ChIP-Seq Tools

J Fass
UCD Genome Center Bioinformatics Core
*Wednesday* September 16, 2015
What’s the Question?

Where do Transcription Factors (TFs) bind genomic DNA$^1$?

(Where do other things bind DNA or RNA? ... e.g. histones$^2$, ribosomes$^3$, chromatin$^4$)

$^1$Furey 2012 Nat Rev Genet 13:840
$^2$Zentner 2013 Nat Struct Mol Biol 20:259
$^3$Ingolia 2014 Nat Rev Genet 15:205
$^4$Dekker 2013 Nat Rev Genet 14:390
Where? = Sequence?

ChIP-Seq

(TF-bound) Chromatin, ImmunoPrecipitated, then Sequenced

A question of location ... becomes ... a question of sequence
What’s the (Informatics) Problem?

Detection of binding peaks from noisy depth signal.

- Illumina- and other high throughput libraries result in highly variable coverage depth.
- Immnoprecipitation is not a perfect separation process.
MACS

MACS uses the biased positions of reads coming from fragments of average size centered around a binding site, to get through the noise.
MACS

MACS uses the biased positions of reads coming from fragments of average size centered around a binding site, to get through the noise.
MACS

DNA fragmentation …
MACS

… (naturally replicated across many cells) …
MACS

... DNA fragments used to make sequencing libraries ...
MACS

... reads aligned to the genome ...
MACS

... alignments processed to discern peaks.
MACS

Fig 1a Zhang 2008 Genome Biology 9:R137
Cross correlation between read alignment positions on strands (after shift of size $k$, on x-axis).

Fig 4D,E *Landt 2012 Genome Research 22:1813*
MACS

How exactly are alignments processed to find discern peaks?

- alignments in ChIP read set are compared to alignments in Input read set
- windows (twice expected sonication size) with above-threshold enrichment (ChIP / Input) are found
- if mode of “Watson” strand alignments is to the left of the mode of “Crick” strand alignments …
- (outer coordinates) ÷ 2 = binding site!
MACS2

- main documentation (currently) for Python module: https://pypi.python.org/pypi/MACS2
- peak calling (ChIP vs Input)
- has functionality for two-condition differential peak finding
IDR

Irreproducible Discovery Rate

- looks at consistency between replicates
- produces correspondence plots
- produces IDR values for peaks - can be used like FDR for differential expression (collect all peaks with IDR < $x$, and you’ll have a list that has at most $x$ false, or *irreproducible*, peaks)
Plots of “fraction of peaks shared between replicates” versus “fraction of peaks considered.” Linear $\sim y=x$ line means all considered peaks are shared.

Parabolic means “fraction shared” = “fraction considered”$^2$ … i.e. they’re only shared by chance.
Once you have your peaks, what do you do with them?

- How similar is the bound sequence?
  - motif finding - MEME-Suite.

- What genes are being regulated by this binding pattern?
  - who knows ... uh, I mean, try the closest gene - CEAS (Cis-regulatory element Annotation System)
MEME, DREME, CentriMo …

“Multiple EM (Expectation Maximization) for Motif Elicitation.” Find sequence motifs by aligning sequences under peaks to each other.

“Discriminative Regular Expression Motif Elicitation.” Finds shorter sequence motifs in DNA only; much faster than MEME, but less sensitive (?), will find shorter, core motifs.

CentriMo looks at motif placement within supplied sequences.
Logos - Motif Representation

- sequence logos are a way to represent a consensus sequence from multiple alignments
- relative heights at one position reflect base frequencies; most common base on top
- total height of all bases at one position reflects information content … related to conservation
What does it all mean?

ChIP-Seq gives you patterns on genomes, maybe correlated with genes, under different conditions, at different times …

What does that mean for expressional or biochemical regulation?

How does your system resemble other explored systems?

● *The ENCODE Project*, & *modENCODE*

● *The Cistrome Project*
Questions?