19AdvInformatics Project

ATAC-seq Analysis Pipeline

Heidi Liang

Location of Scripts: Github

https://github.com/heidiiliang/19AdvInformaticsProject

- atacalignment.sh
- bamCoverage.sh
- homerscript.sh
- run.entire.idrPipeline.sh (still testing)
- Report error files

Practice ATAC-seq Samples: GM12878 cell line

- An ENCODE Project common cell type
- B-lymphocyte, a lymphoblastoid cell line
- Suspension cell culture
- Perform Omni-ATAC-seq on 3 replicates: H37, H38, H39

Pipeline Part1 - Trim reads, align and sort reads

1. Trim the reads: 30bp using Trimmomatic
2. Map reads to mitochondrial chromosome using Bowtie2 and sort with Samtools
3. Map unaligned reads (non-mitochondrial) using Bowtie2 and sort with Samtools
4. Remove multi-mappers using samtools
5. Sort the merged BAM file using Picard tools
6. Remove PCR duplicates using Picard tools
7. Adjust the read start sites to represent the center of the transposon binding event using Samtools
8. Create Tag Directory (homer-tags) using Homer
9. Build BAM index with Picard tools
10. Generate bigWig file using deeptools

- Par1 went well except not able to display on UCSC genome browser and not sure how to solve because my other bigWig files were able to showed in that public directory.
- I saw no error in the error reporting file, but I do not understand some of the reported stuff mean. May be there is problem with the bigWig file generated.
- Files and part of the report from hpc: (more at github)

```
-rw-rw-r-- 1 yhliang bio   42M Mar 15 15:15 H37.bigWig
-rw-r--r-- 1 yhliang bio  1.1K Mar 15 15:14 bam.log.e6078067.1
-rw-rw-r-- 1 yhliang bio   20M Mar 15 15:14 signal.bigWig
-rw-rw-r-- 1 yhliang bio   726 Mar 15 15:11 bamCoverage.sh
-rw-r--r-- 1 yhliang bio   15K Mar 15 13:07 atacalignment.e6077350.1
-rw-rw-r-- 1 yhliang bio  5.4M Mar 15 13:06 shifted_reads.bai
drwxrwsr-x 2 yhliang bio    30 Mar 15 13:05 homer-tags
-rw-rw-r-- 1 yhliang bio  354M Mar 15 13:02 shifted_reads.bam
-rw-rw-r-- 1 yhliang bio  2.8K Mar 15 13:01 duplicates_metric.txt
-rw-rw-r-- 1 yhliang bio  598M Mar 15 13:01 mappings.nodup.bam
-rw-rw-r-- 1 yhliang bio  654M Mar 15 12:58 mappings.sorted.bam
-rw-rw-r-- 1 yhliang bio  853M Mar 15 12:56 mappings.uniq.bam
