ACT GENOMICS ™

Tool comparison study - gene fusion

2024/12/30

Sandy Teng

Contents

Workflow comparison

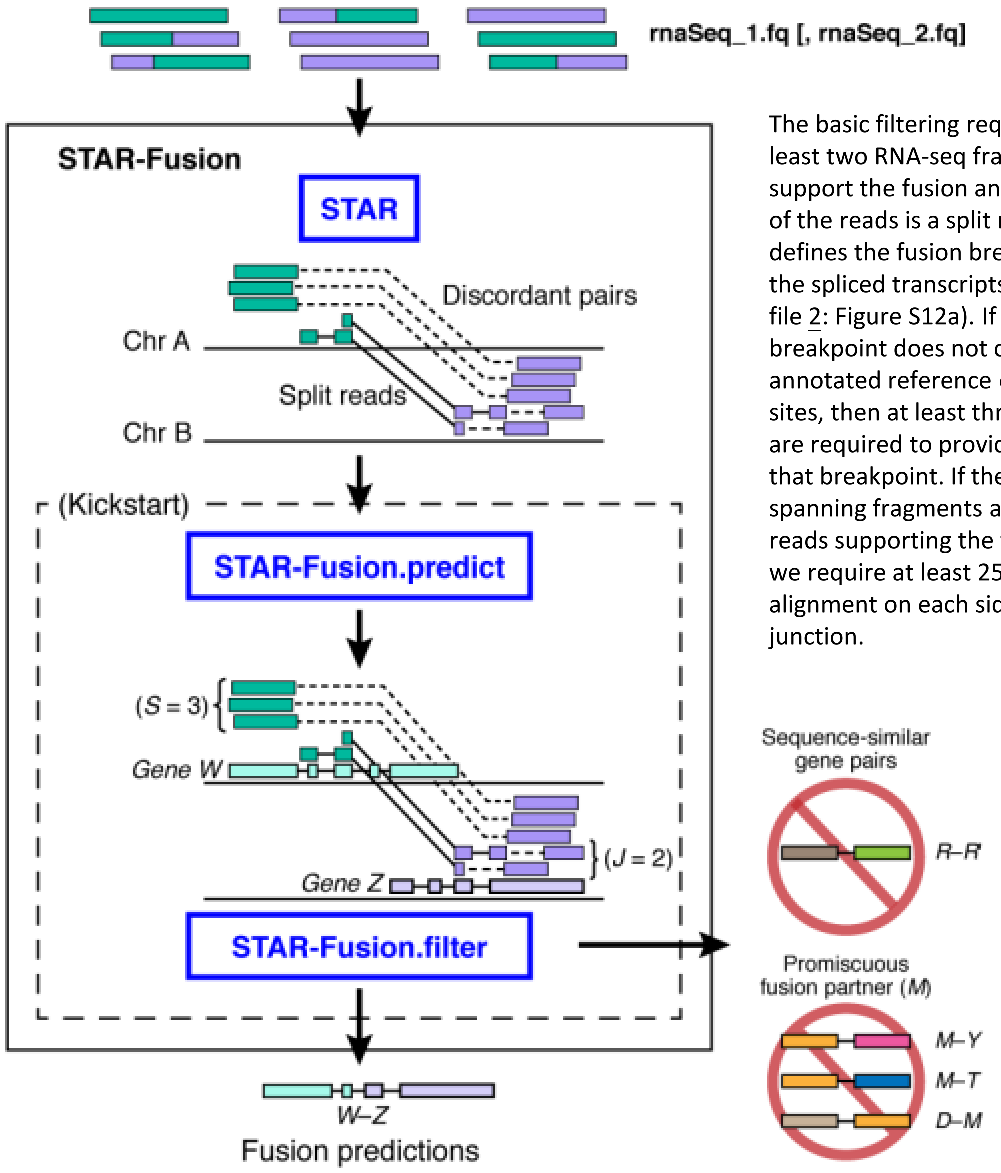
Concordance analysis

Overview

- Workflow comparison (for gene fusion detection)
- Concordance analysis (**assumed hybrid capture data ~ amplicon based data**)
    - Test data: (IVTALL-1 on illumina)
        - AANB02_184_IDD705504_IVTALL-1-AA-21-02200
    - Concordance analysis
    - Discordance analysis

STAR-Fusion workflow

- STAR-Fusion workflow
  - (R1.fq.gz + R2.fq.gz) illumina reads
    → **STAR**
    → **STAR-Fusion** (fusion candidate detection, STAR-Fusion main module)
    → **Trinity de novo transcriptome assembly**
    ("--denovo_reconstruct", STAR-Fusion submodule)
    → **FusionInspector**
    ("--FusionInspector validate“, STAR-Fusion submodule)
- Algorithm
  - Default filters
    - Minimum read filter for fusion breakpoint detection
      - 2 support reads (at least 1 split read) => for annotated spliced sites
      - 3 split reads => for unknown spliced sites
      - 25 bp for each spliced sites => for breakpoint without spanning read support
    - Minimum FFPM (‘STAR-Fusion --min_FFPM’, default = 0.1) (meaning at least 1 fusion-supporting rna-seq fragment per 10M total reads)
    - Annotation filter ('STAR-Fusion --no_annotation_filter’) (filtering reads with certain annotation)



The basic filtering requires that at least two RNA-seq fragments support the fusion and at least one of the reads is a split read that defines the fusion breakpoint within the spliced transcripts (Additional file 2: Figure S12a). If the fusion breakpoint does not correspond to annotated reference exon splice sites, then at least three split reads are required to provide evidence for that breakpoint. If there are no spanning fragments and only split reads supporting the fusion, then we require at least 25 base length alignment on each side of the splice junction.

Ref:

- https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1842-9 (Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods)

ACT GENOMICS ™

CeGaT (customized STAR-Fusion)

