Part-II: Statistical analysis of ChIP-seq data

Outline

- ChIP-seq data, features, detailed modeling aspects (today).
- Other ChIP-seq related problems - overview (next lecture).
- IDR (next lecture)
Why study protein-DNA interactions?

Why study protein-DNA interactions?

A tale of GATA2 (Bresnick Lab, UW Madison).
Why study protein-DNA interactions?

A tale of GATA2 (Bresnick Lab, UW Madison).

---

### Diagram

- **Gata2** gene with a Sac I restriction site.
- **CTATCCGGACATCTGCAGCCGGTAGATAA** sequence.
- Enhancer region labeled:
  - +9.5 enhancer
- pfloxA2 insertion
- Pac I, Xba I, Nhe I restriction sites
- PGK:neo, HSV-TK elements
- +/+ embryonic stem cell image
Why study protein-DNA interactions?

A tale of GATA2 (Bresnick Lab, UW Madison).
Why study protein-DNA interactions?

A tale of GATA2 (Bresnick Lab, UW Madison).
Why study protein-DNA interactions?

A tale of GATA2 (Bresnick Lab, UW Madison).
Why study protein-DNA interactions?

A tale of GATA2 (Bresnick Lab, UW Madison).

Causal deletion: We identified a MonoMAC (a rare genetic disorder) patient with a heterozygous deletion that disrupts the +9.5 composite element. Median age of survival: 36.5 years.
High throughput ChIP assay (ChIP-seq): Chromatin immunoprecipitation combined with high throughput sequencing

- ChIP: Chromatin immunoprecipitation (in vivo).
- seq: High throughput sequencing (mainly Illumina for us).
- ChIP-seq: ChIP followed by high throughput sequencing.
ChIP assay

Target protein =
ChIP assay

1. Crosslink DNA and protein \textit{in vivo} by exposing cells to formaldehyde.

Target protein = \blue{\textbullet}
ChIP assay

1. Crosslink DNA and protein *in vivo* by exposing cells to formaldehyde.

2. Extract the chromatin from cells and fragment by sonication (~ 500-1000 bps).

Target protein =

1. Crosslink DNA and protein *in vivo* by exposing cells to formaldehyde.

2. Extract the chromatin from cells and fragment by sonication (~ 500-1000 bps).
ChIP assay

1. Crosslink DNA and protein \textit{in vivo} by exposing cells to formaldehyde.

2. Extract the chromatin from cells and fragment by sonication (~ 500-1000 bps).

3. Immunoprecipitate using a target protein-specific antibody.

Target protein = 

\begin{itemize}
  \item Selectively binds to the target protein.
\end{itemize}
ChIP assay

1. Crosslink DNA and protein \textit{in vivo} by exposing cells to formaldehyde.

2. Extract the chromatin from cells and fragment by sonication (~ 500-1000 bps).

3. Immunoprecipitate using a target protein-specific antibody.

4. Reverse the cross-links and purify DNA.

Target protein =

\begin{itemize}
  \item Selectively binds to the target protein.
\end{itemize}
ChIP assay

1. Crosslink DNA and protein in vivo by exposing cells to formaldehyde.

2. Extract the chromatin from cells and fragment by sonication (~ 500-1000 bps).

3. Immunoprecipitate using a target protein-specific antibody.

4. Reverse the cross-links and purify DNA.

5. Find the identity of the isolated DNA fragments.
Variations: MNase-seq for nucleosome occupancy

In high throughput experiments measuring nucleosome occupancy, an enzyme called "Micrococcal nuclease" is used to digest nucleosome free regions instead of sonication + immunoprecipitation.
Traditional ChIP assay

- The identity of the DNA fragments isolated in complex with the protein of interest can then be determined by polymerase chain reaction (PCR) using primers specific for the DNA regions that the protein in question is hypothesized to bind.
- One experiment per hypothesized region.
- Identify the identity of all the immunoprecipitated regions? Sequencing (ChIP-Seq) (previously with ChIP-chip).
ChIP-seq data generation
ChIP-seq data generation

