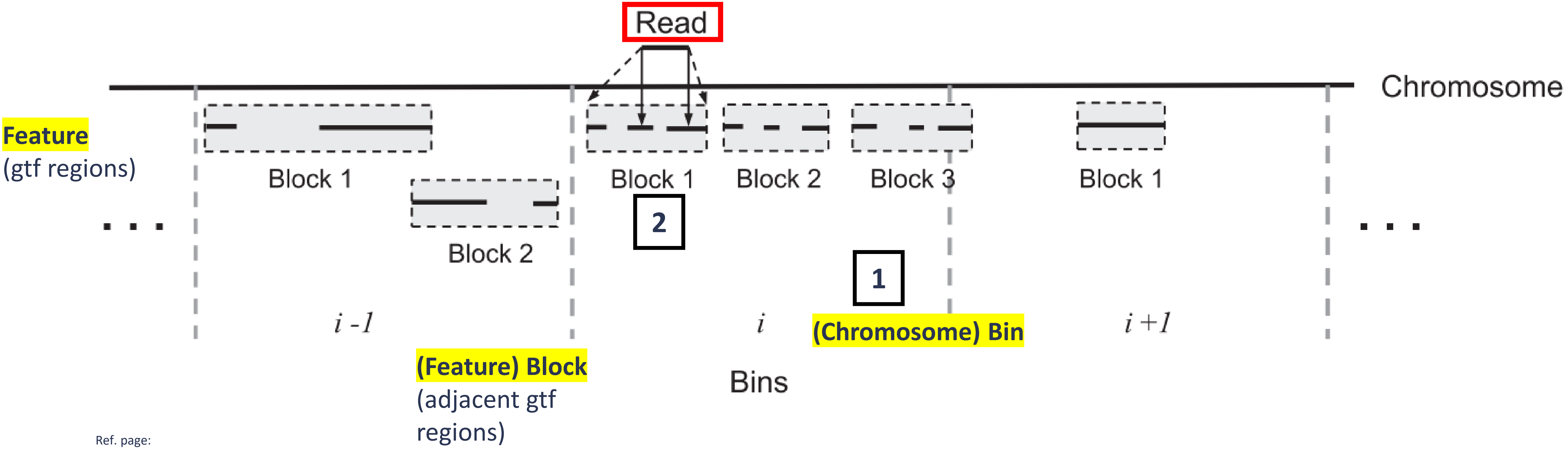
- featureCounts: an efficient general purpose program for assigning sequence reads to genomic features

FeatureCounts

- FeatureCounts (algorithm)

Genomic bins & Feature blocks

- Same number of consecutive features are grouped into a block
- The number of features in a block is nearly equal to the number of blocks in a bin
  => # of blocks in a bin = sqrt(# of features in a bin)



Ref. page:

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

Reference file:

- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_sampleQC.NextSeq/output/genes_featureCounts.count.summary

FeatureCounts Output – count summary

- genes_featureCounts.count.summary

Category	Description
Unassigned Unmapped	Unmapped reads that cannot be assigned.
Unassigned MultiMapping	Alignments reported for multi-mapping reads (indicated by the 'NH' tag).
Unassigned NoFeatures	Alignments that do not overlap any feature.
Unassigned Ambiguity	Alignments that overlap two or more features (for feature-level summarization) or meta-features (for meta-feature-level summarization).

```
Status  /mnt/RD_Develop/sandyteng/ACTFusionV5/20250122_TwistBioscience/testresult/arriba_grch38/AANB02_202_AD02_AA-23-08153/Aligned.sortedByCoord.out.bam
Assigned       1656798
Unassigned_Unmapped     464124
Unassigned_Read_Type    0
Unassigned_Singleton    0
Unassigned_MappingQuality       0
Unassigned_Chimera      0
Unassigned_FragmentLength       0
Unassigned_Duplicate    0
Unassigned_MultiMapping 2976182
Unassigned_Secondary    0
Unassigned_NonSplit     0
Unassigned_NoFeatures   77794
Unassigned_Overlapping_Length   0
Unassigned_Ambiguity    51670
```

Reference file:
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_sampleQC.NextSeq/output/genes_featureCounts.count

FeatureCounts Output – count table

- genes_featureCounts.count

Available columns:
- annotation columns (‘Geneid’, ‘Chr’, ‘Start’, ‘End’, ‘Strand’ and ‘Length’)
- data columns (eg. read counts for genes for each library)

‘Geneid’

```
# MET
MET   7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7;7
116672196;116672196;116672196;116672196;116699071;116699071;116699071;116731668;116731668;116731668;116731668;116739950;116739950;116739950;116739950;1167
40852;116740852;116740852;116740852;116755355;116755355;116755355;116755355;116757437;116757437;116757437;116757437;116757638;116757638;116757638;11675763
8;116758459;116758459;116758459;116758459;116759337;116759391;116759391;116759391;116763050;116763050;116763050;116763050;116769645;116769645;116769645;11
6769645;116771498;116771498;116771498;116771849;116771849;116771849;116774881;116774881;116774881;116777389;116777389;116777389;116778776;116778776;116778
776;116781988;116781988;116781988;116783304;116783304;116783304;116795655;116795655;116795655;116795887;116795887;116795887
116672577;116672577;116672577;116672577;116700284;116700284;116700284;116731859;116731859;116731859;116731859;116740084;116740084;116740084;116740084;1167
41025;116741025;116741025;116741025;116755515;116755515;116755515;116755515;116757539;116757539;116757539;116757539;116757774;116757774;116757774;11675777
4;116758620;116758620;116758620;116758620;116759490;116759490;116759490;116759490;116763268;116763268;116763268;116763268;116769791;116769910;116769791;11
6769791;116771654;116771654;116771654;116771989;116771989;116771989;116775111;116775111;116775111;116777469;116777469;116777469;116778957;116778957;116778
957;116782097;116782097;116782097;116783469;116783469;116783469;116795791;116795791;116795791;116798377;116798377;116798377
+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+;+ 6995   16663
```

Read counts

‘Length’ (# of non-overlapping bases)

```
# ALK
ALK   2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2;2
29192774;29192774;29196770;29196770;29197542;29197542;29207171;29207171;29209786;29209786;29213984;29213984;29220706;29220706;29222344;29222344;29222517;29
222517;29223342;29223342;29225461;29226922;29227574;29228884;29232304;29233565;29239680;29251105;29275099;29275402;29296888;29318304;29320751;29328350;2938
3732;29531915;29694850;29717578;29919993
29193922;29193922;29196860;29196860;29197676;29197676;29207272;29207272;29209878;29209878;29214081;29214081;29220835;29220835;29222408;29222408;29222607;29
222607;29223900;29223528;29225565;29227074;29227672;29229066;29232448;29233696;29239830;29251267;29275227;29275496;29297057;29318404;29320882;29328481;2938
3859;29532116;29695014;29717697;29921586   -;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-;-   6612   143
```

ACT GENOMICS ™

Summary

- Gene count correlation ~99% => the counts obtained by the 2 tools are similar.