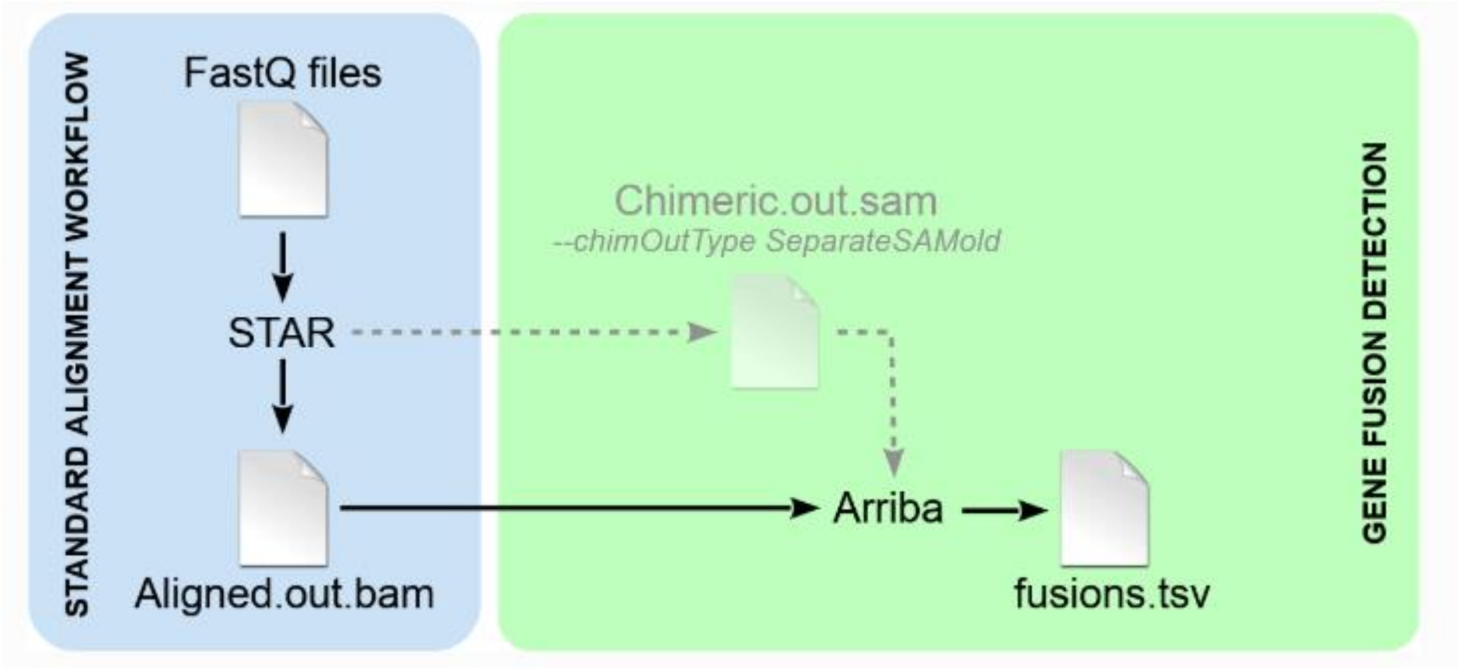
- Calling result files (description obtained from the website)
  - **fusions.tsv**
    - A tabular listing of all detected fusions. This file is produced by STARfusion and is described in detail on the STARfusion wiki. The most important columns are:
      (1) FusionName (The detected fusion, e.g. GNB4–ETV1) and
      (9) FFPM (The number of fragments per million supporting this fusion)
  - **intragene_events.tsv**
    - A tabular listing of all detected intra-gene (exon-skipping) events. This file has 6 columns:
      (1) Fusion Name, e.g. EGFR_VIII
      (2) HGNC symbol (gene name) of the affected gene
      (3)-(5) Genomic location of the skipping event, with respect to the **hg19** reference genome
      (6) FFPM, the number of fragments per million supporting this event
  - **all_reads.bam (+bai)**
    - An alignment of all sequenced reads to the hg19 reference genome
  - **fusions_evidence_mapped.bam (+bai)**
    - Alignments of only the reads supporting fusion events
  - **fusion_evidence_details.html**
    - A self-contained website with visualizations of the detected fusions

Gap analysis summary:

- Only support hg19 annotation
- Exon-level break point annotation is not available
- Limited splicing variants detection
  - EGFR del ex2-22 (mLEEK), EGFR del ex25-26 (EGFRvIVb), EGFR del ex25-27 (EGFRvIVa), EGFR del ex26-27, EGFR del ex14-15 (vII), **EGFR del ex2-7 (vIII)**, FGFR2IIIb, **MET ex14 skipping**, NFE2L2 ex2 skipping, PDGFRA del ex8-9
- Few variant types supported

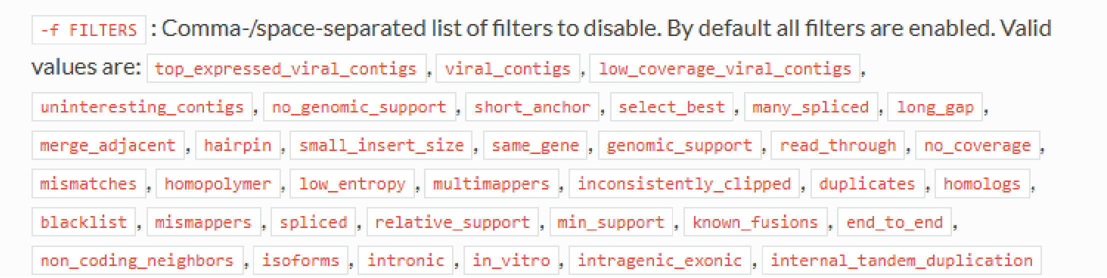
ACT GENOMICS ™

Arriba workflow



- Arriba workflow
  - (R1.fq.gz + R2.fq.gz) illumina reads
    → **STAR**
    → **Arriba**
- Algorithm
  - Default filter
    - Read level filters (see docs)
    - Event level filters (see docs)
  - Available filters &
    Arriba's arguments (-f, -k, -t, -b)
    - fine-tune arriba command
      => filter removal (can try "read_through", "many_spliced", "duplicates", "many_spliced")
    - **Blacklist** (-b)
      => applying sensitive filtering parameters to known fusions (-k) and
      => tagging known fusions in the "tags" column (-t)
      Example:
      => KIAA1549:15/16-BRAF:9 are within the following breakpoint ranges)

      #KIAA1549 BRAF
      -7:138831381-138981318 -7:140719327-140924928 Mitelman
    - **Whitelist** (-k, -t) => applying sensitive filtering parameters to known fusions (-k) and tagging known fusions in the "tags" column (-t)



