Analyzing ChIP-Seq Data with SICER

Chongzhi Zang, PhD

Department of Biostatistics and Computational Biology
Center for Functional Cancer Epigenetics
Dana-Farber Cancer Institute
Harvard T.H. Chan School of Public Health

NCI BTEP Workshop on ChIP-seq Analysis
May 17, 2016
Outline

• ChIP-seq overview
• Characteristics of histone ChIP-seq data
• SICER algorithm
• Hands-on SICER tutorial
ChIP-seq overview
ChIP-seq is used to study the *in vivo* genome-wide location of a transcription factor or a histone modification.
ChIP-seq profiles reveal gene regulatory functions of histone modifications

![Graph showing differential expression and fractions of enhancers](image-url)

Wang, Zang et al. *Nat Genet* 2008
Public ChIP-seq data are skyrocketing
We are entering the “Big Data” era

Number of ChIP-seq datasets on GEO
How ChIP-Seq is done
ChIP-seq data analysis

• Where in the genome do these sequence reads come from? - Sequence alignment and quality control

• What does the enrichment of sequence reads mean? - Peak calling (e.g. SICER, MACS)

• What can we learn from these data? – Downstream analysis and integration
ChIP-Seq data analysis overview: basic processing

• alignment of each sequence read: **bowtie** or **BWA**

  - cannot map to the reference genome
  - can map to multiple loci in the genome
  - can map to a unique location in the genome

• redundancy control:

  ![Redundancy Control Diagram]

ChIP-Seq data analysis overview: basic processing

- DNA fragment size estimation

- pile-up profiling

- Data visualization:
  - UCSC genome browser
  - IGV
  - WashU Browser

\[
C(r) = \frac{1}{X} \int_x \left( T_+(x) - \bar{T}_+ \right) \left( T_-(x + r) - \bar{T}_- \right)
\]
ChIP-Seq data analysis overview: peak calling

- **Sharp peaks**
  - transcription factor binding,
  - DNase HS

- **Broad peaks**
  - histone modifications,
  - “super-enhancers”
  - Diffuse

**MACS** (Zhang, 2008)

**SICER** (Zang, 2009)
Spatial clustering of localized weak signal and integrative Poisson model

Wang, Zang et al. 2014
Characteristics of histone ChIP-seq data

In other words, how to call “peaks” from such diffuse ChIP-seq data?
Histone modification patterns are diffuse

Characteristics:
• Noisy
• Unlike transcription factors
• Enriched regions are spread out
• Lack saturation

Why?
Histone modification tends to spread out

Domain formation model for repressive marks

- Yeast: HP1 H3K9me3
- Drosophila: PC1/PC2 H3K27me3
SICER: Motivation

- To detect broad/diffuse signals from ChIP-Seq
- Make use of the underlying biology
  - domain formation of histone modifications
- Account for background biases and provide statistical significance
SICER: Spatial-clustering method for Identification of ChIP-Enriched Regions
SICER: Definition of Island

- Eligible and ineligible windows
  \[ \sum_{l=l_0}^{\infty} P(l, \lambda) \leq p_0 \]
- Eligible windows are separated by gaps of ineligible windows.
- Island: cluster of eligible windows separated by gaps of size at most \(g\) windows.

Example islands for \(l_0 = 2\) and \(g = 2\)

Zang et al. *Bioinformatics* 2009
SICER: Scoring islands

• The scoring function is based on the probability of finding the observed tag count in a random background.

• For a window with \(m\) reads,
  - The probability of finding \(m\) reads is Poisson \(P(m, \lambda)\)
  - \(\lambda = wN/L\) is the average number of reads in each window

• Scoring function for an eligible window:

\[ S = -\ln P(m, \lambda) \]

• Key quantity: the score of an island
  - Aggregate score of all eligible windows in the island
  - It corresponds to the background probability of finding the observed pattern

Zang et al. Bioinformatics 2009
SICER: Island score statistics

• Probability distribution of scores for a single window in a random background model:

\[ \rho(s) = \sum_{l \geq l_0} \delta(s - s(l))P(l, \lambda) \]

• Probability of a window being ‘ineligible’:

\[ t = P(0, \lambda) + P(1, \lambda) + \cdots + P(l_0 - 1, \lambda) \]

• Gap factor:

\[ G = 1 + t + t^2 + \cdots + t^g \]
SICER: Island score statistics

- Recursion relation

\[ \tilde{M}(s) = G(\lambda, l_0, g) \int_{s_0}^{s} ds' \tilde{M}(s - s') \rho(s') \]

