Fusion V5 db-v3.1 summary

Bioinformatics Development

Sandy

2025.05.08

Documentations

- Fusion db_prep.steps_db-v3.1.docx (Detailed description/overview for db v3.1)
- Fusion db_ref_transcript_v5_draft.docx (Executed commands for db v3.1 steps 1-9)
- Configuration file v3.1
    - /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.v3-1.config
- Executed runs
    - IVTALL 300x sample (JIRA issue: ABIE-1012)
    - AANB01_504 16 samples (JIRA issue: ABIE-1021)

ACT GENOMICS ™

Steps 1-4: Source data preparation

- download required data set from GENCODE and MANE (wget, rsync, zcat, samtools)
    - Input:
        - MANE.GRCh38.v1.4.summary.txt.gz, MANE.GRCh38.v1.4.ensembl_genomic.gff.gz
        - GRCh38.p14.genome.fa.gz
    - Output
        - MANE.GRCh38.v1.4.summary.txt, GRCh38.p14.genome.fa
- generate namemap file manually (awk, cat)
    - Input: MANE.GRCh38.v1.4.summary.txt.gz
    - Output: MANE.GRCh38.v1.4.select.and.plus.clinical.namemap
- retrieve transcript gff file (zgrep, awk, filter_mane_gff.py)
    - Input: MANE.GRCh38.v1.4.ensembl_genomic.gff.gz
    - Output: MANE.GRCh38.v1.4.ensembl_genomic.transcript.gff
- gff to bed file conversion (covert2bed)
    - Input: MANE.GRCh38.v1.4.ensembl_genomic.transcript.gff
    - Output: MANE.GRCh38.v1.4.ensembl_genomic.transcript.bed
- obtain fasta file (bedtools getfasta)
    - Input: MANE.GRCh38.v1.4.ensembl_genomic.transcript.bed
    - Output: MANE.GRCh38.v1.4.ensembl_genomic.transcript.corrected.strand.fasta

