A Brief Introduction on DNase-Seq Data Analysis

Hashem Koohy, Thomas Down, Mikhail Spivakov and Tim Hubbard
Spivakov’s and Fraser’s Lab

September 13, 2014
1 An Introduction to DNase-Seq Data Analysis
DNaseI is an enzyme which cuts duplex DNA at a rate that depends strongly on its chromatin environment.

In combination with HTS technology, it can be used to infer genome-wide landscape of open chromatin regions.

Using this technique, systematic identification of hundreds of thousands of DNaseI Hypersensitive Sites (DHS) per cell type has been possible (ENCODE).

In what follows, we provide a brief introduction about this technique with a particular emphasis on peak calling step.
Currently Existing DNase-Seq Protocols

Double Hit Protocol
Developed in John Stam Lab in University of Washington, and has been used greatly for detection of DHS in ENCODE project.

Single Hit Protocol
Developed in Greg Crawford Lab in Duke University. It has been used for detection DHS in ENCODE. This protocol is also in great use by some other researchers world-wide.

ATAC-Seq
Developed in Greenland Lab in Stanford University. This is a very new protocol (published 2013) and has been reported to be very fast and very efficient.
Two Commonly Used Protocols “End Capture” (Duke) and “Double Hit” (UW) Protocol

End Capture Protocol: Greg Crawford Lab, Duke

Double Hit Protocol: John Stam Lab, UW
An Introduction to DNase-Seq Data Analysis

ChIP-Seq vs DNase-Seq

Note that DNase HS is different from its sister DNase Footprinting

ChIP-Seq: Nature Reviews, Peter J. Park, 2009

DNase HS: Duke Protocol
TF ChIP-Seq vs DNase-Seq

Some key differences between TF ChIP-Seq and DNase-Seq:

- In ChIP-Seq data, a protein is usually in "bound" or "unbound" position, whereas DNase shows a more generic behaviour, representing the openness of the chromatin to any regulatory feature;

- DNase HS are strand-independent and therefore no need to shift size or tag extension;

- DNase HS data sometimes shows less enrichment over wider regions (a kind of Mixed-Source).
DNase-Seq Data Analysis

Similar to any other ChIP-Seq data, there are various steps:

<table>
<thead>
<tr>
<th>Sequencing, Mapping and Quality Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing is getting cheaper, providing us with more data! Mapping possibly is still the most computationally expensive part.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Peak Calling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gauging the statistical significance of reads’ enrichment which is generally known as &quot;Peak Calling&quot; is very central to ChIP-Seq data analysis.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Post Peak Calling Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different directions and purposes, including differential binding analysis, motif discovery, detection of regulatory regions, Genome segmentation and so on · · ·</td>
</tr>
</tbody>
</table>
## Context-Specificity of ChIP-Seq Data

Different protein classes have distinct mode of interactions:

### Point-Source

These factors and chromatin marks are localised specifically and have high signal-to-noise ratio.

### Broad-Source

These factors are associated with wide genomic domains, generating broad but more noisy signals; e.g. H3K9me3, H3K36me3.

### Mixed-Source

These factors show a point-source style signal at some regions whereas more broader in other regions e.g. RNA Pol II.
DNaseI Peak Calling Algorithms

- Note that, a peak caller is an algorithms and/or a model that is applied to gauge the statistical significance of enrichment of the short read tags.
- Context specificity of ChIP data –among some other reasons– has give rise to numerous peak callers.
- MACS and F-SEQ are considered among the best reported peak callers for the DNaseI-Seq.
- Throughout this tutorial, We will try to illustrate how to use these two peak callers for detection statistically significant open chromatin regions.
- We also show you very briefly how to visualise your data as the very fist step of validating your results.
F-Seq

- An histogram-based (number of tags per bin) approach is, possibly, the most naivest for gauging the enrichment of short read tags;
- However, it suffers from some problems including boundary effects and selection of bin width;
- To overcome, F-Seq suggested in which a Kernel Density Estimator (with mean 0 and variance 1) is applied to obtain the distribution of reads:

\[
p(x) = \frac{1}{nb} \sum_{i=1}^{i=n} K(\frac{x - x_i}{b})
\]

- F-Seq has been implemented in Java, easy to use, though, doesn’t support some commonly used file formats.
MACS

- The most used peak callers for ChIP-Seq data;
- It has been reviewed and benchmarked in different studies;
- At the time of development, the emphasis was on handling shift size and local biases from sequencability and mappability;
- A Poisson model is employed for identification of statistically significant enriched regions;
- MACS has been implemented in python and is relatively fast. It is user friendly and fairly well-supported.