

# Fusion V5 (MANE 0.95 to 1.4)

Bioinformatics Development

Sandy

2025.04.15

# Transcriptome db update

- Steps (description & executed commands)
- Files summary

# DB preparation steps (I)

- Steps & Tools

No.	Steps	Description	Tool
1	download required data set from GENCODE and MANE	see sheet "Fusionv4 DB select (18,587)"	wget, rsync, zcat, samtools
2	generate namemap file manually	see sheet "Fusionv4 DB select (18,587)"	awk, cat
3	retrieve transcript gff file	gff file preprocessing (retrieve "transcript" label from gff column 3)	zgrep, awk
4	gff to bed conversion with "bedops_2.4.39/bin/convert2bed"	bed file generation (convert the information in gff to bed for transcript region extraction)	convert2bed (alternative way: bedtools)
4	get fasta via "bedtools getfasta"	fasta file generation (generate the fasta file for the selected regions in bed)	bedtools getfasta
5 (1)	generate annotation file via RefFusion.py	<MANE: 18,583> empty pseudo intron annotation table + pseudo N (10N) fasta generation	RefFusion.py
5 (2)	generate annotation file via RefFusion.py	<GENCODE: 4 manually curated id> empty pseudo intron annotation table + pseudo N (10N) fasta generation	RefFusion.py
6	ERG One manual curation step (editing annotation tables (in the empty genome + transcript loci files) + transcript fasta: keep 3 exons (in the pseudo fasta file) <- 3 output files of 5-(2)	ERG isotranscript (ENST00000442448.5) addition and adjustment	save only the first 3 exons
7	File concatenation with GENCODE (ERG gene adjustment) and MANE output file	concatenation of the curated transcript information to the existing MANE info files obtained from previous steps (5-(1) + 6 = 7)	cat
8	generate annotation with GSP information	perform primer sequence alignment	blastn
8	generate annotation with GSP information	parse the alignment information and add it into the annotation tables	blastnparser.py

Ref repo:  
[https://bitbucket.org/actgenomics/reference\\_preparation\\_pipeline/src/master/](https://bitbucket.org/actgenomics/reference_preparation_pipeline/src/master/)

# Input files/parameters (for pipeline execution)

- Configuration file

Bwa index based on fasta "refFile"

1 => Gencode + MANE (db step 5), v0.28.0

```
params {
  refFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224.fasta"
  ambFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224.fasta.amb"
  annFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224.fasta.ann"
  bwtFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224.fasta.bwt"
  pacFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224.fasta.pac"
  saFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224.fasta.sa"
  annoFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224.corrected.transcript.loci"
  gannoFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224/PA043_ACTFusionV4_PseudoIntron_MANE-v0.95_GENCODE-r38_mPCR-v9.0_GRCh38.20231224.genome.5_3_UTR.loci"
  isoformfaFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.gencode.v38.transcripts.protein.coding.22.isoforms.fa"
  isoformambFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.gencode.v38.transcripts.protein.coding.22.isoforms.fa.amb"
  isoformannFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.gencode.v38.transcripts.protein.coding.22.isoforms.fa.ann"
  isoformbwtFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.gencode.v38.transcripts.protein.coding.22.isoforms.fa.bwt"
  isoformpacFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.gencode.v38.transcripts.protein.coding.22.isoforms.fa.pac"
  isoformsaFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.gencode.v38.transcripts.protein.coding.22.isoforms.fa.sa"
  isoformmetaFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.gencode.v38.transcripts.protein.coding.22.isoforms.fa.catted.meta"
  isoformfilteringflag = "1"
  truncatedmode = "0"
  truncatedseq_min_aligned_len = "12"
  inSpikeinFastqR1 = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/AAS502_953_IonXpress_001_1_MET_14_read-spike-in.fq.gz"
  pdbFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/protein_26db.v1.fasta"
  pdbmFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/protein_26db.v1.mapping.meta.txt"
  adapFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/torrentadapters.corrected.20230704.csv"
  qcconfigFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/filter_internal.QC9.0.mgsp.qcr.0.5.241206.config"
  readqcconfigFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/filter_fuscall_12aa_oratio_read.QC.json"
  qcconfigFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/filter_fuscall_12aa_oratio_read.QC.json"
  incqctemplateFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/Blank_template.fusioncalling.QC.json"
  boundaryqcFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/boundaryqcthresholds.v4.default.config"
  fusion_container = "actgenomics/torrent_fusion_pipeline:v0.23.0"
}

process {
  withLabel: process_low {
    cpus = 1
    maxForks = 1
  }
  executor = 'local'
  container = 'actgenomics/torrent_fusion_pipeline:v0.23.0'
  cpus = '2'
  memory = '16.G'
}
```

1

Ref:

- Repo: (Ref. files: ./params/fus\_torrent\_bam.14\_testdata.json, ./params/fus\_illumina\_fastq.json)  
[https://bitbucket.org/actgenomics/torrent\\_fusion\\_pipeline\\_nextflow/src/master/](https://bitbucket.org/actgenomics/torrent_fusion_pipeline_nextflow/src/master/)

Modified steps from V4 → V5  
transcriptome db update

# DB

preparation  
steps (V5)

- Steps & Tools

Ref. image:  
• actgenomics/fusion\_dev:v0.6

No.	Steps	Description	Tool
1	download required data set from GENCODE and MANE	see sheet "Fusionv4 DB select (18,587)" => 19,226 v1.4 MANE	wget, rsync, zcat, samtools
2	generate namemap file manually	see sheet "Fusionv4 DB select (18,587)" => 19,226 v1.4 MANE	awk, cat
3	retrieve transcript gff file	gff file preprocessing (retrieve "transcript" label from gff column 3) => Inclusion criteria: chr1-22, X, Y, protein_coding, MANE_Select (summary.txt => MANE Select; manually curated)	zgrep, awk, filter_mane_gff.py (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
4	gff to bed conversion with "bedops_2.4.39/bin/convert2bed"	bed file generation (convert the information in gff to bed for transcript region extraction)	convert2bed (alternative way: bedtools) (/tools/Fusion/convert2bed)
4	get fasta via "bedtools getfasta"	fasta file generation (generate the fasta file for the selected regions in bed)	bedtools getfasta
5 (1)	generate annotation file via RefFusion.py → RefFusion.v2.py	<MANE: 18,583> empty pseudo intron annotation table + pseudo N (10N) fasta generation => 19,226 v1.4 MANE	RefFusion.py => RefFusion.v2.py (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
5 (2)	generate annotation file via RefFusion.py	<GENCODE: 4 manually curated id> empty pseudo intron annotation table + pseudo N (10N) fasta generation	RefFusion.py
6	ERG One manual curation step (editing annotation tables (in the empty genome + transcript loci files) + transcript fasta: keep 3 exons (in the pseudo fasta file) ← 3 output files of 5- (2)	ERG isotranscript (ENST00000442448.5) addition and adjustment	save only the first 3 exons
7	File concatenation with GENCODE (ERG gene adjustment) and MANE output file	concatenation of the curated transcript information to the existing MANE info files obtained from previous steps (5- (1) + 6- 7)	cat
8-0-a	convert v5 probe regions to the regions on pseudo transcriptome (MANE v1.4)	1039 probe regions are converted to 533 mapped (probe) exons (on pseudo-transcriptome v1.4)	candidate_exons_mapping.sh (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
8-0-b	modify the header of the probe fasta file (replace with primer-like header)	Header conversion step for blastnparser.py & blastn result	Probe_faheader_converter.py (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
8-0-c	generate annotation with GSP information	perform primer sequence alignment	blastn (/tools/Fusion/ncbi-blast/bin/blastn -task blastn-short -dust no)
8-0-c	generate annotation with GSP information	parse the alignment information and add it into the annotation tables	blastnparser.py (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
9	generate index files for bwa	generate the 5 indices required for "bwa mem"	/tools/Fusion/bwa index

