ATAC-seq Data Analysis Pipeline

For mouse samples, use reference sequence mm10 male from ENCODE.
ENCODE accession ENCSR425FOI, fasta file for mm10 male genome reference, https://www.encodeproject.org/references/ENCSR425FOI/

Use a bowtie index generated on the above reference. (We plan to submit this index to the DCC, and then we can refer to the unique id of the index file.)

1. Run FastQC v0.10.1 for basic information on the fastq file (GC content, duplication rate, etc.)
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

2. Trim reads to 30bp
Python trimfastq.py *inputfile.fastq* 30 > *outputfile.fastq*

3. Map reads using Bowtie version 1.0.0
http://bowtie-bio.sourceforge.net/index.shtml
with the following parameters:
--chunkmbs 1024 (set memory usage to whatever works on your system)
-y -v 2 -p 7 --best --strata –m 3 -k 1
--sam-nohead --sam (set output format)
Give locations of indexes, reads, and output file.

4. Run Samtools version 0.1.18
http://sourceforge.net/projects/samtools/files/samtools/
Convert to bam with samtools view –b –T <genome> -o <output> -S <input>
Sort with samtools sort
Count reads with samtools flagstat

5. Sample 10M reads and compute "complexity"
(algorithm from Georgi; we plan to submit the script from Belinda to the DCC).

6. Run Samtools to filter to mapped reads only
samtools view -b -F 4 -o <output> <input>

7. Filter out reads mapped to mitochondria
samtools view *input.bam* | egrep -v chrM | samtools view -bT *fasta_reference* --o *output.bam*

8. Use Bedtools version 2.16.2
https://github.com/arq5x/bedtools2/releases
to convert to bed format for Fseq
bedtools bamtobed –i <input> (output redirected to file)

9. Call nuclease accessible **regions** using FSeq version 1.85

http://fureylab.web.unc.edu/software/fseq/

Input is bed file from step 6.

Create wiggle: fseq -v -of wig -f 0 -d <outdirectory> -o <outfile> <input>

Create regions: fseq -v -of npf -f 0 -d <outdirectory> -o <outprefix> <input>

Output files are by chromosome, concatenate to single bed and wiggle file.

10. Call nuclease accessible **peaks** using Homer version 4.7

http://homer.salk.edu/homer/chipseq/peaks.html

Input is bam file from step 3.

Setup: makeTagDirectory <outdirectory> <input>

Peaks: findPeaks <tags> -o <output> -localSize 50000 -size 150 -minDist 50 –fragLength 0

11. Finish peaks and regions, changing formats and filtering (perl script; will be submitted to DCC).

Format homer peaks to bed (cut wanted columns from table)

These are the unfiltered peaks, and are provided for users who want the full set.

12. Finish wiggles by combining wiggles per chromosome to one file and convert to bigWig.

perl script OR cat <wigs> | grep –v track > <output>

wigToBigWig <input> <chrom.sizes> <output>

13. Optional Filtering: Remove from ATAC accessible **regions** and **peaks** the genomic intervals that intersect the intervals on the combined blacklist.

bedtools intersect –v a <peaks> -b <black> > <output>

The combined blacklist (blacklist.full.bed) will be submitted to the DCC.

In addition, accessible regions (from Fseq) can be limited to the top 100,000.

sort –k7,7nr <input> | head -100000 > <output>

14. Compute FRIP on both regions and peaks (perl script will be submitted to DCC.)