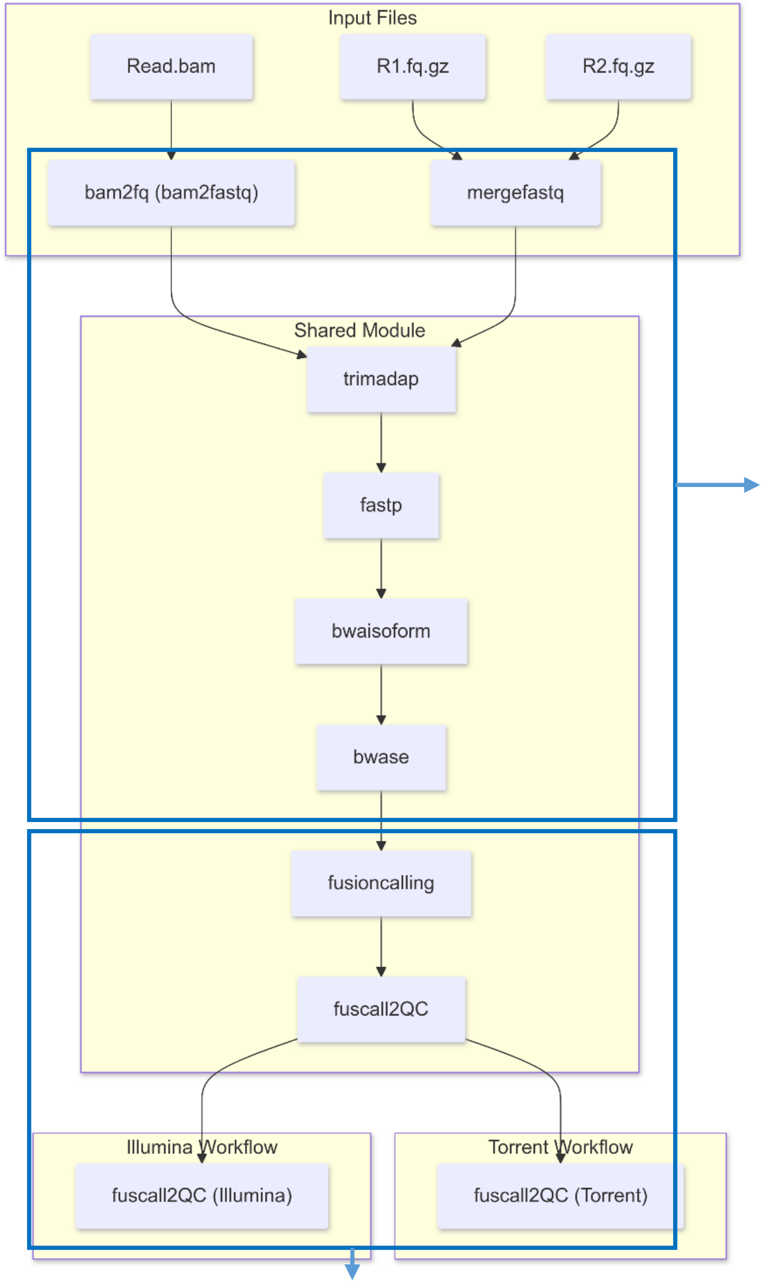
`-f FILTERS` : Comma-/space-separated list of filters to disable. By default all filters are enabled. Valid values are: `top_expressed_viral_contigs`, `viral_contigs`, `low_coverage_viral_contigs`, `uninteresting_contigs`, `no_genomic_support`, `short_anchor`, `select_best`, `many_spliced`, `long_gap`, `merge_adjacent`, `hairpin`, `small_insert_size`, `same_gene`, `genomic_support`, `read_through`, `no_coverage`, `mismatches`, `homopolymer`, `low_entropy`, `multimappers`, `inconsistently_clipped`, `duplicates`, `homologs`, `blacklist`, `mismappers`, `spliced`, `relative_support`, `min_support`, `known_fusions`, `end_to_end`, `non_coding_neighbors`, `isoforms`, `intronic`, `in_vitro`, `intragenic_exonic`, `internal_tandem_duplication`

Ref:
- https://arriba.readthedocs.io/en/latest/workflow/ (Arriba's workflow documentation)

ACT GENOMICS ™

ACTFusion v4 (v0.28.0) workflow

- ACTFusion v4 (v0.28.0) workflow
  - (R1.fq.gz + R2.fq.gz) illumina reads
    → **bwa-mem**
    → **fusioncalling**
- Algorithm
  - Default filters
    - Sample QC filters
      - Minimum raw read count
      - Minimum internal control expression
    - Breakpoint QC filters
      - GSP anchored reads (built-in)
      - Minimum support read count
      - LoD/LoB filter



I. Sequencing read prep

II. Fusion boundary/breakpoint detection

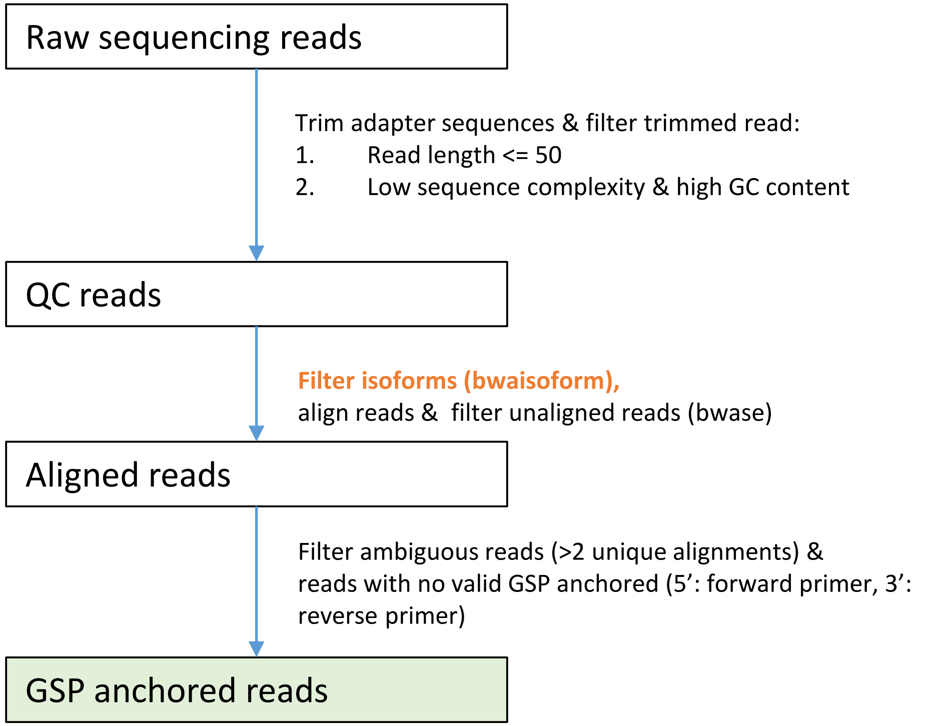
Pipeline v0.24.0 (+ isoform filter (bwaisoform))

Gene fusion detection

The definition of support/functional read:

A read has any in-frame protein product that can be properly aligned to the corresponding kinase sequence, i.e., aligned length >=7 amino acids and >= 0.3 aligned ratio.

