DNA methylation
Epigenetics

http://nihroadmap.nih.gov/EPIGENOMICS/images/epigeneticmechanisms.jpg
Epigenetics

http://nihroadmap.nih.gov/EPIGENOMICS/images/epigeneticmechanisms.jpg
DNA Methylation

- DNA sequence:
  - T T T A C G A T T A C G A
  - A A A G C T A A T G C T

- Methylation sites:
  - Me at C
  - Me at C
Liver

Brain
Liver

Brain

Me → Me

Me → Me
GC counts on the genome

These are counts in 16 basepair bins
CpG are depleted

- These are counts in 16 basepair bins
- We see rate of about 1 in 100
CpG Islands

• But CpGs cluster into *islands* enriched near promoter

Irizarry et al. (2009) Mammalian Genome
Wu et al (2010) Biostatistics,
New illumina CpG array will use our CGI
Gardiner-Garden and Frommer
CpG Island definition

• N > 200
• GC-content > 50%
• obs/exp > 0.6
• Lists contain 20,000 CGI

HMM based definition
• Problems:
  – leaves out many clusters
  – Not applicable to other species
Whole genome view...
Why observed/expected and not counts?
GC content varies
Hidden Markov Model Approach

- Assume that GC content is smooth.
- Estimate and assume known: $p_C(t)$ and $p_G(t)$
- Assume probability of CpG is $\alpha_i p_C(t)p_G(t)$ for two states $i = 0, 1$.
- To avoid correlation problem, assume counts in bins of size $L$ is Poisson with rate is $\alpha_i p_C(t)p_G(t) L$
- We use $L=16$
- Use EM to estimate $\alpha_0$ and $\alpha_1$ from data and fit HMM

Conventional wisdom in 2004

• **Hyper**methylated CpG islands silence tumor suppressor genes

• **Cancer cells are globally hypomethylated**

High throughput measurement permitted us to observe the entire genome:

Irizarry et al. (2008) Genome Research
Aryee el al. (2010) Biostatistics
Finding differentially methylated regions (DMRs)

Irizarry et al. (2008) Genome Research
Aryee el al. (2010) Biostatistics
Jaffe et al (2012) IJE
Genomic traceplot

Microarray data after much preprocessing
General Model

\[ Y_{ij} = \beta_0(l_j) + X_i \beta_1(l_j) + \varepsilon_{ij} \]

Baseline methylation level
Effect at j-th position
Measurement error
Observed Data
Outcome of interest
Do we trust single measurements?

CpG #1

CpG #2

methylated

0.0

0.2

0.4

0.6

0.8

1.0

normal

tumor

normal

tumor
Do we trust single measurements?

Note $X$ is 1 (cancer) or 0 (normal)
CpG #1

![CpG plot]

Mean Methylation

Chromosome 19

51689000 - 51691500

Methylation

Mean Diff

51689000 - 51691500

Cancer vs Normal
CpG #2
Current general approach
Beware of batch effects

Methylation

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry
Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry
Next generation sequencing

Hansen et al. (2011) Nature Genetics
Bisulfite Treatment

<table>
<thead>
<tr>
<th>After</th>
<th>Before</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>Me -&gt; C</td>
<td>C — Me</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>
Whole Genome Bisulfite Sequencing

CTGCACTTGCTGCTTCTGCGCTCGCTATGCAACGATGATCCGG
Whole Genome Bisulfitate Sequencing

CTTGCTGCTTTCTGCCTCGCTCGCTATGCAACGATGAT
CTGCTTCTGCCTCGCTCGCTATGCAACGATGATCCGGCT
TTGCTGCTTTCTGCCTCGCTCGCTATGCAACGATGATCCGGCTGC
CTTGCTGCTTTCTGCCTCGCTCGCTATGCAACGATGATCCGGCTGC
TTGCTGCTTTCTGCCTCGCTCGCTATGCAACGATGATCCGGCTGC
TTGCTGCTTTCTGCCTCGCTCGCTATGCAACGATGATCCGGCTGC
TTGCTGCTTTCTGCCTCGCTCGCTATGCAACGATGATCCGGCTGC
TTGCTGCTTTCTGCCTCGCTCGCTATGCAACGATGATCCGGCTGC
TTGCTGCTTTCTGCCTCGCTCGCTATGCAACGATGATCCGGCTGC
ACTTGCTGCTTTCTGCCTCGCTCGCTATGCAACGATGATGA
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCC
Count Cs and Ts at CpG location

CTTGCTGCTTCTGCGCTCGCTATGCAACGATGAT
CTGCTTCTGCCTCGCTATGCAACGATGATCCGGCT
TTGCTGCTTCTGCCTCGCTATGCAACGATGATCCGGCT
ACTTGCTGCTTCTGCGCTCGCTATGCAACGATGAG
TTGCTGCTTCTGCGCTCGCTATGCAACGATGATCC
CTGCTTCTGCGCTCGCTATGCAACGATGATCCCG
TGCTGCTTCTGCGCTCGCTATGCAACGATGATTC
CTGCTTCTGCGCTCGCTATGCAACGATGATCCCG
TGCTGCTTCTGCGCTCGCTATGCAACGATGATTC
TTGCTGCTTCTGCGCTCGCTATGCAACGATGATCCCG
CTGCACTTTGCTGCTTCTGCTCGCTATGCAACGATGATCCGG
Quantitative Measurement: 80%

CTGCACTTTGCTGCTTCTGCGCTCGCTATGCAACGATGATCCGG
The cost of 30x

• We need biological replicates

• $3 \times 10^9 \times \text{bases} \times (\$ \text{per base}) \times \# \text{samples} = \text{more $ than collaborator has}$

• Can we smooth to save $?
M-bias plots for sequencing

![Graphs showing M-bias plots for sequencing](image)

The graphs illustrate the M-bias plots for sequencing in different positions, comparing normal and cancer samples. Each graph shows the ratio $M/(M+U)$ at various positions, with distinct markers for normal and cancer samples. The plots are labeled with corresponding positions and sample types, aiding in the visualization of sequencing biases across different conditions.
The Data
logit2 (Methylation)
Methylation

0.2

0.5

0.8
Smoothing on 4x vs 30x
Smoothing on 4x vs capture data
Two levels
Differentially methylated region
Hypomethylated blocks

(a) chr1: 175,889,691–178,389,661

(b) chr6: 16,512,239–17,912,058

Methylation

0.2 0.5 0.8

100kb

PMD
LOCKS
LAD
Islands
Genes
End