Introduction to Bioinformatics of Bisulfite Sequencing Methylation Data

Garrett Jenkinson, PhD
Lead Informatics Specialist
Biomedical Statistics and Informatics
Department of Health Sciences Research
Genetics versus Epigenetics

• Which is more different at cellular level phenotypically?
  - Your heart cells from your brain cells
  - A monkey’s heart cells from your heart cells

• How about the old nature versus nurture question?
  - Two healthy but unrelated peoples’ livers
  - Identical twins’ livers when only one is alcoholic

• Regulation of gene expression can be as critical as underlying genomic sequence
  - A gene can be turned off by regulation which can be functionally the same as obliteration of the genomic sequence
  - Not all regulation is epigenetic and causality rarely understood…don’t be oversold
DNA methylation is crucial if you want to understand:

- Developmental biology, stem cells, differentiation
- Carcinogenesis, imprinting disorders
- Aging, environmental exposures
Biology of DNA methylation in mammals

- Covalent addition of a methyl group to the 5’ carbon of cytosine residues (5mC)
- Predominantly at CG dinucleotides (CpG sites)
CpG sites

- The “p” represents phosphate backbone to distinguish between CpG and C—G hydrogen bonding between the strands of DNA
- Only positions in human genome with known mechanisms for epigenetic inheritance past cell division (DNMT enzymes)
- Dense regions of CpG sites referred to as CpG islands which are flanked by shores, shelves and then CpG depleted open seas
- Methylated islands in promoters linked to repressed gene expression
  - Methylation has complicated relationships to chromatin structure and gene expression
  - Mechanistic understanding of DNA methylation in gene regulation is incomplete
Agouti Mouse Model

- Genetically identical, phenotype differences driven by difference in methylation at agouti gene
- Expose pregnant mice to bisphenol A (BPA in plastic products)
  - Disproportionate number of yellow, obese progeny than would normally be expected
  - DNA methylation at the agouti gene sites is decreased (hypomethlyated)
- Need sequencing methods to probe the state of DNA methylation
Detailed View of Bisulfite Sequencing

Legend:
- CpG’s in red = original sequence
- CpG’s in blue = converted
- Positions corresponding to original C’s in CpG underlined

In IGV:
- OT = original top strand
- CTOT = complementary to OT
- OB = original bottom strand
- CTOB = complementary to OB

For OT, CTOT: C>T; C>C
For CTOB, OB: G>A; G>G

ATATCGCGTATT-3’
TATAAGCGCATAA-5’

m

ATATCGUGTATT-3’
TATAAGCGUATAA-5’

m

mol

Bisulfite conversion

ATATCGTGATT-3’
TATAAGCATATAA-5’

CTOT

ATATTGTGTATT-3’
TATAAGCATAA-5’

CTOT

OT

Polymerase chain reaction

ATATCGCATATT-3’
TATAAGCTATAA-5’

CTOB

ATATCACATATT-3’
TATAAGGTATAA-5’

CTOB

OB

Sequencing, alignment, visualization in IGV with bisulfite mode (CG)

ATATCGCGTATT-3’

reference

ATATCGTATT-3’ OT, CTOT
ATATCGCATATT-3’ CTOB, OB

ATATTGTATT-3’ OT, CTOT
ATATCACATATT-3’ CTOB, OB

QC and Alignment of BS-seq data

• Need specialized algorithms/tools to deal with “heavily mutated” BS-seq data

• trimgalore! is a package that wraps cutadapt and allows for the trimming of low quality bases and adapters from sequencing reads

• Bismark is a bisulfite-aware aligner using bowtie2
  • Can also produce QC and methylation summarization information
Post-alignment Data in IGV
Common BS-seq methods

- WGBS completely unfocused
  - Comprehensive ~13 million CpG sites profiled
  - Gold standard
  - ~$5K per sample
- RRBS, 1% of genome with 1.5 million CpGs
  - Most common BS-seq
  - Restriction enzymes chop DNA and results in enrichment for CGIs
  - ~$500 per sample
- “Capture” protocols (e.g., EPIC TruSeq), 3 million CpGs
  - Least common
  - Looks more like “focused” WGBS
"Raw" Data

unmethylated site
methylated site
no data
Methylation status not as “fixed” as genetic