- Built-in read filters (htseq and featureCounts)
  - Duplicates can be excluded by FeatureCounts “−−ignoreDup” argument in featureCounts
  - Multi-mapping reads are recognized via “NH” tag (“__alignment_not_unique” in htseq / “Unassigned MultiMapping” in featureCounts)
    => Not applicable for caller that does not produce “NH” tag. (bwa => produce “XA” tag for multi-mapping reads)

- samtools depth (no output) & bedtools coverage -d (0 depth)
  - Output depth for each “position” => Hard to interpret (Fusion is report on exon-level)

Sample QC metrics re-visit

Arriba (STAR-based)

Fusionv4 (bwa-based)

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)

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)

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

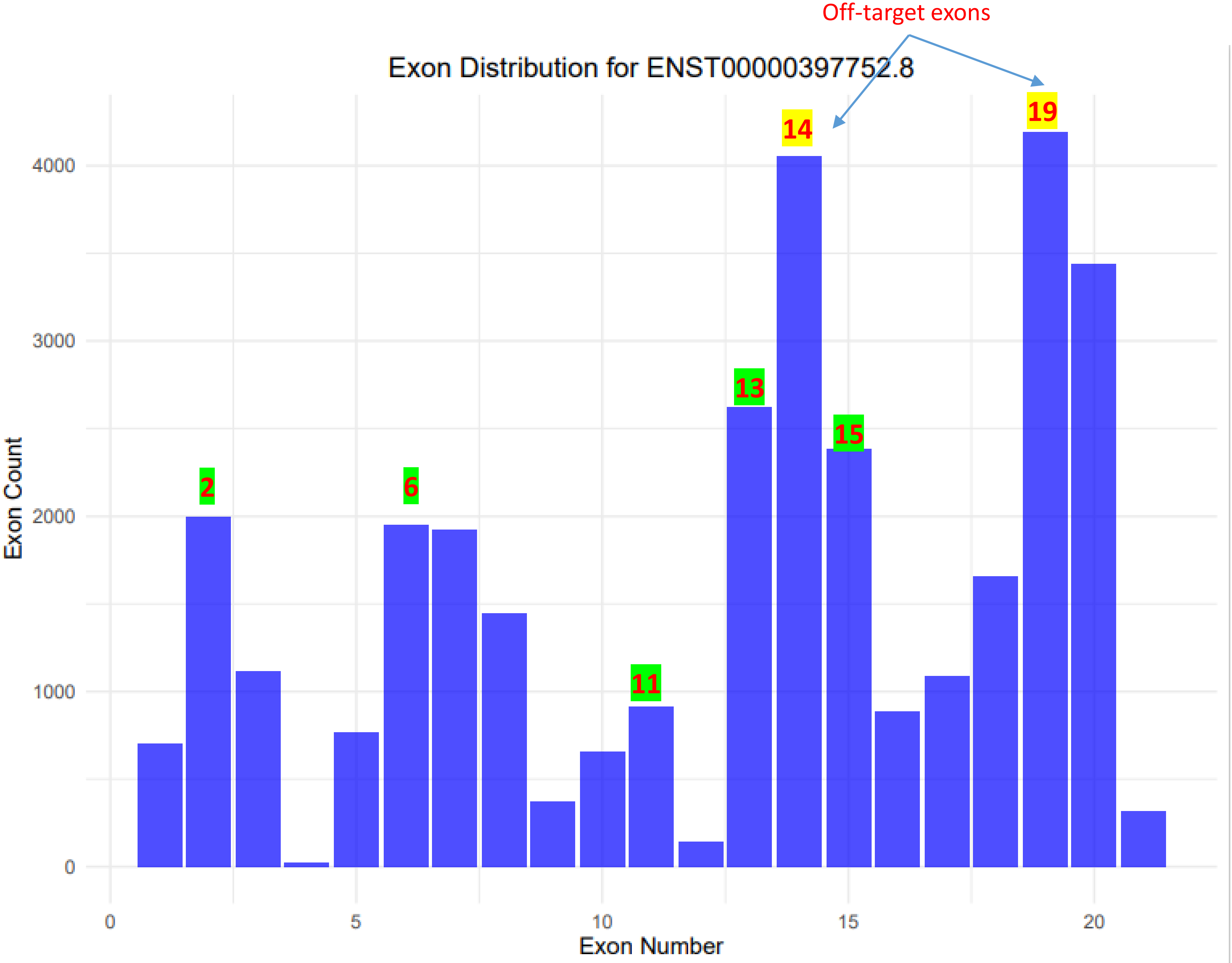
Script:
- /mnt/RD_Develop/sandyteng/ACTFusionV5/code/plot_exon_distribution_test/get_ENST00000397752.8.MET.exon.plot.sh

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

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



Off-target exons

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%

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)

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)

On-Target %

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

NextSeq, Twist 8 RNA data

- % 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%

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)
```

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

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

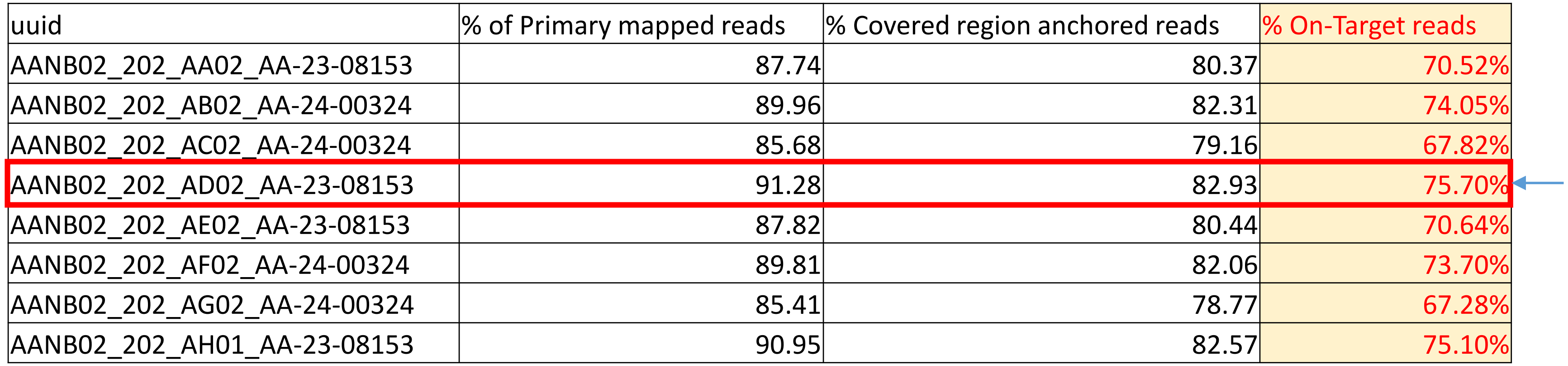
Source files:
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_analysis.NextSeq/STAR-arriba/<uuid>.flagstats.txt
- /mnt/RD_Develop/sandyteng/workdir/bed_intersect/Probe_analysis.NextSeq/STAR-arriba/<uuid>_probe_report.txt

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)