- Probability of finding an island of score \( s \):

\[ M(s) = t^{g+1} \tilde{M}(s) t^{g+1} \]

Zang et al. *Bioinformatics* 2009
SICER: Island score statistics

- Asymptotics of island score distribution in the random background
  \[ \tilde{M}(s) = \alpha \exp(-\beta s) \]

  \[ G(\lambda, l_0, g) \sum_{l \geq l_0} P(l, \lambda)^{1-\beta} = 1 \]

- Statistic: \( E \)-value
  - Expected number of islands with score above \( s_T \) in the background
    \[ \sum_{s \geq s_T} LM(s) \leq e \]

Zang et al. *Bioinformatics* 2009
SICER: Significance determinations

• Significance determination with random background model:
  – \( E \)-value determines an island score threshold

• Significance determination with control sample
  – Identify candidate islands using random background
  – For each candidate island, compare sample with control
  – \( P \)-value \( \sum_{n=n_s}^{\infty} P(n_s, cn_c) \)
  – False Discovery Rate (FDR)

Zang et al. *Bioinformatics* 2009
SICER: Choosing parameters

- Fragment size
- Window size: data resolution
- Gap size:

Zang et al. *Bioinformatics* 2009
SICER: evaluation

- Compared with other methods, SICER focuses on the clustered enrichment rather than local enrichment.
- A schematic illustration:

- SICER can identify clustered enriched regions from diffuse data.
SICER: Work flow

1. **ChIP-Seq library**
   - Given a window size and chromatin fragment size, bin adjusted reads.

2. **summary/wig files**
   - Set gap size, identify and score Islands

3. **Islands**

4. **control library**
   - Use a lenient E-value based on random background for pre-screening
   - Use control library as background to calculate p-value and q-value for each candidate island.
   - Set a p-value or FDR threshold.

5. **Y**
   - **candidate Islands**
   - Determine significance of islands based on a random background model. Set an E-value threshold.

6. **N**
   - Island summary

7. **significant Islands**

---

Zang et al. *Bioinformatics* 2009
SICER: Installation

- Download source code:
  
  http://home.gwu.edu/~wpeng/Software.htm

  Requirements: python and scipy
  (www.scipy.org)

- Galaxy

  https://usegalaxy.org/

- Genomatrix
ChIP-seq data examples

- [http://cistrome.org/~czang/chipseqdata.htm](http://cistrome.org/~czang/chipseqdata.htm)

- Data format requirement:
  Mapped reads, BED format, 6 columns

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Score</th>
<th>Strandedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr11</td>
<td>10344210</td>
<td>10344260</td>
<td>255</td>
<td>0</td>
</tr>
<tr>
<td>chr4</td>
<td>76649430</td>
<td>76649480</td>
<td>255</td>
<td>0</td>
</tr>
<tr>
<td>chr3</td>
<td>77858754</td>
<td>77858804</td>
<td>255</td>
<td>0</td>
</tr>
<tr>
<td>chr16</td>
<td>62688333</td>
<td>62688383</td>
<td>255</td>
<td>0</td>
</tr>
<tr>
<td>chr22</td>
<td>33031123</td>
<td>33031173</td>
<td>255</td>
<td>0</td>
</tr>
</tbody>
</table>

Mapped to reference genome: hg19, hg18, mm10, mm9, ...
BAMtools
Break

Install SICER, download test data
Run SICER

• Case study 1: without input control
  SICER-rb.sh

• Case study 2: with input control
  SICER.sh

• Case study 3: Differential calling
  SICER-df.sh
1. Run SICER without input control

- Data file: H3K27ac_act.bed
- Script: SICER-rb.sh
- Parameters:
  - InputDir
  - bed file: H3K27ac_act.bed
  - OutputDir
  - species: hg19
  - redundancy threshold: 1
  - window size (bp): 200
  - fragment size: 150
  - effective genome fraction: 0.74
  - gap size (bp): 600
  - E-value: 1000
  - ..
## Result output

<table>
<thead>
<tr>
<th>Output file name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K27ac_act-1-removed.bed</td>
<td>Non-redundant reads</td>
</tr>
<tr>
<td>H3K27ac_act-W200.graph</td>
<td>Raw data profile: bedGraph</td>
</tr>
<tr>
<td>H3K27ac_act-W200-normalized.wig</td>
<td>Raw data profile: wiggle</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-E1000.scoreisland</td>
<td>Identified islands</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-E1000-islandfiltered.bed</td>
<td>Island-filtered reads</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-E1000-islandfiltered-normalized.wig</td>
<td>wiggle profile on identified islands</td>
</tr>
</tbody>
</table>
2. Run SICER with input control

