DB preparation steps (db-v3.1)

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	convert2bed (alternative way: bedtools) (/tools/Fusion/convert2bed)

		information in gff to bed for transcript region extraction)	
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/)
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
10	update kinase files (pdb*File): 1. Generate ENST ID – Gene – UniProt ID map 2. Obtain the corresponding sequences	manually retrieved from UniProt 26 MANE Select (v1.4) target transcripts	manually curated ensure the ENST ID – Gene – UniProt ID is included in the transcriptome (protein sequences are manually queried via ensemble website)
11	update whitelist (gsp pair => probe pair)	Include potential probe pair for the following variants: BRAF:1-BRAF:9 BRAF:1-BRAF:12 BRAF:3-BRAF:9 BRAF:19-BRAF:11 EGFR:1-EGFR:8	Get_shifted_boundary.py (/mnt/RD_Develop/sandyteng/ACTFusionV5/code/) update_qcconfig_with_tsv.py (/mnt/BI3/Team_workdir/sandyteng_workdir/ACTFusionV4_Torrent/code/)

		EGFR:24-EGFR:18 EGFR:25-EGFR:18 EGFR:26-EGFR:18 MET:13-MET:15	

Folders for intermediate files:

Step 10:

- /mnt/RD_Develop/sandyteng/ACTFusionV5/test/20250416_kinasedb_v1.4/

Step 11:

- /mnt/RD_Develop/sandyteng/ACTFusionV5/test/20250422_fusionread_generator/ # target splicing reads generation
  # Splicing variants
  - /mnt/RD_Develop/sandyteng/ACTFusionV5/test/20250422_fusionread_generator/testfiles/splicingvariants.shifted.2.exons.v1.4.fastq.gz
  - /mnt/RD_Develop/sandyteng/ACTFusionV5/test/20250422_fusionread_generator/testfiles/splicingvariants.shifted.2.exons.v1.4.R2.fastq.gz
- /mnt/RD_Develop/sandyteng/ACTFusionV5/test/20250423_fusionv42v5_whitelist_gsppair/ # configuration update
  # Variant pair inclusion list
  - /mnt/RD_Develop/sandyteng/ACTFusionV5/test/20250423_fusionv42v5_whitelist_gsppair/data/gsppairs_inclusion_v1.4.txt
  # Blank & updated configuration file
  - /mnt/RD_Develop/sandyteng/ACTFusionV5/test/20250423_fusionv42v5_whitelist_gsppair/testconfigs/filter_internal.QC9.0.mgsp.qcr.0.5.blank.config
  - /mnt/RD_Develop/sandyteng/ACTFusionV5/test/20250423_fusionv42v5_whitelist_gsppair/testconfigs/filter_internal.QC9.0.mgsp.qcr.0.5-dbv3.v1.4.config

Updated files (db v3.1)

Parameter in Config ("params")	Description	File/parameter to update	Note	**v3.1**
refFile	Fasta of the preferred transcripts (pseudo 10*N => intron) (MANE v0.95 + GENCODE-r38)	*		*** (MANE v1.4 + GENCODE-r47)**
ambFile	bwa index file (.amb file) derived from refFile	*		*
annFile	bwa index file (.ann file) derived from refFile	*		*
bwtFile	bwa index file (.bwt file) derived from refFile	*		*
pacFile	bwa index file (.pac file) derived from refFile	*		*
saFile	bwa index file (.sa file) derived from refFile	*		*
annoFile	The annotation file derived from refFile and primerlabelFile	*		*
gannoFile	The annotation file derived from GENCODE-r38	*		*
isoformfaFile	Fasta of the isoforms of the 26 target genes (RefSeq + GENCODE-r38)		(same target genes => no need to update)	
isoformambFile	bwa index file (.amb file) derived from isoformfaFile		(same target genes => no need to update)	
isoformannFile	bwa index file (.ann file) derived from isoformfaFile		(same target genes => no need to update)	
isoformbwtFile	bwa index file (.bwt file) derived from isoformfaFile		(same target genes => no need to update)	
isoformpacFile	bwa index file (.pac file) derived from		(same target genes => no	

	isoformfaFile		need to update)	
isoformsaFile	bwa index file (.sa file) derived from isoformfaFile		(same target genes => no need to update)	
isoformmetaFile	The annotation file derived from isoformfaFile		(same target genes => no need to update)	
isoformfilteringflag	isoform filtering step switch (1: enable filtering, 0: disable filtering) (process "bwaisoform")		(same target genes => no need to update)	
truncatedmode	truncated mode for funcational count summary (process "fuscall2QC")			
truncatedseq_min_aligned_len	minimum aligned length for the truncated sequence (default = 12 a.a.) (process "fuscall2QC")			
inSpikeinFastqR1	spike-in sequence to prevent pipeline termination (process "mergefastq"/"bam2fastq")			
pdbFile	26 protein sequences (fasta) of the corresponding target transcripts	*		*
pdbmFile	The ENST ID to UniProt ID map for pdbFile	*	(manually queried via ensembl website)	*
adapFile	adapter sequence to trim (for universal primer removal) (process "trimadap")	* (TBC)	may not affect calling result	
qcconfigFile	Adjustable QC settings (default settings designed for amplicon based assay)	* (TBC)	may need to adjust qc values for hybrid capture	*** (rebuild white list using MANE v1.4**

			assay	**transcripts)**
readqcconfigFile	Adjustable QC settings (default settings designed for amplicon based assay)	* (TBC)	may need to adjust qc values for hybrid capture assay	
primerlabelFile	the designed primer region and the corresponding meta data	*	need to replace (design probe for target exons)	*
incqctemplateFile	Adjustable QC settings (default settings designed for amplicon based assay)	* (TBC)	may need to adjust qc values for hybrid capture assay	
boundaryqcFile	Adjustable QC settings (default settings designed for amplicon based assay)	* (TBC)	may need to adjust qc values for hybrid capture assay	
fusion_container	latest pipeline image (actgenomics/torrent_fusion_pipeline:v0.23.0 for pipeline v0.29.0)	* (TBC)	(may need to update if we include additional tools)	