SigSeeker: An Ensemble for Analysis of Epigenetic Data

Jens Lichtenberg
Elisabeth F. Heuston
David M. Bodine
Background - Epigenetics

- Heritable changes in gene activity not attributed to changes in DNA
- DNA Methylation
  - Punctuated modification with often direct impact on gene expression
- Histone modification
  - Long range modifications that impacts the accessibility of DNA for transcription
Background - Motivation

- Large variety of epigenetic approaches that produce sequencing data
- Few approaches exist that attempt the large scale comprehensive analysis of these data
- Common threads in the evaluation process
- Large number of tools for each process
Next Generation Sequencing

● De Novo Sequencing and Resequencing
  ○ Single nucleotide polymorphisms, insertions deletions, copy number variations

● Transcript Profiling and Discovery
  ○ High specificity, sensitivity and read coverage

● Bisulfite and ChIP Sequencing
  ○ Mapping accuracy, high sensitivity, low sample input
Discovery of regulatory binding sites in sequencing data

Discovery of functional correlations within omics data

Discovery of regulatory signatures for cell types and experimental conditions

Analysis - Objectives

- Quality Control
- Read Alignment
- Peak Calling
- Motif Discovery
- Omics Correlation
- Signature Assignment
Epigenetic analysis approaches discover regions in which mapped reads are over represented in a sample compared to a control.

Excess of 75 different epigenetic analysis approaches exist geared towards different research questions:
- Transcription factor binding site discovery
  - Slender regions containing specific sequence motifs
- Histone modification localization
  - Broad regions
- Methylation site discovery
  - Regions defined by contained methylated CpG sites
Questions
- Which approach is the best for the analysis?
- Is there a correlation between different approaches?
- Is there a correlation with existing data?

Hypotheses
- Different approaches have different advantages and disadvantages
- Integration of tools will provide better results for the evaluation of sequencing data
- Correlation of with established information provides deeper understanding
Framework - Overview

Aligned Reads

Format Conversion

Tool₁

Peaklist₁

Correlation

Correlation Report

Tool₂

Peaklist₂

Filtering

Filtering Report

Toolₙ

Peaklistₙ

Partitioning

Partitioning Report
Framework - Correlations

- **Structural Correlation**
  - Genomic sites detected by more than 1 tool

- **Intensity Correlation**
  - Co-localized peaks often differ greatly in their intensity
  - Intensities are subjected to linear regression model and outliers are removed
System

- Hematopoiesis is the formation of blood cellular components from a stem cell
- Biological system containing roughly 60 different cell types
- Cell types can be enriched through flow cytometry
- Foundation for important medical insights (e.g. Leukemia research)

Experiment

- DNA Methylation via MBD2 pulldown
- 4 different cell types (HSC, CMP, ERY, MEG)
- 2 replicates each
- 1 supernatant samples as control
Case Study - Results

- Observation of HSC > CMP > ERY (Hogart 2012)
- Generally MEG > ERY
- MACS > MACS2 > ERANGE > SICER
- Sole-Search produces strong outlier for ERY
# Case Study - Correlations

<table>
<thead>
<tr>
<th>HSC</th>
<th>ERY</th>
<th>MACS</th>
<th>SICER</th>
<th>Peaks</th>
<th>Upstream</th>
<th>Promoter</th>
<th>RefSeq</th>
<th>Downstream</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>86733</td>
<td>10433</td>
<td>1278</td>
<td>55566</td>
<td>14325</td>
</tr>
<tr>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>61196</td>
<td>7191</td>
<td>992</td>
<td>38793</td>
<td>10123</td>
</tr>
<tr>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>60452</td>
<td>7096</td>
<td>991</td>
<td>38289</td>
<td>9996</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td>47202</td>
<td>5622</td>
<td>681</td>
<td>29528</td>
<td>8031</td>
</tr>
<tr>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>36712</td>
<td>4368</td>
<td>531</td>
<td>22907</td>
<td>6321</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>35099</td>
<td>4160</td>
<td>512</td>
<td>21770</td>
<td>5982</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td>44794</td>
<td>5294</td>
<td>621</td>
<td>27482</td>
<td>7506</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>34566</td>
<td>4030</td>
<td>476</td>
<td>21105</td>
<td>5837</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>33351</td>
<td>3900</td>
<td>466</td>
<td>20271</td>
<td>5589</td>
</tr>
</tbody>
</table>
Framework - Benchmarks