I. Sequencing read preprocessing



Raw sequencing reads

Trim adapter sequences & filter trimmed read:
1. Read length <= 50
2. Low sequence complexity & high GC content

QC reads

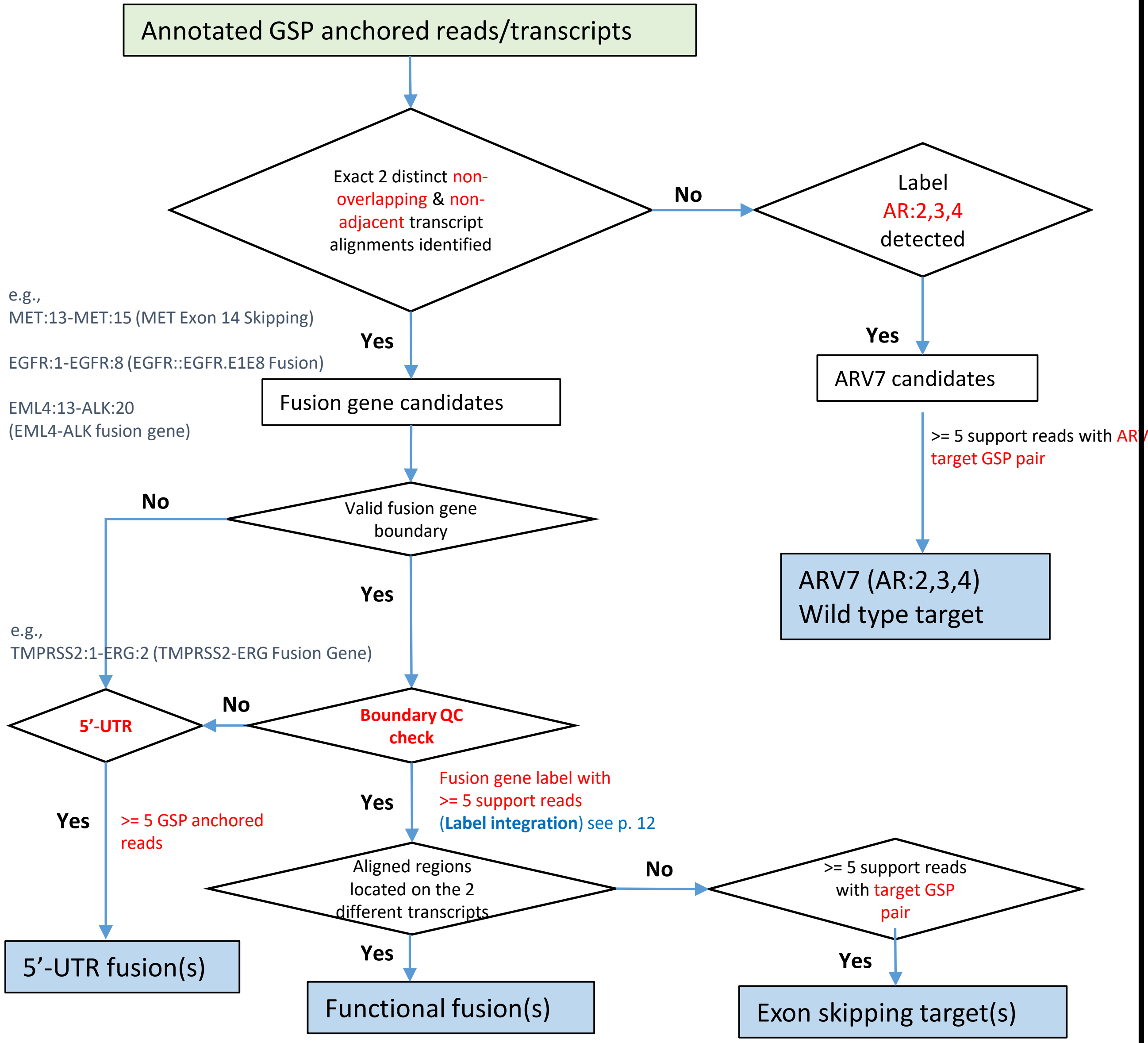
Filter isoforms (bwaisoform),
align reads & filter unaligned reads (bwase)

Aligned reads

Filter ambiguous reads (>2 unique alignments) & reads with no valid GSP anchored (5’: forward primer, 3’: reverse primer)

GSP anchored reads

II. Fusion boundary/breakpoint detection



Annotated GSP anchored reads/transcripts

Exact 2 distinct non-overlapping & non-adjacent transcript alignments identified

No

Label AR:2,3,4 detected

Yes

ARV7 candidates

>= 5 support reads with AR
target GSP pair

ARV7 (AR:2,3,4)
Wild type target

e.g.,
MET:13-MET:15 (MET Exon 14 Skipping)

EGFR:1-EGFR:8 (EGFR::EGFR.E1E8 Fusion)

EML4:13-ALK:20
(EML4-ALK fusion gene)

Yes

Fusion gene candidates

Valid fusion gene boundary

No

Yes

e.g.,
TMPRSS2:1-ERG:2 (TMPRSS2-ERG Fusion Gene)

Boundary QC check

No

5’-UTR

Yes

>= 5 GSP anchored reads

5’-UTR fusion(s)

Yes

Fusion gene label with >= 5 support reads (**Label integration**) see p. 12

Aligned regions located on the 2 different transcripts

No

>= 5 support reads with target GSP pair

Yes

Functional fusion(s)

Yes

Exon skipping target(s)

Comparison study (overview)

- Analyze IVTALL-1 sample (AANB02_184_IDD705504_IVTALL-1-AA-21-02200) using the following tools
  - v0.28.0 (v4), Fusion v4 pipeline
  - Arriba (2.4.0)
  - STAR-Fusion
  - CeGaT (customized STAR-Fusion)

- Comparison
  - Side-by-side comparison for the 4 tools (v0.28.0 (v4), Fusion v4 pipeline, Arriba (2.4.0), STAR-Fusion, CeGaT (customized STAR-Fusion))
  - Accuracy/precision computation
    - Accuracy = (TP + ~~TN~~) / (TP + ~~TN~~ + FP + ~~FN~~) ~ TP / TP + FP => precision
  - Recall computation
    - Recall = TP / (TP + FN))