@HWI-EAS00184:1:1:8:1368#0/1
CGTCACATCCCATACACCACAATGCNNACCCAGGGCCTTA
+HWI-EAS00184:1:1:8:1368#0/1
Y\Wb`QYPW`GFRT]_\aXaBBB BBBB BBB BB BBBB BBBB
@HWI-EAS00184:1:1:8:1763#0/1
TAAACCATCATTCTTAGCAAACTATNNCGAGTACATAAAA
+HWI-EAS00184:1:1:8:1763#0/1
`a`bbbabbbbbbb]aa_^ababaHDD[\bNH]bbLb_bb
ChIP-seq data generation

@HWI-EAS00184:1:1:8:1368#0/1
CGTCACATCCCATACACCACAATGCNNACCCAGGGCCTTA
+HWI-EAS00184:1:1:8:1368#0/1
Y\Wb`QYPW`GFRT]__aXaBBBBBBBBBBBBBBBBBBBBB
@HWI-EAS00184:1:1:8:1763#0/1
TAAACCATCATTCTTAGCAAACTATNNCGAGTACATAAAA
+HWI-EAS00184:1:1:8:1763#0/1
`a`bbbbbbbbbb]aa_^ababaHDD[\bNH]bbLb_bb

Align to reference genome

@HWI-EAS00184:1:1:8:1368#0/1 - chr12 29078411 CGTCACATCCCATACACCACAATGCNNACCCAGGGCCTTA
@HWI-EAS00184:1:1:8:1763#0/1` + chr5 129332046 TAAACCATCATTCTTAGCAAACTATNNCGAGTACATAAAA
ChIP-seq data generation

Machine output

Align to reference genome

Create count data over 100-200 bps non-overlapping genomic intervals

chr7  812800  0
chr7  813000  20
chr7  813200  32
chr7  813400  3
chr7  813600  0
ChIP-seq data generation

---

**Machine output**

@HWI-EAS00184:1:1:8:1368#0/1
CGTCACATCCATACACCACAATGCNNACCGCCTTA
+HWI-EAS00184:1:1:8:1368#0/1
Y`Wb`QYPW`GFRT]\_\_aXaBBBBBBBBBBBBBBBBB
@HWI-EAS00184:1:1:8:1763#0/1
TAAACCATTCTTCTTAGCAAACCTATNNCGAGTAATAA
+HWI-EAS00184:1:1:8:1763#0/1
`a`bbbabbbbbbbbaa_^ababaHDD[\bNH]bbL_b_bb

---

**Align to reference genome**

**Only keep uniquely aligning reads**

@HWI-EAS00184:1:1:8:1368#0/1  -  chr12  29078411  CGTCACATCCATACACCACAATGCNNACCGCCTTA
@HWI-EAS00184:1:1:8:1763#0/1  +  chr5  129332046  TAAACCATTCTTCTTAGCAAACCTATNNCGAGTAATAA

**Create count data over 100-200 bps non-overlapping genomic intervals**

<table>
<thead>
<tr>
<th>chr</th>
<th>Count</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr7</td>
<td>812800</td>
<td>0</td>
</tr>
<tr>
<td>chr7</td>
<td>813000</td>
<td>20</td>
</tr>
<tr>
<td>chr7</td>
<td>813200</td>
<td>32</td>
</tr>
<tr>
<td>chr7</td>
<td>813400</td>
<td>3</td>
</tr>
<tr>
<td>chr7</td>
<td>813600</td>
<td>0</td>
</tr>
</tbody>
</table>
ChIP-seq data structure

Protein

ACGCGTCACGTCAGTCACGCGCTAGATAGATAGCCGCTAGCGAGAGAGA
ChIP-seq data structure

Total tag counts in one "bin"

Protein

ACGCGTCAAGTCAAAACGCGCTAGATAGATAGCCGCTAGCGAGAGAGA

ACGCGTCAAGTCAAAACGCGCTAGATAGATAGCCGCTAGCGAGAGAGA
ChIP-seq data structure

Protein

Total tag counts in one “bin”

ACGCGTCACGTCAAAAACGCCTAGATAGATAGC CGCTAGCGAGAGAGA

Protein

ACGCGTCACGTCAAAAACGCCTAGATAGATAGCCGCTAGCGAGAGAGA
ChIP-seq data structure
ChIP-seq data structure

