Transcriptional Circuitry and the Regulatory Conformation of the Genome

I-CORE for Gene Regulation in Complex Human Disease.

Feb. 1st, 2015
The Hebrew University
Downstream tools for ChIP-seq analysis

Tommy Kaplan
The Hebrew University

I-CORE for Gene Regulation in Complex Human Disease.

Feb. 1st, 2015, The Hebrew University
In a nutshell

- Galaxy ([hm-02.cs.huji.ac.il/galaxy](hm-02.cs.huji.ac.il/galaxy))
- Quality check ([FASTQC](FASTQC))
- **Map reads and display in genome browser**
  - Peak calling ([MACS](MACS))
  - Find motifs ([WEEDER, DECOD, STAMP](WEEDER, DECOD, STAMP))
  - Compare to other genome data ([GEO, UCSC](GEO, UCSC))
- Differential analysis of peaks ([edgeR](edgeR))
- Find nearby genes ([GREAT](GREAT)), test their expression
Lamentably

- Existing web tools can help you “find the story” in your data, but won’t bring you to publication-ready level.
Welcome to galaxy!

This site is best viewed using Firefox or Chrome browsers. In Internet Explorer some of the functions may not work properly.

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
Obtaining data

• Upload to Galaxy (slower)

• Copy fastq files to FTP folder: /cs/icore/$USER/galaxy_upload

• And “chmod 777 <files>”
The following job has been successfully added to the queue:

1: H3K27ac.fastq.gz

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
The following job has been successfully added to the queue:

1: H3K27ac.fastq.gz

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
H3K27ac.fastq FastQC Report
Fri 23 Jan 2015
H3K27ac.fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per base N content
- Per sequence GC content
- Per base content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filename</td>
<td>H3K27ac.fastq</td>
</tr>
<tr>
<td>File type</td>
<td>Conventional base calls</td>
</tr>
<tr>
<td>Encoding</td>
<td>Sanger / Illumina 1.9</td>
</tr>
<tr>
<td>Total Sequences</td>
<td>14675970</td>
</tr>
<tr>
<td>Filtered Sequences</td>
<td>0</td>
</tr>
<tr>
<td>Sequence length</td>
<td>51</td>
</tr>
<tr>
<td>%GC</td>
<td>45</td>
</tr>
</tbody>
</table>
Genomic mapping of sequenced data using BOWTIE

Software
Ultrafast and memory-efficient alignment of short DNA sequences to the human genome
Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Abstract

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Bowtie is open source http://bowtie.ccb.cbrc.umd.edu.

• http://bowtie-bio.sourceforge.net
Running BOWTIE

- `bowtie -c -q -n 2 -m 1 -5 3 -p 32 hg19 <fastq> <bw>`
- 14,675,970 reads in few minutes
- Notable parameters:
  - `-q` = input in FASTQ format
  - `-n 2` = max of 2 mismatches
  - `-m 1` = report only reads w/ unique match
  - `-5 x / -3 x` = trim x bases from 5’ or 3’ end of reads
  - `-p 32` = run on multiple CPUs (faster)
Running BOWTIE

• bowtie -c -q -n 2 -m 1 -5 3 hg19 <.fastq> <.bw>

• Input

```
@FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
AAATGTGAATCTGAANAGCTGGAATCCAGTCTGGTGTTTGTA
+FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
BACCBBCCBBCBB:=!1=CAB7B@BCCA<C?>A8B>@AAAC
```

• Output

```
FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
 - chr15 62798646
TACAAACACCAGACTGGATTCCAGCTNTTCAGATTAACA
CAAA@B8A?>C<ACCBB@B7BAC=1!=:BBCCBBCB
0 12:C>N
```
To run bowtie, Galaxy requires converting the data to fastqsanger format.
Is this library mate-paired?
- Single-end

FASTQ file:
- 1: H3K27ac.fastq

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Write unaligned reads to separate file(s):
- [ ]

Will you select a reference genome from your history or use a built-in index?:
- [ ] Use a built-in index
  - Built-ins were indexed using default options

Select a reference genome:
- Mouse July 2007 (NCBI37/mm9) (mm9)
  - If your genome of interest is not listed, contact the Galaxy team

Specify the read group for this file?:
- [ ]

Parameter Settings:
- [ ] Use defaults
  - You can use the default settings or set custom values for any of Bowtie's parameters.