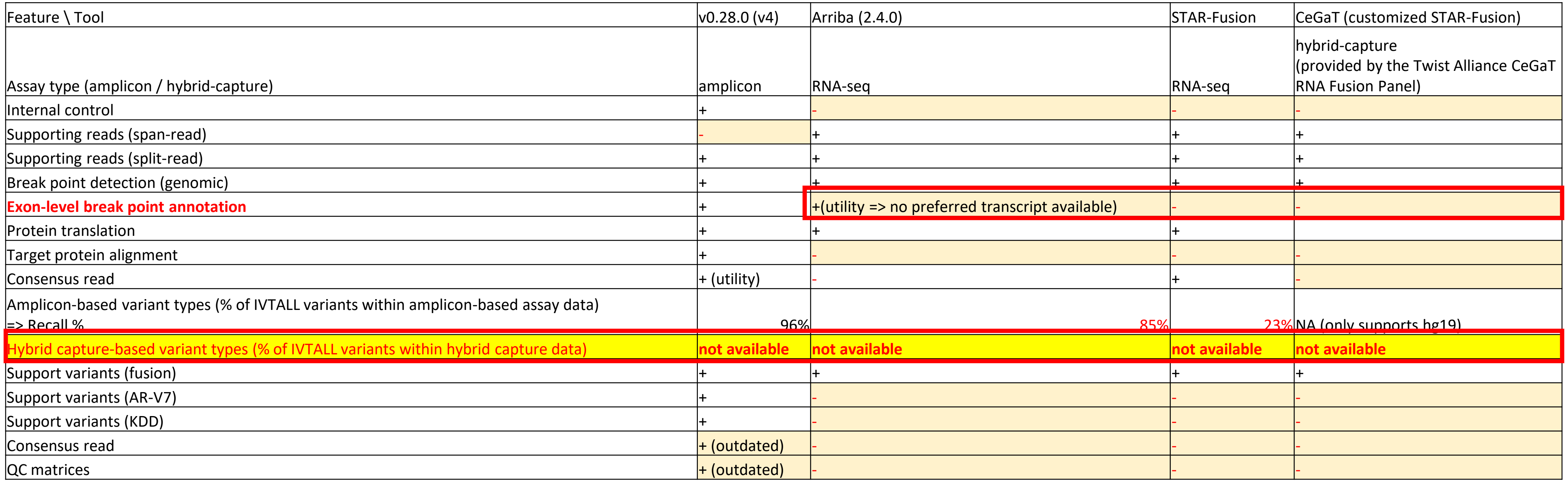
ACT GENOMICS ™

Side-by-side comparison for the 4 tools

- v0.28.0 (v4), Fusion v4 pipeline
- Arriba (2.4.0)
- STAR-Fusion
- CeGaT (customized STAR-Fusion)

Detailed comparison table can be found in the tool comparison table:
https://actg.atlassian.net/browse/ABIE-907

The exon annotator provided by **Arriba ("annotate_exon_numbers.sh") may be incorrect**.
=>
Since it does not utilize preferred transcripts (i.e. **it may randomly pick one transcript to annotate**).



Feature \ Tool	v0.28.0 (v4)	Arriba (2.4.0)	STAR-Fusion	CeGaT (customized STAR-Fusion)
Assay type (amplicon / hybrid-capture)	amplicon	RNA-seq	RNA-seq	hybrid-capture (provided by the Twist Alliance CeGaT RNA Fusion Panel)
Internal control	+	-	-	-
Supporting reads (span-read)	-	+	+	+
Supporting reads (split-read)	+	+	+	+
Break point detection (genomic)	+	+	+	+
Exon-level break point annotation	+	+(utility => no preferred transcript available)	-	-
Protein translation	+	+	+	
Target protein alignment	+	-	-	-
Consensus read	+ (utility)	-	+	-
Amplicon-based variant types (% of IVTALL variants within amplicon-based assay data) => Recall %	96%	85%	23%	NA (only supports hg19)
Hybrid capture-based variant types (% of IVTALL variants within hybrid capture data)	**not available**	**not available**	**not available**	**not available**
Support variants (fusion)	+	+	+	+
Support variants (AR-V7)	+	-	-	-
Support variants (KDD)	+	-	-	-
Consensus read	+ (outdated)	-	-	-
QC matrices	+ (outdated)	-	-	-

