Genome-scale technologies 2/ Algorithmic and statistical aspects of DNA sequencing

ChIP-Seq data analysis

Ewa Szczurek
szczurek@mimuw.edu.pl

Instytut Informatyki
Uniwersytet Warszawski

Model-based Analysis of ChIP-Seq data (MACS)

Input parameters:

bandwidth a sonication size, 0.5 size of a sliding window

mFold tag enrichment

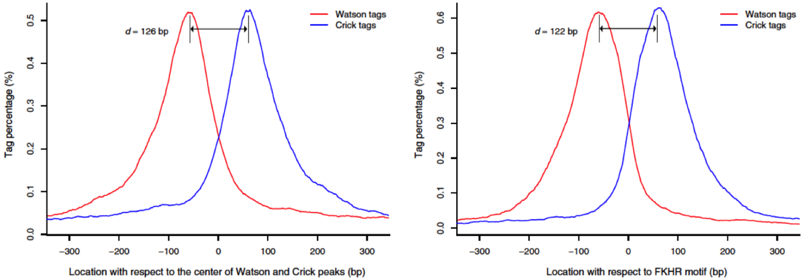
Steps:

1. slide 2*bandwidth* windows across the genome
2. find peaks: regions with tags $>$ *mfold* enriched to random
3. randomly sample 1,000 of these high-quality peaks and
   - separate their Watson and Crick tags
   - align them by the midpoint (if the Watson tag center is left of the Crick center)
4. let $d =$ the distance between the Watson and Crick modes
5. shift all the tags by $d/2$ toward the 3' ends

Output:

Shifted tags are at the most likely protein-DNA binding sites.

MACS model for FoxA1 ChIP-Seq.



- 5’ ends of strand-separated tags from a random sample of 1,000 model peaks, aligned by:
  a) the center of their Watson and Crick peaks
  b) the FKHR motif (precise FoxA1 binding place)

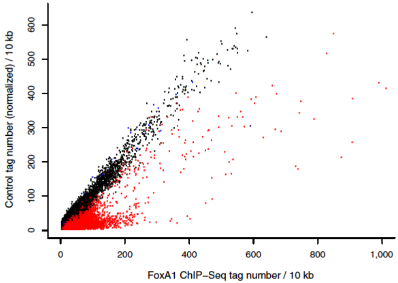
Finding peaks in MACS

- For experiments with a control
  - linearly scale the total control tag count to be the same as the total ChIP tag count.
  - remove duplicate tags in excess of what is warranted by the sequencing depth
- Model tag counts with Poisson distribution (λ_{BG})
- Peaks: significant deviation of counts from $Poiss(\lambda_{BG})$
- Shift tags by $d/2$
- Merge overlapping peaks
- Summit: fragment with the highest tag pileup $\leftrightarrow$ precise prediction of binding site

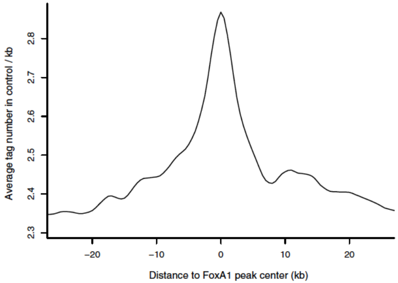
Significance of peaks in MACS

- Tag distribution in control
  - has local biases and correlates with ChIP samples

Tag count in ChIP versus control
(10 kb windows across genome)



Tag density in control samples
around FoxA1 ChIP-Seq peaks



red dots: windows containing ChIP peaks

black dots: windows containing control peaks

Significance of peaks in MACS

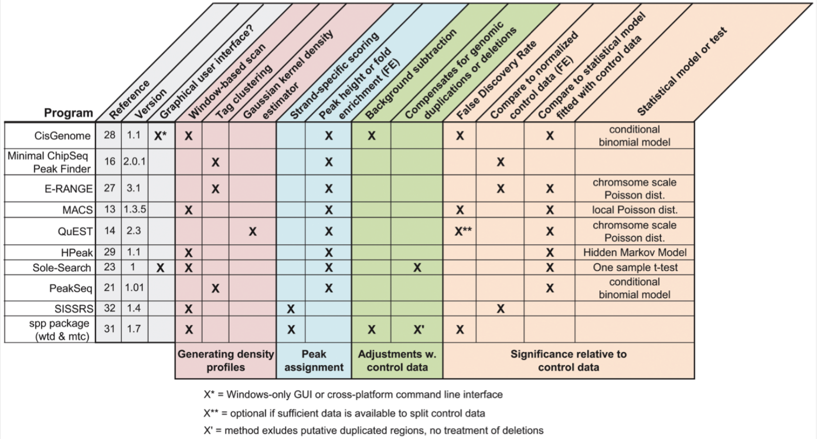
- The uniform, whole-genome λ_{BG} not used
- Instead, λ_{local} (estimated from c.a. $5KB$ around the peak in the control)

Definition (Empirical FDR)

For each detected peak, MACS uses the same parameters to find ChIP peaks over control and control peaks over ChIP (that is, a sample swap). The empirical FDR is defined as:

$$\frac{\text{Number of control peaks}}{\text{Number of ChIP peaks}}$$

ChIP Peak calling algorithms: a comparison[1]



Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

[1] Wilbanks et al., Plos One (2010)

Bibliography

- Y. Zhang et al., *Model-based Analysis of ChIP-Seq (MACS)*. Genome Biology 2008.