Statisitcal Methods and Software for ChIP-seq Data Analysis

CBB Journal Club
Mengting Gu

March 30, 2016

Introduction

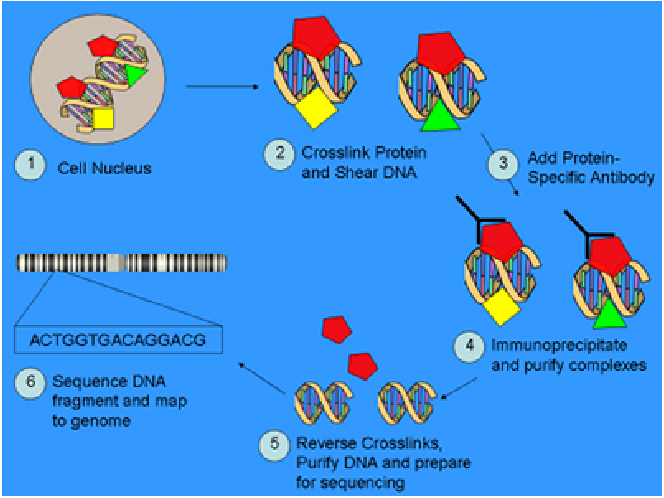

1 Cell Nucleus
2 Crosslink Protein and Shear DNA
3 Add Protein-Specific Antibody
4 Immunoprecipitate and purify complexes
5 Reverse Crosslinks, Purify DNA and prepare for sequencing
ACTGGTGACAGGACG
6 Sequence DNA fragment and map to genome

Figure: ChIP-Seq experiment capture millions of DNA fragments (150 $\sim$ 250 bp in length) that the protein under study interacts with, using a protein-specific antibody

Introduction

- High throughput sequencing of one or both ends of each fragment generates millions of reads.
- Standard preprocessing of ChIP-Seq data involves mapping reads to reference genome and retaining the uniquely mapping reads.
- Genomic regions with large number of aligning reads are identified as binding sites using one or more of many available statistical approaches.

Identify Protein-DNA Interaction Sites in Repetitive Regions

Motivation:

- Discarding multi-reads poses a significant challenge for identifying binding locations residing in genomic regions that have been duplicated over evolutionary time since these regions will not have many uni-reads.
- In some cases discarding multi-reads leads to inaccurate estimation of expression of genes that reside in repetitive regions.

Statistical Framework

Consider the multi-read problem as a non-parametric estimation problem of mixing density and derive an Expectation-Maximization-Smoothing algorithm.

- Let m be the total number of genomic locations. Use j to index the position on the genome, $j = 1, ..., m$.
- Let n be the total number of reads. Use i to index the reads, $i = 1, ..., n$.
- Define π as the density function for generating reads, π_j denote the value of π at j-th position, which is the probability that a read is generated from j-th position.

Statistical Framework

- Let Z_i be a random variable indicating the true origin of i-th read.
- $Z_i = j$ if the i-th read is generated from j-th position.
- $Z_i \in \{1, ..., m\}$, $Z \sim \pi$ and $P(Z = j) \equiv \pi(j)$.
- Observe $Y_i = (Y_{i1}, ..., Y_{im})$ as the mapping result of i-th read.
- $Y_{ij} = 1$ if i-th read aligns to j-th position on the genome, and 0 otherwise.

Goal: Estimate density π and test for significant sites

Statistical Framework

- Define $Z_{ij} = 1$ if $Z_i = j$.
- Assuming a read can be originated from only one location on the genome, $\sum_{j=1}^{m} Z_{ij} = 1, \forall i = 1, ..., n$.
- $(Z_{i1}, ..., Z_{im})$ independently follows a multinomial distribution with parameters $(\pi_1, ..., \pi_m)$
- let A_i be the set of positions that i-th read aligns to, and so $j \in A_i$ if i-th read aligns to j-th position. Then $Y_{ij} = 1$ if $j \in A_i$ and $Y_{ij} = 0$ otherwise

The CSEM algorithm

E-step:

$$z_{ij}^{(t)} = \frac{\pi_j^{(t)}}{\sum_{j' \in A_i} \pi_{j'}^{(t)}} 1(j \in A_i) \tag{1}$$

