

**Table S1: OPTIMAL CASE STUDY: Short Sequencing Reads Alignment and Visualization**

**Scientific question, analysis methods, expected outputs**

ChIP-seq is a technique used to measure transcription factor (TF) binding by sequencing DNA fragments that have been pulled down by an antibody specific to the TF. The number of reads sequenced at a genomic region is proportional to the likelihood of the TF binding to that region. In this session, we will instruct ChatGPT to align the sequencing reads back to a reference genome and generate a file in bigwig format for visualizing the distribution of the reads across the genome.

We generated two sequencing files of different depths for a CTCF ChIP-seq library in human ESC H1 cell line from ENCODE, one containing one million reads and the other containing ten million reads. We explored the impact of sequencing depth on TF prediction for the CTCF protein. Results will be compared to the visualization of the distribution generated by ENCODE.

Tools to be used: Bowtie2, samtools, and deepTools.

**Coding language(s)**

Bash

**PREREQUISITE KNOWLEDGE**

- 1: Have a good understanding of what is ChIP-Seq and key steps involved in this specific data analysis. ChatGPT may be used to provide some hints on this by prompts such as "What is ChIP-Seq?" and "Starting from a fastq sequencing file with a good quality for a ChIP-Seq library, illustrate the key steps of analysis leading to visualize the read distribution across a reference genome." If use ChatGPT as an assistant for this, instructor should guide students how to identify relevant response, figure out missing part, and filter out irrelevant response.
- 2: Ensure necessary packages are installed, including Bowtie2, samtools, and deepTools.
- 3: Ensure relevant input files are available: two fastq files (available upon request) and bowtie2 index file for the human genome.
- 4: Knows how to read and run bash commands

**DEFINE CHATBOT'S BEHAVIORS**

Act as an experienced bioinformatician proficient in ChIP-Seq data analysis, you will assist me by writing code with number of lines as minimal as possible. Rest the thread if asked to. Reply "YES" if understand.

**ITERATION 0**

**Prompts**

I have two fastq files in current folder from single-end sequencing of a ChIP-Seq library: ENCFF000AVS\_1m.fastq.gz, and ENCFF000AVS\_10m.fastq.gz. For each fastq file, align reads to the human reference genome, save to bam file, and then convert it to bigwig file. Tools to use: bowtie2, samtools, and deepTools. The index for bowtie2 is in the folder "../data/indx/bowtie2\_whole\_genome/" with "hg38" as the prefix. Use 24 CPU for the alignment. Please draft the code in bash.

**Code**

```
#!/bin/bash

# Define variables
INDEX="../data/indx/bowtie2_whole_genome/hg38"
CPUS=24
```

```
FASTQ1="ENCFF000AVS_1m.fastq.gz"
FASTQ2="ENCFF000AVS_10m.fastq.gz"

# Align reads to reference genome using bowtie2
bowtie2 -p $CPUS -x $INDEX -U $FASTQ1 | samtools sort -@ $CPUS -o ENCFF000AVS_1m.bam
bowtie2 -p $CPUS -x $INDEX -U $FASTQ2 | samtools sort -@ $CPUS -o ENCFF000AVS_10m.bam

# Convert BAM files to bigWig format using deepTools
bamCoverage -b ENCFF000AVS_1m.bam -o ENCFF000AVS_1m.bw --binSize 10 --normalizeUsing CPM
bamCoverage -b ENCFF000AVS_10m.bam -o ENCFF000AVS_10m.bw --binSize 10 --normalizeUsing CPM
```

## Result

```
1000000 reads; of these:
  1000000 (100.00%) were unpaired; of these:
    340421 (34.04%) aligned 0 times
    465276 (46.53%) aligned exactly 1 time
    194303 (19.43%) aligned >1 times
55.96% overall alignment rate
[bam_sort_core] merging from 0 files and 24 in-memory blocks...
10000000 reads; of these:
  10000000 (100.00%) were unpaired; of these:
    3398797 (33.99%) aligned 0 times
    4721200 (47.21%) aligned exactly 1 time
    1880003 (18.80%) aligned >1 times
56.01% overall alignment rate
[bam_sort_core] merging from 0 files and 24 in-memory blocks...
[E::idx_find_and_load] Could not retrieve index file for 'ENCFF000AVS_1m.bam'
'ENCFF000AVS_1m.bam' does not appear to have an index. You MUST index the file first!
[E::idx_find_and_load] Could not retrieve index file for 'ENCFF000AVS_10m.bam'
'ENCFF000AVS_10m.bam' does not appear to have an index. You MUST index the file first!
```

## Assessment\*

The alignment worked well. However, error message indicates that bam files are not indexed. This error message was passed to ChatGPT and see its response.