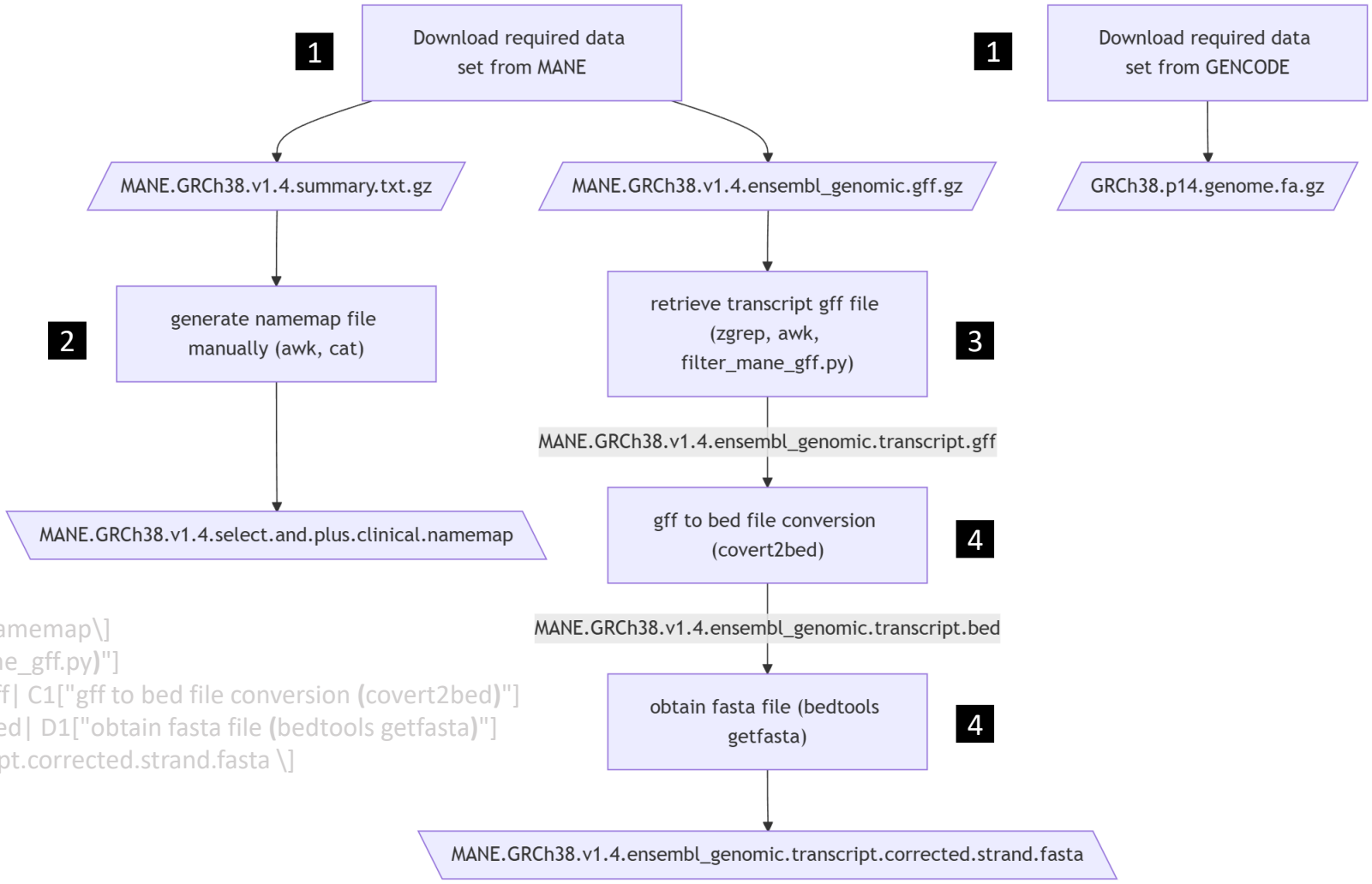
ACT GENOMICS ™

Steps 1-4: Source data preparation

- Code
- Plot

```
graph TD;
  %% Initial Inputs
  I1[/MANE.GRCh38.v1.4.ensembl_genomic.gff.gz/]
  I2[/MANE.GRCh38.v1.4.summary.txt.gz/]
  I3[/GRCh38.p14.genome.fa.gz/]
  %%I4[/probe.bed/]

  %% Source data preparation (steps 1-4)
  S1["Download required data set from MANE"] --> I1
  S1 --> I2
  S2["Download required data set from GENCODE"] --> I3
  I2 --> A1["generate namemap file manually (awk, cat)"]
  A1 --> MAP[\MANE.GRCh38.v1.4.select.and.plus.clinical.namemap\]
  I1 --> B1["retrieve transcript gff file (zgrep, awk, filter_mane_gff.py)"]
  B1 -->|MANE.GRCh38.v1.4.ensembl_genomic.transcript.gff| C1["gff to bed file conversion (covert2bed)"]
  C1 -->|MANE.GRCh38.v1.4.ensembl_genomic.transcript.bed| D1["obtain fasta file (bedtools getfasta)"]
  D1 --> O1[\MANE.GRCh38.v1.4.ensembl_genomic.transcript.corrected.strand.fasta \]
```