Basic inference problem in ChIP-seq data: Identifying "peaks"

ChIP experiments are typically accompanied by a control experiment that measures "background".
Outline

- Q1: Basic inference for ChIP-seq data: Finding peaks.
- Q3: Power (sequencing depth) calculations.
- Q4: Differential occupancy: Comparing peaks across conditions.
- Q5: Detecting binding events within close proximity of each other: Deconvolving peaks.
- Q6: Joint analysis of multiple ChIP-seq datasets.
FASTQ format

@HWUSI-EAS1789_0000:5:1:1049:9966#GGCTAN/1
CAGAAGTGACATCAACATGATTTAGCTTGTAT
+HWUSI-EAS1789_0000:5:1:1049:9966#GGCTAN/1
'ddadc^aYc\b\Ybc^d~~cdd\cdddcacc\cc\d

- The first line begins with an @ symbol and is followed by the sequence name.
  @lane:tile:x_coordinate_on_tile:y_coordinate_on_tile:quality_filter

- The second line contains the base call (in this case for each of 36 nucleotides).

- The third line begins with a + symbol and may (or may not) repeat the sequence name.

- The fourth line contains a symbol that measures the quality score for the corresponding base call as listed on the second line. There should be one symbol for each base call. The symbol on the fourth line uses an ASCII character (American Standard Code for Information Interchange) to encode the quality score.
Mapping reads to reference genome: Tools

- Eland
- Bowtie (*)
- MAQ
- Bwa (*)
Mapping reads to reference genome

Sam file format:

AMELIA:402:C53G2ACXX:1:1101:1778:2111 0 chr6 100308392 255 101M * 0 0
CTGCTTTAGCTGTGTCCAGAGATCTGATACATTGTGTCCTTGTGTTCTGACATTGGTTTTGTTTTATTTCTGCTTTAATTTACTTATGTACCCAA
@@@DDDDADHHHDAG?FHEF;F+AF:A?4<A999:?EEDC
GGF@GC9D?)6)9BDFGC<B8=)8FF78.=C@GEE>CE;?EDBDECDC:@6>;6C##### XA:i:1 MD:Z:47T53 NM:i:1

The read aligns to chr 6 position 100308392 with 1 mismatch [MD:Z:47T53].
101M: CIGAR string - used to indicate which bases align (either a match/mismatch) with the reference, are deleted from the reference, and are insertions that are not in the reference.
NM:i:1: Refers to edit distance to the genome.
XA:i:1: Alternative hits.

As a result of mapping, we obtain the "observations/measurements" that we can do statistical inference on.
Practical challenges of dealing with Next Gen data

- Useful to know some scripting language, e.g. perl, python.
- Storage of the data is a big problem.
  - An aligned read file with 95 million reads is around 15-20GB. This is typically "one sample" for us.
  - Then for each treatment sample we typically have a control sample.
  - If we are going beyond identifying peaks, e.g., differential binding etc, we have at least 2 reps per treatment. A simple study adds up to 8 samples (or to 200 GB).
  - The good news is that, often, we can build statistical inference on data extracted from the aligned files.
  - In our work, we call these bin files. We partition the genome into non-overlapping intervals of 200 bps and count the number of reads falling into each interval – reduces size to 200MB.

<table>
<thead>
<tr>
<th>chr5</th>
<th>152001400</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr5</td>
<td>152001600</td>
<td>12</td>
</tr>
<tr>
<td>chr5</td>
<td>152001800</td>
<td>20</td>
</tr>
<tr>
<td>chr5</td>
<td>152002000</td>
<td>20</td>
</tr>
<tr>
<td>chr5</td>
<td>152002200</td>
<td>13</td>
</tr>
<tr>
<td>chr5</td>
<td>152002400</td>
<td>6</td>
</tr>
</tbody>
</table>
ChIP-Seq vs. Control experiments: Sono-Seq (Input-Seq), Naked-DNA-Seq

<table>
<thead>
<tr>
<th>ChIP-Seq and Sono-Seq</th>
<th>Naked DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crosslink cells</td>
<td>No crosslinking</td>
</tr>
<tr>
<td>Isolate and sonicate nuclei</td>
<td>Degrade protein &amp; RNA, sonicate, size select and purify DNA</td>
</tr>
</tbody>
</table>