## ITERATION 1

### Prompts

```
[E::idx_find_and_load] Could not retrieve index file for 'ENCFF000AVS_1m.bam'
```

### Code

```
#!/bin/bash

# Define variables
INDEX="../../data/idx/bowtie2_whole_genome/hg38"
CPUS=24
BAM1="ENCFF000AVS_1m.bam"
BAM2="ENCFF000AVS_10m.bam"

# Sort BAM files using samtools
samtools sort -@ $CPUS -o ENCFF000AVS_1m.sorted.bam $BAM1
samtools sort -@ $CPUS -o ENCFF000AVS_10m.sorted.bam $BAM2

# Index sorted BAM files using samtools
samtools index ENCFF000AVS_1m.sorted.bam
samtools index ENCFF000AVS_10m.sorted.bam

# Convert BAM files to bigWig format using deepTools
bamCoverage -b ENCFF000AVS_1m.sorted.bam -o ENCFF000AVS_1m.bw --binSize 10 --normalizeUsing CPM
```

```
bamCoverage -b ENCF000AVS_10m.sorted.bam -o ENCF000AVS_10m.bw --binSize 10 --normalizeUsing CPM
```

## Result

NA

## Assessment\*

This time I noticed that "samtools index" was inserted ahead of bamCoverage. Also notice that the code assumed that bowtie alignment was done successfully, and therefore can not be directly applied to address our needs without adding back the code for the alignment.

## ITERATION 2

### Prompts

Wait, I saw that you have "samtools index" before "bamcoverage". Does bamcoverage as bam to be sorted before using as input?

### Code

```
#!/bin/bash

# Define variables
INDEX="../data/indx/bowtie2_whole_genome/hg38"
CPUS=24
FASTQ1="ENCF000AVS_1m.fastq.gz"
FASTQ2="ENCF000AVS_10m.fastq.gz"

# Align reads to reference genome using bowtie2
bowtie2 -p $CPUS -x $INDEX -U $FASTQ1 | samtools sort -@ $CPUS -o ENCF000AVS_1m.sorted.bam
bowtie2 -p $CPUS -x $INDEX -U $FASTQ2 | samtools sort -@ $CPUS -o ENCF000AVS_10m.sorted.bam

# Index sorted BAM files using samtools
samtools index ENCF000AVS_1m.sorted.bam
samtools index ENCF000AVS_10m.sorted.bam

# Convert sorted BAM files to bigWig format using deepTools
bamCoverage -b ENCF000AVS_1m.sorted.bam -o ENCF000AVS_1m.bw --binSize 10 --normalizeUsing CPM
bamCoverage -b ENCF000AVS_10m.sorted.bam -o ENCF000AVS_10m.bw --binSize 10 --normalizeUsing CPM
```

## Result

```
1000000 reads; of these:
  1000000 (100.00%) were unpaired; of these:
    340421 (34.04%) aligned 0 times
    465276 (46.53%) aligned exactly 1 time
    194303 (19.43%) aligned >1 times
65.96% overall alignment rate
[bam_sort_core] merging from 0 files and 24 in-memory blocks...
10000000 reads; of these:
  10000000 (100.00%) were unpaired; of these:
    3398797 (33.99%) aligned 0 times
    4721200 (47.21%) aligned exactly 1 time
    1880003 (18.80%) aligned >1 times
66.01% overall alignment rate
[bam_sort_core] merging from 0 files and 24 in-memory blocks...
normalization: CPM
bamFilesList: ['ENCFF000AVS_lm.sorted.bam']
binLength: 10
numberOfSamples: None
blackListFileName: None
skipZeroOverZero: False
bed_and_bin: False
genomeChunkSize: None
defaultFragmentLength: read length
numberOfProcessors: 1
verbose: False
region: None
bedFile: None
minMappingQuality: None
ignoreDuplicates: False
chrsToSkip: []
stepSize: 10
center_read: False
samFlag_include: None
samFlag_exclude: None
minFragmentLength: 0
maxFragmentLength: 0
zerosToNans: False
smoothLength: None
save_data: False
out_file_for_raw_data: None
maxPairedFragmentLength: 1000
normalization: CPM
bamFilesList: ['ENCFF000AVS_10m.sorted.bam']
binLength: 10
numberOfSamples: None
blackListFileName: None
skipZeroOverZero: False
bed_and_bin: False
genomeChunkSize: None
defaultFragmentLength: read length
numberOfProcessors: 1
verbose: False
region: None
bedFile: None
minMappingQuality: None
ignoreDuplicates: False
chrsToSkip: []
stepSize: 10
center_read: False
samFlag_include: None
samFlag_exclude: None
```



**Figure legend:** Distribution of CTCF ChIP-Seq reads on human H1-hESC across a genomic region of 198 kb. Y-axis for the first two tracks (ENCF000AVS\_1m.bw and ENCF000AVS\_10m.bw) are normalized reads density (CPM: count per million), generated by running codes drafted by chatGPT. Y-axis for the last track represent fold change over control from ENCODE using all the sequencing reads (ENCF332TNJ.bigWig; 24.6 million).

**Assessment\***

The chatbot said that “*bamCoverage* requires that the input BAM file is sorted by genomic position” and generate code to satisfy this requirement. I confirmed that the missing of index error message was issued by *bamCoverage*. The code ran successfully and generated bigwig files for both fastq files.