# Input files for probe region to MANE v1.4 loci conversion

## Step 0

- Source files

- Probe (region) bed file provided by Lucy
  - 1039 v5 probe regions  
/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/captureprobe\_250401/ACTFusionv5\_target-region\_PartAB\_individual\_1039.bed
- Genome/Transcriptome DB (downloaded from websites)
  - MANE v1.4 DB ([https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE\\_human/release\\_1.4/](https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE_human/release_1.4/))  
(/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/mane\_v1.4/OpenDB\_MANE\_human\_v1.4/release\_1.4/MANE.GRCh38.v1.4.summary.txt.gz)
  - Gencode V47 ([http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\\_human/release\\_47/](http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_47/))  
(/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/gencode\_v47/OpenDB\_GENCODE\_human\_r47/GRCh38.p14.genome.fa.gz)  
(/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/gencode\_v47/OpenDB\_GENCODE\_human\_r47/gencode.v47.annotation.gff3.gz)
- Namemap  
(19,338 transcript – RefSeq ID map derived from MANE v1.4 “MANE.GRCh38.v1.4.summary.txt.gz” => “MANE Select” in column 10)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.summary.name  
map
- Derived files generated via steps 1-5(1) (blank loci files)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.transcript.loci (Blank Loci file)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.genome.loci (Genomic Loci file)

# Source gff file

- Enter image “actgenomics/fusion\_dev:v0.6” => samtools 1.9 from base image

```
./OpenDB_GENCODE_human_r47/derived/  
|-- GRCh38.p14.genome.fa  
|-- GRCh38.p14.genome.fa.fai  
|-- gencode.v47.annotation.gff3  
`-- gencode.v47.annotation.transcript.gff3
```

0 directories, 4 files

# commands

```
awk -F"\t" '{if($3=="transcript") print}' > ./OpenDB_GENCODE_human_r38/derived/gencode.v38.annotation.transcript.gff3  
root@184f56381e9e:/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/gencode_v47# zgrep -v  
"^#" ./OpenDB_GENCODE_human_r47/gencode.v47.annotation.gff3.gz | awk -F"\t" '{if($3=="transcript") print}'  
> ./OpenDB_GENCODE_human_r47/derived/gencode.v47.annotation.transcript.gff3
```

# Derived output summary

- Derived output files

# MANE v1.4 derived files

/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/

|-- MANE.GRCh38.v1.4.ensembl\_genomic.gff

|-- MANE.GRCh38.v1.4.summary.namemap

`-- MANE.GRCh38.v1.4.summary.txt

0 directories, 3 files

# GENCODE v47 derived files

/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/gencode\_v47/OpenDB\_GENCODE\_human\_r47/derived/

|-- GRCh38.p14.genome.fa

|-- GRCh38.p14.genome.fa.fai

|-- gencode.v47.annotation.gff3

`-- gencode.v47.annotation.transcript.gff3

0 directories, 4 files



# Input files for probe region to MANE v1.4 loci conversion

- Source files
  - Probe (region) bed file provided by Lucy
    - 1039 v5 probe regions  
/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/captureprobe\_250401/ACTFusionv5\_target-region\_PartAB\_individual\_1039.bed
  - Genome/Transcriptome DB (downloaded from websites)
    - MANE v1.4 DB ([https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE\\_human/release\\_1.4/](https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE_human/release_1.4/))  
(/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/mane\_v1.4/OpenDB\_MANE\_human\_v1.4/release\_1.4/MANE.GRCh38.v1.4.summary.txt.gz)
    - Gencode V47 ([http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\\_human/release\\_47/](http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_47/))  
(/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/gencode\_v47/OpenDB\_GENCODE\_human\_r47/GRCh38.p14.genome.fa.gz)  
(/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/gencode\_v47/OpenDB\_GENCODE\_human\_r47/gencode.v47.annotation.gff3.gz)
- Namemap  
(19,338 transcript – RefSeq ID map derived from MANE v1.4 “MANE.GRCh38.v1.4.summary.txt.gz” => “MANE Select” in column 10)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.summary.name  
map
- Derived files generated via steps 1-5(1) (blank loci files)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.transcript.loci (Blank Loci file)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.genome.loci (Genomic Loci file)

## Step 2

# Namemap construction for 19,292 transcripts

- 66 missing IDs found (=> 66 “MANE\_Plus\_Clinical” IDs)

```
grep -Fvx -f /mnt/RD_Develop/sandyteng/workdir/namemap_fusionv5_2_manev1.4/MANE.GRCh38.v1.4.summary.namemap.19338.namelist
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/Output_Final/MANE.GRCh38.v1.4.ensembl_genomic.transcript.list
```

- Generate namemap for the 66 MANE Plus Clinical transcripts

```
more /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.summary.txt | grep -wvf
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/Output_Final/MANE.GRCh38.v1.4.summary.namemap.66.missing.namelist | awk -F"\t" '{if($10
== "MANE Plus Clinical")print $8"\t"$2"\t"$4"\t"$6}' >
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.summary.plus.clinical.namemap
```

- Generate namemap for the 19338 (MANE Select) + 66 (MANE Plus Clinical) transcripts

```
cat /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.summary.namemap
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.summary.plus.clinical.namemap >
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.select.and.plus.clinical.namemap
```

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE
_human_v1.4/derived/
├── MANE.GRCh38.v1.4.ensembl_genomic.gff
├── MANE.GRCh38.v1.4.ensembl_genomic.transcript.bed
├── MANE.GRCh38.v1.4.ensembl_genomic.transcript.fasta
├── MANE.GRCh38.v1.4.ensembl_genomic.transcript.gff
├── MANE.GRCh38.v1.4.select.and.plus.clinical.namemap => 19404
├── MANE.GRCh38.v1.4.summary.namemap => 19338 (MANE Select)
├── MANE.GRCh38.v1.4.summary.plus.clinical.namemap => 66 (MANE Plus Clinical)
└── MANE.GRCh38.v1.4.summary.txt
```

0 directories, 8 files

# GFF → bed → fasta conversion (conversion log)

- /mnt/RD\_Develop/sandyteng/ACTFusionV5/code/filter\_mane\_gff.py

# (1) extract gff lines with "transcript" value in the 3<sup>rd</sup> column

# (1) extract gff lines via "filter\_mane\_gff.py" (=> 19,292 protein coding transcripts)

python3 /mnt/RD\_Develop/sandyteng/ACTFusionV5/code/filter\_mane\_gff.py -

i ./mane\_v1.4/OpenDB\_MANE\_human\_v1.4/release\_1.4/MANE.GRCh38.v1.4.ensembl\_genomic.gff.gz -

o ./mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.ensembl\_genomic.transcript.gff

# (2) convert gff to bed (=> 19,292 bed lines)

/tools/Fusion/convert2bed -i gff -d < ./mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.ensembl\_genomic.transcript.gff

> ./mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.ensembl\_genomic.transcript.bed

# (3) extract sequence (gencode.genome.fa) via MANE v1.4 bed file (transcript.bed) (=> 19,292 fasta header lines (format: ">ENSTXXX.X"))

bedtools getfasta -name -s -fi ./gencode\_v47/OpenDB\_GENCODE\_human\_r47/derived/GRCh38.p14.genome.fa -

bed ./mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.ensembl\_genomic.transcript.bed -

fo ./mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.ensembl\_genomic.transcript.fasta

# GFF → bed → fasta conversion

- Convert MANE transcript gff to bed

**# (1) gff file filtering** (includes transcript, protein coding, chr1-22, X, Y)

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/
```

```
|-- MANE.GRCh38.v1.4.ensembl_genomic.gff
|-- MANE.GRCh38.v1.4.ensembl_genomic.transcript.bed
|-- MANE.GRCh38.v1.4.ensembl_genomic.transcript.gff
|-- MANE.GRCh38.v1.4.summary.namemap
`-- MANE.GRCh38.v1.4.summary.txt
```

0 directories, 5 files

**# (2) gff to bed**

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/
```

```
|-- MANE.GRCh38.v1.4.ensembl_genomic.gff
|-- MANE.GRCh38.v1.4.ensembl_genomic.transcript.bed
|-- MANE.GRCh38.v1.4.ensembl_genomic.transcript.gff
|-- MANE.GRCh38.v1.4.summary.namemap
`-- MANE.GRCh38.v1.4.summary.txt
```

0 directories, 5 files

**# (3) bed to fasta**

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/gencode_v47/OpenDB_GENCODE_human_r47/derived/
```

```
|-- GRCh38.p14.genome.fa
|-- GRCh38.p14.genome.fa.fai
|-- gencode.v47.annotation.gff3
`-- gencode.v47.annotation.transcript.gff3
```