ChIP:
- IP with Pol II antibody
- Reverse crosslinks, degrade protein & RNA, size select and purify DNA

Sono-Seq:
- reverse crosslinks, degrade protein & RNA, size select and purify DNA

Auerbach et al.
PNAS 2009; 106:14926-14931.
ChIP-seq QC

Landt et al., (2012), ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia Genome Res. 2012 Sep; 22(9):1813-1831.

\[
\text{NSC} = \frac{cc(\text{fragment length})}{\text{min}(cc)}
\]

\[
\text{RSC} = \frac{cc(\text{fragment length}) - \text{min}(cc)}{cc(\text{read length}) - \text{min}(cc)}
\]
A key plot for ChIP-seq data

Counts

Counts

0
100
200
300
400
r1 Input
r1 ChIP
r1 ChIP vs Input 1.491

Counts

3617187
1407579
547740
213146
82943
32276
12560
4887
1902
740
288
112
44
17
7
3
1

0 100 200 300 400
0
100
200
300
400
r1 Input
A key plot for ChIP-seq data

r2 ChIP vs Input 6.665

Counts

990983
418131
176424
74440
31409
13252
5592
2359
995
420
177
75
32
13
6
2
1

Counts

Stat 877 (Spring’17) 04/11-04/18 97 / 237
A key plot for ChIP-seq data

r1 ChIP vs r2 ChIP 4.47

Counts

Counts:
- 929749
- 393861
- 166848
- 70680
- 29942
- 12684
- 5373
- 2276
- 964
- 408
- 173
- 73
- 31
- 13
- 6
- 2
- 1

Stat 877 (Spring’17)

04/11-04/18 98 / 237
A key plot for ChIP-seq data

r1 ChIP vs r2 ChIP 4.47

Counts

Counts

929749
393861
166848
70680
29942
12684
5373
2276
964
408
173
73
31
13
6
2
1

Stat 877 (Spring’17)
A key plot for ChIP-seq data

SA−TTX ChIP vs Input 1.5

Counts

SA−TTX Input

SA−TTX ChIP

Counts

1668912
681603
278374
113691
46433
18964
7745
3163
1292
528
215
88
36
15
6
2
1

Count values shown in the legend correspond to the increasing gradient of the plot.

Stat 877 (Spring'17)
ChIP-seq data structure

Basic inference problem in ChIP-seq data: Identifying "peaks"

ChIP experiments are typically accompanied by a control experiment that measures "background". How would the data look like under the null distribution? ⇒ Same threshold for all the peaks? Same null distribution along the genome?
### ChIP-Seq vs. Control experiments: Sono-Seq (Input-Seq), Naked-DNA-Seq

<table>
<thead>
<tr>
<th>ChIP-Seq and Sono-Seq</th>
<th>Naked DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crosslink cells</td>
<td>No crosslinking</td>
</tr>
<tr>
<td>Isolate and sonicate nuclei</td>
<td>Degrade protein &amp; RNA, sonicate, size select and purify DNA</td>
</tr>
<tr>
<td><strong>ChIP:</strong> IP with Pol II antibody</td>
<td></td>
</tr>
<tr>
<td><strong>Sono-Seq:</strong> reverse crosslinks, degrade protein &amp; RNA, size select and purify DNA</td>
<td></td>
</tr>
<tr>
<td>Reverse crosslinks, degrade protein &amp; RNA, size select and purify DNA</td>
<td></td>
</tr>
</tbody>
</table>

Auerbach et al.  
*PNAS 2009; 106:14926-14931.*
ChIP-Seq, Sono-Seq, Naked-DNA-Seq

STAT1

Tag count

Frequency

STAT1 ChIP

Stat 877 (Spring'17)
ChIP-Seq, Sono-Seq, Naked-DNA-Seq

STAT1

Frequency

Tag count

0 9 99
1 10 100 1000 10000 1e+05

Stim Input DNA

Naked DNA

STAT1 ChIP

04/11-04/18 105 / 237
**Mappability and GC biases in ChIP-Seq data**

- **Mappability bias:** due to retaining only uniquely aligning tags. Rozowsky *et al.* (2009).
  - 79.6% of the human genome is uniquely mappable using 30bp tags.
  - (91.1% for 75bp tags).