-rw-rw-r-- 1 yhliang bio  2.5G Mar 15 12:55 mappings.bam
-rw-rw-r-- 1 yhliang bio  2.2G Mar 15 12:47 chrM.bam
-rw-rw-r-- 1 yhliang bio  4.3G Mar 15 12:47 unaligned.2.fastq
-rw-rw-r-- 1 yhliang bio  4.2G Mar 15 12:47 unaligned.1.fastq
-rw-rw-r-- 1 yhliang bio   11M Mar 15 12:36 se_read2.fastq.gz
-rw-rw-r-- 1 yhliang bio  4.9G Mar 15 12:36 pe_read2.fastq
-rw-rw-r-- 1 yhliang bio   19M Mar 15 12:36 se_read1.fastq.gz
-rw-rw-r-- 1 yhliang bio  4.9G Mar 15 12:36 pe_read1.fastq
-rw-r--r-- 1 yhliang bio  4.9K Mar 15 12:32 atacalignment.sh
-rw-r--r-- 1 yhliang bio  986M Mar 14 17:24 H37_S8_R2_001.fastq.gz
-rw-r--r-- 1 yhliang bio 1019M Mar 14 17:24 H37_S8_R1_001.fastq.gz
```

```
24787186 reads; of these:
  24787186 (100.00%) were paired; of these:
    21667231 (87.41%) aligned concordantly 0 times
    3119955 (12.59%) aligned concordantly exactly 1 time
    0 (0.00%) aligned concordantly >1 times
    ----
    21667231 pairs aligned 0 times concordantly or discordantly; of these:
      43334462 mates make up the pairs; of these:
        39408818 (90.94%) aligned 0 times
        3925644 (9.06%) aligned exactly 1 time
        0 (0.00%) aligned >1 times
20.51% overall alignment rate
21667231 reads; of these:
  21667231 (100.00%) were paired; of these:
    17199340 (79.38%) aligned concordantly 0 times
    3928815 (18.13%) aligned concordantly exactly 1 time
    539076 (2.49%) aligned concordantly >1 times
    ----
    17199340 pairs aligned 0 times concordantly or discordantly; of these:
      34398680 mates make up the pairs; of these:
        27807281 (80.84%) aligned 0 times
        5207346 (15.14%) aligned exactly 1 time
        1384053 (4.02%) aligned >1 times
35.83% overall alignment rate
```

```
total 14G
-rw-rw-r-- 1 yhliang bio  43M Mar 15 13:01 H38.bigWig
-rw-r--r-- 1 yhliang bio  14K Mar 15 13:00 atacalignment.e6077351.1
-rw-rw-r-- 1 yhliang bio  20M Mar 15 13:00 signal.bigWig
-rw-rw-r-- 1 yhliang bio 5.4M Mar 15 13:00 shifted_reads.bai
drwxrwsr-x 2 yhliang bio   30 Mar 15 12:59 homer-tags
-rw-rw-r-- 1 yhliang bio 322M Mar 15 12:59 shifted_reads.bam
-rw-rw-r-- 1 yhliang bio 2.8K Mar 15 12:58 duplicates_metric.txt
-rw-rw-r-- 1 yhliang bio 498M Mar 15 12:58 mappings.nodup.bam
