AANB01_507 run analysis

Bioinformatics Development

Sandy

2025.05.14

AANB01_507

- Run analysis plan

ACT GENOMICS ™

Analysis request

- 8 fusion v5
  hybrid capture samples
  - 81 IVT RNA (different spike-in copies)
  - RM-25-001 (different input amount)

SN	Sample	input amount (ng)	Condition
1	AA-23-08153_R1	50	yeast t-RNA blank
2	AA-23-08153_R1	50	81 IVT RNA; 10^2 copies each
3	AA-23-08153_R1	50	81 IVT RNA; 10^3 copies each
4	AA-23-08153_R1	50	81 IVT RNA; 10^4 copies each
5	RM-25-001_R1	**10**	Seraseq FFPE Tumor Fusion RNA v4 Reference Material
6	RM-25-001_R1	**30**	Seraseq FFPE Tumor Fusion RNA v4 Reference Material
7	RM-25-001_R1	**50**	Seraseq FFPE Tumor Fusion RNA v4 Reference Material
8	RM-25-001_R1	**100**	Seraseq FFPE Tumor Fusion RNA v4 Reference Material

- Recall analysis
  - % (detected variant / known (expected) variant)
    - (v0.95) Fusionv4-based workflow: % (# of detected **exon breakpoint (v0.95)** / # of known (expected) exon breakpoint)
      - **Boundary**
    - (v1.4) Fusionv4-based workflow: % (# of detected **exon breakpoint (v1.4)** / # of known (expected) exon breakpoint)
      - **Boundary**
    - Arriba: % (# of detected **genomic breakpoint (locations)** and gene pair / # of known (expected) genomic breakpoint (locations) and gene pair)
      - **Boundary;(5' gene coordinate,3' gene coordinate)**

ACT GENOMICS ™

Boundary;(5' gene coordinate,3' gene coordinate) versus Boundary

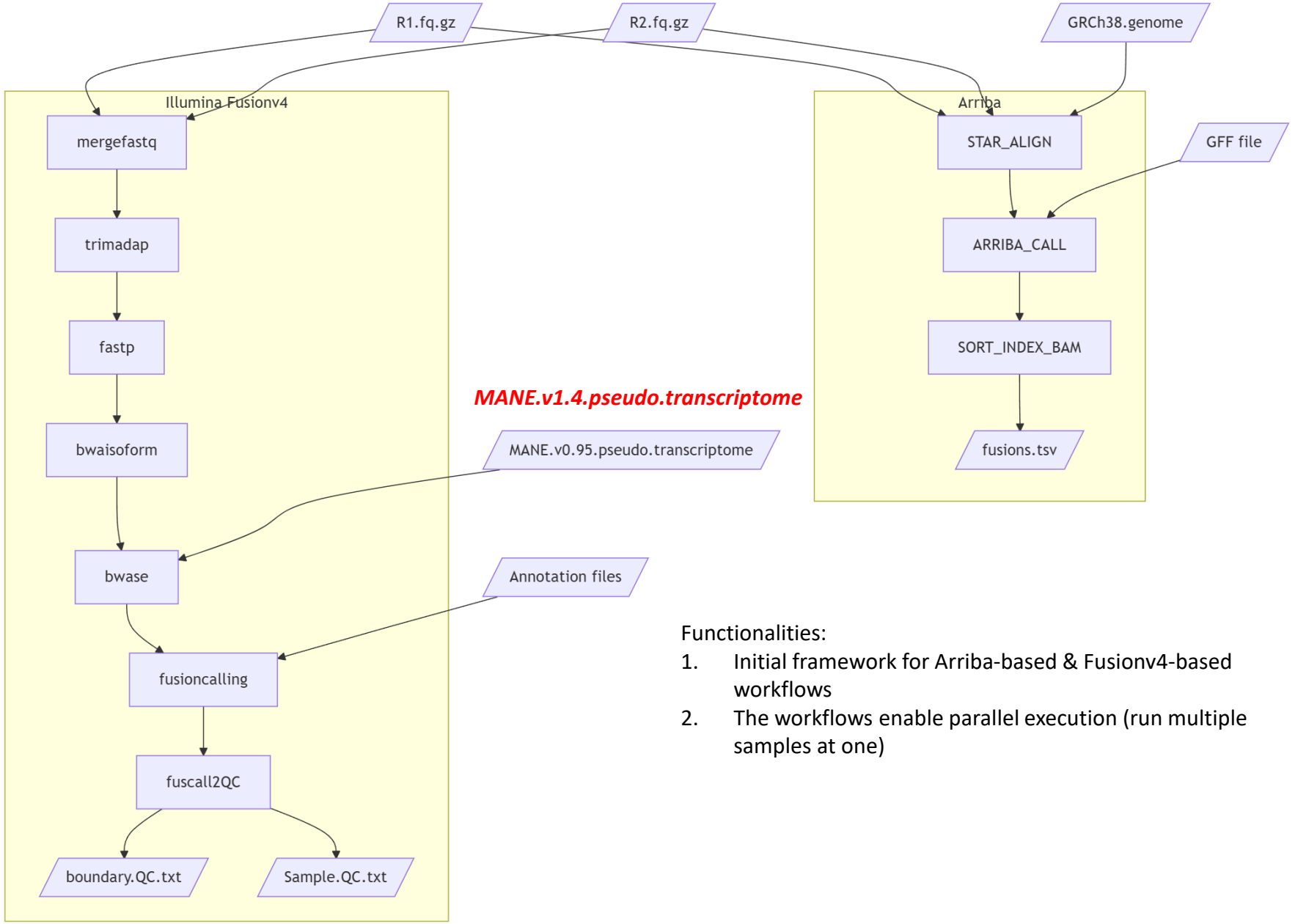
- IVTALL (Group 1) (Ref. file: IVTALL-1_fusionv4-v0.95_calling (blue)_arriba_calling (yellow).xlsx)