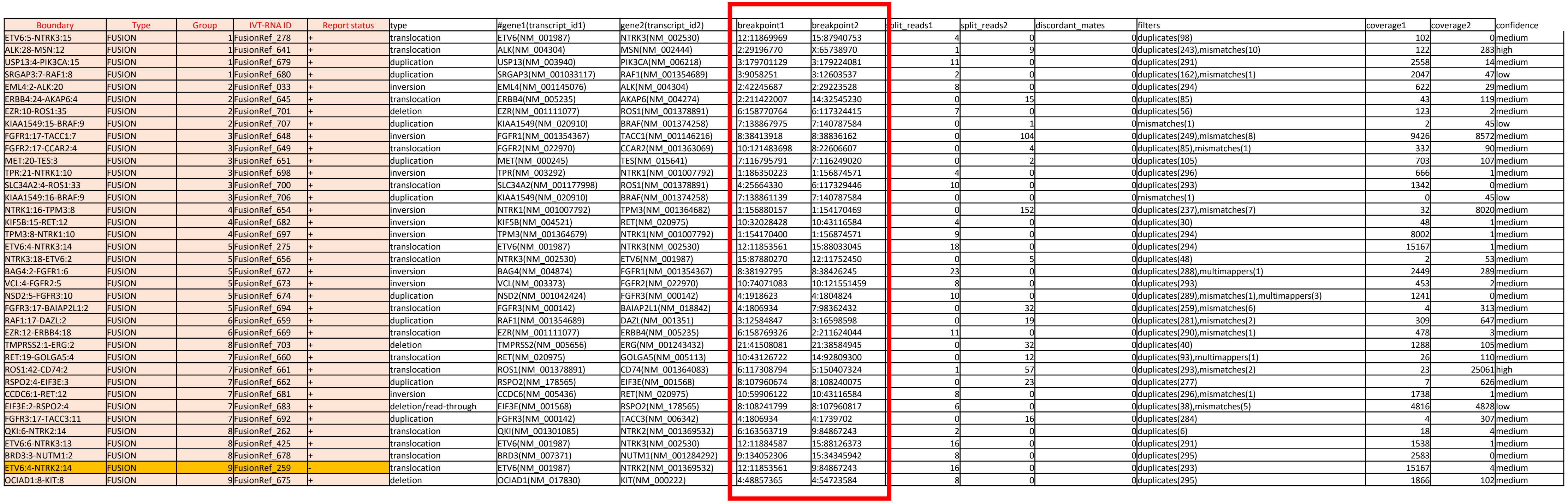
Feature \ Tool	v0.28.0 (v4), Fusion v4 pipeline	Arriba (2.4.0)	STAR-Fusion	CeGaT (customized STAR-Fusion)
Docs	within the pipeline repo	Home - Arriba	Home	https://cegat.com/fusions/
Assay type	amplicon-based	RNA seq	RNA seq	hybrid-capture
Reference genome/transcriptome	Grch38 **MANE v0.95** (+ GENCODE-r38 (NRG1 NTRK3, ERG (first 3 exons), AR (editing) => preferred transcripts Refseq + GENCODE => transcript isoform elimination	Grch38 (GRCh38_**RefSeq_hg38**) (other versions available)	Grch38 (GRCh38_**gencode**_v44) (other versions: https://data.broadinstitute.org/Trinity/CTAT_RESOURCE_LIB/)	hg19
Aligner (sequence alignment tool)	bwa-mem	STAR	STAR	STAR
Tools for fusion detection	Fusion v4 pipeline (v0.28.0) https://bitbucket.org/actgenomics/torrent_fusion_pipeline_nextflow/src/master/	Arriba (2.4.0)	STAR-Fusion	STAR-Fusion
Tools for de novo fusion construction	Consensus read utility (https://bitbucket.org/actgenomics/tool_fusion_consensus_read/src/main/ => for ***v0.24.0*** pipeline (***outdated***))	NA	Trinity de novo transcriptome assembly (include de novo reconstruction: "**--denovo_reconstruct**") (STAR-Fusion submodule)	NA (may need to specify "--denovo_reconstruct")
Tools for read inspection and validation	Utility repo (https://bitbucket.org/actgenomics/torrent_fusion_pipeline_utilities/src/master/)	Utility scripts: (not included in the standard workflow) 1. extract_fusion-supporting_alignments.sh 2. convert_fusions_to_vcf.sh 3. run_arriba_on_prealigned_bam.sh 4. quantify_virus_expression.sh 5. **annotate_exon_numbers.sh => to annotate fusion.tsv => exon-level breakpoint** (Remark: it only annotates breakpoints with annotated NM ID within the .tsv file)	FusionInspector (validate mode: "**--FusionInspector validate**") (STAR-Fusion submodule)	FusionInspector (inspect mode: "**--FusionInspector inspect**") (STAR-Fusion submodule)
Support reads	Functional count, Total read count (filtered), Total read count Decision ("+" for report)	coverage1, coverage2, confidence (several built-in filters applied)	FFPM (fusion fragments per million total reads)	FFPM (fusion fragments per million total reads)
break point resolution	exon-level	exon-level (via "annotate_exon_numbers.sh") Remark: the annotated exons may differ from v4 (since no preferred transcript for annotation) => use provided break point "breakpoint1", "breakpoint2" (Arriba 2.4.0) ~ "5' gene coordinate", "3' gene coordinate" (v0.28.0 v4)	gene-level	gene-level
Variant report files	/Report/{sample name}**_fusioncalling.boundary.QC.txt**	/path_to_output/**fusions.tsv** /path_to_output/fusions.discarded.tsv (discarded variants)	/path_to_output/**star-fusion.fusion_predictions.tsv** /path_to_output/FusionInspector-validate/ (output folder for "**--FusionInspector validate**") /path_to_output/FusionInspector-validate/finspector.mm2_trinity_GG.fusions.fasta (output for "**--denovo_reconstruct**")	/path_to_output/**fusions.tsv** /path_to_output/**intragene_events.tsv**
Internal control files (sample QC)	/Report/{sample name}_fusioncalling.Sample.QC.json	NA	NA	NA
Performance # of detected IVTALL variants / # of IVTALL variants (=81) See sheet "IVTALL-1 comparison"	96%	85%	23%	NA (the genomic location for hg 19 and Grch 38 are not compatiable)
Pros	Most of the required columns are built-in One can adjust LoB/LoD, detection threshold based on assay design	Faster runtime many built-in filters applied for background noise correction	De novo fusion construction is supported Built-in Trinity module via "**--FusionInspector validate**"	Some of the predefined splicing variants
Cons	No spanning read support	Lacking of the required columns for report purpose Need to fine-tune the arguments to rescue some clinical relavant variants	Lacking of the required columns for report purpose Limited detected variants	Lacking of the required columns for report purpose Limited detected variants
Decision based on RM (hybrid-capture) Wait for sequencing data	**TBC**	**TBC**	**TBC**	**TBC**

ACT GENOMICS ™

Detailed comparison table can be found in the tool comparison table:
https://actg.atlassian.net/browse/ABIE-907

Concordance study (fusion v4 vs Arriba)

- **37**/81 => identical exon-level boundary (81-37 = **44 variants missing exon-level boundary**)
  gene + exon number (provided by Arriba's utility)



