**ORIGINAL**

Preliminary quality control analysis of FastQ files was performed using FastQC v.0.11.2 ([**32**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-32)). The three prime adapters were trimmed from the reads using Scythe v0.981 ([**33**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-33)). Quality trimming was carried out using Sickle v1.33 ([**34**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-34)) with a Phred score of 30 as the quality threshold; reads with a length of less than 45 bp were discarded. To remove contaminating cellular reads, reads aligning to African green monkey (*Chlorocebus sabeus* 1.1) or human (hg19) genomes were removed. The remaining unmapped reads were then aligned to the SV40 genome (RefSeq accession no. NC\_001669.1) cut at nt 2666 using Bowtie2 v2.2.4 ([**35**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-35)) for wild-type SV40 or, in the case of cs1085, the wild-type sequence with the deletion. Duplicate fragments were removed using the Picard Tools (Broad) mark duplicates function. Bam files were filtered to contain only fragments between 100 and 150 bp using an awk script. Replicate bam files were merged using SAMtools v1.3.1 ([**36**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-36)). Bed graphs normalized to 1× coverage were generated from filtered, deduplicated reads using Deeptools v2.5.4 ([**37**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-37)). Heat maps were made using the Z scores of the normalized coverage and displayed using IGV v2.3.52 ([**38**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-38)).

**EDITED**

Preliminary quality control analysis of FastQ files was performed using FastQC v0.11.8([**32**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-32)). Three prime adapters of the sequence from New England Biolabs NBNext library prep kit (NEB #E7335) were trimmed from the reads using Scythe v0.991([**33**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-33)). Quality trimming of reads was performed using Sickle v1.33 ([**34**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-34)) with a minimum Sanger phred+33 quality score cutoff of 30 and minimum read length of 45bp. Contaminating host reads were removed by aligning to the human (hg19) using Bowtie2 v2.3.4.3 ([**35**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-35)) using the “—un-conc” flag to collect reads that fail to align, then repeating the process by aligning those reads with the African green monkey (Chlorocebus sabeus 1.1) genome. The resultant un-concordantly aligned reads were aligned to the SV40 genome (RefSeq accession no. NC\_001669.1). Resulting sam files were converted to bam files and sorted using SAMtools v0.1.19 ([**36**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-36)). Duplicate fragments were marked for removal using Picard Tool’s (Broad) “MarkDuplicates” function, and resultant marked bam files were indexed and subsequently filtered using the “0x400” bitwise flag with SAMtools. Bed graph files were generated using resultant reads with Deeptools’ v3.2.0 ([**37**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-37)) “bamCoverage” function with RPGC 1x normalization as well as minimum and maximum fragment sizes filtered using the “minFragmentLength” and “maxFragmentLength” flags respectively. Heat maps of the resultant bed graph files were made by importing them into IGV v2.5.0 ([**38**](https://jvi.asm.org/content/93/4/e01678-18.long#ref-38)) with the SV40 genome.

**CODE – Bash Script run on Torque Cluster**

**Suggested that you create a conda environment to install all packages**

**Possible implementation could be:**

*conda config –add channels bioconda conda-forge faircloth-lab*

*conda create -n chipseqenv scythe sickle-trim samtools deeptools bowtie2 picard*

**For questions/comments email** [**lswhiteh@iu.edu**](mailto:lswhiteh@iu.edu) **or visit** [**https://github.com/Lswhiteh**](https://github.com/Lswhiteh)

**You will need to edit the numbers around the ${label} variable to match the identifiers in the sequencing reads. The “for ###” should match the number around the read ID, but any of this section can be changed to match file structure/naming conventions.**

