Bulk Isoseq pipeline

Wenchao Zhang

07/09/2025

Pacbio recommended bulk Isoseq workflow- version1

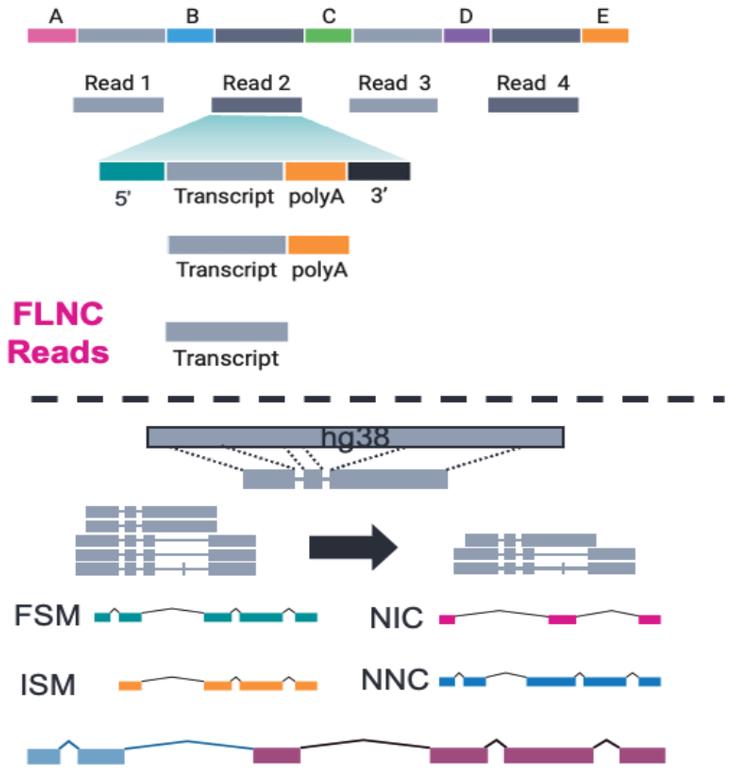
https://isoseq.how/getting-started.html

Recommended bulk Iso-Seq workflow

Command	Description	Output format
lima	Remove cDNA primers	`fl.bam`
isoseq refine	Remove polyA tail and artificial concatemers	`flnc.bam`
isoseq cluster2	*De novo* isoform-level clustering scalable to large number of reads (e.g. 40-100M FLNC reads)	`clustered.bam`
pbmm2	Align to the genome	`mapped.bam`
isoseq collapse	Collapse redundant transcripts based on exonic structures	`collapsed.gff`
pigeon classify	Classify transcripts against annotation	GFF and TXT files
pigeon filter	Filter transcripts for potential artifacts	GFF and TXT files

Pacbio recommended bulk Isoseq workflow- version2

https://github.com/RhettRautsaw/StJude_PacBio-WDL-tutorial/tree/main/Kinnex_IsoSeq_Pipelines