Boundary	Type	Group	IVT-RNA ID	Report status	type	#gene1(transcript_id1)	gene2(transcript_id2)	breakpoint1	breakpoint2	split_reads1	split_reads2	discordant_mates	filters	coverage1	coverage2	confidence
ETV6:5-NTRK3:15	FUSION	1	FusionRef_278	+	translocation	ETV6(NM_001987)	NTRK3(NM_002530)	12:11869969	15:87940753	4	0	0	duplicates(98)	102	0	medium
ALK:28-MSN:12	FUSION	1	FusionRef_641	+	translocation	ALK(NM_004304)	MSN(NM_002444)	2:29196770	X:65738970	1	9	0	duplicates(243),mismatches(10)	122	283	high
USP13:4-PIK3CA:15	FUSION	1	FusionRef_679	+	duplication	USP13(NM_003940)	PIK3CA(NM_006218)	3:179701129	3:179224081	11	0	0	duplicates(291)	2558	14	medium
SRGAP3:7-RAF1:8	FUSION	1	FusionRef_680	+	duplication	SRGAP3(NM_001033117)	RAF1(NM_001354689)	3:9058251	3:12603537	2	0	0	duplicates(162),mismatches(1)	2047	47	low
EML4:2-ALK:20	FUSION	2	FusionRef_033	+	inversion	EML4(NM_001145076)	ALK(NM_004304)	2:42245687	2:29223528	8	0	0	duplicates(294)	622	29	medium
ERBB4:24-AKAP6:4	FUSION	2	FusionRef_645	+	translocation	ERBB4(NM_005235)	AKAP6(NM_004274)	2:211422007	14:32545230	0	15	0	duplicates(85)	43	119	medium
EZR:10-ROS1:35	FUSION	2	FusionRef_701	+	deletion	EZR(NM_001111077)	ROS1(NM_001378891)	6:158770764	6:117324415	7	0	0	duplicates(56)	123	2	medium
KIAA1549:15-BRAF:9	FUSION	2	FusionRef_707	+	duplication	KIAA1549(NM_020910)	BRAF(NM_001374258)	7:138867975	7:140787584	0	1	0	mismatches(1)	2	45	low
FGFR1:17-TACC1:7	FUSION	3	FusionRef_648	+	inversion	FGFR1(NM_001354367)	TACC1(NM_001146216)	8:38413918	8:38836162	0	104	0	duplicates(249),mismatches(8)	9426	8572	medium
FGFR2:17-CCAR2:4	FUSION	3	FusionRef_649	+	translocation	FGFR2(NM_022970)	CCAR2(NM_001363069)	10:121483698	8:22606607	0	4	0	duplicates(85),mismatches(1)	332	90	medium
MET:20-TES:3	FUSION	3	FusionRef_651	+	duplication	MET(NM_000245)	TES(NM_015641)	7:116795791	7:116249020	0	2	0	duplicates(105)	703	107	medium
TPR:21-NTRK1:10	FUSION	3	FusionRef_698	+	inversion	TPR(NM_003292)	NTRK1(NM_001007792)	1:186350223	1:156874571	4	0	0	duplicates(296)	666	1	medium
SLC34A2:4-ROS1:33	FUSION	3	FusionRef_700	+	translocation	SLC34A2(NM_001177998)	ROS1(NM_001378891)	4:25664330	6:117329446	10	0	0	duplicates(293)	1342	0	medium
KIAA1549:16-BRAF:9	FUSION	3	FusionRef_706	+	duplication	KIAA1549(NM_020910)	BRAF(NM_001374258)	7:138861139	7:140787584	0	0	0	mismatches(1)	0	45	low
NTRK1:16-TPM3:8	FUSION	4	FusionRef_654	+	inversion	NTRK1(NM_001007792)	TPM3(NM_001364682)	1:156880157	1:154170469	0	152	0	duplicates(237),mismatches(7)	32	8020	medium
KIF5B:15-RET:12	FUSION	4	FusionRef_682	+	inversion	KIF5B(NM_004521)	RET(NM_020975)	10:32028428	10:43116584	4	0	0	duplicates(30)	48	1	medium
TPM3:8-NTRK1:10	FUSION	4	FusionRef_697	+	inversion	TPM3(NM_001364679)	NTRK1(NM_001007792)	1:154170400	1:156874571	9	0	0	duplicates(294)	8002	1	medium
ETV6:4-NTRK3:14	FUSION	5	FusionRef_275	+	translocation	ETV6(NM_001987)	NTRK3(NM_002530)	12:11853561	15:88033045	18	0	0	duplicates(294)	15167	1	medium
NTRK3:18-ETV6:2	FUSION	5	FusionRef_656	+	translocation	NTRK3(NM_002530)	ETV6(NM_001987)	15:87880270	12:11752450	0	5	0	duplicates(48)	2	53	medium
BAG4:2-FGFR1:6	FUSION	5	FusionRef_672	+	inversion	BAG4(NM_004874)	FGFR1(NM_001354367)	8:38192795	8:38426245	23	0	0	duplicates(288),multimappers(1)	2449	289	medium
VCL:4-FGFR2:5	FUSION	5	FusionRef_673	+	inversion	VCL(NM_003373)	FGFR2(NM_022970)	10:74071083	10:121551459	8	0	0	duplicates(293)	453	2	medium
NSD2:5-FGFR3:10	FUSION	5	FusionRef_674	+	duplication	NSD2(NM_001042424)	FGFR3(NM_000142)	4:1918623	4:1804824	10	0	0	duplicates(289),mismatches(1),multimappers(3)	1241	0	medium
FGFR3:17-BAIAP2L1:2	FUSION	5	FusionRef_694	+	translocation	FGFR3(NM_000142)	BAIAP2L1(NM_018842)	4:1806934	7:98362432	0	32	0	duplicates(259),mismatches(6)	4	313	medium
RAF1:17-DAZL:2	FUSION	6	FusionRef_659	+	duplication	RAF1(NM_001354689)	DAZL(NM_001351)	3:12584847	3:16598598	0	19	0	duplicates(281),mismatches(2)	309	647	medium
EZR:12-ERBB4:18	FUSION	6	FusionRef_669	+	translocation	EZR(NM_001111077)	ERBB4(NM_005235)	6:158769326	2:211624044	11	0	0	duplicates(290),mismatches(1)	478	3	medium
TMPRSS2:1-ERG:2	FUSION	8	FusionRef_703	+	deletion	TMPRSS2(NM_005656)	ERG(NM_001243432)	21:41508081	21:38584945	0	32	0	duplicates(40)	1288	105	medium
RET:19-GOLGA5:4	FUSION	7	FusionRef_660	+	translocation	RET(NM_020975)	GOLGA5(NM_005113)	10:43126722	14:92809300	0	12	0	duplicates(93),multimappers(1)	26	110	medium
ROS1:42-CD74:2	FUSION	7	FusionRef_661	+	translocation	ROS1(NM_001378891)	CD74(NM_001364083)	6:117308794	5:150407324	1	57	0	duplicates(293),mismatches(2)	23	25061	high
RSPO2:4-EIF3E:3	FUSION	7	FusionRef_662	+	duplication	RSPO2(NM_178565)	EIF3E(NM_001568)	8:107960674	8:108240075	0	23	0	duplicates(277)	7	626	medium
CCDC6:1-RET:12	FUSION	7	FusionRef_681	+	inversion	CCDC6(NM_005436)	RET(NM_020975)	10:59906122	10:43116584	8	0	0	duplicates(296),mismatches(1)	1738	1	medium
EIF3E:2-RSPO2:4	FUSION	7	FusionRef_683	+	deletion/read-through	EIF3E(NM_001568)	RSPO2(NM_178565)	8:108241799	8:107960817	6	0	0	duplicates(38),mismatches(5)	4816	4828	low
FGFR3:17-TACC3:11	FUSION	7	FusionRef_692	+	duplication	FGFR3(NM_000142)	TACC3(NM_006342)	4:1806934	4:1739702	0	16	0	duplicates(284)	4	307	medium
QKI:6-NTRK2:14	FUSION	8	FusionRef_262	+	translocation	QKI(NM_001301085)	NTRK2(NM_001369532)	6:163563719	9:84867243	2	0	0	duplicates(6)	18	4	medium
ETV6:6-NTRK3:13	FUSION	8	FusionRef_425	+	translocation	ETV6(NM_001987)	NTRK3(NM_002530)	12:11884587	15:88126373	16	0	0	duplicates(291)	1538	1	medium
BRD3:3-NUTM1:2	FUSION	8	FusionRef_678	+	translocation	BRD3(NM_007371)	NUTM1(NM_001284292)	9:134052306	15:34345942	8	0	0	duplicates(295)	2583	0	medium
ETV6:4-NTRK2:14	FUSION	9	FusionRef_259	-	translocation	ETV6(NM_001987)	NTRK2(NM_001369532)	12:11853561	9:84867243	16	0	0	duplicates(293)	15167	4	medium
OCIAD1:8-KIT:8	FUSION	9	FusionRef_675	+	deletion	OCIAD1(NM_017830)	KIT(NM_000222)	4:48857365	4:54723584	8	0	0	duplicates(295)	1866	102	medium

ACT GENOMICS ™

Tool comparison (break point)