- Populations of genetically homogeneous cells can and do differ in methylation
- Maternal and paternal alleles can and do differ (e.g., imprinting)
- At a given time, each cell’s DNA is either methylated (1) or unmethylated (0), but state can change during life of cell
- End result: we talk of probability that a CpG site is methylated in a given tissue/sequencing run
Marginal Estimation

\[ X_n = 1 \text{ if } \text{nth site methylated} \]

\[ X_n = 0 \text{ if unmethylated} \]

\[ P_n(1) = \Pr[X_n=1] \]

unmethylated site

methylated site

no data
Marginal Estimation

\( X_n = 1 \) if nth site methylated
\( X_n = 0 \) if unmethylated

\( P_n(1) = \Pr[X_n=1] \)

\( p_n(1) = \frac{8}{14} \)
Marginal Estimation

\[ X_n = 1 \text{ if nth site methylated} \]
\[ X_n = 0 \text{ if unmethylated} \]

\[ P_n(1) = \Pr[X_n=1] \]

\[ p_n(1) = \frac{8}{14} \]
\[ \frac{7}{14} \]
Marginal Estimation

\(X_n = 1\) if nth site methylated
\(X_n = 0\) if unmethylated

\(P_n(1) = \Pr[X_n=1]\)

\(p_n(1) = \frac{8}{14}, \frac{7}{14}, \frac{7}{15}\)
Marginal Estimation

\[ X_n = 1 \text{ if nth site methylated} \]
\[ X_n = 0 \text{ if unmethylated} \]

\[ P_n(1) = \Pr[X_n=1] \]

\[ p_n(1) = \frac{8}{14} \quad \frac{7}{14} \quad \frac{7}{15} \]
Marginal Estimation

\[ X_n = 1 \text{ if nth site methylated} \]
\[ X_n = 0 \text{ if unmethylated} \]

\[ P_n(1) = \Pr[X_n=1] \]

\[ p_n(1) \]

\[
\begin{align*}
8/14 & & 7/14 & & 7/15 & & 8/14 & & 11/14
\end{align*}
\]
Marginal Estimation

\(X_n = 1\) if nth site methylated
\(X_n = 0\) if unmethylated

\(P_n(1) = \Pr[X_n=1]\)

\(p_n(1)\)

\[
\begin{array}{cccccc}
8/14 & 7/14 & 7/15 & 8/14 & 11/14 & 10/14
\end{array}
\]
Marginal Estimation

\( X_n = 1 \) if nth site methylated
\( X_n = 0 \) if unmethylated

\[ p_n(1) = \Pr[X_n=1] \]

\( p_n(1) \)
\[
\begin{array}{ccccccc}
8/14 & 7/14 & 7/15 & 8/14 & 11/14 & 10/14 & 11/15
\end{array}
\]
Marginal Estimation

\(X_n = 1\) if nth site methylated
\(X_n = 0\) if unmethylated

\(P_n(1) = \Pr[X_n = 1]\)

<table>
<thead>
<tr>
<th>(p_n(1))</th>
<th>(8) (7)</th>
<th>(7) (8)</th>
<th>(11) (10)</th>
<th>(11) (11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\frac{14}{14})</td>
<td>(\frac{14}{15})</td>
<td>(\frac{14}{15})</td>
<td>(\frac{14}{14})</td>
<td>(\frac{15}{15})</td>
</tr>
</tbody>
</table>
Marginal Estimation

\[ X_n = 1 \text{ if nth site methylated} \]
\[ X_n = 0 \text{ if unmethylated} \]

\[ P_n(1) = \Pr[X_n=1] \]

\[ p_n(1) = \begin{array}{c}
\frac{8}{14} & \frac{7}{14} & \frac{7}{15} & \frac{8}{14} & \frac{11}{14} & \frac{10}{14} & \frac{11}{15} & \frac{11}{15} & \frac{10}{13} \\
\end{array} \]
### Marginal Estimation

- $X_n = 1$ if nth site methylated
- $X_n = 0$ if unmethylated

$P_n(1) = \Pr[X_n=1]

