Epigenetics

• The study of heritable changes in gene function that occur without a change in the sequence of the DNA.

• Critical for development and differentiation

• Epigenetic dysregulation is involved in developmental diseases, cancers, certain infectious diseases, and more
Epigenetics

• Chemical changes to the DNA and histone proteins that affect packaging

• DNA methylation
  (Typically) occurs at CpGs in mammals:

\[
\begin{align*}
5' & \ldots \text{CATTTCGACCGT}\ldots 3' \\
3' & \ldots \text{GTAAGCTGCA}\ldots 5'
\end{align*}
\]
Identical genetic sequence, but very different gene expression...
DNA packaging

6ft of DNA packed into each nucleus
DNA methylation can lead to silencing of gene expression.

Unmethylated DNA
Loosely packed
Active gene expression

Methylated DNA
Tightly packed
Silenced gene expression
DNA methylation is heritable through cell division

Rafael Irizarry
DNA methylation is heritable through cell division
DNA methylation is heritable through cell division
DNA methylation is heritable through cell division
CpG Islands

CG counts in non–overlapping 16 basepair window

- CpGs are depleted
- Remaining CpGs cluster into islands enriched near promoters
- There are tens of thousands of CpG Islands
There is significant information stored in the packaging of the genome that impacts development and disease.
Measuring DNA methylation

• Goals
  • Identify methylated regions within samples
  • Compare methylation between samples
  • Estimate percentage methylation

• Challenges:
  • How to make different parts of the genome comparable
  • How to make different samples comparable
Quantifying epigenetic marks

- Compare marks within a sample

- Compare marks between samples

Epigenetic marks to be discussed

Histone modifications
DNA methylation
DNA methylation profiling

• Gold standard: Bisulfite-Seq
  - Bisulfite treatment followed by sequencing
  - Single base pair resolution
  - Expensive on genome-wide scale (but becoming cheaper)

Human DNA methylomes at base resolution show widespread epigenomic differences

Ryan Lister¹*, Mattia Pelizzola¹*, Robert H. Dowen¹, R. David Hawkins², Gary Hon², Julian Tonti-Filippini⁴, Joseph R. Nery¹, Leonard Lee², Zhen Ye², Que-Minh Ngo², Lee Edsall², Jessica Antosiewicz-Bourget⁵,⁶, Ron Stewart⁵,⁶, Victor Ruotti⁵,⁶, A. Harvey Millar⁴, James A. Thomson⁵,⁶,⁷,⁸, Bing Ren²,³ & Joseph R. Ecker¹

Nature, 2009
DNA methylation is spatially correlated
WHOLE GENOME EPIGENETIC ASSAYS

Epigenetic mark sensitive assay

- Epigenetic mark-dependent restriction enzymes
- Sodium bisulfite treatment
- Pull-down of enriched DNA

DNA fragment detection

- Microarrays
- High-throughput sequencing
WHOLE GENOME EPIGENETIC ASSAYS

Epigenetic mark sensitive assay
- Epigenetic mark-dependent restriction enzymes
- Sodium bisulfite treatment
- Pull-down of enriched DNA

DNA fragment detection
- Microarrays
- High-throughput sequencing
McrBC Fractionation

1. Shear DNA (1.5 kb - 3.0 kb)
2. Mock digestion (UT) and McrBC digestion (MD)
3. Agarose gel size fractionation
4. Differentially label input and methyl-depleted fractions
5. Cohybridization to CHARM microarray

TOTAL INPUT

METHYL-DEPLETED
McRBC

Input

Cuts at $A^mCG$ or $G^mCG$

No Methylation
McRBC

Methylation
CHARM microarray

- Custom 2 million probe array targeting CpG rich regions
- Unbiased by annotation (e.g. CpG islands)
- Design steps:
  - Divide non-repetitive genome into segments
  - Rank segments by CpG density
  - Fill array starting with CpG-rich regions
- Covers 3.5M CpGs (≈20%)
CHARM microarray: Methylation signal

\[ M_i = \log_2 \left( \frac{G_i}{R_i} \right) \]

- \( G_i \): total input fluorescence at probe \( i \)
- \( R_i \): methyl-depleted fluorescence at probe \( i \)
- \( M_i \): methylation signal at probe \( i \)
Background signal

\[ M_i = \log_2 \left( \frac{G_i}{R_i} + b^g_i + b^r_i \right) \]

Estimate using anti-genomic probes
Waves

Control probe signal in unmethylated regions
MA plots

\[ M_i = \log_2 \left( \frac{G_i}{R_i} \right) \]

\[ A_i = \log_2 \left( \frac{G_i + R_i}{2} \right) \]

- Standard microarray data assumption: True log ratio is independent of signal strength
- True methylation log ratio is correlated with CpG/GC Density and hence signal strength
Control probes

- Probes from known unmethylated regions
- Long CpG-free stretches of the genome
- 5000 probes
Probe effects

- Signal strength depends on probe sequence

Control probes from unmethylated regions
Microarray data: Control probe Loess normalization

- Log ratio is independent of signal strength in unmethylated regions
- Fit curve through unmethylated control probes

Unmethylated control probes

Unmethylated regions
Setting the zero level

Raw data

After normalization

Methylation

Methylation

Zhijin Wu
Between sample normalization

- 4 healthy tissues + 1 cancer
  
liver    brain    spleen    colon    colon cancer
Between sample normalization

- Normalization relies on assumptions about things that should be the same between samples
- **Quantile normalization**: Assumes the distribution of intensity should be the same across samples

Katagiri et al., Current Protocols in Molecular Biology, 2009
Quantile normalization

Colon: cancer & normal
Quantile normalization leads to bias in (some) methylation data

- Assumption of equal methylation levels often violated
e.g. Cancer vs normal, Stem cells vs differentiated cells
True signal (simulated)
Observed data

![Observed data graph](image-url)

**Location**

**Methylation**

- 0.0
- 0.2
- 0.4
- 0.6
- 0.8
- 1.0

- 1250000
- 1252000
- 1254000
- 1256000
- 1258000
- 1260000

Rafael Irizarry
Observed data and true signal
What is methylated above 50%?
Naive approach
Many false positives (FP)
Smooth
No FP, but one false negative
Smooth less? No FN, lots of FP
We prefer this!
**WHOLE GENOME EPIGENETIC ASSAYS**

**Epigenetic mark sensitive assay**
- Epigenetic mark-dependent restriction enzymes
- Sodium bisulfite treatment
- Pull-down of enriched DNA

**DNA fragment detection**
- Microarrays
- High-throughput sequencing