```
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)
```

samtools flagstats

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

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%

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%

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

Aligned.bam 1

```
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)
```

```
samtools view -b -F 0x900
```

Filter out secondary and supplementary alignments from the input BAM

Primary.bam 2

```
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)
```

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

Probed.bam 3

```
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)
```

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

% 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 ™

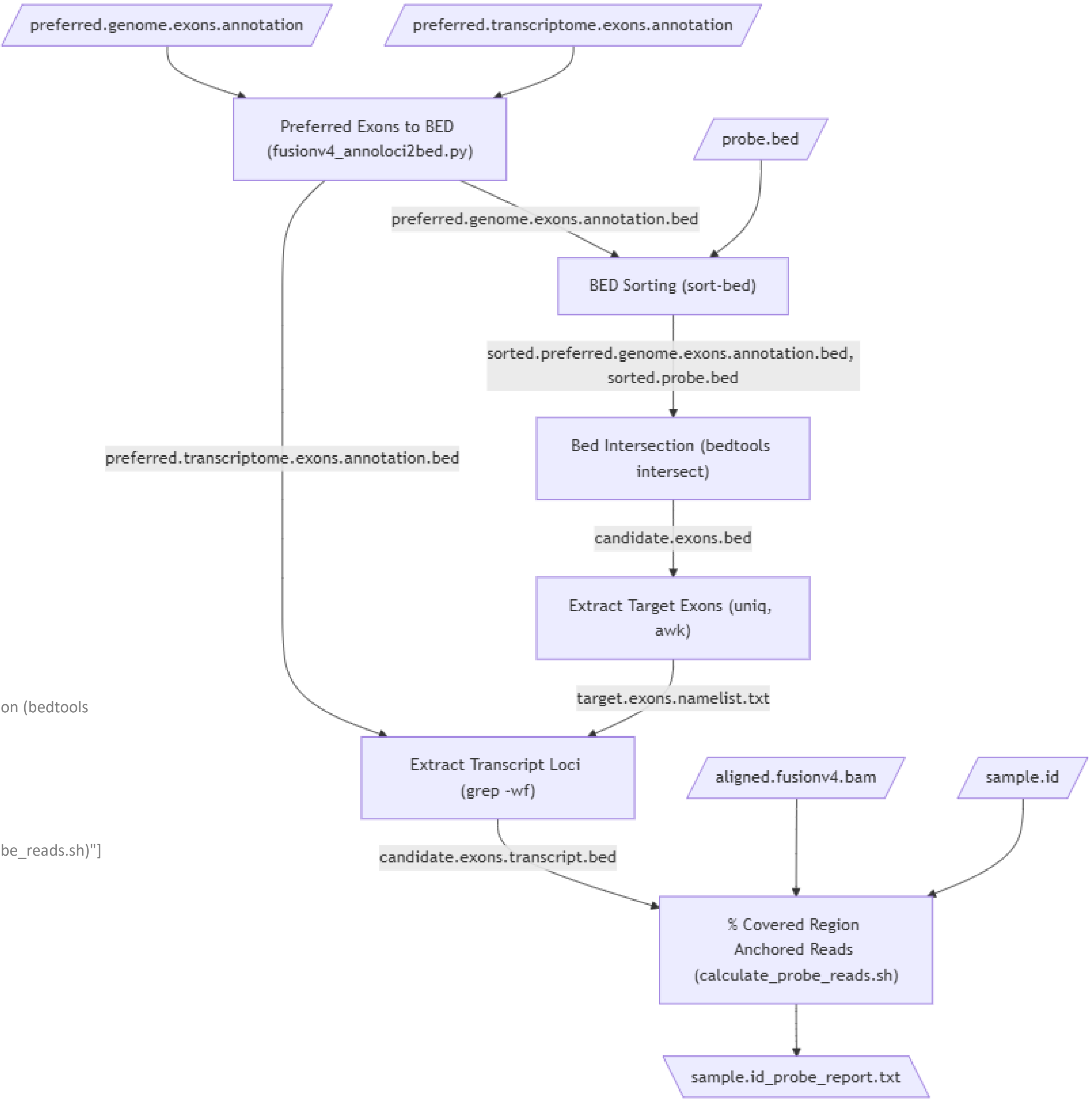
% Covered region anchored reads calculation (arriba)

- 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

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\]
```



preferred.genome.exons.annotation

preferred.transcriptome.exons.annotation

Preferred Exons to BED (fusionv4_annoloci2bed.py)

probe.bed

preferred.genome.exons.annotation.bed

BED Sorting (sort-bed)

sorted.preferred.genome.exons.annotation.bed, sorted.probe.bed

preferred.transcriptome.exons.annotation.bed

Bed Intersection (bedtools intersect)

candidate.exons.bed

Extract Target Exons (uniq, awk)

target.exons.namelist.txt

Extract Transcript Loci (grep -wf)

aligned.fusionv4.bam

sample.id

candidate.exons.transcript.bed

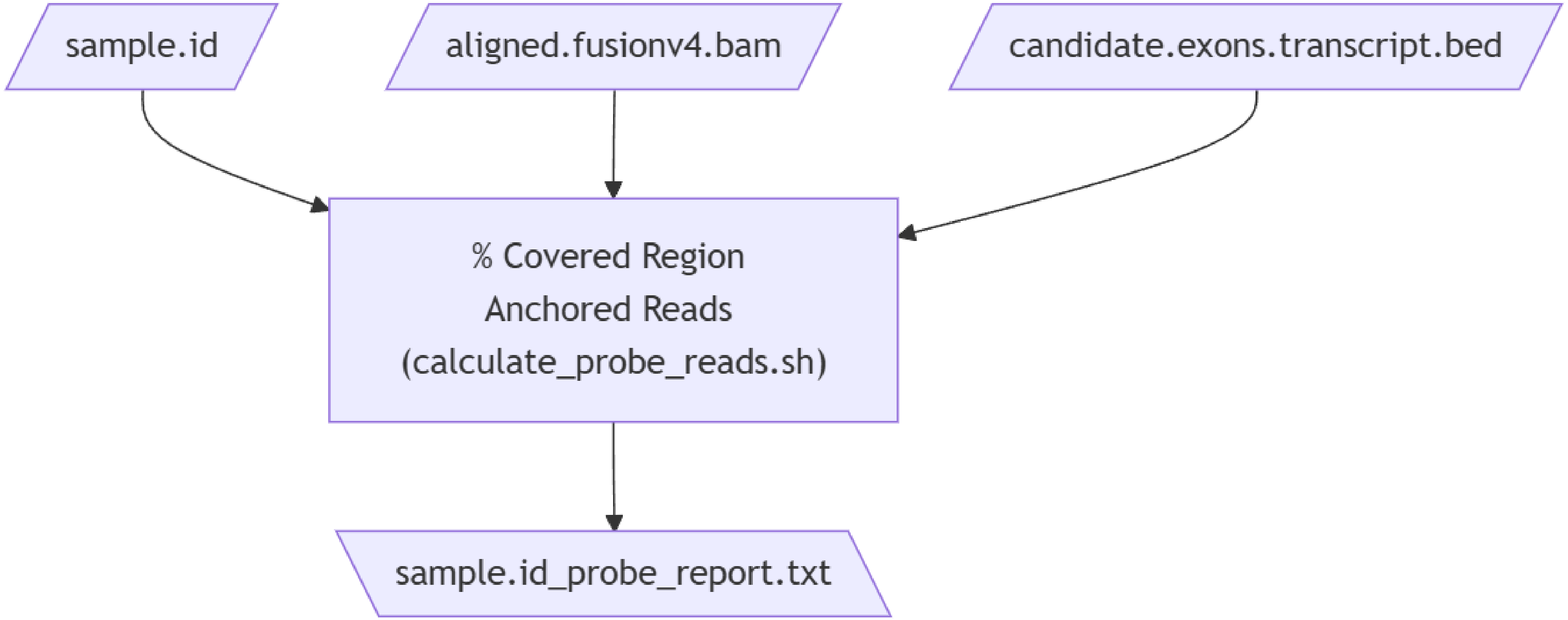
% Covered Region Anchored Reads (calculate_probe_reads.sh)