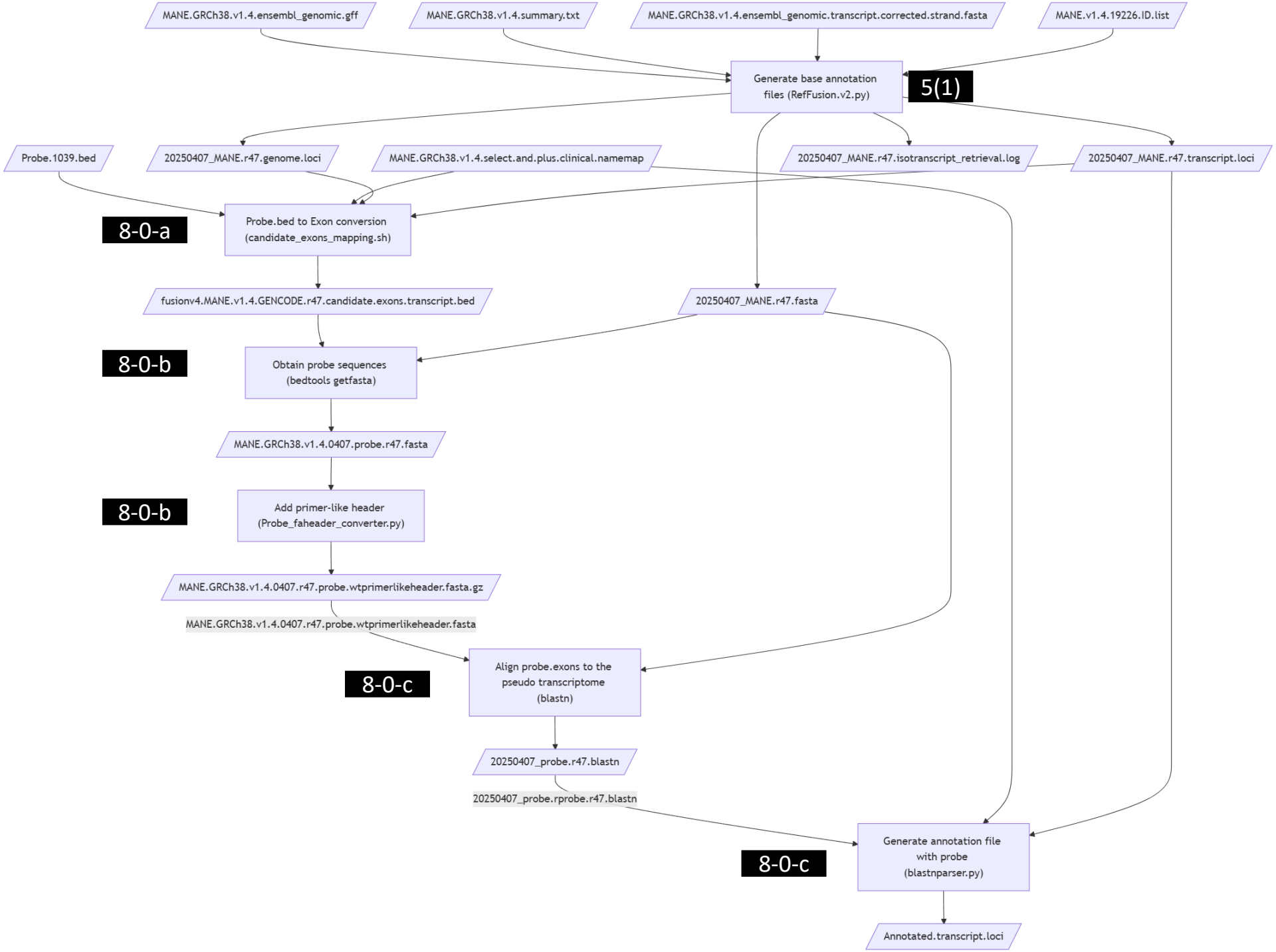


ACT GENOMICS ™

Steps 5 & 8: Annotation files generation

- generate annotation file via in-house script (RefFusion.v2.py)
  - Input:
    MANE.GRCh38.v1.4.ensembl_genomic.gf,
    MANE.GRCh38.v1.4.summary.txt,
    MANE.GRCh38.v1.4.ensembl_genomic.transcript.corrected.strand.fasta,
    PA053_ACTFusionV5_PseudoIntron_MANE-v1.4_GENCODE-r47_capture-v1.0_GRCh38.20250407.transcript.MANE.only.list
  - Output:
    20250407_MANE.r47.* (*= .genome.loci, .transcript.loci, .fasta, .isotranscript_retrieval.log)
- convert v5 probe regions to the regions on pseudo transcriptome (MANE v1.4) (candidate_exons_mapping.sh)
  - Input:
    ACTFusionv5_target-region_PartAB_individual_1039.bed,
    20250407_MANE.r47.genome.loci,
    20250407_MANE.r47.transcript.loci,
    MANE.GRCh38.v1.4.select.and.plus.clinical.namemap
  - Output:
    fusionv4.MANE.v1.4.GENCODE.r47.candidate.exons.transcript.bed
- obtain probe sequence (bedtools getfasta)
  - Input: 20250407_MANE.r47.fasta, fusionv4.MANE.v1.4.GENCODE.r47.candidate.exons.transcript.bed
  - Output:
    MANE.GRCh38.v1.4.0407.probe.r47.fasta
- modify the header of the probe fasta file (replace with primer-like header) (Probe_faheader_converter.py)
  - Input: MANE.GRCh38.v1.4.0407.probe.r47.fasta, MANE.GRCh38.v1.4.select.and.plus.clinical.namemap
  - Output: MANE.GRCh38.v1.4.0407.r47.probe.wtprimerlikeheader.fasta.gz (manually decompress to *.fasta file)
- align probe fasta file to the pseudo transcriptome (blastn)
  - Input: MANE.GRCh38.v1.4.0407.r47.probe.wtprimerlikeheader.fasta (query), 20250407_MANE.r47.fasta (subject),
  - Output: 20250407_probe.r47.blastn
- generate annotation file with GSP information (blastnparser.py)
  - Input: (cat 20250407_probe.r38.blastn 20250407_rprobe.r38.blastn =>) 20250407_probe.rprobe.r47.blastn, MANE.GRCh38.v1.4.select.and.plus.clinical.namemap, 20250407_MANE.r47.transcript.loci
