Charles Girardot, Furlong Lab

MACS, CisGenome, SISSRs and other peak calling algorithms: differences and practical use
ChIP-Seq signal properties

- Only 5’ ends of ChIPed fragments are sequenced
  - Shifted read distribution
  - Expected symmetry between Watson/Crick read distributions

Figure source: Valouev et al. Nat. Methods Sept 2008
Peak finding overview

1. Build strand-specific profiles
   • How (window-scan, KDE…)?
   • Filter duplicates?

2. Combine profiles (shift/extension)
   • Shift/extension estimation?

1. Define enriched regions/peaks
   • Statistics used
   • What boundaries should be reported?
   • What score to use (ratio, p-val, q-val)?
   • Compute/estimate a FDR?

Figure source: Valouev et al. Nat. Methods Sept 2008
# Main aspects of peak finders

<table>
<thead>
<tr>
<th>Profile</th>
<th>Peak criteria</th>
<th>Tag shift</th>
<th>Control data</th>
<th>Rank by</th>
<th>FDR</th>
<th>User input parameters</th>
<th>Artifact filtering: strand-based/ duplicate</th>
<th>Refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CisGenome v1.1</td>
<td>Strand-specific window scan</td>
<td>1: Number of reads in window, 2: Number of ChIP reads minus control reads in window</td>
<td>Average for highest ranking peak pairs</td>
<td>Conditional binomial used to estimate FDR</td>
<td>Number of reads under peak</td>
<td>1: Negative binomial, 2: conditional binomial</td>
<td>Target FDR, optional window width, window interval</td>
<td>Yes / Yes 10</td>
</tr>
<tr>
<td>ERANGE v3.1</td>
<td>Tag aggregation</td>
<td>1: Height cutoff for high quality peak estimate, per-region estimate, or input</td>
<td>High quality peak estimate, per-region estimate, or input</td>
<td>Used to calculate fold enrichment and optionally ( p ) values</td>
<td>( p ) value</td>
<td>1: None, 2: # control ChIP</td>
<td>Optional peak height, ratio to background</td>
<td>Yes / No 4,18</td>
</tr>
<tr>
<td>FindPeaks v3.1.9.2</td>
<td>Aggregation of overlapped tags</td>
<td>Height threshold</td>
<td>Input or estimated</td>
<td>NA</td>
<td>Number of reads under peak</td>
<td>1: Monte Carlo simulation, 2: NA</td>
<td>Minimum peak height, subpeak valley depth</td>
<td>Yes / Yes 19</td>
</tr>
<tr>
<td>F-Seq v1.82</td>
<td>Kernel density estimation (KDE)</td>
<td>( s ) s.d. above KDE for 1: random background, 2: control</td>
<td>Input or estimated</td>
<td>KDE for local background</td>
<td>Peak height</td>
<td>1: None, 2: None</td>
<td>Threshold s.d. value, KDE bandwidth</td>
<td>No / No 14</td>
</tr>
<tr>
<td>GLITR</td>
<td>Aggregation of overlapped tags</td>
<td>Classification by height and relative enrichment</td>
<td>User input tag extension</td>
<td>Multiply sampled to estimate background class values</td>
<td>Peak height and fold enrichment</td>
<td>2: # control ChIP</td>
<td>Target FDR, number nearest neighbors for clustering</td>
<td>No / No 17</td>
</tr>
<tr>
<td>MACS v1.3.5</td>
<td>Tags shifted then window scan</td>
<td>Local region Poisson ( P ) value</td>
<td>Estimate from high quality peak pairs</td>
<td>Used for Poisson fit when available</td>
<td>( p ) value</td>
<td>1: None, 2: # control ChIP</td>
<td>( p )-value threshold, tag length, mfold for shift estimate</td>
<td>No / Yes 13</td>
</tr>
<tr>
<td>PeakSeq</td>
<td>Extended tag aggregation</td>
<td>Local region binomial ( P ) value</td>
<td>Input tag extension length</td>
<td>Used for significance of sample enrichment with binomial distribution</td>
<td>( q ) value</td>
<td>1: Poisson background assumption 2: From binomial for sample plus control</td>
<td>Target FDR</td>
<td>No / No 5</td>
</tr>
<tr>
<td>QuEST v2.3</td>
<td>Kernel density estimation</td>
<td>2: Height threshold, background ratio</td>
<td>Mode of local shifts that maximize strand cross-correlation</td>
<td>KDE for enrichment and empirical FDR estimation</td>
<td>( q ) value</td>
<td>1: NA, 2: # control ChIP</td>
<td>KDE bandwidth, peak height, subpeak valley depth, ratio to background</td>
<td>Yes / Yes 9</td>
</tr>
<tr>
<td>SICER v1.02</td>
<td>Window scan with gaps allowed</td>
<td>( p ) value from random background model, enrichment relative to control</td>
<td>Input</td>
<td>Linearly rescaled for candidate peak rejection and ( P ) values</td>
<td>( q ) value</td>
<td>1: None, 2: From Poisson ( P ) values</td>
<td>Window length, gap size, FDR (with control) or ( E )-value (no control)</td>
<td>No / Yes 15</td>
</tr>
<tr>
<td>SISSRs v1.4</td>
<td>Window scan</td>
<td>( N_+ - N_+ ) sign change, ( N_+ / N_+ ) threshold in region</td>
<td>Average nearest paired tag distance</td>
<td>Used to compute fold-enrichment distribution</td>
<td>( p ) value</td>
<td>1: Poisson, 2: control distribution</td>
<td>1: FDR, 1,2: ( N_+ / N_+ ) threshold</td>
<td>Yes / Yes 11</td>
</tr>
<tr>
<td>spp v1.0</td>
<td>Strand specific window scan</td>
<td>Poisson ( P ) value (paired peaks only)</td>
<td>Maximal strand cross-correlation</td>
<td>Subtracted before peak calling</td>
<td>( p ) value</td>
<td>1: Monte Carlo simulation, 2: # control ChIP</td>
<td>Ratio to background</td>
<td>Yes / No 12</td>
</tr>
<tr>
<td>Useq v4.2</td>
<td>Window scan</td>
<td>Binomial ( P ) value</td>
<td>Estimated or user specified</td>
<td>Subtracted before peak calling</td>
<td>( q ) value</td>
<td>1, 2: binomial, # control ChIP</td>
<td>Target FDR</td>
<td>No / Yes 20</td>
</tr>
</tbody>
</table>

