Techniques for epigenetic analysis
How to apply them to human and epidemiology studies

Andrea Baccarelli, MD, PhD, MPH
Laboratory of Environmental Epigenetics
Harvard School of Public Health

abaccare@hsph.harvard.edu
How to detect epigenetic marks?

**DNA methylation**
Methyl marks added to certain DNA bases repress gene activity

**Histone modifications**
A combination of different molecules can attach to the ‘tails’ of proteins called histones. These alter the activity of the DNA wrapped around them

**microRNAs (snRNAs)**
Small non-coding RNAs that cause mRNA degradation or impair translation into protein
DNA methylation & histone modification analysis

• Molecular biology: largely related to genetics
  – methods that analyze the DNA sequence

• Epigenetic marks: do not modify the underlying DNA sequence

• Workarounds:
  – Bisulfite treatment
  – Antibody-based methods
    (or alternatively methyl-binding proteins)
Bisulfite treatment

- Modifies non methylated cytosines
  - Differentiation of methylated and non methylated cytosines
    - Any method that can analyze sequence
      - DNA methylation

Antibodies

- Bind modified or methylated cytosines, modified histones
  - DNA enriched with the mark of interest (Ab specificity)
    - Any method that can quantify enrichment
      - Histone modifications

Workaround

Analysis
Some Lab Nomenclature

- DNA Methylation
  - Gene-specific analysis
    - How much methylation at or nearby a candidate gene
  - Global methylation content
    - How much methylation in a test DNA, regardless of the position
  - Genome-wide scans
    - Microarrays, Next Generation Sequencing

- Histone Modifications
  - Gene-specific analysis
  - Global modification content
  - Genome-wide scans
Bisulfite modification of DNA

- Prior to PCR, DNA is treated with sodium bisulfite
- Non-methylated C is permanently modified to U
- In PCR, U and T are equivalent
DNA methylation methods

• DNA Methylation Techniques
  – Gene-specific analysis
    • Qualitative
    • Quantitative
  – Global methylation content
    • How much methylation in a test DNA, regardless of the position
  – High-coverage methods
    • Genome-wide scans, Arrays
Gene-Specific Analysis
Workflow for DNA methylation analysis by bisulfite-pyrosequencing

Biospecimen

DNA isolation

Bisulfite treatment

PCR amplification

Pyrosequencing

Data analysis

%5mC = \frac{\text{methylated C}}{\text{(methylated C) + (unmethylated C)}} \times 100\%

Abbreviations:
%5mC, percentage of 5-methylcytosine;
C, cytosines
Pyrosequencing
Pyrogram

A1 : TTC/TGTGGTGC/TGTC/TG

84%
76%
76%
Pyrosequencing analysis

• Provides overview of the methylation pattern
• Allows for the measurement of the methylation percentage of individual CpG dinucleotides
• Flexibility in sequencing primer position to analyze any CpG sites you like
Pyrosequencing & other methods

- **Pyrosequencing**
  - Pro: Highly quantitative, single site resolution
  - Cons: dedicated equipment
- **Methylation Specific PCR**
  - Pro: inexpensive and easy to perform
  - Cons: qualitative/semiquantitative, no single site resolution
- **Real-time PCR**
  - Pro: equipment easily accessible
  - Cons: low precision, no single site resolution
- **Maldi-TOF (Sequenom Mass Array)**
  - Pro: Quantitative, single site resolution, extended sequence (amplicon)
  - Cons: dedicated equipment (high costs), high costs/gene
DNA Methylation Content
Figure 7.13 Genomes 3 (© Garland Science 2007)
Global Methylation

• Most methylation in repeated elements:
  – LINE-1 elements: >500K/haploid genome
  – Alu elements: >1,100K/haploid genome

• LINE-1/Alu methylation is correlated with global content (Weisenberger, 2005)

• Function of repeated elements?:
  – Chromosomal structure
  – Repeat sequence transcription
  – miRNA
How to confuse one reader’s mind

• Global DNA methylation
• Genomic methylation content
• Genome-wide methylation content
• Genome-wide methylation
• Global cell methylation
Lower Global Methylation

• Tissue DNA
  – Cancer (Feinberg & Vogelstein, 1982)
  – Atherosclerotic lesions (Hiltunen, 2002)

• Blood DNA
  – Cancer (Hsiung, 2007)
  – Cardiovascular Disease (Castro, 2003)
  – Folate deficiency (Choi, 2005)
  – Inflammatory states (Stenvinkel, 2007)
  – Aging (Fuke 2004)
Analysis of Global Methylation Content

• Direct measurement of methyl group content
  – GC/MS (Rossella F et al., Rapid Comm MS 2009)
  – Immunostaining with anti-5mC
  – Digestion with methylation sensitive enzymes (MspI, HpaII)