-rw-rw-r-- 1 yhliang bio 541M Mar 15 12:55 mappings.sorted.bam
-rw-rw-r-- 1 yhliang bio 701M Mar 15 12:54 mappings.uniq.bam
-rw-rw-r-- 1 yhliang bio 1.9G Mar 15 12:53 mappings.bam
-rw-rw-r-- 1 yhliang bio 1.9G Mar 15 12:46 chrM.bam
-rw-rw-r-- 1 yhliang bio 3.1G Mar 15 12:46 unaligned.2.fastq
-rw-rw-r-- 1 yhliang bio 3.1G Mar 15 12:46 unaligned.1.fastq
-rw-rw-r-- 1 yhliang bio 6.3M Mar 15 12:35 se_read2.fastq.gz
-rw-rw-r-- 1 yhliang bio 4.0G Mar 15 12:35 pe_read2.fastq
-rw-rw-r-- 1 yhliang bio  18M Mar 15 12:35 se_read1.fastq.gz
-rw-rw-r-- 1 yhliang bio 4.0G Mar 15 12:35 pe_read1.fastq
-rw-r--r-- 1 yhliang bio 4.9K Mar 15 12:33 atacalignment.sh
-rw-r--r-- 1 yhliang bio 1.2K Mar 15 12:32 atacalignment.e6077349.1
-rw-r--r-- 1 yhliang bio  14K Mar 15 12:09 atacalignment.e6077296.1
-rw-r--r-- 1 yhliang bio 9.3K Mar 15 11:37 atacalignment.e6077291.1
-rw-r--r-- 1 yhliang bio 817M Mar 14 17:25 H38_S9_R2_001.fastq.gz
-rw-r--r-- 1 yhliang bio 828M Mar 14 17:25 H38_S9_R1_001.fastq.gz
```

```
-rw-r--r-- 1 yhliang bio  15K Mar 15 13:11 atacalignment.e6077352.1
-rw-rw-r-- 1 yhliang bio  25M Mar 15 13:11 signal.bigWig
-rw-rw-r-- 1 yhliang bio 5.6M Mar 15 13:11 shifted_reads.bai
drwxrwsr-x 2 yhliang bio   30 Mar 15 13:10 homer-tags
-rw-rw-r-- 1 yhliang bio 421M Mar 15 13:08 shifted_reads.bam
-rw-rw-r-- 1 yhliang bio 2.8K Mar 15 13:05 duplicates_metric.txt
-rw-rw-r-- 1 yhliang bio 594M Mar 15 13:05 mappings.nodup.bam
-rw-rw-r-- 1 yhliang bio 653M Mar 15 13:01 mappings.sorted.bam
-rw-rw-r-- 1 yhliang bio 850M Mar 15 12:59 mappings.uniq.bam
-rw-rw-r-- 1 yhliang bio 2.1G Mar 15 12:58 mappings.bam
-rw-rw-r-- 1 yhliang bio 2.3G Mar 15 12:51 chrM.bam
-rw-rw-r-- 1 yhliang bio 3.2G Mar 15 12:51 unaligned.2.fastq
-rw-rw-r-- 1 yhliang bio 3.2G Mar 15 12:51 unaligned.1.fastq
-rw-rw-r-- 1 yhliang bio 9.1M Mar 15 12:37 se_read2.fastq.gz
-rw-rw-r-- 1 yhliang bio 4.7G Mar 15 12:37 pe_read2.fastq
-rw-rw-r-- 1 yhliang bio  23M Mar 15 12:37 se_read1.fastq.gz
-rw-rw-r-- 1 yhliang bio 4.7G Mar 15 12:37 pe_read1.fastq
-rw-r--r-- 1 yhliang bio 4.9K Mar 15 12:33 atacalignment.sh
-rw-r--r-- 1 yhliang bio  15K Mar 15 12:14 atacalignment.e6077295.1
-rw-r--r-- 1 yhliang bio 9.3K Mar 15 11:38 atacalignment.e6077292.1
```