The bigwig files were uploaded to IGV for visualization. We assessed the results by comparing to the bigwig from ENCODE. The results are generally consistent. The more sequencing depth, the more peaks will be observed.

**WRAPPING UP**

**Review the final codes**

Review the code comments added by ChatGPT and identify key steps, as represented by functions and parameters, that are missing in the initial prompts.

**Prompts**

I need to insert line-by-line comments to the below code which works well to address the needs for the data analysis task. Wait for my code.

**Code with explanations**

```

#!/bin/bash

# Set the path to the bowtie2 index
INDEX="../data/idx/bowtie2_whole_genome/hg38"

```

```

# Set the number of CPUs to use for alignment
CPUS=24

# Set the path to the first fastq file
FASTQ1="ENCFF000AVS_1m.fastq.gz"

# Set the path to the second fastq file
FASTQ2="ENCFF000AVS_10m.fastq.gz"

# Align the reads from the first fastq file using bowtie2, then sort the resulting SAM file and convert it to BAM using samtools
bowtie2 -p $CPUS -x $INDEX -U $FASTQ1 | samtools sort -@ $CPUS -o ENCFF000AVS_1m.sorted.bam

# Align the reads from the second fastq file using bowtie2, then sort the resulting SAM file and convert it to BAM using samtools
bowtie2 -p $CPUS -x $INDEX -U $FASTQ2 | samtools sort -@ $CPUS -o ENCFF000AVS_10m.sorted.bam

# Index the BAM file from the first fastq file using samtools
samtools index ENCFF000AVS_1m.sorted.bam

# Index the BAM file from the second fastq file using samtools
samtools index ENCFF000AVS_10m.sorted.bam

# Convert the sorted BAM file from the first fastq file to bigwig using deepTools, with a bin size of 10 and normalized using CPM
bamCoverage -b ENCFF000AVS_1m.sorted.bam -o ENCFF000AVS_1m.bw --binSize 10 --normalizeUsing CPM

# Convert the sorted BAM file from the second fastq file to bigwig using deepTools, with a bin size of 10 and normalized using CPM
bamCoverage -b ENCFF000AVS_10m.sorted.bam -o ENCFF000AVS_10m.bw --binSize 10 --normalizeUsing CPM

```

### Missing details in the initial prompts

1. Index the bam file, which needs to be sorted in prior
2. The bamCoverage has an option "--normalizeUsing" to specify the type of normalization to facilitate comparison between different libraries.

### Crafting the final prompts

The initial prompts were revised to include the missing details (in bold) discovered during code review.

Act as an experienced bioinformatician proficient in ChIP-Seq data analysis, you will assist me by writing code with number of lines as minimal as possible. Rest the thread if asked to. Reply "YES" if understand.

I have two fastq files in current folder from single-end sequencing of a ChIP-Seq library: ENCFF000AVS\_1m.fastq.gz, and ENCFF000AVS\_10m.fastq.gz. For each fastq file, align reads to the human reference genome, save to bam file, **sorted it, index it**, and then convert it to bigwig file with **CPM normalization**. Tools to use: bowtie2, samtools, and deepTools. The index for bowtie2 is in the folder "../data/indx/bowtie2\_whole\_genome/" with "hg38" as the prefix. Use 24 CPU for the alignment. Please draft the code in bash.

### Summary\*

#### Aims

The goal of this session is to optimize the prompts which will guide ChatGPT to generate code for aligning sequencing files (fastq) and then generate a summarized file (bigwig) for the distribution of reads across the genome.

#### Materials and Methods

The original fastq sequencing file was obtained from ENCODE (ENCFF000AVS). In this analysis session, we extracted the first one million reads and the first ten million reads from

the original file. In this way, we could tell how sequence depth impacts the finding TF bindings sites even from a visual inspection.

We aligned the sequencing file to the human reference (hg38) genome using Bowtie2 (v 2.4.2)<sup>1</sup> and converted the output to sorted and indexed BAM files using Samtools (v 1.10)<sup>2</sup>. We then used the *bamCoverage* (v 3.5.1) function from deepTools<sup>3</sup> to convert the indexed BAM file to a bigwig file. The bigwig files were visualized by IGV (v2.9.2)<sup>4</sup>.

### Results and Discussions

Visual inspection of the distribution of CTCF ChIP-Seq reads across the genome revealed a significantly higher number of peaks in the ten million reads dataset compared to the one million reads dataset. Data analysis, such as examining the signal at known CTCF target sites in ESC, conducting motif enrichment analysis, and comparing to other public CTCF ChIP-Seq datasets can be used as additional quality control for the current library.

### Additional Comments\*

Students are encouraged to test the robust of these prompts with new ChatGPT sessions. Alternatively, students could drop out certain details in the prompts to check what alternative solutions ChatGPT may come up with.

\*The writing has been polished by ChatGPT after an initial human draft

### References

- 1 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
- 2 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
- 3 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).
- 4 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).