The E–step includes the following two special cases:

- **Special case 1:** If *i*-th read does not align to *j*-th position, then $z_{ij}^{(t)} = 0$
- **Special case 2:** If *i*-th read is an uni-read and aligns to *j*-th position, then $z_{ij}^{(t)} = 1$

The CSEM algorithm

M-step:

- First obtain initial ML estimates for $\pi_j^{(t+1)}$, denoted as $\mu_j^{(t+1)}$
- Maximizing the log likelihood function $logL_c(\pi)$ with respect to μ_j
- constraint: $\sum_{j=1}^{m} \mu_j = 1$

$$\mu_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} Z_{ij}^{(t)}. \tag{2}$$

The CSEM algorithm

S-step:
S-step of the CSEM algorithm smooth $\mu^{(t+1)}$ to obtain $\pi^{(t+1)}$.
This step can accommodate multiple choice of smoothing algorithms.

- *Bin smoother* (Hastie and Tibshirani, 1990)
  Partition the genome into fixed non-overlapping bins
  Assume π is constant within each bin

$$\pi_j^{(t+1)} = \frac{1}{w} \sum_{j' \in B_j} \mu_{j'}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{w} \sum_{j' \in B_j} z_{ij'}^{(t)} \tag{3}$$

The CSEM algorithm

S-step:

- *Moving average* (Hastie and Tibshirani, 1990)
  Instead of fixing non-overlapping bins, move a window by one base each time
  Estimate π at each position using moving average
  Approached was used in peak calling algorithms for ChIP-ship experiments (Kuan et al., 2008)

$$\pi_j^{(t+1)} = \frac{1}{2w+1} \sum_{j'=j-w}^{j+w} \mu_{j'}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2w+1} \sum_{j'=j-w}^{j+w} z_{ij'}^{(t)} \qquad (4)$$

where w is size of a half window.

The CSEM algorithm

S-step:

- *Kernel regression estimator* (Fan and Gijbels, 1996)
  Using *Nadaraya-Watson estimator* (Nadaraya, 1964; Watson, 1964), one of the simplest types of kernel regression estimators, $\pi(j)$ can be estimated as

$$\pi_j^{(t+1)} = \frac{\sum_{j'=1}^{m} K_w(j'-j)\mu_j'^{(t+1)}}{\sum_{j'=1}^{m} K_w(j'-j)} \tag{5}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sum_{j'=1}^{m} K_w(j'-j)} \sum_{j'=1}^{m} K_w(j'-j) z_{ij}^{(t)} \tag{6}$$

where $K_w(j'-j)$ is a kernel that assigns a non-negative weight to j' based on the distance between j' and j, with a bandwidth w.
$\int K_w(u)du = 1$ and K is an even function.

Algorithm

Multi-read allocation algorithm (with moving average)

1. A short-read alignment tool is used to establish a set of candidate alignments for each read against the reference genome.
2. Initialize $\pi_j = 1/m$ for all positions $j = 1, ..., m$ and $z_{ij} = 0$, $\forall i = 1, ..., n$, $\forall j = 1, ..., m$.
3. Until convergence,
    - For each read $i = 1, ..., n$, if i-th read has a_i possible starting positions, $s_1, ..., s_{a_i}$, then update z_{is_t} as $\pi_{s_t} / \sum_{k=1}^{a_i} \pi_{s_k}$, $l = 1, 2, ..., a_i$.
    - For each position $j = 1, ..., m$, update π_j as$\sum_{i=1}^{n} \sum_{j'=j-w}^{j+w} z_{ij'} / n(2w+1)$

The choice of w controls the degree at which multi-read allocation is affected by uni-reads.

Setting $2w + 1 \approx L$, where L is the expected fragment size ensures that uni-reads and multi-reads within a given window corresponding to the same binding event.

