**Notes/Overview of DS\_PE\_Unified.sh**

Last modified by Dana Nachmanson 9/2/16

1. Raw FASTQ file 🡪 SSCS 🡪 DCS

* Scott’s unified consensus maker
* **“”\_read1\_sscs.fq.gz**
* **“”\_read2\_sscs.fq.gz**
* **“”\_read1\_dcs.fq.gz**
* **“”\_read2\_dcs.fq.gz**
  + - “”.tagstats.txt
    - “”.temp.sort.bam
    - “”fam\_size\_relation.png
    - “”family\_size.png

1. Align to reference genome

* Either BWA-mem or BWA-sw, current algorithm produces both alignments. Further testing needs to determine which is optimal. Mem is currently recommended by BWA but sw might perform better with large alignment gaps
  + - “”.sscs.sort.bam
    - “”\_mem.sscs.sort.bam
    - **“”.dcs.sort.bam**
    - “”\_mem.dcs.sort.bam

\*\*\*USING DCS FILES CREATED FROM SW FILES FROM THIS POINT… will need to test SSCS files and MEM aligned fragments later on

1. Filter out unmapped reads

* “”.dcs.sort.mapped.bam

1. Filter out reads with less than perfect mapping scores (<60)
   * + “”.dcs.sort.mapped.q60.bam
2. Filter out reads with iSize less than Read length – adaptor length – spacer length

(Explanation at end of file)

* “”.dcs.filt.bam

6. Clip ends of reads – Maybe not needed after filtering reads for size, needs testing

\*\*\* Note that the read sequence remains the same but the cigar is adjusted

* “”.dcs.filt.clipped.bam

7. Local realignment – Critical for Soft-clipping

* “”.dcs.filt.clipped.realign.bam

8. Soft-clipping of overlapping reads

* “”.dcs.filt.clipped.realign.no\_overlap.bam
* “”.overlapping-reads.stats

9. Create pileup files

* “”.dcs.clipped.no\_overlap.pileup

10. Trim pileup to genomic coordinates

* “”.dcs.clipped.no\_overlap.region.pileup

11. Create a version of pileup file with & without SNPs

* “”.dcs.clipped.no\_overlap.region.noSNPs.pileup
* “”.dcs.clipped.no\_overlap.region.ONLYSNPs.pileup

12. Create final mutation files with SNPs:

* “”.DCS-muts.txt (Min # of mutations is 1)
* “”.DCS.pileup.mutpos

13. Create final mutation files without SNPs:

* “”.DCS.noSNPs.pileup.countmuts (unique only)
* ‘’.DCS-muts.noSNPs.txt (Min # of mutations is 1)
* “”.noSNPs.pileup.mutpos

14. Create final mutation files with only SNPs:

* “”.DCS.ONLYSNPs.pileup.mutpos

15. Plot iSize histogram

* “”.iSize\_Metrics.txt
* “”.iSize\_Histogram.pdf

16. Plot mutations by read cycle (Unclipped data)

* “”.ErrorRatePerCycle.txt
* “”.Muts\_by\_Cycle.png

17. Plot DCS depth by genomic position

* “”.DCS\_Depth\_by\_pos.png

18. Print reads statistics

* “”.flagstats.stats

**Tools necessary:**

* Picard 2.2.1
* pysam 0.9.0 (I used 0.9.1 and it worked fine)
* samtools 1.3.1
* htslib 1.3.1
* BWA
* X-code (text editor)
* Matplotlib
* bamUtil

**I-Size filtering thoughts:**

Minimum fragment length

Adaptor + spacer length

Read length

90

84

89

83

101bp

100bp

11

11

17

17

After trimming the adaptors + spacers off of the reads, the minimum fragment length should be 83 bp

83bp

R1

R2

Minimum fragment length is thus 83bp

Before trimming the adaptor + spacers off of the individual reads, minimum DNA fragment before reading into adaptors is 117bp

83 bp

100bp

117bp

17bp

17bp

R2

R1