0 directories, 4 files

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/
```

```
|-- MANE.GRCh38.v1.4.ensembl_genomic.gff
|-- MANE.GRCh38.v1.4.ensembl_genomic.transcript.bed
|-- MANE.GRCh38.v1.4.ensembl_genomic.transcript.fasta
|-- MANE.GRCh38.v1.4.ensembl_genomic.transcript.gff
|-- MANE.GRCh38.v1.4.summary.namemap
`-- MANE.GRCh38.v1.4.summary.txt
```

0 directories, 6 files

# (Rerun) Command (V5 probe → V1.4 exons) (19,292 – 66 = 19,226 MANE only transcripts)

- Executed commands

```
# generate the 19,226 id list
grep -Fvx -f /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/Output_Final/MANE.GRCh38.v1.4.summary.namemap.66.missing.namelist
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/Output_Final/MANE.GRCh38.v1.4.ensembl_genomic.transcript.list >
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/Output_Final/PA053_ACTFusionV5_Pseudointron_MANE-v1.4_GENCORE-r47_capture-
v1.0_GRCh38.20250407.transcript.MANE.only.list
```

```
# generate the 19,226 loci (transcript, genomic)
mkdir -p /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select
python3 /mnt/RD_Develop/sandyteng/ACTFusionV5/code/RefFusion.v2.py \
-g /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.ensembl_genomic.gff \
-m /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.summary.txt \
-f /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.ensembl_genomic.transcript.fasta \
-p /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/Output_Final/PA053_ACTFusionV5_Pseudointron_MANE-v1.4_GENCORE-r47_capture-
v1.0_GRCh38.20250407.transcript.MANE.only.list \
-o /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE

# create & enter container
docker run --rm -v /mnt:/mnt/ -it actgenomics/fusion_dev:v0.6

# create output folder
mkdir -p /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/

# obtain mapping exons (pseudo locations on 10*N transcriptome)
bash /mnt/RD_Develop/sandyteng/ACTFusionV5/code/candidate_exons_mapping.sh /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.genome.loci
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.transcript.loci
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.select.and.plus.clinical.namemap
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/captureprobe_250401/ACTFusionV5_target-region_PartAB_individual_1039.bed fusionv4.MANE.v1.4.GENCORE.r47
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/ /tools/Fusion
```

# Get Fasta (Probe anchored exon sequences)

- Target (mapped to 1.4) bed  
/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/InhouseDB\_Probe/captureprobe\_250407\_MANE\_Select/fusionv4.MANE.v1.4.GENCODE.r47.candidate.exons.transcript.bed (bed using pseudo FASTA file generated via “RefFusion.v2.py”)

# create output folder

```
mkdir -p /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq
```

# create & enter container

```
docker run --rm -v /mnt:/mnt/ -it actgenomics/fusion_dev:v0.6
```

# generate index for fasta file (for bedtools getfasta)

```
samtools faidx /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.fasta
```

# extract mapped exons (candidate.exons.transcript.bed) sequences from gencode fasta file (gencode.genome.fa)

```
bedtools getfasta -name -s -fi /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.fasta -bed
```

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/fusionv4.MANE.v1.4.GENCODE.r47.candidate.exons.transcript.bed -fo
```

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/MANE.GRCh38.v1.4.0407.probe.fasta
```

# FASTA header conversion

- Source files
- Primer fasta file (403 primers)
  - /mnt/RD\_Develop/yfhuang/ACTFusionV4DB/20240507/InhouseDB\_mPCR/v9.0/20231105\_mPCR9.0.fasta
- Probe fasta file (533 mapped exons)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/InhouseDB\_Probe/captureprobe\_250407\_MANE\_Select/probeseq/MANE.GRCh38.v1.4.0407.probe.fasta
- Namemap file (19,404 MANE Select + MANE Plus Clinical transcripts (MANE, v1.4))
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.select.and.plus.clinical.namemap
    - 19,338 MANE Select transcripts
      - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.summary.namemap
    - 66 MANE Plus Clinical transcripts
      - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/mane\_v1.4/OpenDB\_MANE\_human\_v1.4/derived/MANE.GRCh38.v1.4.summary.plus.clinical.namemap

# Probe\_faheader\_converter.py

- usage: Probe\_faheader\_converter.py [-h] -f INPUTPROBEFASTA -n MANENAMEMAPTABLE -o OUTPUTPROBEFASTA

```
>ENST00000373237.4;PSMB2-exon-fusionv4-2()
```

```
GCCGGGCATGGTGGTGGGCAACTGTAATCCCAGCTACTCGGGAGGCTAAGGCAAGAGAATTGCTTGAACCCAGAGGCGGAGGTTGCAGTGAGCTGAAATCGTGCCACTGCACTCCAGCCTGGG
```

```
>ENST00000373237.4;PSMB2-exon-fusionv4-3()
```

```
AAACAGAGTGACTATAGTTAATAATAGCTTAACTGTATATTTTAAAATAACTAAAAGAGTATAATTGGGTT
```

```
>ENST00000373237.4;PSMB2-exon-fusionv4-5()
```

```
CGCCATTGCACTCCAGCCTGGGCAACAGAGCGAGACTCCGTCTCAAAAAA
```

```
>Probe-mane001|PSMB2|NM_002794.5|ENST00000373237.4|2|F|123
```

```
GCCGGGCATGGTGGTGGGCAACTGTAATCCCAGCTACTCGGGAGGCTAAGGCAAGAGAATTGCTTGAACCCAGAGGCGGAGGTTGCAGTGAGCTGAAATCGTGCCACTGCACTCCAGCCTGGG
```

```
>Probe-mane002|PSMB2|NM_002794.5|ENST00000373237.4|3|F|71
```

```
AAACAGAGTGACTATAGTTAATAATAGCTTAACTGTATATTTTAAAATAACTAAAAGAGTATAATTGGGTT
```

```
>Probe-mane003|PSMB2|NM_002794.5|ENST00000373237.4|5|F|50
```

```
CGCCATTGCACTCCAGCCTGGGCAACAGAGCGAGACTCCGTCTCAAAAAA
```

```
python3 Probe_faheader_converter.py \
```

```
-f /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/MANE.GRCh38.v1.4.0407.probe.fasta \
```

```
-n /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.select.and.plus.clinical.namemap \
```

```
-o MANE.GRCh38.v1.4.0407.probe.wtprimerlikeheader.fasta.gz
```



## Step 8 (v1.4) – blastn (probe – transcriptome)

Description:

1. Convert Probe.fasta headers
2. Align probe sequence to transcriptome with pseudo-intron (10\*N) via blastn

- Commands

# probe fasta generation

```
python3 /mnt/RD_Develop/sandyteng/ACTFusionV5/code/Probe_faheader_converter.py \
-f /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/MANE.GRCh38.v1.4.0407.probe.fasta \
-n /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.select.and.plus.clinical.namemap \
-o
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/MANE.GRCh38.v1.4.0407.probe.wtprimerli
keheader.fasta.gz
```

# unzip fasta.gz

```
gunzip
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/MANE.GRCh38.v1.4.0407.probe.wtprimerli
keheader.fasta.gz
```

# create & enter container

```
docker run --rm -v /mnt:/mnt/ -it actgenomics/fusion_dev:v0.6
```

# folder to store blastn results

```
mkdir -p /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/blastn/
```

# add on Probe information with BLASTN result

```
/tools/Fusion/ncbi-blast/bin/blastn -query
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/MANE.GRCh38.v1.4.0407.probe.wtprimerli
keheader.fasta -subject /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.fasta -outfmt 6 -task blastn-short >
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/blastn/20250407_probe.blastn
```

## Step 8 (v1.4) – blastn results

- Result files

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/  
|-- MANE.GRCh38.v1.4.0407.probe.fasta  
`-- MANE.GRCh38.v1.4.0407.probe.wtprimerlikeheader.fasta
```

0 directories, 2 files

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/blastn/  
`-- 20250407_probe.blastn
```

0 directories, 1 file

## Step 8 (v1.4) – generate loci file with probe info

- Loci file generation

# create & enter container