Execute
Number (and percent) of mapped reads display at UCSC (main)
Valouev et al, 2008
How to extend reads (from ~50 to ~150)?
Change type of mapped data to bed.
Create bedGraph of genome coverage
Peak calling

• This should be enough to be able to “see the data”
• Browse the genome. Look near your favorite genes.
• Zoom out and in. Find the story!
• Peak calling algorithms can tell you where to look.
Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li and X. Shirley Liu

Published: 17 September 2008


Abstract

We present Model-based Analysis of ChIP-Seq data, MACS, which analyzes data generated by short read sequencers such as Solexa’s Genome Analyzer. MACS empirically models the shift size of ChIP-Seq tags, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome, allowing for more robust predictions. MACS compares favorably to existing ChIP-Seq peak-finding algorithms, and is freely available.
MACS

• Infers avg. length of DNA fragments

• Allows usage of control IP for local normalization

• Uses Poisson distribution to assign p-value per position (given surroundings)

• Tweak peak calling by playing with
  --shiftsize : ⇒ peak width

  --local : ⇒ sensitivity to peak’s surrounding

  --pvalue : ⇒ number of peaks
--shiftsize

--pvalue
Grizzly Peak Fitting

- For more accurate calls of TF binding events, use model-based, iterative, multi-peak algorithm
- Untangles "binding" within MACS broader regions

Capaldi et al, 2008; Harrison et al, 2011
Experiment Name:
MACS2 in Galaxy

ChIP-Seq Tag File:
3: Bowtie2 on data 1: aligned reads

ChIP-Seq Control File:
Selection is Optional

Select action to be performed:
Peak Calling

Effective genome size:
270000000
Human: 3.3e+9, Mouse: 3.0e+9, Fly: 1.9e+8, Worm: 1.3e+8 (---size)

Band width:
300
(---bw)

Parse xls files into into distinct interval files:

Save fragment pileup, control lambda, -log10pvalue/qvalue in bedGraph:

Select p-value or q-value:
q-value
default uses q-value

q-value cutoff for peak detection:
0.05
default: 5e-2 (---qvalue)

Display advanced options:
Hide

Build Model:
Build the shifting model (---nomodel disabled)

Execute

What it does
With the improvement of sequencing techniques, chromatin immunoprecipitation followed by high throughput sequencing (ChIP-Seq) is getting popular to study genome-wide protein-DNA interactions. To address the lack of powerful ChIP-Seq analysis method, we present a novel algorithm, named Model-based Analysis of ChIP-Seq (MACS), for identifying transcription factor binding sites from DNA sequencing data.
Galaxy allows to:
- Sort peaks by height
- Select top ones
- Get sequence
The following job has been successfully added to the queue:
16: Extract Genomic DNA on data 15

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
• Download, process elsewhere
• Use built-in tools
Motif analysis

- Should be discriminative (vs. genome, intergenic, etc).
- Motifs (e.g. `GGTCANNNTGACC`) or PWMs
- **DECOD** - [www.sb.cs.cmu.edu/DECOD](http://www.sb.cs.cmu.edu/DECOD)

- Bound by which TF?
- **STAMP** - [www.benoslab.pitt.edu/stamp](http://www.benoslab.pitt.edu/stamp)
Get more data

Data deposited to GEO

- Raw reads (SRA or FASTQ format)

\[
\begin{align*}
@\text{FoxA1.1} & \text{ HWUSI-EAS582}_229:6:1:1:1235 \ length=42 \\
& \text{AAATGTTAATCTGAANAGCTGGAATCCAGTCTGGTGTTTGT} \\
+\text{FoxA1.1} & \text{ HWUSI-EAS582}_229:6:1:1:1235 \ length=42 \\
& \text{BACCBBCBCCBB}::=1=\text{CAB7B}@\text{BBCCA}<\text{C?>A8B}>@\text{AAAC} \\
@\text{FoxA1.2} & \text{ HWUSI-EAS582}_229:6:1:1:569 \ length=42 \\
& \text{CAGTATGGAGGTGAATAAACAGCAGATGGCCTGGAAGATACA} \\
+\text{FoxA1.2} & \text{ HWUSI-EAS582}_229:6:1:1:569 \ length=42 \\
& \text{A?AB>CA@AB}:833;>:2A@)@@?>6(<9@?;135B4>\text{A??}}
\end{align*}
\]

- Read coverage along genome (WIG)

- Called peaks

- To convert SRA format to FASTQ, use

Get data from the UCSC genome browser

- For ENCODE and other projects that deposit data:
  
  Download as bigwig
Comparative analysis

• From http://hgdownload.cse.ucsc.edu/admin/exe/

• bigWigAverageOverBed -minMax
  Find values of called peaks in input bigwig files

• Collect in table

• Differential binding using edgeR package
  [in R language, assign each peak a p-value for change]
Basic “functional” questions

- Identify bound regions along genome
- Quantify ChIP-signal
- Estimate peaks, identify DNA motif
- Where along the gene? Promoter, etc. How?
- Compare to other genomic data (time point, condition, cell line, other TF, etc)
- Enriched near which genes? Of specific function?
Web server for predicting the functions of cis-regulatory regions.
• Assign peaks to nearest gene
• Look for enriched annotations