Boundary;(5' gene coordinate,3' gene coordinate)	Boundary	Type	Group	IVT-RNA ID	Report status (v0.24.0)
CD74-ROS1;(chr5:150404680,chr6:117329446)	CD74:6-ROS1:33	FUSION	1	FusionRef_077	+
NACC2-NTRK2;(chr9:136013199,chr9:84744973)	NACC2:5-NTRK2:11	FUSION	1	FusionRef_260	+
ETV6-NTRK3;(chr12:11869969,chr15:87940753)	ETV6:5-NTRK3:15	FUSION	1	FusionRef_278	+
AKT3-HEATR1;(chr1:243545510,chr1:236566876)	AKT3:12-HEATR1:30	FUSION	1	FusionRef_640	+
ALK-MSN;(chr2:29196770,chrX:65738970)	ALK:28-MSN:12	FUSION	1	FusionRef_641	+
BRAF-ATG7;(chr7:140734617,chr3:11340645)	BRAF:19-ATG7:12	FUSION	1	FusionRef_642	+
EGFR-VOPP1;(chr7:55200413,chr7:55521130)	EGFR:24-VOPP1:2	FUSION	1	FusionRef_643	+
USP13-PIK3CA;(chr3:179701129,chr3:179224081)	USP13:4-PIK3CA:15	FUSION	1	FusionRef_679	+
SRGAP3-RAF1;(chr3:9058251,chr3:12603537)	SRGAP3:7-RAF1:8	FUSION	1	FusionRef_680	+

ACT GENOMICS ™

V5 Workflows

- Nextflow repo
  (hybridcapture_fusion_pipeline_nextflow)
- Pipeline utility repo
  (fusion_pipeline_env)



Functionalities:

1. Initial framework for Arriba-based & Fusionv4-based workflows
2. The workflows enable parallel execution (run multiple samples at one)