sample.id_probe_report.txt

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\]
```



sample.id

aligned.fusionv4.bam

candidate.exons.transcript.bed

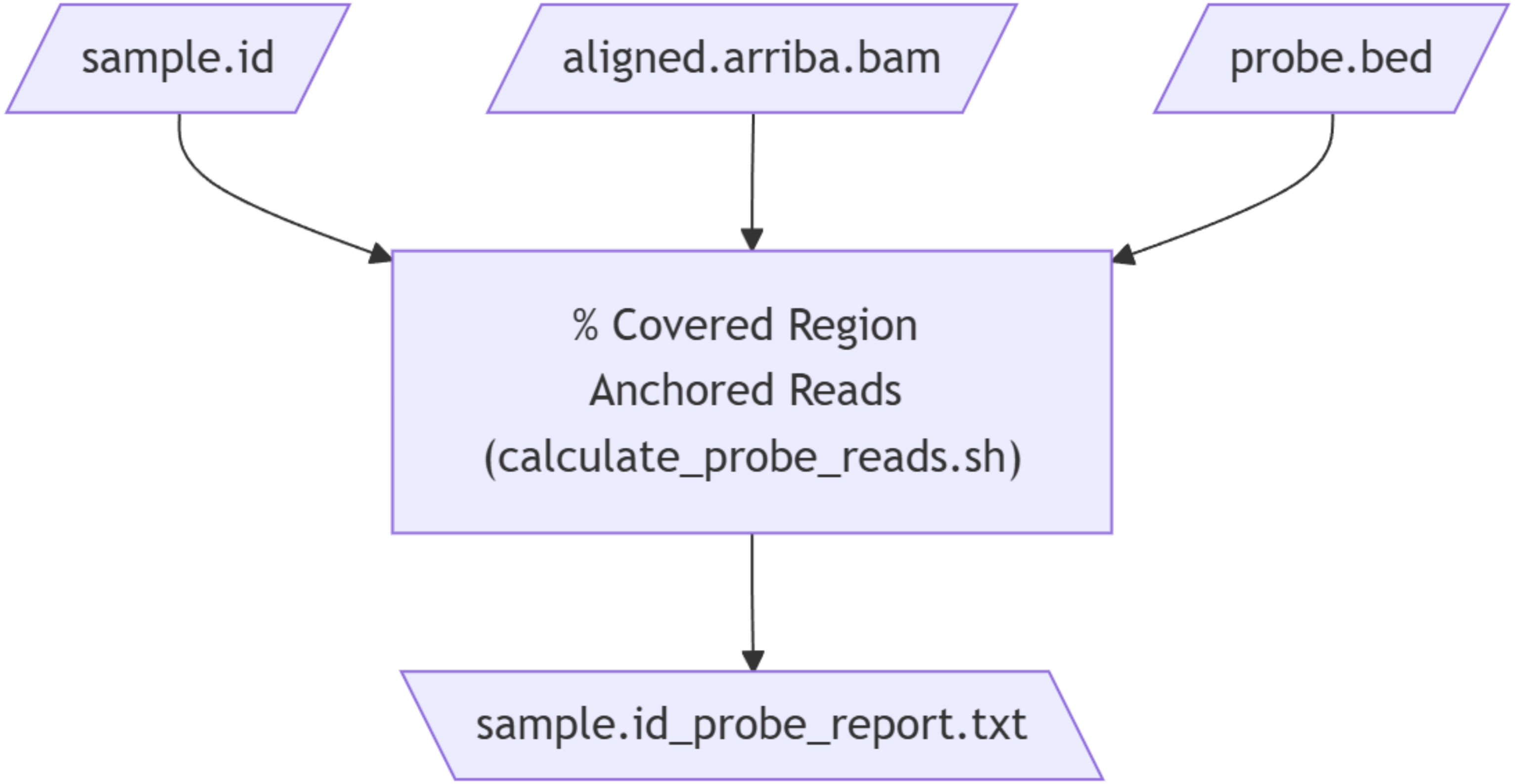
% Covered Region Anchored Reads (calculate_probe_reads.sh)

sample.id_probe_report.txt

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\]
```


sample.id
aligned.arriba.bam
probe.bed
% Covered Region Anchored Reads (calculate_probe_reads.sh)
sample.id_probe_report.txt

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)

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)