```
docker run --rm -v /mnt:/mnt/ -it actgenomics/fusion_dev:v0.6
```

# folder to stored db output

```
mkdir -p /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_Loci/250407/
```

# blastn parser

```
python3 /mnt/RD_Develop/sandyteng/ACTFusionV5/code/blastnparser.py \
```

```
-if /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/blastn/20250407_probe.blastn \
```

```
-mp /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.select.and.plus.clinical.namemap \
```

```
-lf /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.transcript.loci >
```

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_Loci/250407/PA053_ACTFusionV5_Pseudointron_MANE-v1.4_GENCODE-r47_capture-v1.0_GRCh38.20250407.transcript.MANE.only.loci
```

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/blastn/`-- 20250407_probe.blastn
```

0 directories, 1 file

0 directories, 1 file

## Step 8 (v1.4) – loci file

- ALK (29 exons)

```
>ENST00000389048.8  ENSG00000171094.18  ALK  NM_004304.5  chr2  29192774  29921586  -  29  728813 6520  T
1  1  1594 1  1594 1  F  1|1594|F|Probe-mane046
2  203890 204009 1605 1724 1  F  1605|1724|F|Probe-mane047
3  226573 226737 1735 1899 1  F  1735|1899|F|Probe-mane048
4  389471 389672 1910 2111 1  F  1910|2111|F|Probe-mane049
5  537728 537855 2122 2249 1  F  2122|2249|F|Probe-mane050
6  593106 593237 2260 2391 1  F  2260|2391|F|Probe-mane051
7  600705 600836 2402 2533 1  F  2402|2533|F|Probe-mane052
8  603183 603283 2544 2644 1  F  2544|2644|F|Probe-mane053
9  624530 624699 2655 2824 1  F  2655|2824|F|Probe-mane054
10 646091 646185 2835 2929 1  F  2835|2929|F|Probe-mane055
11 646360 646488 2940 3068 1  F  2940|3068|F|Probe-mane056
12 670320 670482 3079 3241 1  F  3079|3241|F|Probe-mane057
13 681757 681907 3252 3402 1  F  3252|3402|F|Probe-mane058
14 687891 688022 3413 3544 1  F  3413|3544|F|Probe-mane059
15 689139 689283 3555 3699 1  F  3555|3699|F|Probe-mane060
16 692521 692703 3710 3892 1  F  3710|3892|F|Probe-mane061
17 693915 694013 3903 4001 1  F  3903|4001|F|Probe-mane062
18 694513 694665 4012 4164 1  F  4012|4164|F|Probe-mane063
19 696022 696126 4175 4279 1  F  4175|4279|F|Probe-mane064
20 698059 698245 4290 4476 1  F  4290|4476|F|Probe-mane065
21 698980 699070 4487 4577 1  F  4487|4577|F|Probe-mane066
22 699179 699243 4588 4652 1  F  4588|4652|F|Probe-mane067
23 700752 700881 4663 4792 1  F  4663|4792|F|Probe-mane068
24 707506 707603 4803 4900 1  F  4803|4900|F|Probe-mane069
25 711709 711801 4911 5003 1  F  4911|5003|F|Probe-mane070
26 714315 714416 5014 5115 1  F  5014|5115|F|Probe-mane071
27 723911 724045 5126 5260 1  F  5126|5260|F|Probe-mane072
28 724727 724817 5271 5361 1  F  5271|5361|F|Probe-mane073
29 727665 728813 5372 6520 1  F  5372|6520|F|Probe-mane074
```

Inspection:

Only 531 probes are annotated.

Missing probes:

Probe-mane081, Probe-mane501

>Probe-mane501|TMPRSS2|NM\_005656.4|ENST00000332149.10|2|F|71  
AGTGAGGGGGTGAGGGCATGAGTGAGGAGGTGAGGGGGTGAGTGAGGAGGT  
GAGGGGGTGAGTGAGGAGGT

>Probe-mane081|ERBB4|NM\_005235.3|ENST00000342788.9|7|F|142  
ATATGTTATATATGTTATATATAATATATGTTATATGTTATATATGTTATATGTTATATAT  
GTTATATGTTATATATGTTATATGTTATATATATTATATATATTATATAAGTTATATATTATA  
TATTATATATTATG

# The 2 missing probes

- >Probe-mane081|ERBB4|NM\_005235.3|ENST00000342788.9|7|F|142
- >Probe-mane501|TMPRSS2|NM\_005656.4|ENST00000332149.10|2|F|71

```
>ENST00000342788.9 ENSG00000178568.16 ERBB4 NM_005235.3 chr2 211375717 212538802 - 28 1163086 12367 T
1 1 354 1 354 1 F 1|354|F|Probe-mane075
2 413900 414051 365 516 1 F 365|516|F|Probe-mane076
3 591187 591373 527 713 1 F 527|713|F|Probe-mane077
4 750644 750778 724 858 1 F 724|858|F|Probe-mane078
5 788099 788164 869 934 1 F 869|934|F|Probe-mane079
6 813609 813727 945 1063 1 F 945|1063|F|Probe-mane080
7 816269 816410 1074 1215 0
8 825155 825268 1226 1339 1 F 1226|1339|F|Probe-mane082
9 826627 826753 1350 1476 1 F 1350|1476|F|Probe-mane083
10 833412 833485 1487 1560 1 F 1487|1560|F|Probe-mane084
11 834609 834699 1571 1661 1 F 1571|1661|F|Probe-mane085
12 836637 836836 1672 1871 1 F 1672|1871|F|Probe-mane086
13 859619 859751 1882 2014 1 F 1882|2014|F|Probe-mane087
14 865546 865639 2025 2118 1 F 2025|2118|F|Probe-mane088
15 873326 873480 2129 2283 1 F 2129|2283|F|Probe-mane089
16 880975 881049 2294 2368 1 F 2294|2368|F|Probe-mane090
17 908209 908341 2379 2511 1 F 2379|2511|F|Probe-mane091
18 914759 914881 2522 2644 1 F 2522|2644|F|Probe-mane092
19 919528 919626 2655 2753 1 F 2655|2753|F|Probe-mane093
20 976715 976900 2764 2949 1 F 2764|2949|F|Probe-mane094
21 1107703 1107858 2960 3115 1 F 2960|3115|F|Probe-mane095
22 1110320 1110395 3126 3201 1 F 3126|3201|F|Probe-mane096
23 1114502 1114648 3212 3358 1 F 3212|3358|F|Probe-mane097
24 1116699 1116796 3369 3466 1 F 3369|3466|F|Probe-mane098
25 1118192 1118362 3477 3647 1 F 3477|3647|F|Probe-mane099
26 1150811 1150858 3658 3705 1 F 3658|3705|F|Probe-mane100
27 1151653 1151950 3716 4013 1 F 3716|4013|F|Probe-mane101
28 1154743 1163086 4024 12367 1 F 4024|12367|F|Probe-mane102
```

```
>ENST00000332149.10 ENSG00000184012.14 TMPRSS2 NM_005656.4 chr21 41464305 41508158 - 14 43854 3580 T
1 1 78 1 78 1 F 1|78|F|Probe-mane500
2 9970 10040 89 159 0
3 13581 13803 170 392 1 F 170|392|F|Probe-mane502
4 18566 18652 403 489 1 F 403|489|F|Probe-mane503
5 19646 19765 500 619 1 F 500|619|F|Probe-mane504
6 27557 27683 630 756 1 F 630|756|F|Probe-mane505
7 28877 28987 767 877 1 F 767|877|F|Probe-mane506
8 31539 31582 888 931 1 F 888|931|F|Probe-mane507
9 34663 34834 942 1113 1 F 942|1113|F|Probe-mane508
10 36178 36353 1124 1299 1 F 1124|1299|F|Probe-mane509
11 37416 37511 1310 1405 1 F 1310|1405|F|Probe-mane510
12 39621 39763 1416 1558 1 F 1416|1558|F|Probe-mane511
13 40273 40425 1569 1721 1 F 1569|1721|F|Probe-mane512
14 42006 43854 1732 3580 1 F 1732|3580|F|Probe-mane513
```

# Try different blastn command (to rescue the 2 missing probes)

- Test commands => the 2 missing probes are rescued!

# create & enter container

