Analysis Procedure – No-Amp Data Preparation and Repeat Analysis

Data preparation for repeat analysis requires three steps, all of which are available in SMRT Link v7.0 or on the command line via pbbioconda. From subreads, data need the following steps:

1. Demultiplex
2. Generate CCS
3. Map to Reference

Steps 1 and 2 can be interchanged with equivalent final results. For mapping, we recommend some parameter changes to account for mapping long expanded alleles.

Outputs from the analysis scripts include high-accuracy (≥QV20) CCS sequences for target regions so that results can be analyzed using third party tools, as necessary.

Requirements

Sequence Processing requires the following, available from SMRT Link v7.0 or pbbioconda.

- Demultiplex (lima)
- CCS
- Aligner (pbmm2)

Raw Sequence Processing

For SMRT Link users, the analysis can be completed in two application workflows:

1. Demultiplex
2. CCS and Map

Modified Advanced Analysis Parameters for step 2:

- Consolidate.bam (ON)
- Minimum Concordance (0)
- Override Options (-L 0.1 -E 0)

The resulting aligned .bam file(s) from step 2 is used as input for the repeat analysis tools described below. See next section for an example workflow on the command line using tools from pbbioconda.

Example Workflow

Below are the steps for repeat analysis on the command line. The analysis scripts used are provided in this repository. Note that we use GNU parallel to simplify the processing of multiple samples at once.

Create Working Directories

```
for d in demux ccs align fastq reports ; do
  mkdir ${d}
done
```

Demultiplex Subreads

The program *lima* is provided for demultiplexing.

```
lima --same \
     --min-score 26 \
     --split-bam-named \
     --peek-guess \
     -j 20 \
     m54006_190802_093121.subreadset.xml \
     pacbio.barcodeset.xml \
     demux/m54006_190802_093121.subreadset.xml
```

Call CCS Reads

We recommend calling CCS consensus and filtering to ≥QV20 for further analysis.

```
parallel -j 2 ccs -j 12 --minPredictedAccuracy 0.99 {} ccs/{/.}.ccs.bam ::: demux/*bam
```

Alignment to Reference

We provide two bash scripts (in the repository) to parameterize alignment of CCS reads to the reference. Both scripts use minimap2 as the aligner and have parameters set such that reads with extended repeat motifs are correctly mapped to the reference.

```
parallel -j 6 pbmm2_extension.sh human_hs37d5.fasta {} align/{/.}.aligned.bam {/.} ::: ccs/*bam
```

Extract Target Region

Target regions are identified and clipped out using sequence from the reference which flanks the target region. The sequence extracted from each CCS read is located between the mapped positions of the two flank sequences for that read.

```
parallel python extractRegion.py {} \
                              human_hs37d5.fasta \
                              'X:146993569-146993629' \
                              -o fastq/{/.}.extracted_FMR1.fastq ::: align/*bam
```

Generate Waterfall Plots

A useful and clear visualization.

```
parallel python waterfall.py -m CGG,AGG -i {} -o reports/{/.}.waterfall.png ::: fastq/*fastq
```

Generate Count Histograms

Another useful visualization.

```
parallel python plotCounts.py -m CGG -i {} -o reports/{/.} ::: fastq/*fastq
```

Generate Motif Count Table

Counts of exact string matches, as well as total length of sequence.

```
parallel python countMotifs.py -m CGG,AGG -i {} -o reports/{/.}.counts.csv ::: fastq/*fastq
```