- **GC bias:** Dohm *et al.* (2008), Vega *et al.* (2009).

> Background for ChIP-Seq data is not uniform $\implies$ need for region/location specific cut-offs for calling peaks.

> Naked DNA sequencing data provides an excellent platform to investigate and model these effects.
Sequence bias in ChIP-Seq data: Mean tag count vs Mappability (bin-level) and GC content: HeLa S3 Naked-DNA-Seq (GSE14022)

Same relationships hold for data from mouse.
Existing methods:
- Poisson distribution.
- Negative binomial distribution (CisGenome, Ji et al. (2008)).

Excess zeroes and over-dispersion.

Mappability and GC bias $\rightarrow$ bin specific distributions.
Poisson vs. Negative Binomial Distributions

\[ Y \sim \text{Poisson}(\lambda), \]
\[ \implies E[Y] = \lambda, \text{var}[Y] = \lambda. \]

\( a \): shape, \( b \): scale.

\[ Y \sim \text{NegBin}(a, b) \]
\[ \implies E[Y] = a/b, \text{var}[Y] = a(1 + b)/b^2. \]

Alternative parametrization (in \( \mathbb{R} \))

\[ Y \sim \text{NegBin}(\rho, \mu) \]
\[ \implies E[Y] = \mu, \text{var}[Y] = \mu + \mu^2/\rho, \]

where \( \rho = a, \mu = a/b. \)

1/\( \rho \) is referred to as the "dispersion" parameter.
Background/Null model for one sample ChIP-Seq data

\( Y_j \): observed tag counts for bin \( j \).
\( N_j \): background tag counts for the bin.
\( M_j \): average mappability score.
\( GC_j \): average GC content.

Non-homogeneous background

\[ Y_j \sim N_j \]
\[ N_j | \mu_j \sim g(\mu_j) \]

Candidate models for \( \mu_j \)

1. \( \mu_j = \exp(\beta_0) \)
2. \( \mu_j = \exp(\beta_0 + \beta M \log_2(M_j + 1)) \)
3. \( \mu_j = \exp(\beta_0 + \beta_{GC} GC_j) \)
4. \( \mu_j = \exp(\beta_0 + \beta M \log_2(M_j + 1) + \beta_{GC} GC_j) \)
5. \( \mu_j = \exp(\beta_0 + \beta_{GC} Sp(GC_j)) \)
6. \( \mu_j = \exp(\beta_0 + \beta M \log_2(M_j + 1) + \beta_{GC} Sp(GC_j)) \)

Candidate models for \( g(\mu_j) \)

1. \( g(\mu_j) \sim Po(\mu_j) \) (Poi Reg)
2. \( g(\mu_j) \sim NegBin(a, a/\mu_j) \) (NegBin Reg)

CisGenome (Ji et al. (2009))

\[ Y_j \sim NegBin(a, b) \]
Goodness of Fit

HeLa S3 NakedDNA

Black line: Actual data

Stat 877 (Spring'17)
Goodness of Fit: CisGenome NegBin

HeLa S3 NakedDNA

Black line: Actual data  
Red line: Simulated data from the fitted CisGenome NegBin

Stat 877 (Spring'17)
Goodness of Fit: CisGenome NegBin

HeLa S3 NakedDNA

- Actual data
- Sim: CisGenome NegBin

Over-estimated background!

Stat 877 (Spring’17)
Goodness of Fit: Poisson Reg \((M, GC)\)

Black line: Actual data
Red line: Simulated data from the fitted Poisson Reg\((M, GC)\)
Goodness of Fit: NegBin Reg (GC)

Black line: Actual data
Red line: Simulated data from the fitted NegBin Reg (GC)
Goodness of Fit: NegBin Reg ($M$)

HeLa S3 NakedDNA

Black line: Actual data
Red line: Simulated data from the fitted NegBin Reg ($M$)
Goodness of Fit: NegBin Reg \((M, GC)\)

HeLa S3 NakedDNA

- Black line: Actual data
- Red line: Simulated data from the fitted NegBin Reg \((M, GC)\)