```
docker run --rm -v /mnt:/mnt -it actgenomics/fusion_dev:v0.6
```

# add on Probe information with BLASTN result

```
/tools/Fusion/ncbi-blast/bin/blastn -query
```

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/MANE.GRCh38.v1.4.0407.probe.wtprimerlikeheader.fasta -subject /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.fasta -outfmt 6 -task blastn-short -dust no >  
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/blastn/20250407_probe.default.blastn
```

# blastn parser (loci annotation)

```
python3 /mnt/RD_Develop/sandyteng/ACTFusionV5/code/blastnparser.py \
```

```
-if /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/blastn/20250407_probe.default.blastn \
```

```
-mp /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.select.and.plus.clinical.namemap \
```

```
-lf /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.transcript.loci >
```

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_Loci/250407/PA053_ACTFusionV5_PseudoIntron_MANE-v1.4_GENCODE-r47_capture-v1.0_GRCh38.20250407.transcript.MANE.only.blastn.default.loci
```

# BWA indices generation

- Source FASTA (19,226 MANE Select v1.4 transcripts)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.fasta

# create & enter container

docker run --rm -v /mnt:/mnt/ -it actgenomics/fusion\_dev:v0.6

# generate bwa indices for fusion calling module ("ACTGfuscall.py")

/tools/Fusion/bwa index /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.fasta

```
root@d550999a19b:/tools# /tools/Fusion/bwa index /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.fasta
[bwa_index] Pack FASTA... 0.84 sec
[bwa_index] Construct BWT for the packed sequence...
[BWTIncCreate] textLength=145800268, availableWord=22258764
[BWTIncConstructFromPacked] 10 iterations done. 36716444 characters processed.
[BWTIncConstructFromPacked] 20 iterations done. 67829580 characters processed.
[BWTIncConstructFromPacked] 30 iterations done. 95479004 characters processed.
[BWTIncConstructFromPacked] 40 iterations done. 120049852 characters processed.
[BWTIncConstructFromPacked] 50 iterations done. 141884412 characters processed.
[bwt_gen] Finished constructing BWT in 52 iterations.
[bwa_index] 35.79 seconds elapse.
[bwa_index] Update BWT... 0.86 sec
[bwa_index] Pack forward-only FASTA... 0.57 sec
[bwa_index] Construct SA from BWT and Occ... 20.03 sec
[main] Version: 0.7.17-r1188
[main] CMD: /tools/Fusion/bwa index /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.fasta
[main] Real time: 61.155 sec; CPU: 58.087 sec
```

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select
|-- /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.fasta
|-- /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.fasta.fai
|-- /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.genome.loci
|-- /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.isotranscript_retrieval.log
`-- /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.transcript.loci
```

0 directories, 5 files

# output directory

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/
|-- 20250407_MANE.fasta
|-- 20250407_MANE.fasta.amb
|-- 20250407_MANE.fasta.ann
|-- 20250407_MANE.fasta.bwt
|-- 20250407_MANE.fasta.fai
|-- 20250407_MANE.fasta.pac
|-- 20250407_MANE.fasta.sa
|-- 20250407_MANE.genome.loci
|-- 20250407_MANE.isotranscript_retrieval.log
`-- 20250407_MANE.transcript.loci
```

0 directories, 10 files

# File summary

- Steps
- Generated files
- Updated configuration file



# DB preparation steps (V5) (clean)

• Steps & Tools

Ref. image:  
• actgenomics/fusion\_dev:v0.6

No.	Steps	Description	Tool
1	download required data set from GENCODE and MANE	see sheet "Fusionv4 DB select <del>(18,587)</del> " => 19,226 v1.4 MANE	wget, rsync, zcat, samtools
2	generate namemap file manually	see sheet "Fusionv4 DB select <del>(18,587)</del> " => 19,226 v1.4 MANE	awk, cat
3	retrieve transcript gff file	gff file preprocessing (retrieve "transcript" label from gff column 3) => Inclusion criteria: chr1-22, X, Y, protein_coding, MANE_Select (summary.txt => MANE Select; manually curated)	zgrep, awk, filter_mane_gff.py (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
4	gff to bed conversion with "bedops_2.4.39/bin/convert2bed"	bed file generation (convert the information in gff to bed for transcript region extraction)	convert2bed (alternative way: bedtools) (/tools/Fusion/convert2bed)
4	get fasta via "bedtools getfasta"	fasta file generation (generate the fasta file for the selected regions in bed)	bedtools getfasta
5 (1)	generate annotation file via RefFusion.py → RefFusion.v2.py	<MANE: <del>18,583</del> > empty pseudo intron annotation table + pseudo N (10N) fasta generation => 19,226 v1.4 MANE	RefFusion.py => RefFusion.v2.py (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
8-0-a	convert v5 probe regions to the regions on pseudo transcriptome (MANE v1.4)	1039 probe regions are converted to 533 mapped (probe) exons (on pseudo-transcriptome v1.4)	candidate_exons_mapping.sh (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
8-0-b	modify the header of the probe fasta file (replace with primer-like header)	Header conversion step for blastnparser.py & blastn result	Probe_faheader_converter.py (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
8-0-c	generate annotation with GSP information	perform primer sequence alignment	blastn (/tools/Fusion/ncbi-blast/bin/blastn -task blastn-short -dust no)
8-0-c	generate annotation with GSP information	parse the alignment information and add it into the annotation tables	blastnparser.py (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/)
9	generate index files for bwa	generate the 5 indices required for "bwa mem"	/tools/Fusion/bwa index

# Generated files summary (clean)

- Probe meta file (533 mapped (probe) exons)

Step 8-0-a

- /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/InhouseDB\_Probe/captureprobe\_250407\_MANE\_Select/fusionv4.MANE.v1.4.GENCODE.r47.candidate.exons.transcript.bed

- Transcriptome (with 10\*N)

- /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.fasta (FASTA file for the 19,226 selected transcripts)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.fasta.<bwa index suffix> (suffixes include "sa", "pac", "ann", "amb" and "bwt")

Step 5 (1)

Step 9

- Transcript Loci files

- /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.transcript.loci (Blank Loci file)
  - /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_Loci/250407/PA053\_ACTFusionV5\_PseudoIntron\_MANE-v1.4\_GENCODE-r47\_capture-v1.0\_GRCh38.20250407.transcript.MANE.only.blastn.default.loci (Loci file with 533 probes (mapped exons in v1.4) annotated)

Step 5 (1)

Step 8-0-c

- Genomic Loci file

- /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/Output\_MANE\_Select/20250407\_MANE.genome.loci

Step 5 (1)

# Updated pipeline input files (colored)

- /mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/v1.4\_inputfiles/

# 9 files for v1.4 MANE transcriptome

/mnt/RD\_Develop/sandyteng/ACTFusionV5/db\_fusionv5/v1.4\_inputfiles/

```
└─ 20250407_MANE.fasta
└─ 20250407_MANE.fasta.amb
└─ 20250407_MANE.fasta.ann
└─ 20250407_MANE.fasta.bwt
└─ 20250407_MANE.fasta.pac
└─ 20250407_MANE.fasta.sa
└─ 20250407_MANE.genome.loci
└─ 20250407_MANE.transcript.loci (Blank transcript loci file)
└─ fusionv4.MANE.v1.4.GENCODE.r47.candidate.exons.transcript.bed
└─ PA053_ACTFusionV5_PseudoIntron_MANE-v1.4_GENCODE-r47_capture-v1.0_GRCh38.20250407.transcript.MANE.only.blastn.default.loci
```

0 directories, 10 files

# Updated configuration file

- Configuration file (for db update test)