• Estimated in repeated elements
  – Based on Pyrosequencing (Yang et al. Nucl Ac Res 2004)
  – Based on Methylight (Weisenberger et al. Nucl Ac Res 2005)
Retrotransposons

- Transposition: Movement of gene from one chromosome to another or movement from one site to another; does not require homology
- Transposons: Mobile genetic elements that enable genes to move between non-similar sites
- Retrotransposition: Creates genetic diversity
- Retrotransposons: Replicate and move to other sites on DNA through an RNA intermediate
What do we measure in repeated elements

• A few CpG sites in a specific repeated element sequence
• The sequence (and the CpG sites therein) are repeated throughout one single haploid genome:
  – LINE-1 elements: >500K/haploid genome
  – Alu elements: >1,100K/haploid genome
  – The LINE-1 and Alu assays measure the %mC in those repeated elements, with no distinction about their position in the genome
• Correlation with global content demonstrated only in studies including cancer tissues
  – A marker of global methylation only for cancer tissues?
Genome-wide scans

• Microarrays
  – Illumina Infinium for DNA methylation (bisulfite treatment)
    • 484,000 CpG sites (450K, released in early 2011)
    • $400/samples
  – Nimblegen for DNA methylation or Histone Modifications (Ab-based)
    • 2 Million probes
    • >$2000/samples

• Next Generation sequencing
  – Various platforms
  – Bisulfite treatment (DNA methylation)
  – Ab-based (DNA methylation or Histone modifications)
  – $3000/sample for RRBS
Illumina 450K BeadChip Coverage

CpG shelves, shores & islands classification (UCSC CpGi annotation)
The 450K BeadChip covers a total of 77,537 CpG Islands and CpG Shores (N+S)

<table>
<thead>
<tr>
<th>Region Type</th>
<th>Regions</th>
<th>CpG sites covered on 450K BeadChip array</th>
<th>Average # of CpG sites per region</th>
</tr>
</thead>
<tbody>
<tr>
<td>CpG Island</td>
<td>26,153</td>
<td>139,265</td>
<td>5.08</td>
</tr>
<tr>
<td>N Shore</td>
<td>25,770</td>
<td>73,508</td>
<td>2.74</td>
</tr>
<tr>
<td>S Shore</td>
<td>25,614</td>
<td>71,119</td>
<td>2.66</td>
</tr>
<tr>
<td>N Shelf</td>
<td>23,896</td>
<td>49,093</td>
<td>1.97</td>
</tr>
<tr>
<td>S Shelf</td>
<td>23,968</td>
<td>48,524</td>
<td>1.94</td>
</tr>
<tr>
<td>Remote/Unassigned</td>
<td>-</td>
<td>104,926</td>
<td>-</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>485,553</strong></td>
<td></td>
</tr>
</tbody>
</table>
Infinium: less expensive, most accurate (together with RRBS), easiest for bioinformatic analysis

Bock et al., Nat Genet 2010
MSP: Methylation >25% generated positive results

Histone Modification Analysis

• Step 1:
  – Histone Purification & Isolation

• Step 2:
  – Histone Analysis (Several Methods)
    • ELISA (Enzyme-Linked ImmunoSorbent Assay)
    • ChIP (Chromatin ImmunoPrecipitation)
Types of measure

• ELISA
  – Global genomic content of a certain modification:

• ChIP qPCR
  – Gene specific measure of a certain modification next to a specific gene

• ChIP-on-chip
  – Gene specific measure of a certain modification next to many specific gene at the same time
Examples

• ELISA
  • Histone: H3
  • Modification: Di-Methyl-Histone
  • Position: Lys4 (K4)

• ChIP qPCR
  • Histone: H3
  • Modification: Di-Methyl-Histone
  • Position: Lys4 (K4)
  • Gene: p15

• ChIP-on-chip
  • Histone: H3
  • Modification: Di-Methyl-Histone
  • Position: Lys4 (K4)
  • Genes: all the genes spotted on the chip
ChIP qPCR

1. Formaldehyde Cross Link
2. Cells isolated nuclei
3. Sonicate
4. Immunoprecipitate
5. Reverse crosslink
6. Purify DNA
7. qPCR
ChIP-on-chip

1. Formaldehyde treat cells
2. Isolate nuclei
3. Sonicate
4. Immunoprecipitate
5. Reverse crosslinks
   Purify DNA
6. Identify target genes
7. Grossly enriched for DNA that was bound by the purple transcription factor
8. Label this enriched DNA with red fluorescence
9. Mix with un-enriched genomic DNA labeled with green fluorescence