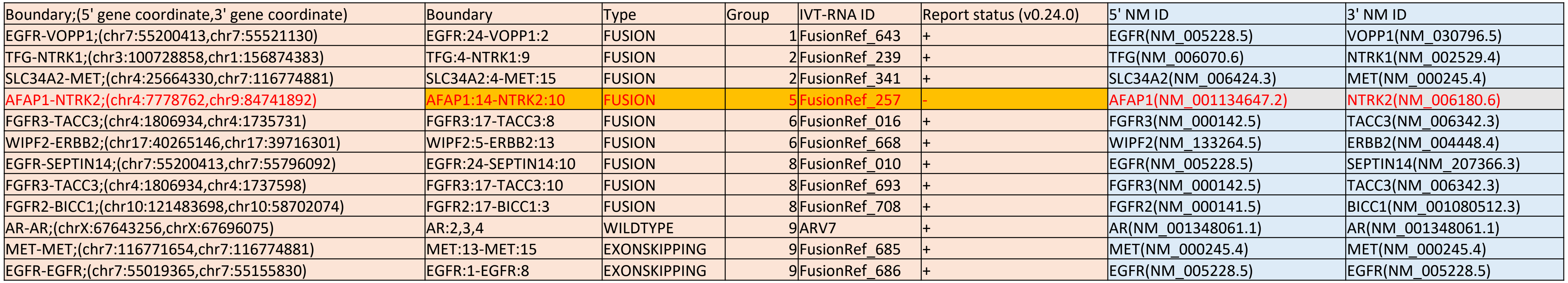
- **Accuracy/precision** computation (accuracy = (TP + ~~TN~~) / (TP + ~~TN~~ + FP + ~~FN~~) ~ TP / TP + FP => precision
  - **v0.28.0 (v4), Fusion v4 pipeline:** 78/87 (Decision = '+') ~ 90% => need LoB/LoD, boundary threshold
  - **Arriba (2.4.0):** 69/71 (all break point found within fusion.tsv) ~ 97% (correct exon-boundary not available)
    => **12 missing variants in Arriba (variants to rescue)**
  - **STAR-Fusion:** 19/19 (all break point found within fusion.tsv) ~ 100% (exon-boundary not available)

- **Recall** computation (recall = TP / (TP + FN))
  - **v0.28.0 (v4), Fusion v4 pipeline:** 78/81 ~ 96%
  - **Arriba (2.4.0):** 69/81 ~ 85% (correct exon-boundary not available)
  - **STAR-Fusion:** 19/81 ~ 23% (exon-boundary not available)

ACT GENOMICS ™

Discordance study (fusion v4 vs Arriba)

- 12 missing variants in Arriba (variants to rescue)
    - Gene annotation different from MANE (Arriba uses GENCODE to annotate detected variants => no preferred transcript applied)
    - Only genomic break points available



Boundary;(5' gene coordinate,3' gene coordinate)	Boundary	Type	Group	IVT-RNA ID	Report status (v0.24.0)	5' NM ID	3' NM ID
EGFR-VOPP1;(chr7:55200413,chr7:55521130)	EGFR:24-VOPP1:2	FUSION	1	FusionRef_643	+	EGFR(NM_005228.5)	VOPP1(NM_030796.5)
TFG-NTRK1;(chr3:100728858,chr1:156874383)	TFG:4-NTRK1:9	FUSION	2	FusionRef_239	+	TFG(NM_006070.6)	NTRK1(NM_002529.4)
SLC34A2-MET;(chr4:25664330,chr7:116774881)	SLC34A2:4-MET:15	FUSION	2	FusionRef_341	+	SLC34A2(NM_006424.3)	MET(NM_000245.4)
AFAP1-NTRK2;(chr4:7778762,chr9:84741892)	AFAP1:14-NTRK2:10	FUSION	5	FusionRef_257	-	AFAP1(NM_001134647.2)	NTRK2(NM_006180.6)
FGFR3-TACC3;(chr4:1806934,chr4:1735731)	FGFR3:17-TACC3:8	FUSION	6	FusionRef_016	+	FGFR3(NM_000142.5)	TACC3(NM_006342.3)
WIPF2-ERBB2;(chr17:40265146,chr17:39716301)	WIPF2:5-ERBB2:13	FUSION	6	FusionRef_668	+	WIPF2(NM_133264.5)	ERBB2(NM_004448.4)
EGFR-SEPTIN14;(chr7:55200413,chr7:55796092)	EGFR:24-SEPTIN14:10	FUSION	8	FusionRef_010	+	EGFR(NM_005228.5)	SEPTIN14(NM_207366.3)
FGFR3-TACC3;(chr4:1806934,chr4:1737598)	FGFR3:17-TACC3:10	FUSION	8	FusionRef_693	+	FGFR3(NM_000142.5)	TACC3(NM_006342.3)
FGFR2-BICC1;(chr10:121483698,chr10:58702074)	FGFR2:17-BICC1:3	FUSION	8	FusionRef_708	+	FGFR2(NM_000141.5)	BICC1(NM_001080512.3)
AR-AR;(chrX:67643256,chrX:67696075)	AR:2,3,4	WILDTYPE	9	ARV7	+	AR(NM_001348061.1)	AR(NM_001348061.1)
MET-MET;(chr7:116771654,chr7:116774881)	MET:13-MET:15	EXONSKIPPING	9	FusionRef_685	+	MET(NM_000245.4)	MET(NM_000245.4)
EGFR-EGFR;(chr7:55019365,chr7:55155830)	EGFR:1-EGFR:8	EXONSKIPPING	9	FusionRef_686	+	EGFR(NM_005228.5)	EGFR(NM_005228.5)

ACT GENOMICS ™

Summary (to-do items)

- Clinical relevant fusions / splicing variants can not be reported by Arriba
- To-do items (modify current v0.28.0 fusion v4 pipeline)
    - Gap analysis
    - Annotation table update (gsp location => probe location)
    - Consensus read (fine-tune)
    - QC matrices (fine-tune)

- To-do items (to build an Arriba based workflow)
    - Exon-level break point annotation
        - (reannotate Arriba's fusion.tsv using the provided genomic coordinates)
    - Fine tune pipeline (Arriba)
    - Variant inclusion
        - (rescue clinical relevant variants)
    - Target protein alignment (not available)
        - (may need another assembly tool before read alignment) => currently not supported by Arriba
        - (Read assembly => Target protein alignment)
    - Consensus read generation (not available)
        - Read assembly tool => need to search (de novo assembly?)
    - QC matrices settings
        - Internal control matrices
        - Variant thresholds (e.g., minimum required coverage for each break point)

Next Steps

- ~~Verification study for amplicon-based pipeline (***pending (must do)***, wait for sequencing run)~~ => deprecated
- ~~QC metrics setting for amplicon-based pipeline (***pending (must do)***, wait for sequencing run)~~ => deprecated
- Gap analysis (***pending (must do)***, wait for hybrid capture probe design and sequencing run)

ACT GENOMICS ™

Make Personalized Medicine Accessible to All