```
params {
  refFile = "/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/v1.4_inputfiles/20250407_MANE.fasta"
  ambFile = "/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/v1.4_inputfiles/20250407_MANE.fasta.amb"
  annFile = "/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/v1.4_inputfiles/20250407_MANE.fasta.ann"
  bwtFile = "/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/v1.4_inputfiles/20250407_MANE.fasta.bwt"
  pacFile = "/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/v1.4_inputfiles/20250407_MANE.fasta.pac"
  saFile = "/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/v1.4_inputfiles/20250407_MANE.fasta.sa"
  annoFile = "/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/v1.4_inputfiles/PA053_ACTFusionV5_Pseudointron_MANE-v1.4_GENCODE-r47_capture-v1.0_GRCh38.20250407.transcript.MANE.only.blastn.default.loci"
  gannoFile = "/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/v1.4_inputfiles/20250407_MANE.genome.loci"
  isoformfaFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.encode.v38.transcripts.protein.coding.22.isoforms.fa"
  isoformambFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.encode.v38.transcripts.protein.coding.22.isoforms.fa.amb"
  isoformannFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.encode.v38.transcripts.protein.coding.22.isoforms.fa.ann"
  isoformbwtFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.encode.v38.transcripts.protein.coding.22.isoforms.fa.bwt"
  isoformpacFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.encode.v38.transcripts.protein.coding.22.isoforms.fa.pac"
  isoformsaFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.encode.v38.transcripts.protein.coding.22.isoforms.fa.sa"
  isoformmetaFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/merged.refseq.20240807.encode.v38.transcripts.protein.coding.22.isoforms.fa.catted.meta"
  isoformfilteringflag = "1"
  truncatedmode = "0"
  truncatedseq_min_aligned_len = "12"
  inSpikeinFastqR1 = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/AAS502_953_IonXpress_001_1_MET_14_read-spike-in.fq.gz"
  pdbFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/protein_26db.v1.fasta"
  pdbmFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/protein_26db.v1.mapping.meta.txt"
  adapFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/torrentadapters.corrected.20230704.csv"
  qcconfigFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/filter_internal.QC9.0.mgsp.qcr.0.5.241206.config"
  readqcconfigFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/filter_fuscall_12aa_Oratio_read.QC.json"
  primerlabelFile = "/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/v1.4_inputfiles/fusionv4.MANE.v1.4.GENCODE.r47.candidate.exons.transcript.bed"
  incqctemplateFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/Blank_template.fusioncalling.QC.json"
  boundaryqcFile = "/mnt/Bishare/sandyteng/ACTFusionV4_Torrent/data/v9.0_inputfiles/boundaryqcthresholds.v4.default.config"
  fusion_container = "actgenomics/torrent_fusion_pipeline:v0.23.0"
}

process {
  withLabel: process_low {
    cpus = 1
    maxForks = 1
  }
  executor = 'local'
  container = 'actgenomics/torrent_fusion_pipeline:v0.23.0'
  cpus = '2'
  memory = '16.G'
}
```

# v1.4 annotation evaluation

- Result summary
- Manual inspection

# Fusionv4 test v0.1 (config test: MANE v0.95 → v1.4)

- 300x IVTRNA

```
sandyteng@tp-rdl77:/mnt/RD_Develop/sandyteng/ACTFusionV5/nextflow_outdir/20250414_ACTFusion_v4_81-IVTRNA-240906-300x_v0.1_fusionv4_MANE_v1.4_config_test$ ls -al
total 8
drwxrwxr-x 2 sandyteng sandyteng 4096 Apr 14 15:12 .
drwxrwxr-x 9 sandyteng sandyteng 4096 Apr 14 15:12 ..
sandyteng@tp-rdl77:/mnt/RD_Develop/sandyteng/ACTFusionV5/nextflow_outdir/20250414_ACTFusion_v4_81-IVTRNA-240906-300x_v0.1_fusionv4_MANE_v1.4_config_test$ nextflow run /mnt/RD_Develop/nextflow/Workflow_docker_illumina.nf \
-params-file /mnt/RD_Develop/sandyteng/ACTFusionV5/nextflow/repo_code_v1.4_dbtest_0414.2025/params/ACTFusion_v4_81-IVTRNA-240906-300x_v0.1_fusionv4_MANE_v1.4_test.json \
-c /mnt/RD_Develop/sandyteng/ACTFusionV5/nextflow/repo_code_v1.4_dbtest_0414.2025/dockerconfigs/fusion_multi_localdocker.v9.20241125.v0.23.0_v1.4.MANE.transcriptome.config
Nextflow 24.10.5 is available - Please consider updating your version to it

NEXTFLOW ~ version 24.04.4

Launching `/mnt/RD_Develop/sandyteng/ACTFusionV5/repo_dev/hybridcapture_fusion_pipeline_nextflow/Workflow_docker_illumina.nf` [modest_venter] DSL2 - revision: ecbd2c7c06

executor > local (8)
[39/989f3e] process > SampleInfo (ACTFusion_v4_81-IVTRNA-240906-300x_R1.fastq.gz ACTFusion_v4_81-IVTRNA-240906-300x_R2.fastq.gz) [100%] 1 of 1 ✓
[26/4be765] process > Fusion:mergefastq (ACTFusion_v4_81-IVTRNA-240906-300x ACTFusion_v4_81-IVTRNA-240906-300x_R1.fastq.gz) [100%] 1 of 1 ✓
[d0/c29fb0] process > Fusion:trimadap (ACTFusion_v4_81-IVTRNA-240906-300x ACTFusion_v4_81-IVTRNA-240906-300x.fastp.zcatted.merged.fastq.gz) [100%] 1 of 1 ✓
[69/15f2ce] process > Fusion:fastp (ACTFusion_v4_81-IVTRNA-240906-300x ACTFusion_v4_81-IVTRNA-240906-300x.fastp.zcatted.merged__trimadap.fq.gz) [100%] 1 of 1 ✓
[8a/965030] process > Fusion:bwa_isoform (ACTFusion_v4_81-IVTRNA-240906-300x ACTFusion_v4_81-IVTRNA-240906-300x.clean.fq.gz) [100%] 1 of 1 ✓
[2c/cl532c] process > Fusion:bwa_se (ACTFusion_v4_81-IVTRNA-240906-300x ACTFusion_v4_81-IVTRNA-240906-300x.isofilered.fq.gz) [100%] 1 of 1 ✓
[aa/fc331a] process > Fusion:bwa_fusioncalling (ACTFusion_v4_81-IVTRNA-240906-300x) [100%] 1 of 1 ✓
[6f/680b36] process > Fusion:fuscall2QC_illumina (ACTFusion_v4_81-IVTRNA-240906-300x) [100%] 1 of 1 ✓
Completed at: 14-Apr-2025 15:20:22
Duration : 5m 26s
CPU hours : 0.2
Succeeded : 8
```