Is that the right way?
<table>
<thead>
<tr>
<th>Region</th>
<th>Gene</th>
<th>Distance to TSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACS_peak_2412</td>
<td>GREB1 (-35,170), E2F6 (-32,775)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_4803</td>
<td>C9orf3 (+56,361), FANCC (+534,636)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_729</td>
<td>SYT8 (-37,790), HCCA2 (-32,383)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_225</td>
<td>FGFR1B (-548,445)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_4647</td>
<td>MYC (+123,952)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_1278</td>
<td>STK24 (-73,396), SLC15A1 (+102,137)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_53</td>
<td>RCC2 (-82,010), ARHGEF10L (-19,263)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_2692</td>
<td>TNP1 (-527,674), TNS1 (+556,340)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_3036</td>
<td>ZNRF3 (-69,738), XBP1 (-13,592)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_2289</td>
<td>ISYNA1 (-8,871), ELL (+75,123)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_1945</td>
<td>IGFBP4 (+4,959), TNS4 (+53,219)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_1143</td>
<td>TMEM120B (+36,403), RHOF (+44,533)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_2202</td>
<td>STK11 (-24,243), SBN02 (-7,273)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_1844</td>
<td>KHLDC4 (-24,301), SLC7A5 (+79,257)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_1048</td>
<td>DYRK2 (-165,170), CAND1 (+214,281)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_4182</td>
<td>KIAA0415 (-92,067), FOXX1 (+1,267)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_1947</td>
<td>CCR7 (-11,276), SMARCE1 (+71,103)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_2977</td>
<td>TMPRSS2 (-145,861), RIPK4 (+161,396)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_2194</td>
<td>ATP9B (-6,190), SALL3 (+82,932)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_3705</td>
<td>LILST (-434,434), MAP3K1 (-856,703)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_1395</td>
<td>ZFP36L1 (+222,939), RAD51L1 (+750,337)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_2348</td>
<td>CYP2B6 (-2,800)</td>
<td></td>
</tr>
<tr>
<td>MACS_peak_863</td>
<td>NARS2 (-485,489), ODZ4 (+380,297)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>Region</th>
<th>Distance to TSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4GALT</td>
<td>MACS_peak_3086</td>
<td>(-14,035)</td>
</tr>
<tr>
<td>ABAT</td>
<td>MACS_peak_1682</td>
<td>(+4,572)</td>
</tr>
<tr>
<td>ABHD2</td>
<td>MACS_peak_1606</td>
<td>(+27,679)</td>
</tr>
<tr>
<td>ACSL4</td>
<td>MACS_peak_5022</td>
<td>(-218,800)</td>
</tr>
<tr>
<td>ACTL7B</td>
<td>MACS_peak_4841</td>
<td>(+377,506)</td>
</tr>
<tr>
<td>ACTN4</td>
<td>MACS_peak_2336</td>
<td>(+66,208)</td>
</tr>
<tr>
<td>ACTR2</td>
<td>MACS_peak_2499</td>
<td>(+72,801)</td>
</tr>
<tr>
<td>ACTR3B</td>
<td>MACS_peak_4392</td>
<td>(-60,191)</td>
</tr>
<tr>
<td>ADA</td>
<td>MACS_peak_2853</td>
<td>(-41,825)</td>
</tr>
<tr>
<td>ADAMTS12</td>
<td>MACS_peak_3680</td>
<td>(+341,842)</td>
</tr>
<tr>
<td>ADAMTS15</td>
<td>MACS_peak_2207</td>
<td>(-450)</td>
</tr>
<tr>
<td>ADCY9</td>
<td>MACS_peak_1679</td>
<td>(-36,649)</td>
</tr>
<tr>
<td>ADD2</td>
<td>MACS_peak_2513</td>
<td>(+151,622)</td>
</tr>
<tr>
<td>ADHFE1</td>
<td>MACS_peak_4513</td>
<td>(+80,520), MACS_peak_4514 (+90,242)</td>
</tr>
<tr>
<td>ADORA3</td>
<td>MACS_peak_195</td>
<td>(-9,976)</td>
</tr>
<tr>
<td>ADRA2C</td>
<td>MACS_peak_3444</td>
<td>(+60,732)</td>
</tr>
<tr>
<td>ADRBK1</td>
<td>MACS_peak_811</td>
<td>(-17,343)</td>
</tr>
<tr>
<td>AFF3</td>
<td>MACS_peak_2550</td>
<td>(+468,876)</td>
</tr>
<tr>
<td>AGRN</td>
<td>MACS_peak_2</td>
<td>(+53,567)</td>
</tr>
<tr>
<td>AHCL1</td>
<td>MACS_peak_189</td>
<td>(-33,761)</td>
</tr>
<tr>
<td>AK3L1</td>
<td>MACS_peak_138</td>
<td>(-164,345)</td>
</tr>
<tr>
<td>ALG8</td>
<td>MACS_peak_858</td>
<td>(-16,945)</td>
</tr>
<tr>
<td>AMZ1</td>
<td>MACS_peak_4178</td>
<td>(+9,213)</td>
</tr>
<tr>
<td>ANLN</td>
<td>MACS_peak_4222</td>
<td>(-116,421)</td>
</tr>
<tr>
<td>ANO1</td>
<td>MACS_peak_833</td>
<td>(-216,777)</td>
</tr>
</tbody>
</table>