Data constrution for fusion v5

- MANE v1.4
  - 19, 226 transcripts (filtered via “filter_mane_gff.py”)
    Inclusion criteria:
    chr1-22, X, Y, protein_coding, MANE_Select (summary.txt => MANE Select; manually curated)
- GENCODE-r47
  - only use its FASTA source file => extract defined transcripts in MANE v1.4 gff via “bedtools getfasta”

The fusion v5 db is derived from the following downloaded files:

1. (MANE v1.4)
   MANE v1.4 DB
   (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***)
   (/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/OpenDB_MANE_human_v1.4/release_1.4/***MANE.GRCh38.v1.4.ensembl_genomic.gff.gz***)

2. (Genome sequence, Grch38, GENCODE-r47)
   GRCh38.p14.genome.fa.gz
   Gencode V47
   (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***)

3. (Probe information file provided by AD team)
   → 1,039 probe regions bed file (genomic locations of the designed/target region

   /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/captureprobe_250401/***ACTFusionv5_target-region_PartAB_individual_1039.bed*** (obtained from AD team Lucy)

<19,292-transcript sequence extraction>

The fasta sequence of the 19,292 transcripts are obtained via the following steps:

1. gff lines extraction (“filter_mane_gff.py”)
   ```
   python3 /mnt/RD_Develop/sandyteng/ACTFusionV5/code/filter_mane_gff.py \
   -i MANE.GRCh38.v1.4.ensembl_genomic.gff.gz \
   -o MANE.GRCh38.v1.4.ensembl_genomic.transcript.gff
   ```
2. gff to bed file conversion (“/tools/Fusion/convert2bed”)
   ```
   /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. bed to fasta file conversion (“bedtools” in image actgenomics/fusion_dev:v0.6)
   ```
   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.corrected.strand.fasta
   ```
4. Additional (-) or (+) strings (within the fasta file) removal
   ```
   sed -i
   's/([+-])//g' ./mane_v1.4/OpenDB_MANE_human_v1.4/derived/MANE.GRCh38.v1.4.ensembl_genomic.transcript.corrected.strand.fasta
   ```

Note: The gff extraction program “filter_mane_gff.py” extracts 19,292 protein_coding transcripts located on chr1-22, X, Y from “***MANE.GRCh38.v1.4.ensembl_genomic.gff.gz***”

<Target transcript selection (ENST ID – RefSeq ID map)>

Among the 19,292 extracted transcripts, 66 transcripts are not labeled “MANE Select” in the “***MANE.GRCh38.v1.4.summary.txt.gz***” file (column: MANE_status). The 66 transcripts are labeled “MANE Plus Clinical” instead. To avoid mapping ambiguity, we only include the 19,226 (=19,292 - 66) transcripts labeled “MANE Select”.

Note:

There are total 19,338 transcripts labeled “MANE Select” and 66 transcripts labeled “MANE Plus Clinical” in the *summary.txt.gz file

The namemap files for the 19,404 (19,338 + 66) transcripts:
/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***

<Empty pseudo intron annotation (loci) table + pseudo N (10N) fasta generation>

The pseudo intron sequences and the corresponding annotation (loci) tables (transcriptome + genome) for the 19,226 transcripts were generated via “RefFusion.v2.py”.

```
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.corrected.strand.fasta \
-p
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/mane_v1.4/Output_Final/PA053_ACTFusionV5_PseudoIntron_MANE-v1.4_GENCODE-r47_capture-v1.0_GRCh38.20250407.transcript.MANE.only.list \
-o
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.r47
```

<Probe anchored (mapping) exons extraction>

The 1,039 probe regions are converted to 533 exons located on the 19,226 extracted transcripts via “candidate_exons_mapping.sh”.

The information of the 533 extracted exons:

- ***fusionv4.MANE.v1.4.GENCODE.r47.candidate.exons.transcript.bed***
- /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/***MANE.GRCh38.v1.4.0407.probe.r47.fasta***

For the fasta header for each probe sequence are converted via “Probe_faheader_converter.py” and decompressed via “gunzip”:

- /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/probeseq/***MANE.GRCh38.v1.4.0407.r47.probe.wtprimerlikeheader.fasta***

```
# sample command (prerequisite: “actgenomics/fusion_dev:v0.6”)
docker run --rm -v /mnt/:/mnt/ -it actgenomics/fusion_dev:v0.6
bash ./candidate_exons_mapping.sh \
        /path/to/genome.loci \
        /path/to/transcript.loci \
        /path/to/namemap \
        /path/to/probe.bed \
        fusionv4.MANE.v0.95.GENCODE.r38 \
        /mnt/RD_Develop/sandyteng/workdir \
        /tools/Fusion
```

```
# 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.r47.genome.loci
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.r47.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.GENCODE.r47
```

```
/mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/ /tools/Fusion
# 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.r47.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.r47.fasta

# 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.r47.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.r47.probe.wtprimerlikeheader.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.r47.probe.wtprimerlikeheader.fasta.gz
```

< Probe/Exon (query) to Pseudo-intron Transcript (subject) alignment via blastn >

To annotate the 533 mapped exons to the selected transcriptome, the raw probe fasta file (***MANE.GRCh38.v1.4.0407.probe.r47.fasta***) is converted to a fasta file (/***MANE.GRCh38.v1.4.0407.r47.probe.wtprimerlikeheader.fasta***) of the following format:

```
>Probe ID|Gene Name|RefSeq ID|ENST ID|exon number|F|probe length
Probe sequence (mapped exon sequence)
```

(e.g.,

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

)

The probe sequences (fasta file) were then aligned against the 19,226 pseudo-intron (10N) containing fast file (***20250407_MANE.r47.fasta***) using the following command:

```
/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.r47.probe.wtprimerlikeheader.fasta -subject /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/Output_MANE_Select/20250407_MANE.r47.fasta -outfmt 6 -task blastn-short -dust no > /mnt/RD_Develop/sandyteng/ACTFusionV5/db_fusionv5/InhouseDB_Probe/captureprobe_250407_MANE_Select/blastn/20250407_probe.r47.blastn
```

After obtaining the “forward” probe alignment result, the “reverse” probe result is manually constructed and combined with the “forward” probe result via the following commands:

```
# create blastn result for “reverse probe” and concatenate all the alignments
cat 20250407_probe.r47.blastn > 20250407_rprobe.r47.blastn
sed -i 's/|F|/|R|/' 20250407_rprobe.r47.blastn
sed -i 's/mane/rmane/' 20250407_rprobe.r47.blastn
cat 20250407_probe.r38.blastn 20250407_rprobe.r38.blastn > 20250407_probe.rprobe.r47.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.rprobe.r47.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.r47.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.r47.loci
```