Pepke et al; Nature Methods 6, S22 - S32 (2009)

6/8/10
PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls

Joel Rozowsky¹, Ghia Euskirchen², Raymond K Auerbach³, Zhengdong D Zhang¹, Theodore Gibson¹, Robert Bjornson⁴, Nicholas Carriero⁴, Michael Snyder¹,² & Mark B Gerstein¹,³,⁴
PeakSeq

- Sequence tags from certain location are not unique in the genome
- Tags that don’t uniquely map are usually discarded

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size (Mb)</th>
<th>Nonrepetitive sequence</th>
<th>Mappable sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Size (Mb)</td>
<td>Percentage</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>100.28</td>
<td>87.01</td>
<td>86.8%</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>168.74</td>
<td>117.45</td>
<td>69.6%</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>2,654.91</td>
<td>1,438.61</td>
<td>54.2%</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>3,080.44</td>
<td>1,462.69</td>
<td>47.5%</td>
</tr>
</tbody>
</table>

For four common model organisms—worm, fruit fly, mouse and human—we have determined the fraction of each genome sequence that is nonrepetitive as well as the fraction that is mappable using 30-nt sequence tags. The genome coverage achievable from genomic tiling arrays corresponds to the nonrepetitive fraction of a genome whereas the mappable coverage is what is achievable by tag-based sequencing approaches. We also determined that as the length of the sequence tags is increased beyond 30, the number of nucleotides in the genomes that are uniquely mappable is 2,452 Mb (79.6%) for 30-nt reads, 2,586 Mb (84.0%) for 40 nt, 2,669 Mb (86.7%) for 50 nt, 2,720 Mb (88.3%) for 60 nt and 2,750 Mb (89.3%) for 70 nt.