<table>
<thead>
<tr>
<th>$p_n(1)$</th>
<th>$\frac{8}{14}$</th>
<th>$\frac{7}{14}$</th>
<th>$\frac{7}{15}$</th>
<th>$\frac{8}{14}$</th>
<th>$\frac{11}{14}$</th>
<th>$\frac{10}{14}$</th>
<th>$\frac{11}{15}$</th>
<th>$\frac{11}{15}$</th>
<th>$\frac{10}{13}$</th>
<th>$\frac{10}{13}$</th>
</tr>
</thead>
</table>

Smoothed Marginals

Use smoothing to improve marginal estimates

unmethylated site
methylated site
no data

raw estimates:

\[
\begin{align*}
8 & \quad 7 & \quad 7 & \quad 8 & \quad 11 & \quad 10 & \quad 11 & \quad 11 & \quad 10 & \quad 10 \\
\frac{14}{14} & \quad \frac{14}{15} & \quad \frac{14}{15} & \quad \frac{14}{14} & \quad \frac{14}{14} & \quad \frac{14}{15} & \quad \frac{15}{15} & \quad \frac{13}{15} & \quad \frac{13}{13} \\
\end{align*}
\]
Smoothed Marginals

raw estimates:

smoothed estimates:
Smoothed Marginals

raw estimates:

smoothed estimates:
Smoothed Marginals

Unmethylated site  methylated site  no data

Raw estimates:

\[
\begin{array}{cccccccccc}
8 & 7 & 7 & 8 & 11 & 10 & 11 & 11 & 10 & 10 \\
14 & 15 & 14 & 14 & 15 & 15 & 13 & 13 & 13 & 13 \\
\end{array}
\]

Smoothed estimates:

\[
\begin{array}{cccccccccc}
15 & 323 & 323 & 323 & 15 & 630 & 630 & 630 & 630 & 630 \\
28 & 630 & 630 & 630 & 28 & 630 & 630 & 630 & 630 & 630 \\
\end{array}
\]

27
Smoothed Marginals

raw estimates:

smoothed estimates:

unmethylated site
methylated site
no data
Smoothed Marginals

raw estimates:

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>7</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

smoothed estimates:

<table>
<thead>
<tr>
<th></th>
<th>15</th>
<th>323</th>
<th>323</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>630</td>
<td>630</td>
<td>630</td>
</tr>
</tbody>
</table>

29
Smoothed Marginals

unmethylated site
methylated site
no data

raw estimates:

\[
\begin{array}{cccccccc}
8 & 7 & 7 & 8 & 11 & 10 & 11 & 11 \\
14 & 14 & 15 & 14 & 14 & 14 & 15 & 15 \\
\end{array}
\]

smoothed estimates:

\[
\begin{array}{cccccccc}
15 & 323 & 323 & 383 & 29 & 67 & 229 & 315 \\
28 & 630 & 630 & 630 & 42 & 90 & \text{no data} & \text{no data} \\
\end{array}
\]
Smoothed Marginals

raw estimates:

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>7</th>
<th>7</th>
<th>8</th>
<th>11</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>14</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>15</td>
</tr>
</tbody>
</table>

smoothed estimates:

<table>
<thead>
<tr>
<th></th>
<th>15</th>
<th>323</th>
<th>323</th>
<th>383</th>
<th>29</th>
<th>67</th>
<th>229</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>630</td>
<td>630</td>
<td>630</td>
<td>42</td>
<td>90</td>
<td>315</td>
<td>585</td>
</tr>
</tbody>
</table>

unmethylated site
methylated site
no data

33
Smoothed Marginals

Unmethylated site •
Methylated site ⊗
No data

Raw estimates:

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>7</th>
<th>7</th>
<th>8</th>
<th>11</th>
<th>10</th>
<th>11</th>
<th>11</th>
<th>10</th>
<th>13</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>14</td>
<td>15</td>
<td>14</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>13</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

Smoothed estimates:

<table>
<thead>
<tr>
<th></th>
<th>15</th>
<th>323</th>
<th>323</th>
<th>383</th>
<th>29</th>
<th>67</th>
<th>229</th>
<th>436</th>
<th>443</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>630</td>
<td>630</td>
<td>630</td>
<td>42</td>
<td>90</td>
<td>315</td>
<td>585</td>
<td>585</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>
Methylation is a coordinated phenomena