skera — READ SEGMENTATION

lima — DEMULTIPLEX / REMOVE PRIMERS

isoseq refine — REMOVE POLY-A & ARTIFACTS

pbmm2 — MAP ISOFORMS

IsoQuant* — COLLAPSE ISOFORMS & QUANTIFY

pigeon (classify, report)* — CLASSIFY ISOFORMS

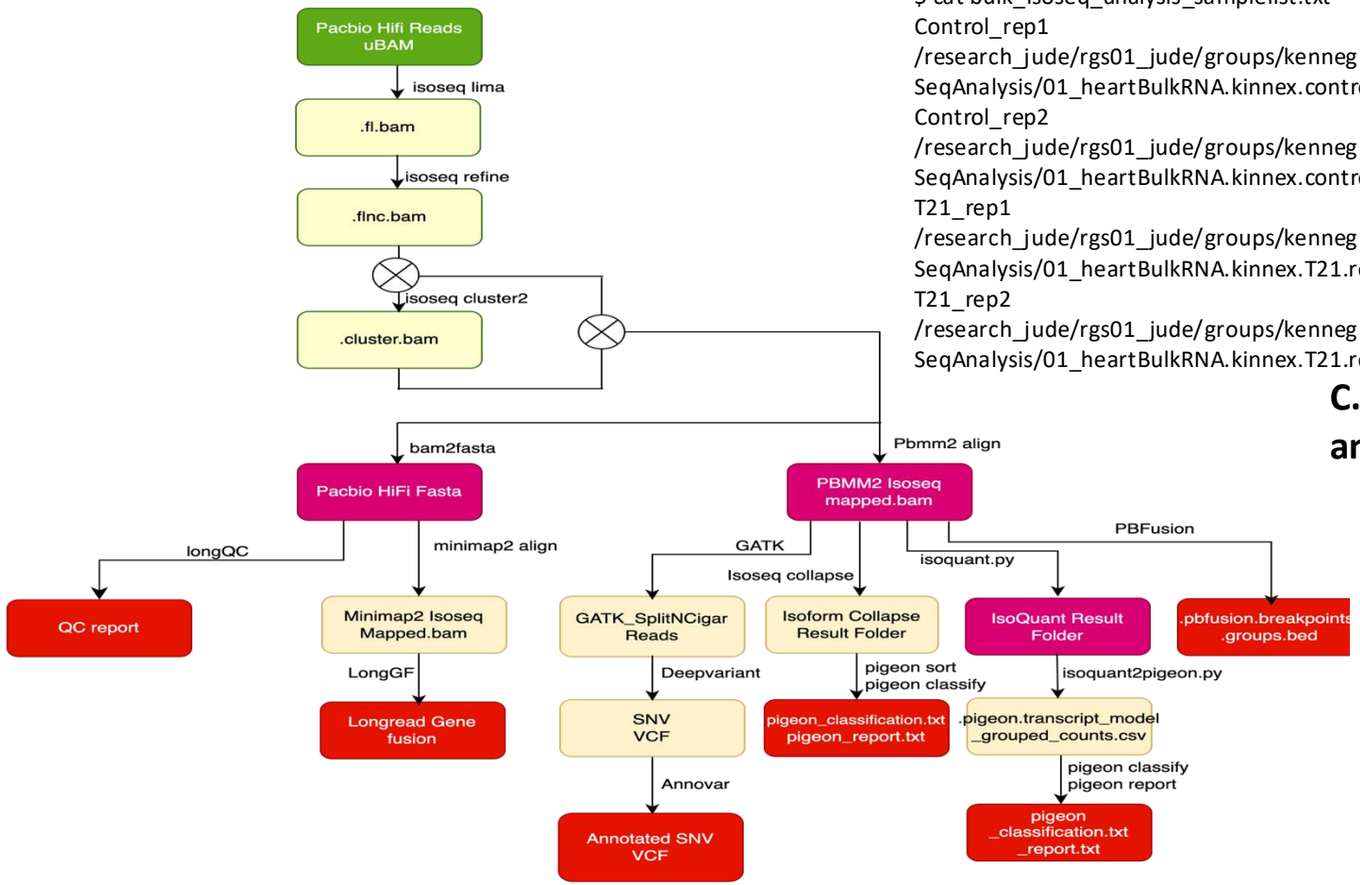
pbfusion — FUSION DETECTION

DESeq2 — DIFFERENTIAL EXPRESSION

Third-party tools (e.g., tappAS)