ACT GENOMICS ™

AANB01_507

- Result
  - 81 IVT RNA (different spike-in copies)
  - RM-25-001 (different input amount)

ACT GENOMICS ™

Expected answer

- 81 IVT RNA
  - IVTALL-1_fusionv4-v0.95_calling (blue)_arriba_calling (yellow).xlsx (Arriba, Fusionv4-MANE-v0.95 => obtained via IVTALL-1 amplicon data)
  - IVTALL_fusionv5.v1.4_calling.xlsx (Fusionv4-MANE-v1.4 => obtained via 300x IVTALL in-silico study)

- RM-25-001
  - Seraseq Fusion RM infomation (Transcript, exon, breakpoint) 20250513.xlsx (Arriba (breakpoint), Fusionv4 (boundary))

Shared folder: report2AD_AANB01_507

ACT GENOMICS ™

Raw calling result summary

- Arriba
  - arriba.merged.fusions.keycols.xlsx
- Fusionv4 (V0.95)
  - AANB01_507_8samples-v0.95.boundaryQC.summary.xlsx
- Fusionv4 (V1.4)
  - AANB01_507_8samples-v1.4.boundaryQC.summary.xlsx

Shared folder: report2AD_AANB01_507

ACT GENOMICS ™

IVTALL (v1.4 versus v0.95 versus Arriba)

- Expected (Theoretical) Recall (on the 300x IVTALL in-silico data)
  - v0.95: 78 / 81 ~ 96.3%
  - v1.4: 67 / 81 ~ 82.7%

Ref:
IVTALL-1_fusionv4-v0.95_calling (blue)_arriba_calling (yellow).xlsx
Recall-IVTALL.tables.xlsx

- Recall for different IVT-RNA copies (# of corrected boundary (same expected boundary provided by AD team, v0.95))

	yeast t-RNA blank	81 IVT RNA; 10^2 copies each	81 IVT RNA; 10^3 copies each	81 IVT RNA; 10^4 copies each	AANB02_184_IDD705504_IVTALL-1-AA-21-02200 (amplicon, fusion v4)
v0.95	ND	61	73	79 (~97.5%)	78 (~96.3%)
v1.4	ND	60	69	71(~87%)	-
arriba	ND	65	77	78(~96.3%)	69 (~85.2%)

- Note:
  - # **AR-V7, MET exon 14 skipping, EGFR VIII can not be reported by Arriba**
  - # **FusionRef_653 (NRG2:7-UBE2D2:5 (4 reads (10^3 copies))) and FusionRef_689 (NPM1:4-ALK:20 (AMBIGUITY)) can not be reported by Fusion V4 (0.95)**

ACT GENOMICS ™

Fusion v4 RM (v1.4 versus Arriba)

- Recall for different input amout
  - Arriba: 15/18 ~ 83.3%
    (**ND: PAX8-PPARG, EGFR:1-EGFR:8, MET ex 14 Skipping**)
  - Fusionv4-based: 17/18 ~ 94%