=> The fraction of the “mappable” genome is usually a parameter of peak finders
PeakSeq

- Background models are usually assumed to follow a Poisson statistics
- Unfortunately, the real background results from multiple effects
  1. Mappability
  2. Chromatin structure (e.g. accessibility/openness)

Enrichments in reference sample is not randomly placed

Chromatin structure is the major factor

N.B.: See also Kharchenko et al. Nat biotech 2008
PeakSeq

Step 1: Signal map(s) construction

1. Tag extension (user input)
2. Signal map: count for each bp
PeakSeq

Step 2: Determination of potential regions using simulated bg

1. Uses Poisson statistics
2. Work per window (1 Mb) and correct signal (of different windows) using mappability maps
3. Given a user-defined target FDR, a threshold is computed
4. Keep regions above threshold
PeakSeq

Step 3: Normalizing control to ChIP-seq sample

1. Count tags in bins along chromosome for ChIP-seq and reference
2. Correct tag counts using slope of linear regression
3. Pf = fraction (i.e. in [0,1]) of potential peaks to exclude
PeakSeq

Step 4: Scoring enriched target regions relative to control

1. Compute fold enrichment for each candidate (defined in step 2)
2. Compute p-value from binomial distribution
3. Correct for multiple testing and call enriched regions
What have we learned so far

- The size of the mappable genome varies with your tag length
- Background is not accurately modeled by Poisson
  - Use of input DNA is recommended
- The scaling factor between ChIP and input sample is not directly the tag ratio
MACS

Method

Model-based Analysis of ChIP-Seq (MACS)
Yong Zhang\textsuperscript{*,*}, Tao Liu\textsuperscript{*,*}, Clifford A Meyer\textsuperscript{*,}, Jérôme Eeckhoute\textsuperscript{†},
David S Johnson\textsuperscript{†}, Bradley E Bernstein\textsuperscript{‡,}, Chad Nussbaum\textsuperscript{†},
Richard M Myers\textsuperscript{‡,}, Myles Brown\textsuperscript{†}, Wei Li\textsuperscript{†} and X Shirley Liu\textsuperscript{‡}


- **Step 1: Modeling the tag shift**

1. Scan genome with a window of user-defined sonication size
2. Keep the best 1000 (or less) peaks having a fold enr. > \textit{mfold} (default 32, relative to random model)
3. Separate Watson/Crick tags
4. Shift size is modeled as the distance between the modes of the Watson and Crick peaks
MACS

- Step 2: Peak detection
  1. Shift every tag by d/2
  2. Slide a 2d window across the genome to find candidate peaks with significant tag enrichment (according to Poisson distribution, default p-value = $10^{-5}$)
  3. Merge of overlapping peaks
  4. Report:
    - fold enrichment for called peaks: ratio between tag counts and expected using Poisson distribution (using input data if provided)
    - Position with highest pile-up is defined as the summit
    - Empiric FDR if control sample is provided (sample swap)
MACS: key aspects

- Adaptive Poisson distribution to model background
  - Usually, this $\lambda$ is computed once i.e. $\lambda_{BG}$
  - Here, they use a dynamic $\lambda_{\text{local}}$ to account for local biases:
    - $\lambda_{\text{local}} = \max(\lambda_{BG}, \lambda_{1K}, \lambda_{5K}, \lambda_{10K})$

- Model the tag shift using the bimodal property of ChIP-seq reads using high confidence peaks (fold cutoff)

- Automatic removal of duplicated tags in excess of what is expected given the sequencing depth (using p-val cutoff of $10^{-5}$, binomial dist.)
  - Always check the default setting for duplicates in your peak finder
CisGenome

- Two pass algorithm, globally similar to MACS
- First pass:
  - scan similar to MACS (100 bp window) to evaluate DNA fragment length i.e. tag shift value
  - FDR estimation (based on non overlapping window of 100 bp) from following statistics:
    - One sample analysis: based on a negative binomial
    - Two sample analysis: tag count in IP bin evaluated against tag sum $n_i$ (IP+ref) using binomial
CisGenome

- Second pass after tag shift: principle similar to first scan (FDR also recomputed):
  - Overlapping windows below user defined FDR are merged (best FDR is kept). In two sample analysis, the best fold change is also reported.
  - Regions that do not exhibit bimodal read distribution (e.g. b/w strands) are filter out (significant strand-specific peak expected).
  - Peak boundaries may be refined using the read distributions (on by default).
SISSRs