Results

Data:

- STAT1 binding in interferon-γ-stimulated-HeLa S3 cells (Rozowsky et al., 2009)
- GATA1 binding in mouse GATA1-null erythroid cells (G1E-ER4) (Cheng et al., 2009)
- Both datasets utilize single end short reads (30mers for STAT1 and 36mers for GATA1)

Results

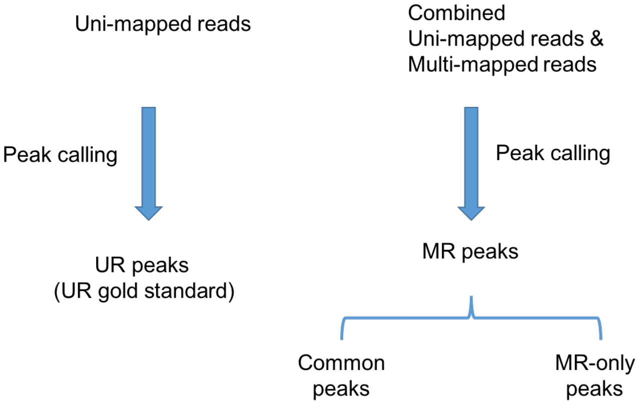
Table: Impact of multi-reads on sequencing depth

Dataset	# of reads	Alignable	Uni-reads	Multi	Rescued
STAT1(C)	76,913,219	36.64	29.92	6.72	22.46
STAT1(I)	49,771,625	47.90	38.31	9.59	25.03
GATA1(C)	33,124,216	79.27	67.81	11.46	16.90
GATA1(I)	20,711,007	82.37	69.38	12.99	18.73
MECP2-SET(C)	15,253,906	79.23	65.06	14.16	21.76
MECP2-SET(I)	21,870,009	90.35	78.14	12.21	15.63
MECP2-PET(C)	18,622,331	68.55	64.24	4.31	6.70
MECP2-PET(I)	18,498,899	84.26	78.92	5.34	6.77

**"(C)" and "(I)" refer to ChIP and input samples, respectively*

** Percentage is calculated in alignable, uni-reads, multi-reads and rescued*

Results



Uni-mapped reads

Peak calling

UR peaks
(UR gold standard)

Combined
Uni-mapped reads &
Multi-mapped reads

Peak calling

MR peaks

Common
peaks

MR-only
peaks

Figure: Peak calling at FDR level 0.005 for both uni-mapped reads and combined uni-reads and multi-reads

Results

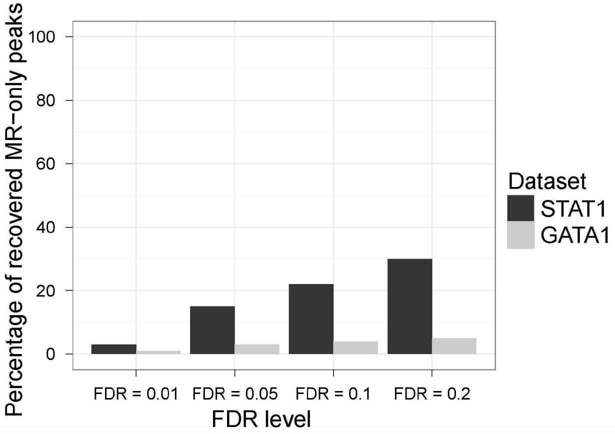


Figure: **Sensitivity analysis.** ”Recovered MR-only peaks” refer to MR-only peaks that are defined at FDR level of 0.005 and are detectable by the UR analysis at higher FDR levels.

Results

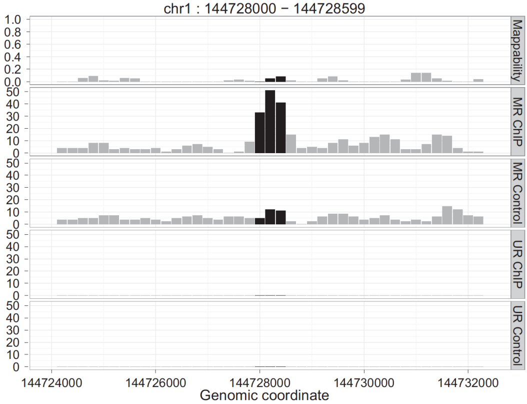


Figure: **Tag count profiles of MR-only peaks with corresponding mappability scores.** STAT1 MR-only peak in a poorly mappable region. Peak regions are depicted with black bars.