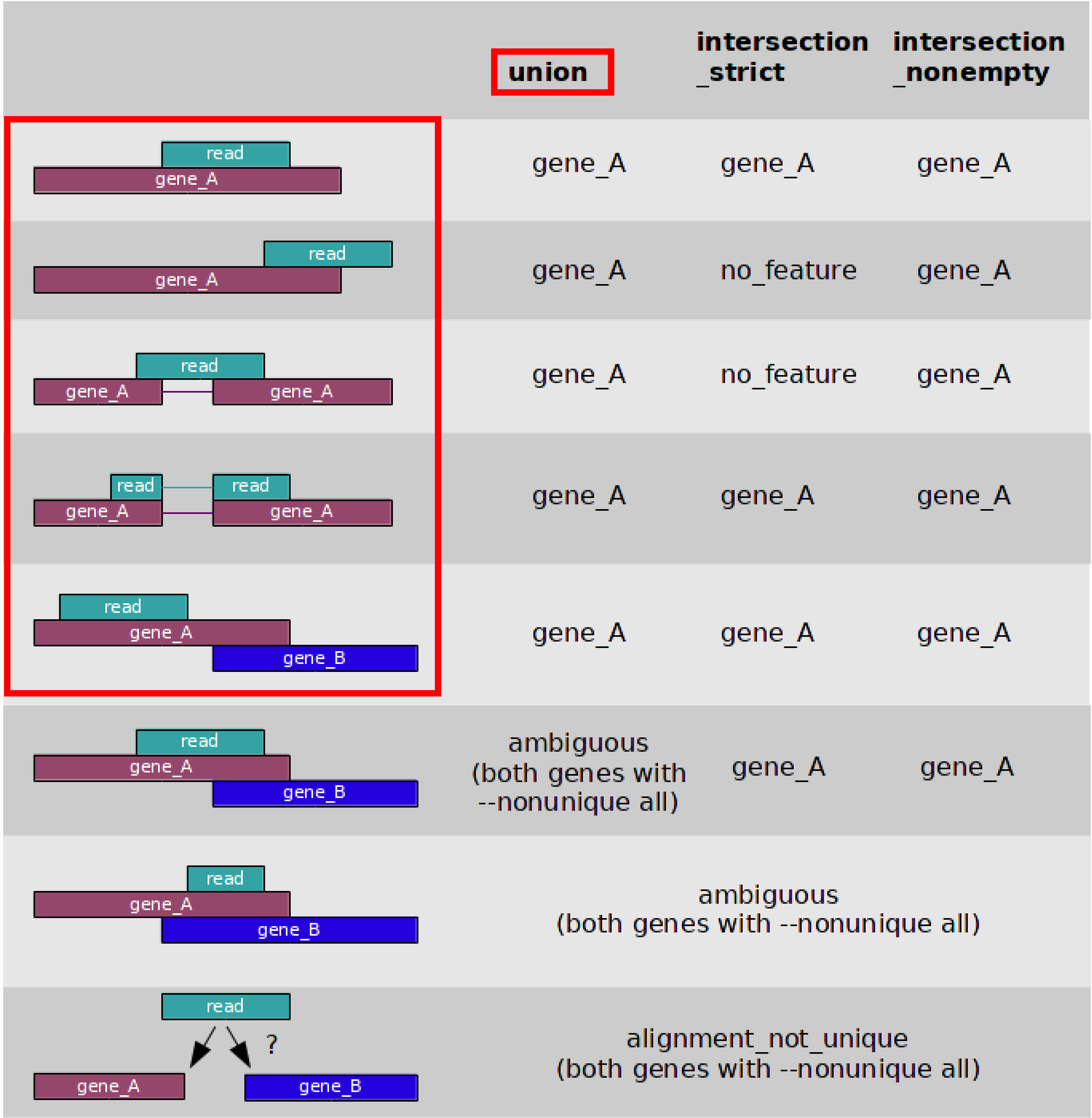
HTSeq-count

- 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



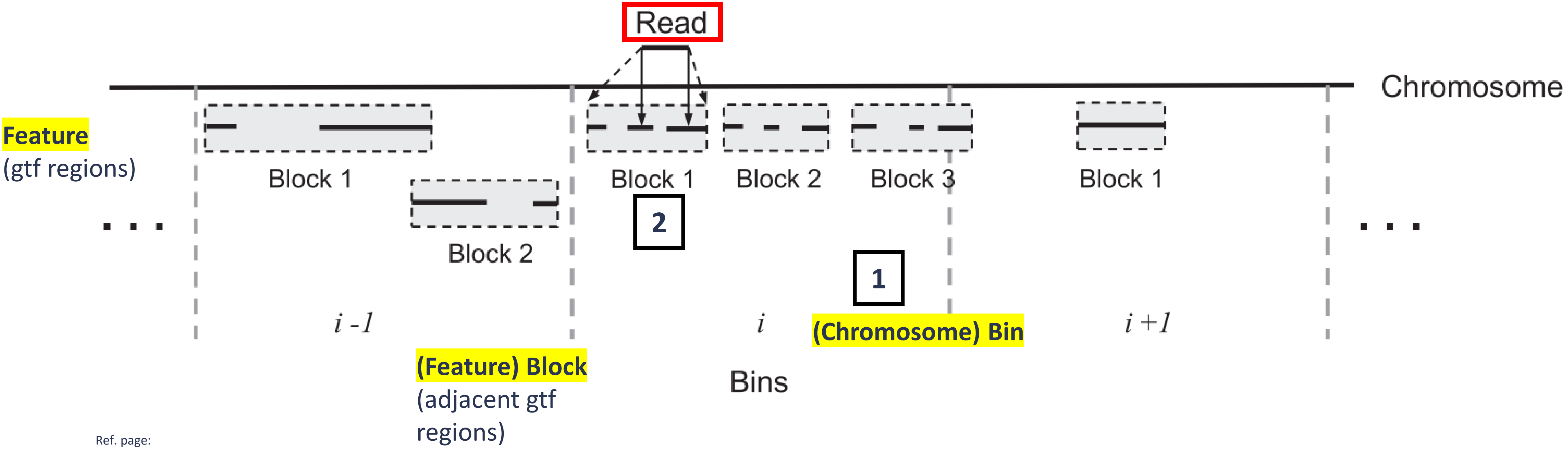
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