Current architecture of nextflow workflow

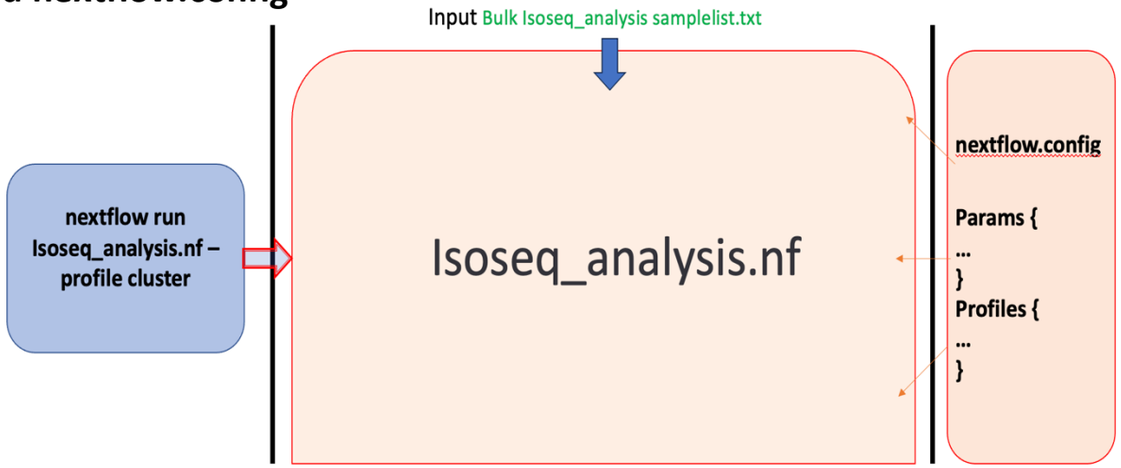
A.) Architecture of Nextflow pipeline



B.) Example Samplesheet

$ cat bulk_Isoseq_analysis_samplelist.txt
Control_rep1
/research_jude/rgs01_jude/groups/kennegrp/projects/kennegrp_cab/common/Pacbio_Isoseq/Shaohua_HiFi_Isoseq_New/pacobio_isoquant/KinnexBulkIsoSeqAnalysis/01_heartBulkRNA.kinnex.control.rep1.segmented.bam
Control_rep2
/research_jude/rgs01_jude/groups/kennegrp/projects/kennegrp_cab/common/Pacbio_Isoseq/Shaohua_HiFi_Isoseq_New/pacobio_isoquant/KinnexBulkIsoSeqAnalysis/01_heartBulkRNA.kinnex.control.rep2.segmented.bam
T21_rep1
/research_jude/rgs01_jude/groups/kennegrp/projects/kennegrp_cab/common/Pacbio_Isoseq/Shaohua_HiFi_Isoseq_New/pacobio_isoquant/KinnexBulkIsoSeqAnalysis/01_heartBulkRNA.kinnex.T21.rep1.segmented.bam
T21_rep2
/research_jude/rgs01_jude/groups/kennegrp/projects/kennegrp_cab/common/Pacbio_Isoseq/Shaohua_HiFi_Isoseq_New/pacobio_isoquant/KinnexBulkIsoSeqAnalysis/01_heartBulkRNA.kinnex.T21.rep2.segmented.bam

C.) Relationship among three files: Isoseq_analysis.nf, samplelist.txt and nextflow.config



D.) CMD to run the Nextflow pipeline

bsub -P Bulk_Isoseq -q standard -M 20000 -e err%J.err -o out%J.out -J Bulk_Isoseq "module load nextflow/21.10.5 && nextflow run Isoseq_analysis.nf -profile cluster -with-report Example_test_Nextflow_Run_Report.html -with-dag example_test_flowchart.png"