**Black line: Actual data**

**Red line: Simulated data from the fitted NegBin Reg \((M, GC)\)**
Goodness of Fit

Bayesian Information Criterion (BIC) for model selection (smaller better):

None > GC > M > M + GC > M + Sp(GC).
Mixture model for one-sample ChIP-Seq data

- \( Z_j \): unknown/latent state, \( Z_j = 1(0) \) if bound (unbound).
- \( Y_j | Z_j = 0 \sim N_j \) \( \text{NegBin Reg}(M_j, GC_j) \)
- \( Y_j | Z_j = 1 \sim N_j + S_j \)
- \( S_j \): protein-binding signal
  1. \( S_j \sim \text{NegBin}(b_1, c_1) \) (1-component)
  2. \( S_j \sim p_1 \text{NegBin}(b_1, c_1) + (1 - p_1) \text{NegBin}(b_2, c_2) \) (2-component)

Estimate unknown parameters with maximum likelihood method using the EM algorithm.
Case Study 1: STAT1 ChIP-Seq data

- STAT1 in IFN-γ-stimulated HeLa S3 ChIP-Seq data (GSE12782).
- 6 lanes of Illumina sequencing data, 23 million mapped reads.

Graphs showing mappability and GC content for HeLa S3 STAT1.
Case Study 1: STAT1 ChIP-Seq data
## Case Study 1: STAT1 ChIP-Seq data

<table>
<thead>
<tr>
<th>Tag count</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>1000</td>
<td>10000</td>
</tr>
<tr>
<td>10000</td>
<td>1e+05</td>
</tr>
<tr>
<td>1e+06</td>
<td></td>
</tr>
</tbody>
</table>

**Background:**

- NegBin
- Reg(M, GC)

---

[Graph showing actual data and background distribution]
Case Study 1: STAT1 ChIP-Seq data

**STAT1 ChIP**

<table>
<thead>
<tr>
<th>Tag count</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>10000</td>
<td>1e+05</td>
</tr>
<tr>
<td>1e+06</td>
<td></td>
</tr>
</tbody>
</table>

- **Actual Data**
- **Background:** NegBin
- **Background:** CisGenome NegBin

Background: NegBin Reg(M, GC)
Case Study 1: STAT1 ChIP-Seq data

<table>
<thead>
<tr>
<th>Tag count</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>10000</td>
<td>1e+05</td>
</tr>
<tr>
<td>1e+06</td>
<td></td>
</tr>
</tbody>
</table>

Actual Data

Background: NegBin Reg(M, GC)

NegBin Reg(M, GC) Background + 1-comp signal
Case Study 1: STAT1 ChIP-Seq data

<table>
<thead>
<tr>
<th>STAT1 ChIP</th>
<th>Tag count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2-9</td>
<td>99</td>
</tr>
<tr>
<td>1-10</td>
<td>100</td>
</tr>
<tr>
<td>11-99</td>
<td>1000</td>
</tr>
<tr>
<td>1000-9999</td>
<td>10000</td>
</tr>
<tr>
<td>10000-99999</td>
<td>1e+05</td>
</tr>
<tr>
<td>1e+05-1e+06</td>
<td>1e+06</td>
</tr>
</tbody>
</table>

Actual Data

Background:
NegBin Reg(M, GC)

Background + 2-comp signal

Background:
NegBin Reg(M, GC)
MOSAiCS with Input-Seq: Two-sample analysis

$Y_j$: tag count from ChIP-Seq;

$X_j$: tag count from Input-Seq.
ChIP-Seq vs Input-Seq: Does Input-Seq account for all the M and GC bias?
ChIP-Seq vs Input-Seq: Does Input-Seq account for all the M and GC bias?