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

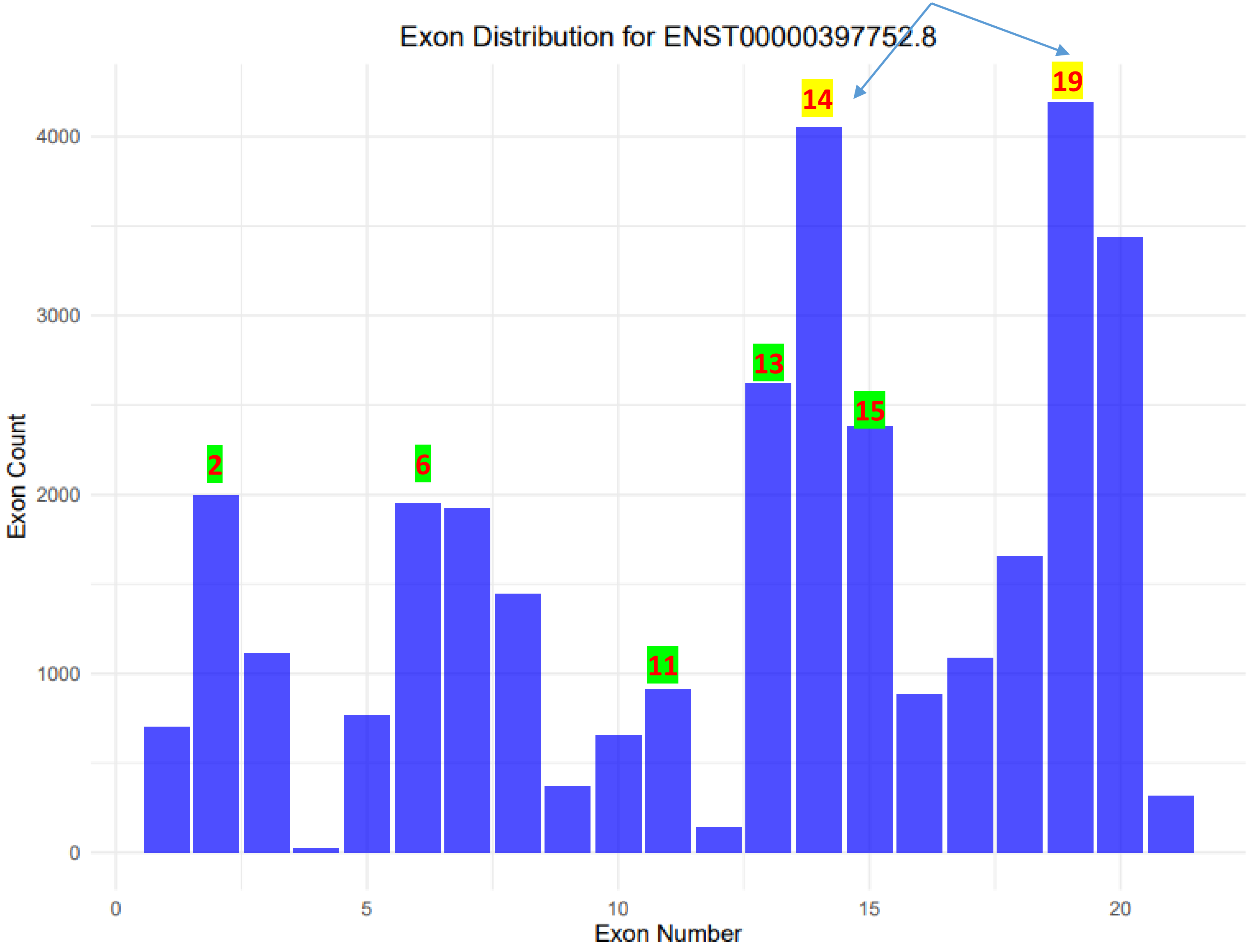
Exon number vs Count

- 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)



Off-target exons

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

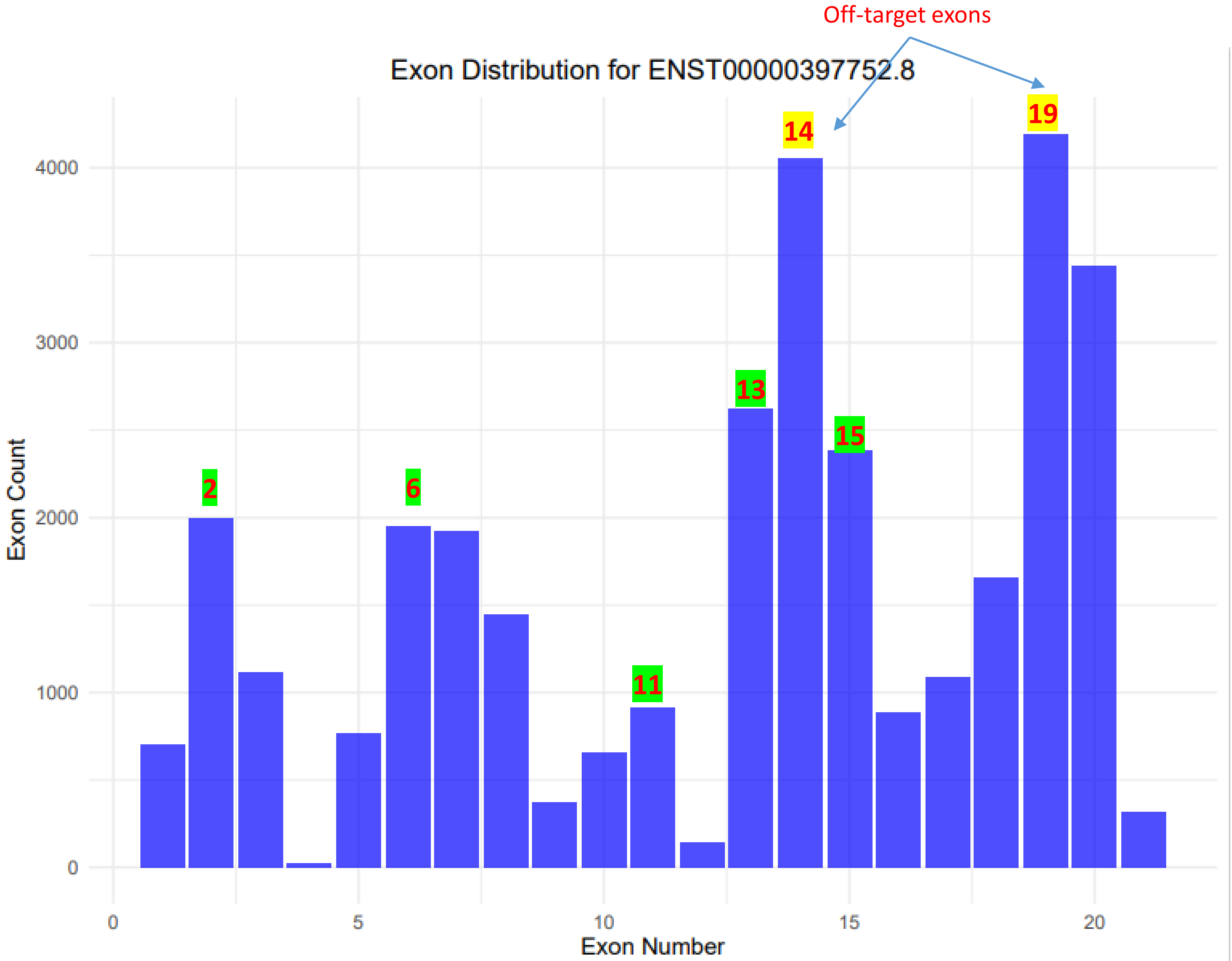
Exon number vs Count

- 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

Gene quantification (Arriba)

