Epigenomics: Some Statistical Applications

Rafael A. Irizarry
Department of Biostatistics
John Hopkins Bloomberg School of Public Health
Acknowledgements

• Tom Albert, Nimblegen
• Benilton Carvalho, JHU Biostatistics
• Andy Feinberg Lab, JHU Medicine
• Todd Richmond, Nimblegen
• Hao Wu, JHU Biostatistics
• Jean Wu, Brown Biostat
• Vasan Yegnasubramanian, JHU Oncology
Outline

• Quick Introduction to Epigenetics
• Introduction to Methylation
• Overview of competing technologies
• Review: Expression arrays lessons
• Comparison
• Role of statisticians
Genetics: the alphabet of life

- Letters of DNA sequence carry the information
Epigenetics
(3.4×10^{-10} \text{ meters/bp}) \times (6\times10^9 \text{ bp/genome}) = \sim 2 \text{ meters/genome}

Radius of the nucleus is \sim 10 \mu\text{M} !!!

Klug and Cummings, 1997
\[
\frac{(6 \times 10^9 \text{ bp/genome})}{(195 \text{ bp/nucleosome})} = \sim 30.8 \times 10^6 \text{ nucleosomes/genome}
\]

\sim 5 \% \text{ of nuclear volume}

(Klug & Cummings 1997)
Human DNA is 2 billion nanometer (nm) long, and 2 nm wide. The diameter of the cell nucleus is typically about 2,000 to 6,000 nm.

Nucleosome, Solenoid Model of Chromatin and Chromosomes

A partially extended (left) and condensed (right) version of the helical chromatin superhelix, according to the solenoid model of Atkin & Fogg. Each turn of the helix has ten nucleosomes, and contains about 1200 base pairs. At this stage the chromatin fiber is 30 nm thick.

A chromosome is formed by the coiling of the solenoid fiber, shown on the right side. Each of the two chromosomes contain approximately 1 million nucleosomes.

http://www.albany.edu/~achm110/solenoidchromatin.html
Epigenetics: the grammar of life
DNA methylation

\[
\text{Observed to expected} = \frac{\text{Pr(CG)}}{\{\text{Pr(C)} \cdot \text{Pr(G)}\}}
\]
DNA methylation can lead to silencing of gene expression

Robertson and Wolffe, Nat Rev Genet, 2000
Expression Array Lessons
Normalization

A  Hypomethylated sample

B  Hypermethylated sample
Probe effect

Intensity = Background + Probe Effect x Quantity x Error
Sequence effect for BG

Wu et al. (2004) JASA 99(468) 909

\[
Affinity = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} 1_{b_k = j} \quad \mu_{j,k} \sim \text{smooth function of } k
\]
Back to Methylation

High throughput of course....
Densities for three methods

HCT116 lots of methylation
DKO very little methylation
Hunh?
MeDIP (like ChIPchip)

1. Crosslink
2. Lyse & Sonicate
3. Other controls for IP (e.g., no antibody, non-specific antibody)
4. Total Reverse crosslinks
5. Amplify Label/hybridize

Other controls for IP (e.g., no antibody, non-specific antibody)
Some Data

log intensity (total genomic DNA)

log intensity (total genomic DNA)

log ratio

CpG counts
Problem: Not specific
HELP: Two enzymes

Cuts at CCGG

Cuts at $\text{C}^{\text{M}}\text{CGG}$

No Methylation
HELP after PCR

No Methylation
HELP

Methylation
Problem with HELP

Cuts at CCGG

Cuts at $C^{M}CGG$

No Methylation
The Problem

Obsered to expected = \( \frac{\text{Pr}(CG)}{\{ \text{Pr}(C) \text{ Pr}(G) \}} \)
Proportion of neighboring CpG also methylated/not methylated
McRBC on Tiling array
ROC now

[Graphs showing ROC curves for different methods]
Problems for Statisticians

• Background Correction + Normalization
• Probability Model for Segments
• Use these to from null and alternative models... we need power!
• Use these to create bump finding algorithms
• Adapt to high-throughput sequencing
Supplemental Slides
McRBC: One enzyme

Cuts at A<sup>m</sup>CG or G<sup>m</sup>CG

Input

No Methylation
McRBC after Gel

No Methylation
McRBC after Gel

No Methylation
McRBC

Methylation
McRBC after GEL

Methylation
McRBC after GEL

Methylation