Pipeline Part2 - IDR peak calling

1. Call peaks from each homer tags using Homer : 150bp and 500bp
2. Copy the peaks for each replicate into one folder
3. Make a combined tag directory for all reps, find peaks for pooled combined using Homer
4. Create pseudoreps for individual reps using pre-existing tool (run_idr.py)
5. Create pseudoreps for pooled using the pre-existing tool
6. Call peaks for individual pseudoreps
7. Call peaks for combined pseudoreps
8. Finally run IDR the pre-existing tool

```
drwxrwsr-x 2 yhliang bio    4 Mar 18 16:47 __pycache__
-rw-r--r-- 1 yhliang bio  764 Mar 18 16:13 t2estentire.IDR.run.e6083615.1
-rw-rw-r-- 1 yhliang bio 3.1K Mar 18 16:11 2testrun.entire.idrPipeline.sh
-rw-r--r-- 1 yhliang bio  758 Mar 18 16:05 t2estentire.IDR.run.e6083542.1
-rw-r--r-- 1 yhliang bio 1.3K Mar 18 16:02 t2estentire.IDR.run.e6083487.1
-rw-r--r-- 1 yhliang bio 1.5K Mar 18 15:58 t2estentire.IDR.run.e6083473.1
-rw-r--r-- 1 yhliang bio 1.8K Mar 18 15:48 t2estentire.IDR.run.e6083455.1
-rw-r--r-- 1 yhliang bio  12K Mar 18 15:48 utils.pyc
-rw-r--r-- 1 yhliang bio 4.1K Mar 18 15:48 idr_caller.pyc
-rw-r--r-- 1 yhliang bio 1.2K Mar 18 15:14 t2estentire.IDR.run.e6083410.1
-rw-r--r-- 1 yhliang bio  11K Mar 18 15:12 t1estentire.IDR.run.e6083398.1
-rw-r--r-- 1 yhliang bio  143 Mar 18 15:12 t1estentire.IDR.run.o6083398.1
drwxrwsr-x 2 yhliang bio   31 Mar 18 15:05 Combined
-rw-r--r-- 1 yhliang bio  14K Mar 18 15:05 utils.py
-rw-r--r-- 1 yhliang bio  20K Mar 18 15:05 run_idr.py
-rw-r--r-- 1 yhliang bio    0 Mar 18 15:05 __init__.py
-rw-r--r-- 1 yhliang bio 4.5K Mar 18 15:05 idr_caller.py
-rw-rw-r-- 1 yhliang bio 4.8K Mar 18 14:45 testrun.entire.idrPipeline.sh
-rw-r--r-- 1 yhliang bio 1.3K Mar 18 14:41 t2estentire.IDR.run.e6083340.1
drwxrwsr-x 6 yhliang bio    6 Mar 18 14:41 peaks
-rw-r--r-- 1 yhliang bio 5.3K Mar 18 14:40 t1estentire.IDR.run.e6083284.1
-rw-r--r-- 1 yhliang bio  143 Mar 18 14:40 t1estentire.IDR.run.o6083284.1
-rw-r--r-- 1 yhliang bio 7.1K Mar 15 16:46 testentire.IDR.run.e6078223.1
-rw-r--r-- 1 yhliang bio 6.6K Mar 15 16:23 testentire.IDR.run.e6078190.1
drwxrwsr-x 4 yhliang bio    4 Mar 15 16:23 pseudoreps
-rw-r--r-- 1 yhliang bio 5.2K Mar 15 16:05 testentire.IDR.run.e6078172.1
drwxr-sr-x 3 yhliang bio   14 Mar 15 15:59 idrCode
-rw-r--r-- 1 yhliang bio 8.0K Mar 15 15:47 homer.e6078159.1
-rw-rw-r-- 1 yhliang bio  22M Mar 15 15:47 H39_150bp.peaks.txt
-rw-rw-r-- 1 yhliang bio  18M Mar 15 15:46 H38_150bp.peaks.txt
-rw-rw-r-- 1 yhliang bio  19M Mar 15 15:45 H37_150bp.peaks.txt
-rw-r--r-- 1 yhliang bio  832 Mar 15 15:42 homerscript.sh
```

Stop at step 4 because not sure how to use source code properly

1. Python code problem - probably the version problem
    1.1. Tried to use the newest enthought_python to run the python code
    1.2. TypeError: super() takes at least 1 argument
    1.3. Found it was written in 2014
    1.4. Solution: python3 is not the right version to use
    1.5. Tried python/2.4.6 python/2.7.10 python/2.7.15 python/2.7.2
        1.5.1. ImportError: No module named pandas
    1.6. Realized a note in the script: IDR code runs only with python3. Use Anaconda3 python by setting path in the .bashrc file
        1.6.1. But this conflict with what I read online
    1.7. Tried anaconda/3.7-5.3.0/
        1.7.1. More different errors
    1.8. Gave up and will solve the problem by asking PhD who gave me the source code

Things learned

1. Learned how to write bash script
2. Learned how to debug by checking error files
3. Learned how to use different softwares and read many manuals
    a. Trimmomatic
    b. Bowtie2
    c. Samtools
    d. Picard tools
    e. Deeptools
    f. Homer
4. Learned how to optimize the options
    a. For example, unique mapping by setting -q. The -q value for unique found under mapping.bam using Samtools.
5. When some script seems not working and do not know what is the problem, separating scripts may solve the problem and easier to debug.
    a. Ex. I separate bamCoverage from atacalignment.sh because in the report it shows the command not working. So at the end of the atacalignment.sh, I wrote qsub bamCoverage.sh.