Setup
- Positive and negative sites established for a subset of predicted peaks in well studied epigenetics data sets
  - Rye et al.
    - Transcription Factor Binding
  - Micsinai et al.
    - Histone Modification

Accuracy
- Closeness of measurements to true value
- Average ACC increases by about 150%
- Similar observations hold true for all benchmarking metrics
Conclusions

- **SigSeeker**
  - Framework for the analysis of epigenetics data
  - Ensemble of peak calling techniques

- **Insights**
  - Adding a tool to the ensemble seems to have a similar power as adding another replicate
  - Different backgrounds have strong impact on number of peaks called

- **Applications**
  - Hematopoiesis case study through large methylation pulldown experiment
  - Global decline in the number of predicted methylated sites during hematopoietic stem cell differentiation
Acknowledgements

Bodine Lab at NHGRI
- David M. Bodine
- Elisabeth F. Heuston
- Amber Hogart
- Stephanie Battle

NISC
- Jim Mullikin
- Baishali Maskeri

Flow Cytometry Core
- Stacie Anderson

Penn State University
- Ross Hardison
- Tejaswini Mishra

Johns Hopkins University
- Michael McDevitt

NHGRI
- Laura N. Elnitski
- Nigel Crawford

Ohio University
- Lonnie R. Welch
Thank you!
Background - Epigenetics

- **Transcription Factors**
  - Punctuated binding events that facilitates the recruitment of RNA pol II and subsequent transcription

- Antibodies have been designed for the recognition of DNA methylation, Histone modification and a large number of Transcription factors
# Analysis - Algorithms

<table>
<thead>
<tr>
<th>Tool</th>
<th>Citation</th>
<th>Peak Finding Categorization</th>
<th>Background Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACS</td>
<td>Zhang2008</td>
<td>Sliding Window</td>
<td>Poisson tag count distribution</td>
</tr>
<tr>
<td>cisGenome</td>
<td>Ji2008</td>
<td>Sliding Window</td>
<td>Negative binomial distribution</td>
</tr>
<tr>
<td>BayesPeak</td>
<td>Spyrou2009</td>
<td>Sliding Window</td>
<td>Negative binomial distribution</td>
</tr>
<tr>
<td>BroadPeak</td>
<td>Wang2013</td>
<td>Sliding Window, Predefined Regions</td>
<td>Poisson tag count distribution</td>
</tr>
<tr>
<td>Sole-Search</td>
<td>Blahnik2009</td>
<td>Sliding Window</td>
<td>Duplicate and Deletion Compensation</td>
</tr>
<tr>
<td>SICER</td>
<td>Zang2009</td>
<td>Binning</td>
<td>Poisson tag count distribution</td>
</tr>
<tr>
<td>E-Range</td>
<td>Mortazavi2008</td>
<td>Clustering Distance XSETs</td>
<td>Fold Enrichment</td>
</tr>
<tr>
<td>FSeq</td>
<td>Boyle2008</td>
<td>Gaussian KDE</td>
<td>Std. Dev. Threshold</td>
</tr>
<tr>
<td>FindPeaks</td>
<td>Fejes2008</td>
<td>Overlapping XSETs</td>
<td>Monte Carlo Simulation</td>
</tr>
</tbody>
</table>
Results - Case Study

- Hematopoiesis is a biological system containing roughly 60 different cell types
- Cell types can be enriched through flow cytometry
- Can be studied in adult specimen
- Good in-vitro vs in-vivo correlation
- Foundation for important medical insights (e.g. Leukemia research)
Novel peaks in ERY rarely conserved

Majority of CpG Island methylation is lost during differentiation

Peaks retained during differentiation discovered by all tools, ERY specific peaks are best described by MACS

ERY methylation leads to lack of transcription factor binding in ERY