RNA-Seq “Best Practices” for *Splice-Break2*

This document describes the methodological considerations and bioinformatics processes to analyze common mtDNA deletions from RNA-Seq data.

Library Preparation Methods

The following RNA-Seq library preparation methods are amenable to mtDNA deletion detection using our workflow:

- Bulk RNA-Seq *without* a ribosomal depletion step (e.g., polyA capture/enrichment)
- LCM-Seq
- Spatial Transcriptomics (10x Genomics Visium platform)

FASTQ Formatting Requirements

- FASTQ files should have appropriate file endings. Paired-end files must have reads identified with one of the following options: *.R1/.R2; .r1/.r2; .READ1/.READ2*; or *.read1/.read2*. FASTQ files must end with one of the following extensions: *.fq, .fastq,* or *.txt*.
  See https://github.com/brookehjelm/Splice-Break2 for more details.
- Paired-end FASTQ reads should be formatted such that the header line ends with “/1” or “/2”, corresponding to read number. For downloading samples from GEO, this format can be retained by using the option --origfmt in the sra-toolkit fastq-dump.

To Process all RNA-Seq reads

- For batch processing, all FASTQ files can be stored in one input directory.
- Command line for paired-end files:
  ./Splice-Break2_paired-end.sh <inputDir> <outputDir> <logDir> <SB_Path> --align=yes --ref=rCRS --fastq_keep=no --skip_preAlign=yes
- Command line for single-end files:
  ./Splice-Break2_single-end.sh <inputDir> <outputDir> <logDir> <SB_Path> --align=yes --ref=rCRS --fastq_keep=no --skip_preAlign=yes

To Process only Uniquely Mapped Reads (chrM)

- Align FASTQ to only the nuclear genome using HiSat2 (https://github.com/infphilo/hisat).
  1. Remove chrM from human genome reference file (e.g., ensembl.GRCh38.103.fa), and build index files:
     hisat2-build chrM-removed.ensembl.GRCh38.103.fa chrM-removed.ensembl.GRCh38.103
  2. Align FASTQ files to nuclear genome with HiSat2 and saved unmapped reads to a separate file:
     Paired-End:
     hisat2 -x path/to/reference/chrM-removed.ensembl.GRCh38.103 -1 Sample1.R1.fastq.gz -2 Sample1.R2.fastq.gz -S Sample1_mapped --un-conc Sample1_unmapped
     Single-End:
     hisat2 -x path/to/reference/chrM-removed.ensembl.GRCh38.103 -1 Sample1.R1.fastq.gz -S Sample1_mapped --un-conc Sample1_unmapped
  3. Add appropriate extension *(.fastq, .fq,* or *.txt*) to unmapped read outputs:
     mv Sample1_unmapped.1 Sample1_unmapped.R1.fastq
     mv Sample1_unmapped.2 Sample1_unmapped.R2.fastq
- Process unmapped reads through Splice-Break2.
- For batch processing, all FASTQ files can be stored in one input directory.
- Command line for paired-end files:
  ./Splice-Break2_paired-end.sh <inputDir> <outputDir> <logDir> <SB_Path> --align=yes --ref=rCRS --fastq_keep=no --skip_preAlign=yes
- Command line for single-end files:
  ./Splice-Break2_single-end.sh <inputDir> <outputDir> <logDir> <SB_Path> --align=yes --ref=rCRS --fastq_keep=no --skip_preAlign=yes

Output File

- Only the “Top 30” output file is suggested for RNA-Seq analysis since these deletions have been validated by Sanger sequencing and were evaluated in our RNA-Seq study. (e.g., *Sample1_unmapped_LargeMTDeletions_DNAorRNA_Top30_NARpub.txt*)
- Other output files normally investigated in DNA sequencing studies are also provided but should be used with caution for RNA-Seq data unless the user performs further validation.

A.A.O. version 1: 2023-11-06