- Data files: H3K27ac_act.bed and input_act.bed
- Script: SICER.sh
- Parameters:
  
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>[InputDir]</td>
<td>..</td>
</tr>
<tr>
<td>[bed file]</td>
<td>H3K27ac_act.bed</td>
</tr>
<tr>
<td>[control file]</td>
<td>input_act.bed</td>
</tr>
<tr>
<td>[OutputDir]</td>
<td>.</td>
</tr>
<tr>
<td>[Species]</td>
<td>hg19</td>
</tr>
<tr>
<td>[redundancy threshold]</td>
<td>1</td>
</tr>
<tr>
<td>[window size (bp)]</td>
<td>200</td>
</tr>
<tr>
<td>[fragment size]</td>
<td>150</td>
</tr>
<tr>
<td>[effective genome fraction]</td>
<td>0.74</td>
</tr>
<tr>
<td>[gap size (bp)]</td>
<td>600</td>
</tr>
<tr>
<td>[FDR]</td>
<td>0.01</td>
</tr>
</tbody>
</table>
## Result output

<table>
<thead>
<tr>
<th>Output file name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K27ac_act-1-removed.bed</td>
<td>Non-redundant reads</td>
</tr>
<tr>
<td>H3K27ac_act-W200.graph</td>
<td>Raw data profile: bedGraph</td>
</tr>
<tr>
<td>H3K27ac_act-W200-normalized.wig</td>
<td>Raw data profile: wiggle</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600.scoreisland</td>
<td>Prescreened islands</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-islands-summary</td>
<td>SICER summary</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-islands-summary-FDR.01</td>
<td>SICER summary on identified islands</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-FDR.01-island.bed</td>
<td>SICER identified islands</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-FDR.01-islandfiltered.bed</td>
<td>Island-filtered reads</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-FDR.01-islandfiltered-normalized.wig</td>
<td>wiggle profile on identified islands</td>
</tr>
</tbody>
</table>
3. Run SICER for differential peak calling

- Data files:
  - H3K27ac_act.bed, input_act.bed
  - H3K27ac_inh.bed, input_inh.bed
- Script: SICER-df.sh
- Parameters:
  - [KO bed file]: H3K27ac_act.bed
  - [KO control file]: input_act.bed
  - [WT bed file]: H3K27ac_inh.bed
  - [WT control file]: input_inh.bed
  - [window size (bp)]: 200
  - [gap size (bp)]: 150
  - [FDR for KO vs KOCONTROL or WT vs WTCONTROL]: 0.01
  - [FDR for WT vs KO]: 0.01
- What it does:
  1. Call peaks for “WT” and “KO” separately (SICER.sh)
  2. Identify union (merged) islands
  3. Compare “KO” vs. “WT” for increased islands
  4. Compare “WT” vs. “KO” for decreased islands
## Output example

<table>
<thead>
<tr>
<th>Output file name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K27ac_act-vs-H3K27ac_inh-W200-G600-E-union.island</td>
<td>Merged islands</td>
</tr>
<tr>
<td>H3K27ac_act-and-H3K27ac_inh-W200-G600-summary</td>
<td>Merged island summary</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-increased-islands-summary-FDR0.01</td>
<td>Identified increased islands</td>
</tr>
<tr>
<td>H3K27ac_act-W200-G600-decreased-islands-summary-FDR0.01</td>
<td>Identified decreased islands</td>
</tr>
</tbody>
</table>
Summary

• ChIP-seq for histone mark/epigenetic profiling
• ChIP-seq “broad peak” calling: SICER
• Use SICER for:
  – Peak calling: with or without input control
  – Differential peak calling
• SICER users group:
  https://groups.google.com/forum/#!forum/sicer-users
Acknowledgments

Weiqun Peng
Wenjing Yang

Keji Zhao
Dustin E. Schones
Zhibin Wang
Kairong Cui
Gang Wei
Tae-Young Roh
Artem Barski
Iouri Chepelev

Chen Zeng

Xiaole Shirley Liu
Clifford Meyer
Tao Liu
Han Xu
Sheng’En Hu
Su Wang
Qian Qin
Sujun Chen

Gary Felsenfeld
Andre Nussenzweig
John O’Shea
Michael Q. Zhang
Nan-Ping Weng
Anand Swaroop

Myles Brown
Jun S Liu
Ramesh Shivdasani
Jon Aster
Warren Pear
Stephen Blacklow

All SICER users!
Thank you very much!