A.) Display of the configuration of Nextflow pipeline

```
Loading nextflow/21.10.5
  Loading requirement: java/17.0.1
N E X T F L O W  ~  version 21.10.6
Launching `./Isoseq_analysis.nf` [lethal_newton] - revision: 9967a7dd2a

Welcome to run Nextflow Pipeline Isoseq_analysis.nf
Your configuration are the following:
  project                 : Benchcompare_Isoseq
  isoseq_filelist         : Benchcompare_Isoseq_Samplelist.txt
  outdir                  : .

  genome_build            : hg38
  Protocol_Polya          : Y
  PRIMERS_FA              : /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/Primers/lima_primer/primers_NEB_Clontech.fasta

  Select_Isoseq_Lima      : Y
  Select_Isoseq_Refine    : Y
  Select_Isoseq_Cluster   : N
  Cluster_Singleton       :

  Select_minimap2         : Y
  Select_PBMM2            : Y

  Select_SplitNCigarReads : Y
  Select_Deepvariant      : N
  Select_SNV_ANNOVAR      : N
  vsc_min_fraction_snp:   : 0.07
  vsc_min_fraction_indel  : 0.07

  Select_LongQC_Isoseq    : N
  Sequencing_platform     : pb-sequel
  LongQC_NSAMPLE          : 1000

  Select_LongGF           : Y
  LongGF_min_overlap_len  : 100
  LongGF_bin_size         : 50
  LongGF_min_map_len      : 200

  Select_PBFusion         : Y
  PBFusion_sorted_gtf     : /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/RefGenomes/Human_hg38_Gencode_v39/gencode.v39.annotation.sorted.gtf

  Select_Isoseq_Collapse  : Y
  Collaspe_Min_Coverage   : 0.98
  Collaspe_Min_Identity   : 0.96

  Select_Transcript_Filtering:Y

  Select_Transcript_Classify: Y
  Classify_cage_refTSS_bed  : /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/RefGenomes/Human_hg38_Gencode_v39/refTSS_v3.3_human_coordinate.hg38.sorted.bed
  Classify_polyA_list       : /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/RefGenomes/Human_hg38_Gencode_v39/polyA.list.txt

  Select_IsoQuant         : Y
  GENECOD_gtf_db          : /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/RefGenomes/Human_hg38_Gencode_v39/gencode.v39.annotation.sorted.gtf.db

  Python_Script_Path      : /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/Python_script
  Select_Pigeon_Classify_Report : Y
  gencode_annotation_sorted_gtf : /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/RefGenomes/Human_hg38_Gencode_v39/gencode.v39.annotation.sorted.gtf
  cage_peak_sorted_bed          : /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/RefGenomes/Human_hg38_Gencode_v39/refTSS_v3.3_human_coordinate.hg38.sorted.bed
  polyA_list_txt                : /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/RefGenomes/Human_hg38_Gencode_v39/polyA.list.txt
  covearge_min_count_10_modified2_sorted_tsv: /research/groups/cab/projects/Control/common/Bulk_Isoseq_Nextflow/RefGenomes/Human_hg38_Gencode_v39/intropolis.v1.hg19_with_liftover_to_hg38.tsv.min_count_10.modified2.sorted.tsv
```

B.) Display of the status of completed run

```
executor >  lsf (16)
[b4/1d0f9d] process > Isoseq_Lima (4)              [100%] 4 of 4, cached: 4 ✔
[bb/09ab60] process > Isoseq_Refine (4)            [100%] 4 of 4, cached: 4 ✔
[-         ] process > Isoseq_Cluster              -
[fb/0fe768] process > Pacbio_BAM_To_Fasta (4)      [100%] 4 of 4, cached: 4 ✔
[-         ] process > LongQC_Isoseq               -
[e2/6ad964] process > minimap2_isoseq_mapping (4) [100%] 4 of 4, cached: 4 ✔
[48/9ebbed] process > LongGF (3)                   [100%] 4 of 4, cached: 2 ✔
[b8/6db679] process > PBMM2_Isoseq_mapping (4)     [100%] 4 of 4, cached: 4 ✔
[a6/a8f078] process > IsoQuant_Buffer (4)          [100%] 4 of 4, cached: 4 ✔
[f0/7150e6] process > IsoQuant                     [100%] 1 of 1, cached: 1 ✔
[5b/47d466] process > Isoquant2Pigeon              [100%] 1 of 1 ✔
[a6/1fbabd] process > Pigeon_Classify_Report       [100%] 1 of 1 ✔
[97/9ea384] process > GATK_SplitNCigarReads (4)    [100%] 4 of 4 ✔
[-         ] process > SNV_Deepvariant             -
[-         ] process > SNV_Variant_Filtering       -
[-         ] process > Variant_SNV_ANNOVAR         -
[cd/18a3f0] process > Isoseq_Collapse (4)          [100%] 4 of 4, cached: 4 ✔
[ec/0cea05] process > Transcript_Classify (3)      [100%] 4 of 4 ✔
[3f/2a6f0d] process > Transcript_Filtering (4)     [100%] 4 of 4 ✔
[ae/cb0d3b] process > PBFusion (4)                 [100%] 4 of 4, cached: 4 ✔

Completed at: 02-Jul-2025 13:04:27
Duration    : 1h 23m 36s
CPU hours   : 76.2 (89.7% cached)
Succeeded   : 16
Cached      : 35
```

C.) Result folder

GATK_SplitNCigarReads	Isoseq_Lima	Pacbio_BAM_To_Fasta	Transcript_Classify
IsoQuant	Isoseq_Refine	PBFusion	Transcript_Filtering
Isoquant2Pigeon	LongGF	PBMM2_IsoSeq_mapping	
Isoseq_Collapse	minimap2_isoseq_mapping	Pigeon_Classify_Report	