```
nextflow run /mnt/RD_Develop/sandyteng/ACTFusionV5/repo_dev/hybridcapture_fusion_pipeline_nextflow/Workflow_docker_illumina.nf \
-params-file /mnt/RD_Develop/sandyteng/ACTFusionV5/nextflow/repo_code_v1.4_dbtest_0414.2025/params/ACTFusion_v4_81-IVTRNA-240906-300x_v0.1_fusionv4_MANE_v1.4_test.json \
-c /mnt/RD_Develop/sandyteng/ACTFusionV5/nextflow/repo_code_v1.4_dbtest_0414.2025/dockerconfigs/fusion_multi_localdocker.v9.20241125.v0.23.0_v1.4.MANE.transcriptome.config
```

# Result summary (v1.4 annotation evaluation)

- Test sample: ACTFusion\_v4\_81-IVTRNA-240906-300x
  - 300 replicates of the 81 IVT sequences
    - Summarized result
      - /mnt/RD\_Develop/sandyteng/ACTFusionV5/nextflow/repo\_code\_v1.4\_dbtest\_0414.2025/summary/
    - Raw result folder
      - /mnt/RD\_Develop/sandyteng/ACTFusionV4/nextflow\_outdir/20241121\_v0.20.0\_image\_v0.25.0\_pipeline\_test/ACTFusion\_v4\_81-IVTRNA-240906\_check/ACTFusion\_v4\_81-IVTRNA-240906/
- Forward primers successfully annotated (for the 23 identified fusions & 3 exon-skipping cases)
- 81-read inspection
  - 17 missing reads
    - Not recognized by “fuscall.py” program, but primary alignments of the 17 reads are found within the aligned.bwa.bam (**unaligned reads** => SAM bitwise FLAG == 4)
  - 64 reads processed by “ACTGfuscall.py”
    - 36 WILDTYPE (inclusive of AR-V7) (22 wild type reads with no primer detected)
    - 23 FUSION
    - 3 EXONSKIPPING (MET, EGFR (2), ~~BRAF~~)
    - 2 AMBIGUITY (FusionRef\_659, FusionRef\_689)
- Recall: 23/81 ~ 28.4%
  - (81-23)/81 ~ 2/3 of the IVT variants designed using v0.95 transcripts cannot be detected when migrating the transcriptome to v1.4 (which is reasonable since v1.4 and v0.95 included different preferred transcripts)

# Result (list of missing fusion reads)

- 17 missing reads

IVT-RNA ID	Boundary	Type	Group	Report status (v0.24.0)
FusionRef_077	CD74:6-ROS1:33	FUSION	1+	
FusionRef_640	AKT3:12-HEATR1:30	FUSION	1+	
FusionRef_680	SRGAP3:7-RAF1:8	FUSION	1+	
FusionRef_646	ERG:6-APP:14	FUSION	2+	
FusionRef_701	EZR:10-ROS1:35	FUSION	2+	
FusionRef_707	KIAA1549:15-BRAF:9	FUSION	2+	
FusionRef_706	KIAA1549:16-BRAF:9	FUSION	3+	
FusionRef_652	NRG1:7-PSAP:11	FUSION	4+	
FusionRef_699	CD74:6-ROS1:35	FUSION	4+	
FusionRef_704	TMPRSS2:2-ERG:2	FUSION	5+	
FusionRef_669	EZR:12-ERBB4:18	FUSION	6+	
FusionRef_709	FGFR2:17-SHTN1:7	FUSION	6+	
FusionRef_662	RSPO2:4-EIF3E:3	FUSION	7+	
FusionRef_683	EIF3E:2-RSPO2:4	FUSION	7+	
FusionRef_684	FKBP5:6-TMPRSS2:7	FUSION	7+	
FusionRef_677	SIL1:3-NRG2:6	FUSION	8+	



# Result (23 FUSION IVT reads)

- 23 Fusions (20 of which are equivalent to the boundaries reported via MANE v0.95)

Fusion Type	T_Type	G_Type	5-gsp_name	5-gsp_coverage	5-Transcript	5-orientation	5-Exon	3-gsp_name	3-gsp_coverage	3-Transcript	3-orientation	3-Exon	Boundary Check
FUSION	ENST00000292431.5:5-ENST00000277120.8:11	NACC1:5-NTRK2:11	-	-	-	-	-	Probe-mane376	0.574257	ENST00000277120.8	F	11	FALSE
FUSION	ENST00000263966.8:4-ENST00000263967.4:15	USP13:4-PIK3CA:15	-	-	-	-	-	Probe-mane118	0.495327	ENST00000263967.4	F	15	TRUE
FUSION	ENST00000240851.9:4-ENST00000524377.7:9	TFG:4-NTRK1:9	-	-	-	-	-	Probe-mane038	0.910714	ENST00000524377.7	F	10	TRUE
FUSION	ENST00000382051.8:4-ENST00000397752.8:15	SLC34A2:4-MET:15	-	-	-	-	-	Probe-mane309	0.251082	ENST00000397752.8	F	15	TRUE
FUSION	ENST00000269571.10:23-ENST00000447079.6:6	ERBB2:23-CDK12:6	Probe-mane479	0.408163265	ENST00000269571.10	F	23	-	-	-	-	-	TRUE
FUSION	ENST00000206249.8:3-ENST00000371998.8:15	ESR1:3-NCOA3:15	Probe-mane243	0.478632479	ENST00000206249.8	F	3	-	-	-	-	-	TRUE
FUSION	ENST00000288135.6:20-ENST00000257290.10:6	KIT:20-PDGFR:6	Probe-mane164	0.537735849	ENST00000288135.6	F	20	-	-	-	-	-	TRUE
FUSION	ENST00000397752.8:20-ENST00000358204.9:3	MET:20-TES:3	Probe-mane314	0.416058394	ENST00000397752.8	F	20	-	-	-	-	-	TRUE
FUSION	ENST00000524377.7:16-ENST00000643579.2:7	NTRK1:16-TPM4:7	Probe-mane044	0.352201258	ENST00000524377.7	F	16	-	-	-	-	-	FALSE
FUSION	ENST00000277120.8:18-ENST00000314328.14:8	NTRK2:18-SPECC1L:8	Probe-mane383	0.389937107	ENST00000277120.8	F	18	-	-	-	-	-	TRUE
FUSION	ENST00000338663.12:2-ENST00000405005.8:6	SLC3A2:2-NRG1:6	-	-	-	-	-	Probe-mane345	0.384615	ENST00000405005.8	F	6	TRUE
FUSION	ENST00000643579.2:7-ENST00000524377.7:10	TPM4:7-NTRK1:10	-	-	-	-	-	Probe-mane038	0.946429	ENST00000524377.7	F	10	FALSE
FUSION	ENST00000508803.6:5-ENST00000440486.8:10	NSD2:5-FGFR3:10	-	-	-	-	-	Probe-mane175	0.376712	ENST00000440486.8	F	10	TRUE
FUSION	ENST00000440486.8:17-ENST00000313288.9:8	FGFR3:17-TACC3:8	Probe-mane182	0.481132075	ENST00000440486.8	F	17	-	-	-	-	-	TRUE
FUSION	ENST00000545606.6:8-ENST00000275493.7:15	CAND1:8-EGFR:15	-	-	-	-	-	Probe-mane263	0.35443	ENST00000275493.7	F	15	TRUE
FUSION	ENST00000323571.9:5-ENST00000269571.10:13	WIPF2:5-ERBB2:13	-	-	-	-	-	Probe-mane469	0.571429	ENST00000269571.10	F	13	TRUE
FUSION	ENST00000355710.8:19-ENST00000163416.7:4	RET:19-GOLGA5:4	Probe-mane421	0.398648649	ENST00000355710.8	F	19	-	-	-	-	-	TRUE
FUSION	ENST00000440486.8:17-ENST00000313288.9:11	FGFR3:17-TACC3:11	Probe-mane182	0.481132075	ENST00000440486.8	F	17	-	-	-	-	-	TRUE
FUSION	ENST00000361752.8:6-ENST00000277120.8:14	QKI:6-NTRK2:14	-	-	-	-	-	Probe-mane379	0.301587	ENST00000277120.8	F	14	TRUE
FUSION	ENST00000440486.8:17-ENST00000313288.9:10	FGFR3:17-TACC3:10	Probe-mane182	0.481132075	ENST00000440486.8	F	17	-	-	-	-	-	TRUE
FUSION	ENST00000396373.9:4-ENST00000277120.8:14	ETV6:4-NTRK2:14	-	-	-	-	-	Probe-mane379	0.301587	ENST00000277120.8	F	14	TRUE
FUSION	ENST00000431975.7:5-ENST00000206249.8:5	AKAP7:5-ESR1:5	-	-	-	-	-	Probe-mane245	0.42446	ENST00000206249.8	F	5	TRUE
FUSION	ENST00000264312.12:8-ENST00000288135.6:8	OCTD1:8-KIT:8	-	-	-	-	-	Probe-mane152	0.469565	ENST00000288135.6	F	8	TRUE

# Result (FUSION IVT read)

- NTRK1:16-TPM3:8 (MANE v0.95) → NTRK1:16-TPM4:7 (MANE v1.4)

# MANE v0.95

```
FusionRef_654 FUSION ENST00000524377.7:16-ENST00000651641.1:8 NTRK1:16-TPM3:8 ENST00000524377.7:16-ENST00000651641.1:8,9 NTRK1:16-TPM3:8,9 1 56
TCTGGGAGATCTTCACCTACGGCAAGCAGCCCTGGTACCAGCTCTCCAACACGGAGGCAGAGACCCGTGCTGAGTTTGCTGAGAGATCGGTAGCCAAGCTGGAAAAGACAATTGATGACCTGGAAGATGAGC
TCTATGCCAGAACTGAAGTACAAGGCCATTAGCGAGGAGCTGGAC FT-mPCR0022 1 ENST00000524377.7 F 16 - - - - -
```

# MANE v1.4