RM_name	Fusion nomenclature 1	Fusion nomenclature 3	Fusion v4-1.4	Note	Fusion v4-1.4 (boundary)
Seraseq Tumor Fusion RNA v4 Reference Materials	CCDC6-RET	CCDC6:1-RET:12	*		CCDC6:1-RET:12
Seraseq Tumor Fusion RNA v4 Reference Materials	CD74-ROS1	CD74:6-ROS1:35	*		CD74:6-ROS1:35
Seraseq Tumor Fusion RNA v4 Reference Materials	EGFR Variant III	EGFR:1-EGFR:8	*		EGFR:1-EGFR:8
Seraseq Tumor Fusion RNA v4 Reference Materials	EGFR-SEPTIN14	EGFR:24-SEPTIN14:10	*		EGFR:24-SEPTIN14:10
Seraseq Tumor Fusion RNA v4 Reference Materials	EML4-ALK	EML4:13-ALK:20	*		EML4:13-ALK:20
Seraseq Tumor Fusion RNA v4 Reference Materials	ETV6-NTRK3	ETV6:5-NTRK3:15	*		ETV6:5-NTRK3:15
Seraseq Tumor Fusion RNA v4 Reference Materials	FGFR3-BAIAP2L1	FGFR3:17-BAIAP2L1:2	*	FGFR3:18-BAIAP2L1:2 is reported as well	FGFR3:17-BAIAP2L1:2
Seraseq Tumor Fusion RNA v4 Reference Materials	FGFR3-TACC3	FGFR3:17-TACC3:11	*		FGFR3:17-TACC3:11
Seraseq Tumor Fusion RNA v4 Reference Materials	KIF5B-RET	KIF5B:24-RET:11	*		KIF5B:24-RET:11
Seraseq Tumor Fusion RNA v4 Reference Materials	LMNA-NTRK1	LMNA:2-NTRK1:11	*		LMNA:2-NTRK1:11
Seraseq Tumor Fusion RNA v4 Reference Materials	MET ex 14 Skipping	MET:13-MET:15	*		MET:13-MET:15
Seraseq Tumor Fusion RNA v4 Reference Materials	NCOA4-RET	NCOA4:7-RET:12	*		NCOA4:7-RET:12
Seraseq Tumor Fusion RNA v4 Reference Materials	PAX8-PPARG	PAX8:9-PPARG:3	ND	PPARG(NM_138711.6), PAX8(NM_003466.4) => same preferred transcripts, yet both partners are non-target genes	-
Seraseq Tumor Fusion RNA v4 Reference Materials	SLC34A2-ROS1	SLC34A2:4-ROS1:35	*		SLC34A2:4-ROS1:35
Seraseq Tumor Fusion RNA v4 Reference Materials	SLC45A3-BRAF	SLC45A3:1-BRAF:8	*	No functional count, BRAF(NM_004333.6) (same), 5' UTR => Reportable	SLC45A3:1-BRAF:8
Seraseq Tumor Fusion RNA v4 Reference Materials	TFG-NTRK1	TFG:5-NTRK1:10	*		TFG:5-NTRK1:10
Seraseq Tumor Fusion RNA v4 Reference Materials	TMPRSS2-ERG	TMPRSS2:1-ERG:2	*	maximum reads < 20	TMPRSS2:1-ERG:2
Seraseq Tumor Fusion RNA v4 Reference Materials	TPM3-NTRK1	TPM3:8-NTRK1:10	*		TPM3:8-NTRK1:10

ACT GENOMICS ™

Summary

- Run summary
- Proposed solutions

Run summary

- Fusion v4-based workflow
    - ND variants
        - Fusion v4 RM
            - PAX8:9-PPARG:3
        - IVTALL
            - FusionRef_653 (NRG2:7-UBE2D2:5 (4 reads (10^3 copies))) and FusionRef_689 (NPM1:4-ALK:20 (AMBIGUITY))
- Arriba
    - ND variants
        - Fusion v4 RM
            - PAX8-PPARG, EGFR:1-EGFR:8, MET ex 14 Skipping
        - IVTALL
            - AR-V7, MET exon 14 skipping, EGFR VIII

ACT GENOMICS ™

Workflow comparison

- Workflow comparison
  - Concerns
    - Arriba workflow is not able to detect ARV7 and exon skipping variants. => Need other tool to assist (e.g., CTAT)
    - Arriba workflow has no in-frame/protein reads support =>
    - Arriba workflow may annotate the same break point with different RefSeq IDs. (Previous run (AANB01_504) : “NACC2-NTRK2”)
    - Fusion V4 workflow may not be able to report all possible boundaries => This issue also appears in Arriba. Not reporting all boundaries
- Possible solutions
  - Fusion-v4 based v0.95 + v1.4 (execute the same workflow with different annotation files)
  - Arriba + CTAT for splicing variant calling + Another tool for ARV7
    + tool for in-frame read & protein generation