- The diagram shows a scatter plot with the x-axis representing GC content and the y-axis representing the mean ChIP tag count.
- The plot includes multiple panels with data points indicating differences between Input-Seq and ChIP-Seq.
- The data suggests that Input-Seq does not fully account for the M and GC bias.
MOSAiCS two-sample background model

\[ Y_j \mid Z_j = 0, X_j, M_j, GC_j \sim \text{NegBin}(a, a/\mu_j) \]
\[ \mu_j = \exp(\beta_0 + f(M_j, GC_j, X_j)) \]

where

\[
\begin{align*}
f(M, GC, X) &= I(X \leq c) \left[ \beta_M \log_2(M + 1) + \beta_C Sp(GC) + \beta_X^1 X^d \right] \\
&\quad + I(X > c) \beta_X^2 X^d
\end{align*}
\]
MOSAiCS two-sample background model

\[ Y_j \mid Z_j = 0, X_j, M_j, GC_j \sim \text{NegBin}(a, a/\mu_j) \]
\[ \mu_j = \exp(\beta_0 + f(M_j, GC_j, X_j)) \]

where

\[ f(M, GC, X) = I(X \leq c) \left[ \beta_M \log_2(M + 1) + \beta_C \text{Sp}(GC) + \beta_X^1 X^d \right] \]
\[ + I(X > c) \beta_X^2 X^d \]
MOSAiCS two-sample model GOF for STAT1 ChIP-Seq

STAT1 ChIP

Actual Data

Tag count

Frequency

0 9 99
1 10 100 1000 10000 1e+05

Actual Data

04/11-04/18 133 / 237
MOSAiCS two-sample model GOF for STAT1 ChIP-Seq

STAT1 ChIP

Tag count

Frequency

Actual Data

Background: NegBin Reg(Input)

Stat 877 (Spring'17)
MOSAiCS two-sample model GOF for STAT1 ChIP-Seq

STAT1 ChIP

Tag count  Frequency
0  9  99
1  10  100  1000  10000  1e+05

Actual Data

Background:
NegBin Reg(Input)

Background:
NegBin Reg(M, GC, Input)

Stat 877 (Spring'17)
04/11-04/18 135 / 237
MOSAiCS two-sample model GOF for STAT1 ChIP-Seq

STAT1 ChIP

- Actual Data
- NegBin Reg(M, GC, Input)
- Background + Signal
- Background: NegBin Reg(Input)
- Background: NegBin Reg(M, GC, Input)

Tag count

Frequency

- MOSAiCS two-sample for GATA1

Stat 877 (Spring’17)
STAT1 motif scans

Proportion of peaks with motif vs. Rank of peaks for different peak calling tools:
- MOSAiCS−OS
- PeakSeq−OS
- CisGenome−OS
- MACS−OS
- MOSAiCS−TS (Input+M+GC)
- MOSAiCS−TS (Input only)
- PeakSeq−TS
- CisGenome−TS
- MACS−TS

Pairwise comparisons

Information content vs. Position

0.4 0.5 0.6 0.7 0.8 0.9
2000 4000 6000 8000 10000

Rank of peaks

Proportion of peaks with motif

MOSAiCS−OS
PeakSeq−OS
CisGenome−OS
MACS−OS
MOSAiCS−TS (Input+M+GC)
MOSAiCS−TS (Input only)
PeakSeq−TS
CisGenome−TS
MACS−TS
CTCF motif scans

ENCODE pipeline evaluation

ENCODE Peak Caller
Evaluation Subgroup – 2013
Courtesy of Gifford Lab.

CTCF-GM12878-Rep0 spatial resolution

2271 motifs shared by 10 methods

Cumulative fraction (%)

Spatial resolution (distance from CTCF motif, bp)
If the sample is deeply sequenced

E.g.,

- *E. coli* ChIP-seq - one lane from the Illumina GA-II.
- Higher eukaryote ChIP-seq sample on Illumina Hi-seq.
MOSAiCS on the FNR ChIP-seq data from the Kiley Lab @ UW Madison

GOF on log scale

Actual data (ChIP)  
Actual data (Control)  
Sim:N  
Sim:N+S1  
Sim:N+S1+S2

Tag count  
Frequency
MOSAiCS on the FNR ChIP-seq data from the Kiley Lab @ UW Madison

If the sample is deeply sequenced,
Input-only model fits well.