```
FusionRef_654 FUSION ENST00000524377.7:16-ENST00000643579.2:7 NTRK1:16-TPM4:7 ENST00000524377.7:16-ENST00000643579.2:7 NTRK1:16-TPM4:7 1 57
TCTGGGAGATCTTCACCTACGGCAAGCAGCCCTGGTACCAGCTCTCCAACACGGAGGCAGAGACCCGTGCTGAGTTTGCTGAGAGATCGGTAGCCAAGCTGGAAAAGACAATTGATGACCTGGAAGATGAGC
TCTATGCCAGAACTGAAGTACAAGGCCATTAGCGAGGAGCTGGAC Probe-mane044 0.3522012578616352 ENST00000524377.7 F 16 - - - - -
```

# Genomic coordinates

# MANE v0.95

NTRK1(NM\_002529.4), chr1:156880157; TPM3(NM\_152263.4), chr1:154170469

# MANE v1.4

NTRK1(NM\_002529.4), chr1:156880157; TPM4(NM\_003290.3), chr19:16093684

# Result (3 EXONSKIPPING IVT reads)

- 3 (/ 5 EXONSKIPPING + ARV7) targets are reported
  - BRAF:19-BRAF:11 are not reported

IVT-RNA ID	Boundary	Type	Group	Report status (v0.24.0)	Fusion Type	G_Type	5-gsp_name	5-gsp_coverage	5-Transcript	5-orientation	5-Exon	3-gsp_name	3-gsp_coverage	3-Transcript	3-orientation	3-Exon
ARV7	AR:2,3,4	WILDTYPE	9+		WILDTYPE	AR:2	Probe-mane527	0.151315789	ENST00000374690.9	F	2	Probe-mane528	1	ENST00000374690.9	F	3
FusionRef_685	MET:13-MET:15	EXONSKIPPING	9+		EXONSKIPPING	MET:13-MET:15	Probe-mane307	0.394904459	ENST00000397752.8	F	13	Probe-mane309	0.251082251	ENST00000397752.8	F	15
FusionRef_686	EGFR:1-EGFR:8	EXONSKIPPING	9+		EXONSKIPPING	EGFR:1-EGFR:8	Probe-mane249	0.166189112	ENST00000275493.7	F	1	Probe-mane256	0.452991453	ENST00000275493.7	F	8
FusionRef_687	BRAF:19-BRAF:11	EXONSKIPPING	9+													
FusionRef_688	EGFR:24-EGFR:18	EXONSKIPPING	9+		EXONSKIPPING	EGFR:24-EGFR:18	Probe-mane272	0.642857143	ENST00000275493.7	F	24	Probe-mane266	0.504065041	ENST00000275493.7	F	18

# Result (WILDTYPE inspection, FusionRef\_665)

- Fusion V4 (MANE v0.95) & Arriba calling results

FusionRef\_665 FUSION **ENST00000307546.14:9-ENST00000673466.1:3** **MAGI3:9-AKT3:3** ENST00000307546.14:9-ENST00000673466.1:3 MAGI3:9-AKT3:3 1 141  
GGGTTGTTTCGAGCATCACTGAAAAAAGCACAAATGGGATTTGGTTTTACTATTATTGGTGGAGATAGACCTGATGAGTTCCTACAAGTGAAAAATGTGCTGAAAGATGGTCCCGCAGCTCAGGATGGGAAAATTGC  
ACCAGGAGAATATATAAAAAACTGGAGGCCAAGATACTTCCTTTGAAGACAGATGGCTCATTCATAGGATA - - - - - FT-mPCR0772 1 ENST00000673466.1 R 3

NM IDs reported by Arriba for the same boundary

MAGI3(NM_001142782)	AKT3(NM_181690)
---------------------	-----------------

Boundary;(5' gene coordinate,3' gene coordinate)	Boundary	Type	Group	IVT-RNA ID	Report status (v0.24.0)	5' NM ID	3' NM ID	5' gene coordinate	3' gene coordinate
MAGI3-AKT3;(chr1:113622994,chr1:243695716)	<b>MAGI3:9-AKT3:3</b>	FUSION	6	FusionRef_665	+	MAGI3(NM_001142782.2)	AKT3(NM_005465.7)	chr1:113622994	chr1:243695716

# Result (WILDTYPE inspection, FusionRef\_665)

- V5 (MANE v1.4) Result / blastn result

## Note:

- Different sets of preferred transcriptome => different alignment result
- If we includes multiple transcripts for the same gene, it may result in ambiguity in terms of the alignment result and the corresponding NM ID(s) to report.

# Calling result

FusionRef\_665 WILDTYPE

ENST00000307546.14:9 MAGI3:9 1 -

GGGTTGTTCGAGCATCACTGAAAAAAGCACAATGGGATTTGGTTTTACTATTATTGGTGGAGATAGACCTGATGAGTTCCTACAAGTGAAAAATGTGCTGAAAGATGGTCCCGCAGCTCAGGATG

GGAAAATTGCACCAGGAGAATATATAAAAACTGGAGGCCAAGATACTTCCTTTGAAGACAGATGGCTCATTCATAGGATA - - - - -

# MANE (RefSeq – Ensembl id map)

ENST00000307546.14 ENSG00000081026.19 MAGI3 NM\_001142782.2 # the same id as the one in fusion v4 (MANE v0.95)

ENST00000673466.1 ENSG00000117020.19 AKT3 NM\_005465.7 # the same id as the one in fusion v4 (MANE v0.95)

Boundary;(5' gene coordinate,3' gene coordinate)	Boundary	Type	Group	IVT-RNA ID	Report status (v0.24.0)	5' NM ID	3' NM ID	5' gene coordinate	3' gene coordinate
MAGI3-AKT3;(chr1:113622994,chr1:243695716)	MAGI3:9-AKT3:3	FUSION	6	FusionRef_665	+	MAGI3(NM_001142782.2)	AKT3(NM_005465.7)	chr1:113622994	chr1:243695716

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
✓	<a href="#">Homo sapiens membrane associated guanylate kinase, WW and PDZ domain containing 3 (MAGI3), tran...</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	7559	<a href="#">NM_152900.3</a>
✓	<a href="#">Homo sapiens membrane-associated guanylate kinase-related MAGI-3 (MAGI-3) mRNA, complete cds</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	3504	<a href="#">AF213259.1</a>
✓	<a href="#">PREDICTED: Homo sapiens membrane associated guanylate kinase, WW and PDZ domain containing 3...</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	6965	<a href="#">XM_054335878.1</a>
✓	<a href="#">PREDICTED: Homo sapiens membrane associated guanylate kinase, WW and PDZ domain containing 3...</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	6965	<a href="#">XM_005270737.4</a>
✓	<a href="#">PREDICTED: Homo sapiens membrane associated guanylate kinase, WW and PDZ domain containing 3...</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	3473	<a href="#">XM_054335879.1</a>
✓	<a href="#">Homo sapiens membrane-associated guanylate kinase MAGI3 mRNA, complete cds</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	4288	<a href="#">AF257238.1</a>
✓	<a href="#">PREDICTED: Homo sapiens membrane associated guanylate kinase, WW and PDZ domain containing 3...</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	6993	<a href="#">XM_054335877.1</a>
✓	<a href="#">Human DNA sequence from clone RP11-473L1 on chromosome 1, complete sequence</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	71464	<a href="#">AL390759.10</a>
✓	<a href="#">Homo sapiens membrane associated guanylate kinase, WW and PDZ domain containing 3 mRNA (cDNA...</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	3580	<a href="#">BC130409.1</a>
✓	<a href="#">PREDICTED: Homo sapiens membrane associated guanylate kinase, WW and PDZ domain containing 3...</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	6993	<a href="#">XM_017000974.2</a>
✓	<a href="#">Homo sapiens cDNA clone IMAGE:9052664, containing frame-shift errors</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	3653	<a href="#">BC144144.1</a>
✓	<a href="#">PREDICTED: Homo sapiens membrane associated guanylate kinase, WW and PDZ domain containing 3...</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	3473	<a href="#">XM_047417371.1</a>
✓	<a href="#">Homo sapiens membrane associated guanylate kinase, WW and PDZ domain containing 3 (MAGI3), tran...</a>	<a href="#">Homo sapiens</a>	257	257	67%	4e-64	100.00%	6874	<a href="#">NM_001142782.2</a>