**#!/bin/bash**

**#PBS -k o**

**#PBS -l nodes=1:ppn=8,walltime=01:00:00,vmem=8gb**

**#PBS -M lswhiteh@iu.edu**

**#PBS -m abe**

**#PBS -N chipseqanalysis**

**#PBS -j oe**

**module load anaconda**

**source activate chipseq**

**for FILE in ~/chipseq/FASTQ/\*.fastq**

**do**

**echo $FILE**

**#Trim and quality threshold**

**scythe -a ~/chipseq/TruSeq3-PE.fa -o ~/chipseq/trimmed/${FILE##\*/} $FILE**

**done**

**for label in 19 20 21**

**do**

**#Quality trimming**

**sickle pe -f ~/chipseq/trimmed/19-035-0${label}\_S${label}\_L001\_R1\_001.fastq \**

**-r ~/chipseq/trimmed/19-035-0${label}\_S${label}\_L001\_R2\_001.fastq \**

**-o ~/chipseq/trimmed/sickled\_19-035-0${label}\_S${label}\_L001\_R1\_001.fastq \**

**-p ~/chipseq/trimmed/sickled\_19-035-0${label}\_S${label}\_L001\_R2\_001.fastq \**

**-s ~/chipseq/trimmed/sickled\_singles.fastq \**

**-t sanger \**

**-q 30 \**

**-l 45**

**done**

**fastqc ~/chipseq/trimmed/\***

**for label in 19 20 21**

**do**

**# Get aligned and unaligned, convert to bam**

**bowtie2 -q -x ~/chipseq/reference\_genomes/homosapiens/GCA\_000001405.15\_GRCh38\_full\_analysis\_set.fna.bowtie\_index \**

**-1 ~/chipseq/trimmed/sickled\_19-035-0${label}\_S${label}\_L001\_R1\_001.fastq \**

**-2 ~/chipseq/trimmed/sickled\_19-035-0${label}\_S${label}\_L001\_R2\_001.fastq \**

**-S ~/chipseq/S${label}homosapiensalignments.sam \**

**--un-conc ~/chipseq/S${label}hsunmappedreads.fq**

**bowtie2 -q -x ~/chipseq/reference\_genomes/vervet/vervet\_1.1\_genomic \**

**-1 ~/chipseq/S${label}hsunmappedreads.1.fq \**

**-2 ~/chipseq/S${label}hsunmappedreads.2.fq \**

**-S ~/chipseq/S${label}vervetalignments.sam \**

**--un-conc ~/chipseq/S${label}vervetunmappedreads.fq**

**bowtie2 -q -x ~/chipseq/reference\_genomes/sv40/sv40index \**

**-1 ~/chipseq/S${label}vervetunmappedreads.1.fq \**

**-2 ~/chipseq/S${label}vervetunmappedreads.2.fq \**

**-S ~/chipseq/S${label}vervetsv40aligned.sam**

**#Convert to bam, filter by read lengths**

**echo "convert to bam"**

**samtools view -bS ~/chipseq/S${label}vervetsv40aligned.sam > ~/chipseq/S${label}vervetsv40aligned.bam**

**echo "sort"**

**samtools sort ~/chipseq/S${label}vervetsv40aligned.bam ~/chipseq/S${label}sortedvervetsv40aligned**

**#Mark duplicates**

**picard MarkDuplicates \**

**INPUT=~/chipseq/S${label}sortedvervetsv40aligned.bam \**

**OUTPUT=~/chipseq/S${label}sv40aligneddupesmarked.bam \**

**ASSUME\_SORTED=true \**

**METRICS\_FILE=~/chipseq/S${label}sv40aligneddupesmarkedmetrics.txt**

**samtools index ~/chipseq/S${label}sv40aligneddupesmarked.bam**

**#Get rid of dupes**

**samtools view -b -F 0x400 ~/chipseq/S${label}sv40aligneddupesmarked.bam > ~/chipseq/S${label}sv40aligneddupesremoved.bam**

**samtools index ~/chipseq/S${label}sv40aligneddupesremoved.bam**

**bamCoverage --bam ~/chipseq/S${label}sv40aligneddupesremoved.bam -o ~/chipseq/S${label}finaloutput\_150plus.bedgraph \**

**--effectiveGenomeSize 5243 \**

**--normalizeUsing RPGC \**

**--outFileFormat bedgraph \**

**--minFragmentLength 100 \**

**--maxFragmentLength 150**

**done**