GOF on log scale
Software implementation: R package mosaics

Available through

- This recent version has more post-processing options.
- MOSAiCS-HMM: Extension to histone ChIP-seq.
R package: mosaics

ChIP file has 95 million reads.
Input file has 14 million reads.
mosaics runs in about 2 hrs using a single CPU.

```r
library(mosaics)
mosaicsRunAll(
    chipDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/",
    chipFileName="GSM746584_tal1_ter119_r2_mapped.txt",
    chipFileFormat="bowtie",
    controlDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/",
    controlFileName="GSM746580_input_ter119_mapped.txt",
    controlFileFormat="bowtie",
    binfileDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/",
    peakDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/peak/",
    peakFileName="tal1_ter119_peak_list_r2.txt",
    peakFileFormat="txt",
    reportSummary=TRUE,
    summaryDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/reports/",
    summaryFileName="mosaics_summary_tal1_ter119_r2.txt",
    reportExploratory=FALSE,
    exploratoryDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/reports/",
    exploratoryFileName="mosaics_exploratory_tal1_ter119_r2.pdf",
    reportGOF=TRUE,
    gofDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/reports/",
    gofFileName="mosaics_GOF_tal1_ter119_r2.pdf",
    byChr=FALSE,
    FDR=0.05,
    fragLen=200,
    binSize=200,
    capping=0,
    analysisType="IO",
)```
R package: mosaics

d=0.25,
signalModel="BIC",
maxgap=200,
minsize=50,
thres=40,
nCore=20)
Info: constructing bin-level files...
------------------------------------------------------------
Info: setting summary
------------------------------------------------------------
Directory of aligned read file: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/
Name of aligned read file: GSM746584_tal1_ter119_r2_mapped.txt
Aligned read file format: Bowtie default
Directory of processed bin-level files: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/
Construct bin-level files by chromosome? N
Fragment length: 200
Bin size: 200
------------------------------------------------------------
Info: setting summary
------------------------------------------------------------
Directory of aligned read file: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/
Name of aligned read file: GSM746580_input_ter119_mapped.txt
Aligned read file format: Bowtie default
Directory of processed bin-level files: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/
Construct bin-level files by chromosome? N
Fragment length: 200
Bin size: 200
------------------------------------------------------------
Info: reading the aligned read file and processing it into bin-level files...
Info: reading the aligned read file and processing it into bin-level files...
Info: done!
Info: processing summary

Directory of processed bin-level files: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/
Processed bin-level file: GSM746580_input_ter119_mapped.txt_fragL200_bin200.txt

Info: done!

Info: processing summary

Directory of processed bin-level files: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/
Processed bin-level file: GSM746584_tal1_ter119_r2_mapped.txt_fragL200_bin200.txt
R package: mosaics

Info: analyzing bin-level files...
Info: fitting MOSAiCS model & call peaks...
Info: reading and preprocessing bin-level data...
Info: data contains more than one chromosome.
Info: done!

Info: background estimation method is determined based on data.
Info: background estimation based on robust method of moment
Info: two-sample analysis (Input only).
Info: use adaptive griding.
Info: fitting background model...
Info: done!
Info: fitting one-signal-component model...
Info: fitting two-signal-component model...
Info: calculating BIC of fitted models...
Info: done!
Info: use two-signal-component model.
Info: calculating posterior probabilities...
Info: calling peaks...
Info: done!
Info: writing the peak list...
Info: peak file was exported in TXT format:
Info: file name = tal1_ter119_peak_list_r2.txt
Info: directory = /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/peak/
Info: generating reports...

> proc.time()
  user  system elapsed
7154.502  114.432  6986.742

Run time ≈ 2 hours.
R package: mosaics - Main output

- A peak list with columns: chrID peakStart peakStop peakSize aveP minP aveChipCount maxChipCount aveInputCount aveInputCountScaled aveLog2Ratio.
- Exploratory plots. Mean read count vs. Input, vs. Mappability, vs. GC content.
- Goodness-of-fit plots.
Summary: MOSAiCS

- implements a model-based approach for ChIP-seq data analysis;
- available as a R package through Bioconductor and as a Galaxy tool from the Galaxy tool shed;
- provides basic pre-processing functions for ChIP-seq data;
- provides functions for generating wig formatted data to upload to UCSC Genome Browser;
- provides plotting functions;
- an HMM extension to accommodate histone modifications;
- a fully ENCODE compatible MOSAiCS pipeline (finds summits, utilizes IDR (irreproducible discovery rate) for deciding on the number of peaks).