Published online 6 August 2008

doi:10.1093/nar/gkn488

Genome-wide identification of *in vivo* protein–DNA binding sites from ChIP-Seq data

Raja Jothis, Suresh Cuddapah, Artem Barski, Kairong Cui and Keji Zhao*

Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20894, USA

- DNA fragment length estimated from the data
- No tag shift / extension
- FDR estimate from Poisson model or from reference dataset
- Reports TFBS location estimation i.e. very small region
SISSRs

Jothi et al.; NAR 36, 16 (2008)
What have we learned more

• The size of the mappable genome varies with your tag length
• Background is not accurately modeled by Poisson
  • Use of input DNA is recommended
  • If no input is available, favor methods using negative binomial (or local poisson)
• The scaling factor between ChIP and input sample is not the tag ratio
• Fragment length can be estimated from top peaks or given as input
• Usually duplicate reads are filtered, a gentler approach might be better or no filtering (?)
• Enrichment is usually reported, sometimes with FDR/q-value; methods vary
Peaks vs enriched regions (TF ChIP-Seq)

MACS summits

Mef2 Known TFBS (in Act57B_-539/ 2 enhancer)

MACS groups 2 TFBS together: might need to use a "peak splitter" or check how other peakfinder behave (here SISSRs)
IGB – Another Mef2 known TFBS

- Check different parameter settings together with positive controls
- Visualize to get a feel
Some options might look great...

- *br* option in cisgenome hts_peakdetectorv2* tool

![Graph showing NO refinement and Boundary refinement with peaks]

Asking for boundary refinement may cause loss of peaks: peak finders assumes a single peak is in the region…
Peak Splitting

- Window based detection (MACS, CisGenome,…) will report unique regions encompassing several binding sites.
- A post processing to split regions into multiple peaks is needed.
- PeakSplitter developed by Mali Salmon in EBI.
- The new beta version of MACS integrate PeakSplitter.
- Tools like SISSR and QuEST implement a different approach (detect summit then extend).
The different types of ChIP-seq signal

1. Proteins binding DNA in a site-specific fashion

=> Narrow peaks, hundreds of bp wide

Pepke et al; *Nature Methods* 6, S22 - S32 (2009)
The different types of ChIP-seq signal

1. Proteins binding DNA in a site-specific fashion

2. RNA Pol II like signal
   => Mixture of strong binding (at TSS) and broader enrichment over several Kb (active transcription)

Pepke et al; *Nature Methods* 6, S22 - S32 (2009)
The different types of ChIP-seq signal

1. Proteins binding DNA in a site-specific fashion

2. RNA Pol II signal

3. Chromatin marks
   H3K4me3, active promoters
   H3K36me3, active genes
   H3K27me3, repressed regions
   => Enrichment from nucleosome size domain to several hundreds of Kb
Example of Histone marks

- Two marks, at same dev. Stage:
  - H3K4me3 : active promoters (~ short mark)
  - H3K36me3 : active genes (~ long mark)

=> Good test case b/c one should see both marks at active genes

- Analyzed with (in progress):
  - SISSRs failed at finding anything
  - CisGenome also (still investigating the pb)
  - Will show MACS and QuEST results

6/8/10
Good agreement between the tools

Missed by QuEST

Too large extension by MACS?
MACS suggests more active promoters and genes: predictions correlate
=> Is QuEST too stringent?
Detection of gene within gene example

=> Would you trust this with only one of the two marks?
Which one to use?

- You might want to run different tools and check how they behave on your datasets
- Do you have reference sample or not?
- Detection method should be adapted to signal type i.e. SISSR certainly has a too strong peak assumption for (long) histone marks?
- Laajala et al compared results with different peak finders – using TF signal only (BMC Genomics 2009, 10:618)
Visualization is important

- Assessment of the data quality e.g. positive controls, background
- Determine cutoffs (looking at positive controls)
- Compare peak finders outputs
- Integration of data / co-visualization
  - Your brain catches aspects that computers can’t: hypothesis generation.
Thanks!

You
Eileen Furlong
Robert Zinzen / Stefan Bonn
Nicolas Delhomme
Ismael Padioleau
Martina Braun
Furlong Lab
GeneCore