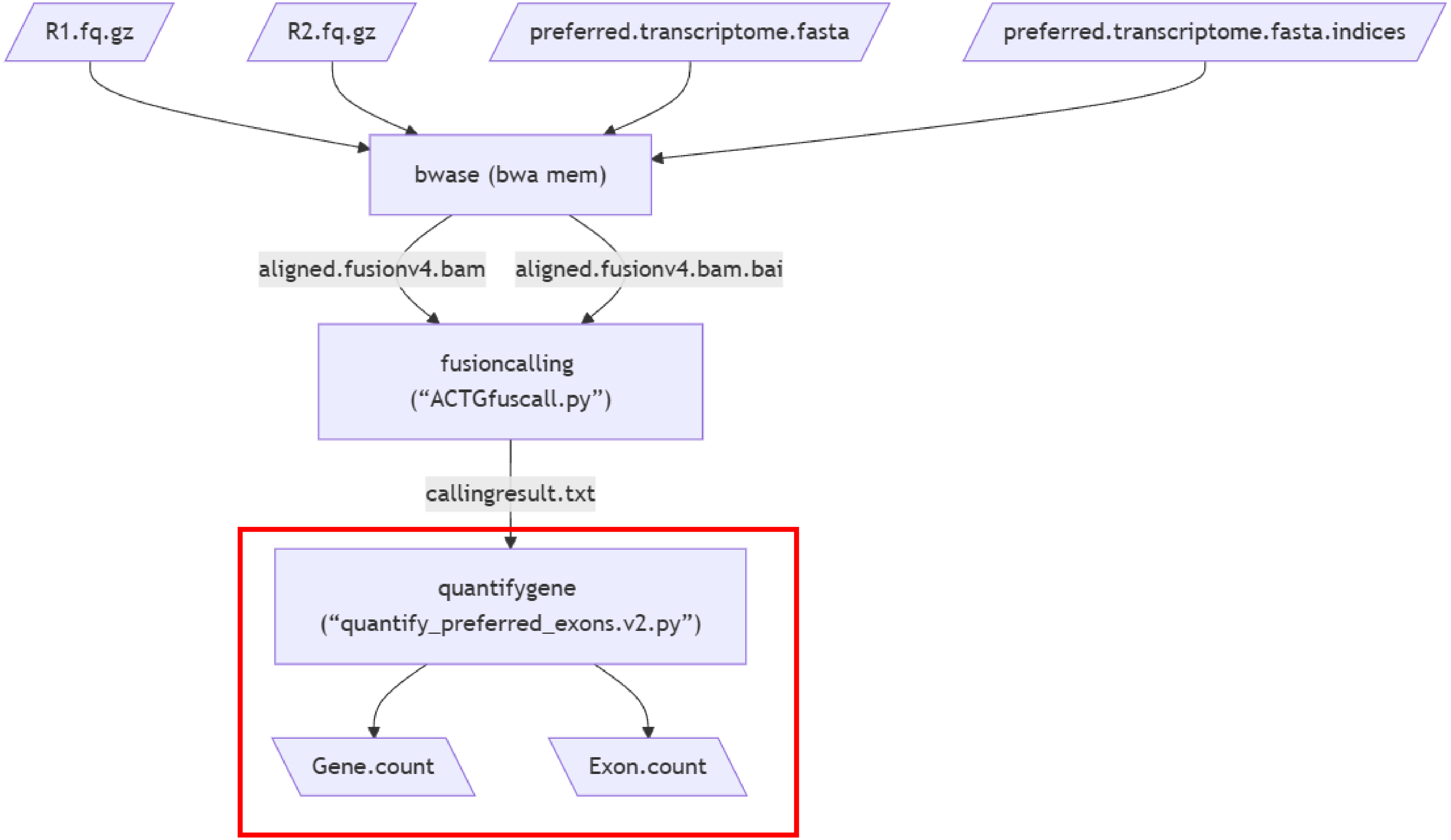
Ref. issue
https://actg.atlassian.net/browse/ABIE-987
https://actg.atlassian.net/browse/ABIE-988

- 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)