Results

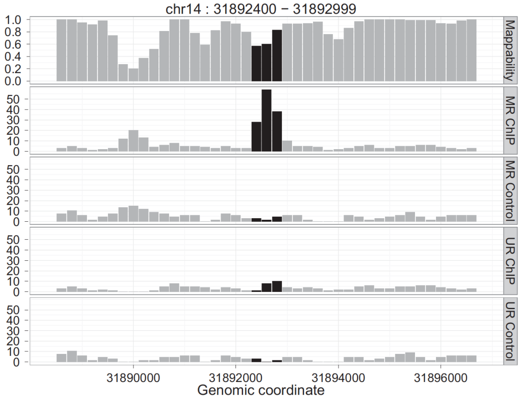


Figure: **Tag count profiles of MR-only peaks with corresponding mappability scores.** GATA1 MR-only peak in a moderately mappable region. Peak regions are depicted with black bars.

Results

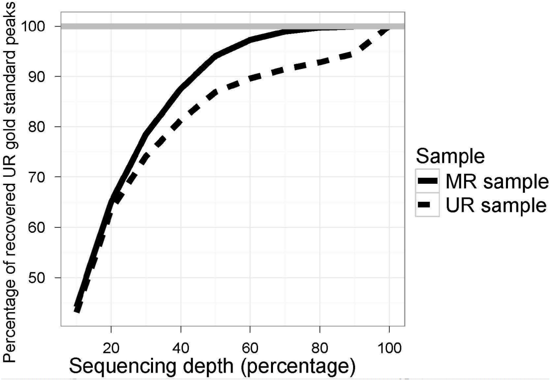


Figure: **Saturation plot of the STAT1 sample.** Percentage of STAT1 UR gold standard peaks recovered using sub-sampled UR and MR samples with lower sequencing depths. *x*-axis refers to the percentage of reads sampled from the full dataset.

Results

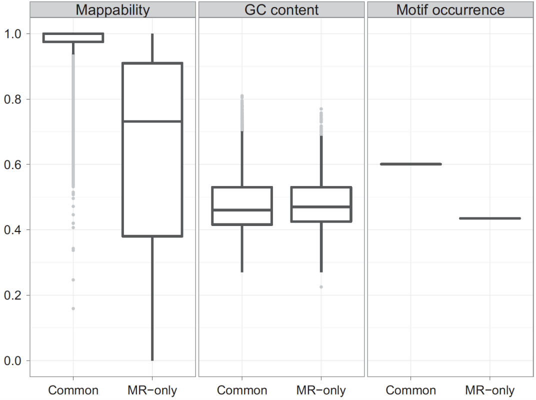


Figure: **Mappability, GC content, and STAT1 motif occurrence of the STAT1 common and MR-only peaks.** "Common" refers to common peaks identified by both the MR and the UR samples; "MR-only" peaks are unique to the MR sample. For the motif occurrence panel, y-axis represents the proportion of peaks with the consensus binding site.

Results

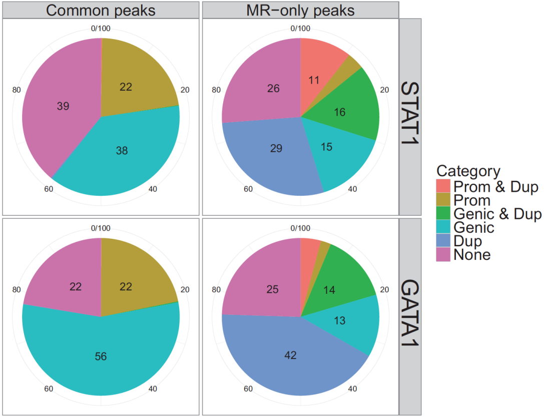


Figure: **Annotation of common and MR-only peaks with respect to TSS and duplicated regions**

Summary

- Investigated the shortcomings of discarding multi-reads in ChIP-Seq analysis; Illustrated how incorporating multi-reads can improve detection of binding sites in highly repetitive regions of genomes.
- Multi-reads lead to identification of novel binding sites that are located in highly repetitive and low mappability regions and are not identifiable with uni-reads alone.
- Effective utilization of uni-reads and multi-reads so that more peaks can be detected with lower sequencing depths.
- Substantial fraction of peaks specific to multi-read analysis are located in segmental duplications of the human and mouse genomes, and attributes to genes that are well associated with immunity and defense.