Week 12, Lecture 23

István Albert

Biochemistry and Molecular Biology and Bioinformatics Consulting Center

Penn State
Protein - DNA interactions
ChIP-Chip and ChIP-Seq studies

• **ChIP** $\rightarrow$ Chromatin Immuno-Precipitation (refers to sample preparation)

• **Chip** $\rightarrow$ microarray technology to detect bound genomic locations

• **Seq** $\rightarrow$ high throughput sequencing to detect bound genomic locations
Chromatin Immuno-Precipitation

It is a well-known methodology to detect protein-DNA interactions:

• transcription factor binding
• polymerase binding
• chromatin structure and modifications
ChIP: Quick Overview

Double stranded DNA

P1

P2

Proteins

Crosslink bound proteins

Fragment/digest DNA around bound locations

Isolate with protein specific antibody

Reverse cross link (separate)

This is the DNA fragment that gets sequenced
Sample origins

Understanding the sample preparation is essential for analysis

• **WGS** whole genome sequencing (shotgun) → random DNA fragments covering the entire genome

• **Chip-Seq** → DNA fragments covering bound locations in the genome
The ChIP output

- a DNA sample enriched for fragments associated with the events under study

- BUT we are measuring an ensemble of cells that may be in different states!

- Coverage depends on the number of sites, efficiency of the IP (precipitation) step.

- Fragment accuracy depends on fragmentation strategy: sonication, MNASE digestion, lambda-nuclease digestion

**Note:** Lots of other DNA fragments can make it through!
Chip-Seq → High throughput sequencing

• Fragments are sequenced

• Aligned against genome (Bowtie is a good choice for Chip-Seq)

• The output is in the form of the intervals (start/end) where each read matches the genome
Sequencing proceeds from the 5’ to 3’. This is where we get reads from.

Original two stranded DNA fragment

the original DNA fragment could be longer/shorter than the length of the read
Other techniques: Chip-Exo

Instead of fragmentation nuclease digestion of the 5’ location.

Note how the reads may be longer than the bound location.
Type of events of interest

- Long-range
  (e.g., histone modifications)

- Mid-range
  (e.g., polymerase binding)

- Punctate
  (e.g., TF binding)
Alignment to the genome

After alignment we get genomic intervals for each read, minimally (SAM, BED or GFF):

\[
\text{chrom, start, end, strand}
\]

The fragment 5’ end locations of each fragment correspond to

the \textit{start} coordinate for the + strand
the \textit{end} coordinate for the – strand

Chip-Seq is about locations. The 5’ end matters!

The rest of the read is needed only for alignment
Other considerations

- For single end sequencing each fragment may correspond to 0, 1 or 2 reads. If it has two reads we don’t know which two formed the fragment.

- For paired and sequencing each fragment corresponds to 0, 1 or 2 reads. (1 no mate). We know which two reads correspond to one another → fragment size estimation.
Peak Calling

- Process of finding the locations enriched due to events of interest

We will need to define

- **Peak Region** - contiguous set of basepairs that belong to a peak

- **Enrichment Level** - read-based measure of supporting evidence
Peak calling: base pair level measurements

- Number of fragments overlapping that position (need to go from reads to fragments)

- Number of reads (fragment ends, midpoints) reported at that position (possibly, taking strandedness into account)

Variation: **kernel-smoothed read density.** This is closer to overlap approach.

(we will cover the peak calling in next lecture)
Practical exercise (also homework)

• We have two datasets → the same binding factor was simulated as if it had a short or a long footprint (long.fq, short.fq)

• We will visualize and investigate these: detect bound locations, fragment size, peak locations etc with one or more tools
The alignment script

```bash
# producing a BAM from a sequence file
# stop on error
set -ue
GENOME=refs/mini.fa
# data target
TARGET=$1
SAI=$TARGET.sai

#
~/bin/bwa aln $GENOME $TARGET > $SAI
~/bin/bwa samse $GENOME $SAI $TARGET > tmp.sam

# convert, sort and index to BAM
~/bin/samtools view -Sb tmp.sam > tmp.bam
~/bin/samtools sort tmp.bam $TARGET
~/bin/samtools index $TARGET.bam

# remove temporary files
rm -f tmp.sam tmp.bam $SAI
```
Visualize in IGV

strand bias
Deduplication

- Distinction needs to be made between **natural** vs **artificial** (PCR) duplicates

- There is no obvious consensus – the more accurate the method the more likely that we have **natural** duplicates

- Look for obvious flaws (strand biases), paired end sequencing helps identifying artificial duplicates

```
ialbert@porthos ~/work/lec23/data
$ samtools rmdup -s short.fq.bam short.fq.uniq.bam
[bam_rmdupse_core] 4451 / 13501 = 0.3297 in library '

ialbert@porthos ~/work/lec23/data
$ samtools index short.fq.uniq.bam

ialbert@porthos ~/work/lec23/data
$```
Deduplicated reads
"Poor" man's peak predictor

```
BEGIN {
    MINCOV = 10
    inside = 0
}

$3 >= MINCOV {
    if (inside == 0) {
        chrom = "$1"
        start = "$2"
        inside = 1
    }
}

$3 < MINCOV {
    if (inside == 1) {
        end = "$2"
        inside = 0
        print chrom, ",", ",", start, end, ",", ",", ",", ",";
    }
}

samtools depth short.fq.uniq.bam | awk -f poor.awk > short-peak.gff
```
“Poor man’s” peak prediction
Homework 23

• Using the data and scripts found in the file hw23.tar.gz on the webpage produce the plot seen on the previous slide.

• How many peaks can you find for a coverage of 10?