Workflow

- FusionV4

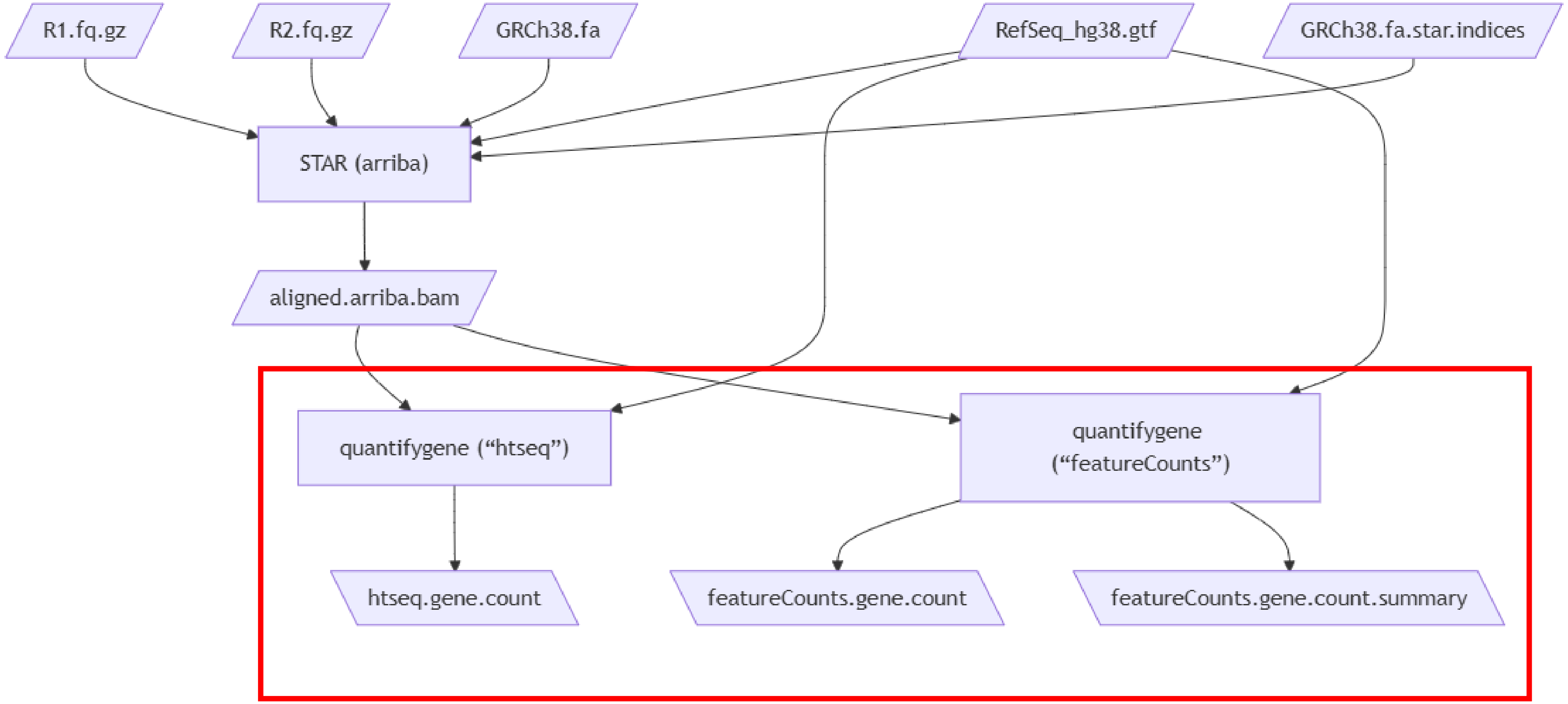
```
graph TD;
  %% Initial Inputs
  I1[/R1.fq.gz/]
  I2[/R2.fq.gz/]
  I7[/preferred.transcriptome.fasta/]
  I8[/preferred.transcriptome.fasta.indices/]
%% FusionV4 Workflow
  I1 --> A1
  I2 --> A1
  I7 --> A1["bwase (bwa mem)"]
  I8 --> A1
  A1 -->|aligned.fusionv4.bam| B1
  A1 -->|aligned.fusionv4.bam.bai|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


R1.fq.gz
R2.fq.gz
preferred.transcriptome.fasta
preferred.transcriptome.fasta.indices
bwase (bwa mem)
aligned.fusionv4.bam
aligned.fusionv4.bam.bai
fusioncalling (“ACTGfuscall.py”)
callingresult.txt
quantifygene (“quantify_preferred_exons.v2.py”)
Gene.count
Exon.count

Workflow

- Arriba

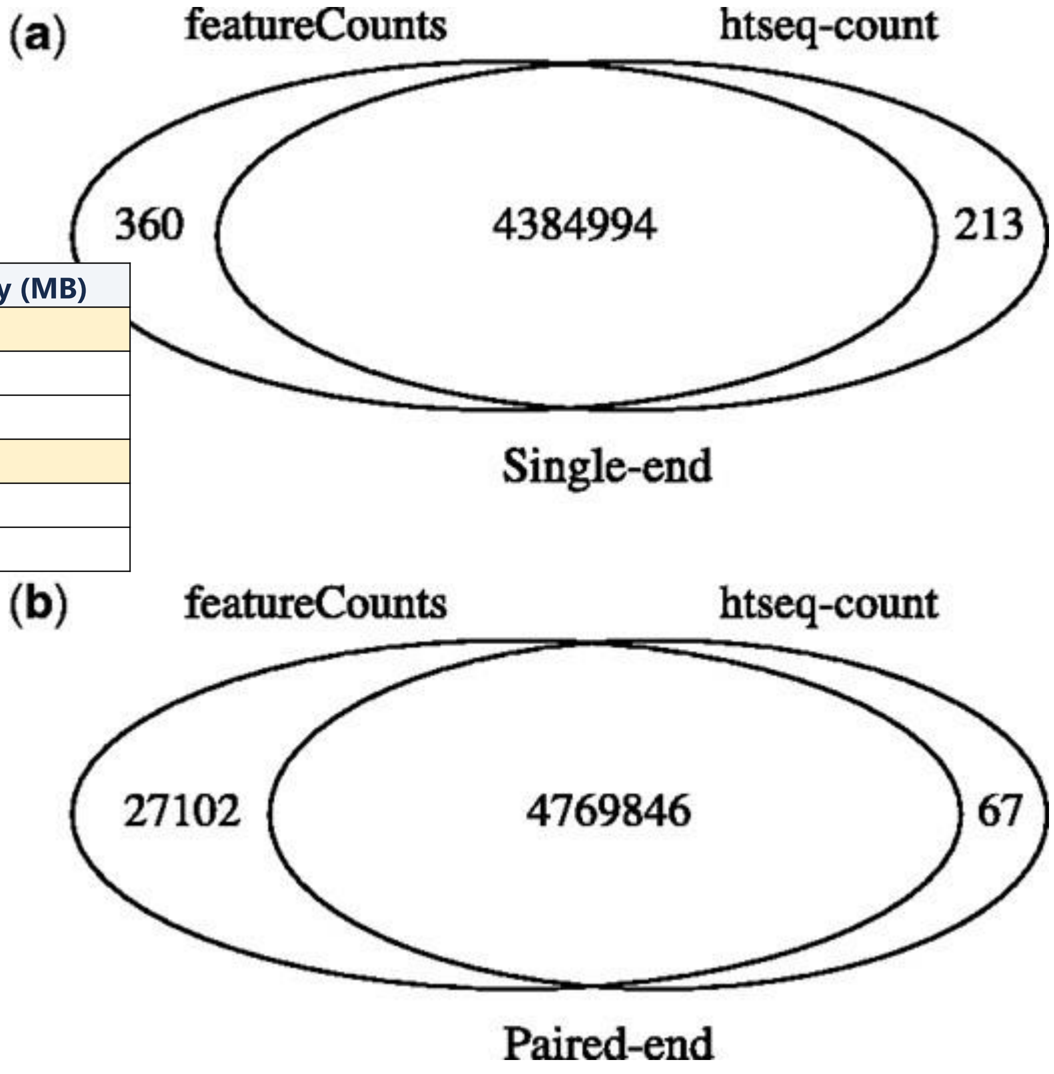
```
graph TD;
  %% Initial Inputs
  I1[/R1.fq.gz/]
  I2[/R2.fq.gz/]
  I6[/GRCh38.fa/]
  I7[/GRCh38.fa.star.indices/]
  I8[/RefSeq_hg38.gtf/]
%% FusionV4 Workflow
  I1 --> A1
  I2 --> A1
  I6 --> A1
  I7 --> A1["STAR (arriba)"]
  I8 --> A1
  I8 --> B1
  I8 --> C1
  A1 --> I9[/aligned.arriba.bam/]
  I9 --> 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\]
```



Gene count comparison (paper)

- HTSeq/FeatureCounts

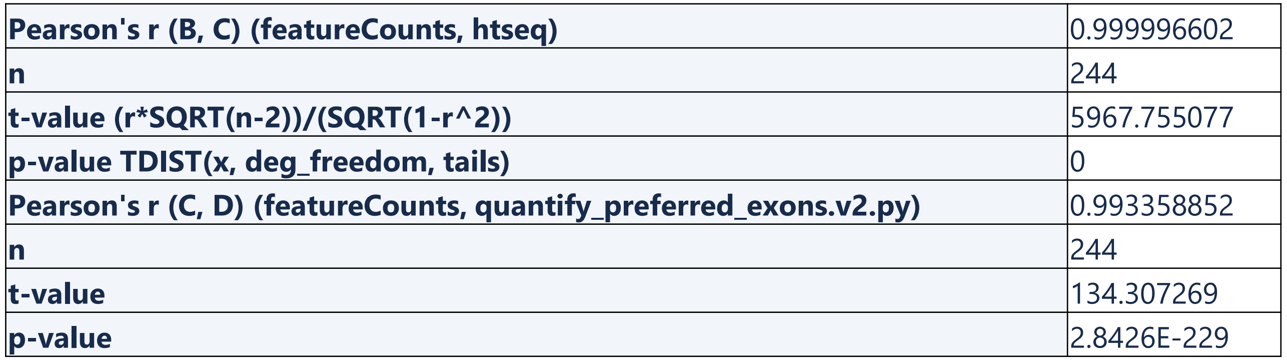
Method	Number of fragments	Time (min)	Memory (MB)
featureCounts	5 392 155	0.9	4
CountOverlaps (whole genome at once)	5 392 155	24.4	7000
CountOverlaps (by chromosome)	5 392 155	36.6	783
htseq-count (union)	4 978 050	36	31
htseq-count (intersection-nonempty)	4 993 644	35.7	31
coverageBED	5 366 902	4.4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



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

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

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 ™

ACT GENOMICS ™

Make Personalized Medicine Accessible to All