  - Output: PA053_ACTFusionV5_PseudoIntron_MANE-v1.4_GENCODE-r47_capture-v1.0_GRCh38.20250407.transcript.MANE.only.blastn.r47.loci

ACT GENOMICS ™


MANE.GRCh38.v1.4.ensembl_genomic.gff
MANE.GRCh38.v1.4.summary.txt
MANE.GRCh38.v1.4.ensembl_genomic.transcript.corrected.strand.fasta
MANE.v1.4.19226.ID.list
Generate base annotation files (RefFusion.v2.py)
5(1)
Probe.1039.bed
20250407_MANE.r47.genome.loci
MANE.GRCh38.v1.4.select.and.plus.clinical.namemap
20250407_MANE.r47.isotranscript_retrieval.log
20250407_MANE.r47.transcript.loci
8-0-a
Probe.bed to Exon conversion (candidate_exons_mapping.sh)
fusionv4.MANE.v1.4.GENCODE.r47.candidate.exons.transcript.bed
20250407_MANE.r47.fasta
8-0-b
Obtain probe sequences (bedtools getfasta)
MANE.GRCh38.v1.4.0407.probe.r47.fasta
8-0-b
Add primer-like header (Probe_faheader_converter.py)
MANE.GRCh38.v1.4.0407.r47.probe.wtprimerlikeheader.fasta.gz
MANE.GRCh38.v1.4.0407.r47.probe.wtprimerlikeheader.fasta
8-0-c
Align probe.exons to the pseudo transcriptome (blastn)
20250407_probe.r47.blastn
20250407_probe.rprobe.r47.blastn
8-0-c
Generate annotation file with probe (blastnparser.py)
Annotated.transcript.loci

Steps 9-11: Input configuration files update

- generate index files for bwa (bwa index)
  - Input: 20250407_MANE.r47.fasta
  - Output:
    20250407_MANE.r47.* (*= .amb, .ann, .bwt, .pac, .sa)

- update kinase files (pdb*File)
  - Input:
    sequences & IDs obtained from “UniProt Website”
  - Output:
    protein.26.v1.4.kinase.fasta
    protein.26.v1.4.kinase.meta.txt

- update whitelist (gsp pair => probe pair) (Get_shifted_boundary.py, update_qcconfig_with_tsv.py)
  - Input:
    filter_internal.QC9.0.mgsp.qcr.0.5.blank.config, gsppairs_inclusion_v1.4.txt
  - Output:
    filter_internal.QC9.0.mgsp.qcr.0.5-dbv3.v1.4.config

ACT GENOMICS ™

Steps 9-11: Input configuration files update