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)	-	-	-

ACT GENOMICS ™

Preferred Transcript list

- 26 target genes

V1.4 MANE Select + V1.4 MANE Plus Clinical

V1.4 MANE Plus Clinical

ENST	ENSG	Gene	RefSeq ID
ENST00000374690.9	ENSG00000169083.18	AR	NM_000044.6
ENST00000646891.2	ENSG00000157764.14	BRAF	NM_004333.6
~~ENST00000288319.12~~	~~ENSG00000157554.20~~	~~ERG~~	~~NM_182918.4~~
~~ENST00000358487.10~~	~~ENSG00000066468.24~~	~~FGFR2~~	~~NM_000141.5~~
ENST00000405005.8	ENSG00000157168.22	NRG1	NM_013964.5
ENST00000629765.3	ENSG00000140538.17	NTRK3	NM_001012338.3
ENST00000251849.9	ENSG00000132155.14	RAF1	NM_002880.4
ENST00000644969.2	**ENSG00000157764.14**	**BRAF**	**NM_001374258.1**
ENST00000457416.7	**ENSG00000066468.24**	**FGFR2**	**NM_022970.4**

v0.95 MANE Select

v0.95 MANE Select

ENST	ENSG	Gene	RefSeq ID
ENST00000644969.2	**ENSG00000157764.14**	**BRAF**	**NM_001374258.1**
~~ENST00000288319.12~~	~~ENSG00000157554.20~~	~~ERG~~	~~NM_182918.4~~
~~ENST00000358487.10~~	~~ENSG00000066468.24~~	~~FGFR2~~	~~NM_000141.5~~
ENST00000442415.7	ENSG00000132155.14	RAF1	NM_001354689.3
ENST00000356819.7	ENSG00000157168.20	NRG1	NM_013957.5
ENST00000394480.6	ENSG00000140538.16	NTRK3	NM_002530.4
ENST00000442448.5	ENSG00000157554.19	ERG	NM_004449.4
ENST00000504326.5	ENSG00000169083.18	AR	NM_001348061.1

- Note:
  - BRAF (**ENST00000644969.2**) => MANE Select (v0.95) → MANE Plus Clinical (v1.4) (v0.95 MANE Select, preferred transcript)
  - FGFR2 (**ENST00000457416.7**) => MANE Plus Clinical (v1.4) (not in MANE Select, transcript to include)

ACT GENOMICS ™

Proposed solution
(to report all MANE Select/Plus Clinical v1.4 “target” transcripts)

- Fusion pipeline + v0.95 db (18,583 MANE Select + 4 additional transcripts (NRG1, AR, NTRK3, ERG) from GENCODE r38)
  # To-do
  => add probe annotation
- Fusion pipeline + v1.4 db (19,226 MANE Select transcripts)
  # To-do
  => use MANE Plus Clinical for FGFR2 (bugfix: BRAF kinase)

Gene_name	V4-ENST_ID	V4-Source	V5-ENST_ID	Current v1.4 db
FGFR2	-	(v1.4 MANE Plus Clinical)	ENST00000457416.7	**Not included**
BRAF	ENST00000644969.2	v0.95 MANE Select (=> v1.4 MANE Plus Clinical)	ENST00000646891.2	v1.4 MANE Select
NRG1	ENST00000356819.7	GENCODE r38	ENST00000405005.8	v1.4 MANE Select
RAF1	ENST00000442415.7	v0.95 MANE Select	ENST00000251849.9	v1.4 MANE Select
AR	ENST00000504326.5	GENCODE r38	ENST00000374690.9	v1.4 MANE Select
NTRK3	ENST00000394480.6	GENCODE r38	ENST00000629765.3	v1.4 MANE Select
ERG	ENST00000442448.5	GENCODE r38, only include the first 3 exons for TMPRSS2-ERG fusion detection	-	**Not included**

ACT GENOMICS ™