Work directory

Primary contents of working directory (not including indices etc)

```
tree -P '*bam|*fastq|*csv|*png'
.
├── align
│   ├── m54006_190802_093121.bc1015--bc1015.ccs.aligned.bam
│   ├── m54006_190802_093121.bc1016--bc1016.ccs.aligned.bam
│   ├── m54006_190802_093121.bc1017--bc1017.ccs.aligned.bam
│   ├── m54006_190802_093121.bc1018--bc1018.ccs.aligned.bam
│   └── m54006_190802_093121.bc1019--bc1019.ccs.aligned.bam
├── ccs
│   ├── m54006_190802_093121.bc1015--bc1015.ccs.bam
│   ├── m54006_190802_093121.bc1016--bc1016.ccs.bam
│   ├── m54006_190802_093121.bc1017--bc1017.ccs.bam
│   ├── m54006_190802_093121.bc1018--bc1018.ccs.bam
│   └── m54006_190802_093121.bc1019--bc1019.ccs.bam
├── demux
│   ├── m54006_190802_093121.bc1015--bc1015.bam
│   ├── m54006_190802_093121.bc1016--bc1016.bam
│   ├── m54006_190802_093121.bc1017--bc1017.bam
│   ├── m54006_190802_093121.bc1018--bc1018.bam
│   └── m54006_190802_093121.bc1019--bc1019.bam
├── fastq
│   ├── m54006_190802_093121.bc1015--bc1015.ccs.aligned.extracted_FMR1.fastq
│   ├── m54006_190802_093121.bc1016--bc1016.ccs.aligned.extracted_FMR1.fastq
│   ├── m54006_190802_093121.bc1017--bc1017.ccs.aligned.extracted_FMR1.fastq
│   ├── m54006_190802_093121.bc1018--bc1018.ccs.aligned.extracted_FMR1.fastq
│   └── m54006_190802_093121.bc1019--bc1019.ccs.aligned.extracted_FMR1.fastq
└── reports
    ├── m54006_190802_093121.bc1015--bc1015.ccs.aligned.extracted_FMR1.counts.csv
    ├── m54006_190802_093121.bc1015--bc1015.ccs.aligned.extracted_FMR1.insertSize.png
    ├── m54006_190802_093121.bc1015--bc1015.ccs.aligned.extracted_FMR1.motifCount.png
    ├── m54006_190802_093121.bc1015--bc1015.ccs.aligned.extracted_FMR1.waterfall.png
    ├── m54006_190802_093121.bc1016--bc1016.ccs.aligned.extracted_FMR1.counts.csv
    ├── m54006_190802_093121.bc1016--bc1016.ccs.aligned.extracted_FMR1.insertSize.png
    ├── m54006_190802_093121.bc1016--bc1016.ccs.aligned.extracted_FMR1.motifCount.png
    ├── m54006_190802_093121.bc1016--bc1016.ccs.aligned.extracted_FMR1.waterfall.png
    ├── m54006_190802_093121.bc1017--bc1017.ccs.aligned.extracted_FMR1.counts.csv
    ├── m54006_190802_093121.bc1017--bc1017.ccs.aligned.extracted_FMR1.insertSize.png
    ├── m54006_190802_093121.bc1017--bc1017.ccs.aligned.extracted_FMR1.motifCount.png
    ├── m54006_190802_093121.bc1017--bc1017.ccs.aligned.extracted_FMR1.waterfall.png
    ├── m54006_190802_093121.bc1018--bc1018.ccs.aligned.extracted_FMR1.counts.csv
    ├── m54006_190802_093121.bc1018--bc1018.ccs.aligned.extracted_FMR1.insertSize.png
    ├── m54006_190802_093121.bc1018--bc1018.ccs.aligned.extracted_FMR1.motifCount.png
    ├── m54006_190802_093121.bc1018--bc1018.ccs.aligned.extracted_FMR1.waterfall.png
    ├── m54006_190802_093121.bc1019--bc1019.ccs.aligned.extracted_FMR1.counts.csv
    ├── m54006_190802_093121.bc1019--bc1019.ccs.aligned.extracted_FMR1.insertSize.png
    ├── m54006_190802_093121.bc1019--bc1019.ccs.aligned.extracted_FMR1.motifCount.png
    └── m54006_190802_093121.bc1019--bc1019.ccs.aligned.extracted_FMR1.waterfall.png
```

Revision History (Description)	Version	Date
Initial Release.	01	August 2019