- Rarely do we care about methylation for a single CpG site…often care about entire island’s coordinated behavior
  - To the extent people care about single sites, it is due to the highly correlated/coordinated behaviors of site with neighbors
- “Marginal” view of methylation as a probability at each site is inadequate to capture the richness and diversity of the underlying biology
Stochasticity: Epipolymorphism/Entropy

Joint Probability Distributions

• Need to talk about probabilities of *patterns* of CpG sites

• From such probabilities, any other quantity of interest is available
  • Epipolymorphism
  • Entropy

• Now possible to detect not just hypo- or hyper-methylation changes in the mean, but any difference in methylation behavior
Empirical Estimation
Empirical Estimation
The 1017 other patterns are assigned zero probability.
Ising model

- Each read is a single-cell measurement even in bulk sequencing
- Means and nearest-neighbor correlations frequently observed
- 1D Ising model is MaxEnt model consistent with these quantities
- Well studied model in statistical physics with many existing computational techniques/results
- Provides full joint distribution
Ising model performance

- Empirical and marginal methods under- and over-estimate heterogeneity
- Ising is accurate even in low data
Ising model specification

\[ P(\mathbf{x}) = \frac{1}{Z} \exp\{-U(\mathbf{x})\} \quad Z = \sum_{\mathbf{x}} \exp\{-U(\mathbf{x})\} \]

\[ U(\mathbf{x}) = - \sum_{n=1}^{N} a_n(2x_n - 1) - \sum_{n=2}^{N} c_n(2x_n - 1)(2x_{n-1} - 1) \]

- All patterns have non-zero probability
- General model requires estimation of \( a_n \) and \( c_n \) parameters; \((2N-1) \ll 2^N\)
- Improve performance further by imposing parametric structure based on the biology
Normalized Methylation Entropy

\[ h = - \left\{ \sum_m \Pr[M = m] \log_2 \Pr[M = m] \right\} / \log_2(L + 1) \]

- Rigorously quantifies stochasticity in DNA methylation using Shannon entropy
- Another degree-of-freedom compared to standard mean analyses
- Shown to have discriminatory power in aging, carcinogenesis and stem cell differentiation
Jensen Shannon distance
Information Theoretic Bioinformatics Software

• informME is an information theoretic package designed to implement the Ising model, NME, JSD
  • Available as a thoroughly used/tested matlab/C++ code base, with bash wrappers and SLURM/SGE submission scripts
  • Or recently informME.jl is released as a trial package in julia language requiring no licensing or complex pipelines
Example Application

Multidimensional, longitudinal assays of the NASA Twins Study. (Left and middle) Genetically identical twin subjects (ground and flight) were characterized across 10 generalized biomedical modalities before (preflight), during (inflight), and after flight (postflight) for a total of 25 months (circles indicate time points at which data were collected). (Right) Data were integrated to guide biomedical metrics across various "-omes" for future missions (concentric circles indicate, from inner to outer, cytokines, proteome, transcriptome, and methylome).
Highlights of DNA methylation in twins study

- Twin astronauts with similar past flight experience studied in detail during longest American spaceflight in history
- Surprising result that space twin globally had less DNA-methylation variability than ground twin; hypotheses why?
Focal changes in DNA methylation

• Less surprising results when looking for focal genes with DNA methylation differences:
  • Regulation of ossification, and cellular response to ultraviolet-B (UV-B), platelet aggregation
  • Somatostatin signaling pathway and regulation of superoxide anion generation
  • Response to platelet-derived growth factor (PDGF) and T cell differentiation and activation pathways
Example Detailed Analysis of NOTCH3

<table>
<thead>
<tr>
<th>CD4: NOTCH3</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr19: NOTCH3 &lt; CGI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TW vs HR</th>
<th>Pre-Flight</th>
<th>Late Flight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dMML</td>
<td>dMML</td>
</tr>
<tr>
<td></td>
<td>dNME</td>
<td>dNME</td>
</tr>
<tr>
<td></td>
<td>JSD</td>
<td>JSD</td>
</tr>
</tbody>
</table>
Papers for more detail or applications
Questions?