*Nature Reviews Immunology 4: 381-386 (2004);
Science, Vol 290, Issue 5500, 2306-2309, 22 December 2000*
MiRNA biogenesis and mechanisms of action

He and Hannon, 2004
miRNA analysis

• Sample collection and processing
  – Same methods as for mRNA
  – Isolation of total RNA

• Methods
  – Candidate miRNAs: Real Time PCR
  – Microarrays
  – nCounter Nanostring Analysis
  – Deep Sequencing
## Microarrays for miRNA analysis

<table>
<thead>
<tr>
<th></th>
<th># of measured human miRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABI TLDA</td>
<td>673</td>
</tr>
<tr>
<td>Affymetrix</td>
<td>847</td>
</tr>
<tr>
<td>Agilent</td>
<td>723</td>
</tr>
<tr>
<td>Exiqon</td>
<td>732</td>
</tr>
<tr>
<td>Illumina</td>
<td>1146</td>
</tr>
</tbody>
</table>

From Susan Hester, EPA

Nanostring nCounter

- Digital counting technology
- Allows for measuring more than 700 miRNAs
- Moderate operating costs
- No need for PCR amplification
Next Generation Sequencing

Analysis

doi: 10.1038/nbt1394

Discovering microRNAs from deep sequencing data using miRDeep

Marc R Friedländer¹, Wei Chen², Catherine Adamidi¹, Jonas Maaskola¹, Ralf Einspanier³, Signe Knespel¹ & Nikolaus Rajewsky¹

The capacity of highly parallel sequencing technologies to detect small RNAs at unprecedented depth suggests their value in systematically identifying microRNAs (miRNAs). However, the identification of miRNAs from the large pool of sequenced transcripts from a single deep sequencing run remains a major challenge. Here, we present an algorithm, miRDeep, which uses a probabilistic model of miRNA biogenesis to score compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor. We demonstrate its accuracy and robustness using published _Caenorhabditis elegans_ data and data we generated by deep sequencing human and dog RNAs. miRDeep reports altogether ~230 previously unannotated miRNAs, of which four novel _C. elegans_ miRNAs are validated by northern blot analysis.
Questions related to study design

• Which tissue can I use?
• How stable are the epigenetic marks within-subject over time?
• Data analysis:
  – Cell type effects
  – Confounders
Human Studies: Which Tissue?

• Target tissues
  – Cancer tissues are easy to obtain
    • They do not necessarily reflect pre-disease status
    • FFPE are not easy to work with
  – For some disease we can get at least close
    • Bladder cancer → Cells in urine sediments
    • Leukemias → White blood cells
      (benzene effects in Bollati et al Cancer Res 2007)
Non-Target Tissues

• How about psychiatric diseases?
  – Embryo layer approach
    (Neuroectoderma→buccal; good only for in-utero exposure?)
  – Uniform effects approach: Do exposures modify epigenetic marks in the same way across all tissues?
    • Problem 1: Epigenetic marks are tissue specific
    • Problem 2: Distribution of pollutants: different doses in different tissues (e.g., particulate matter)
  – Highest dose/first target approach
    • Inhalable pollutants → nasal mucosa
    • Pollutants in food and water/smoking→ buccal cells
A Practical Starting Point

• Existing cohort/studies have collected:
  – Blood/buffy coat
  – Buccal cells
  – Urine

• Storage
  – Many studies extracted DNA and have no more cells available

• Questions:
  – Can we use them?
  – Which information can we get?
How we have been using WBCs

- **Requirement 2:** The gene has a function in WBCs that is related with:
  - Mechanisms of action of the exposure
- E.g.: Particulate matter $\rightarrow$ iNOS $\rightarrow$ Cardiovascular disease
WBC Methylation in Epidemiology

• When analyzing data, which are the confounders we need to control for?
• What do we know about determinants of DNA methylation in human populations?
• To address this question:
  – Pooled analysis of 1,465 subjects from five investigations
  – In all the investigations: DNA methylation analysis of LINE-1 and Alu
# Pooled Analysis Data

<table>
<thead>
<tr>
<th></th>
<th>Study #1</th>
<th>Study #2</th>
<th>Study #3</th>
<th>Study #4</th>
<th>Study #5</th>
<th>All studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boston, USA</td>
<td>Warsaw, Poland</td>
<td>Milan, Italy</td>
<td>Brescia, Italy</td>
<td>Trissino, Italy</td>
<td>n=1,465</td>
</tr>
<tr>
<td>Alu, 5% mC (SD)</td>
<td>26.3 (1.1)</td>
<td>25.1 (1.4)</td>
<td>27.1 (3.4)</td>
<td>25.8 (0.7)</td>
<td>24.9 (1.0)</td>
<td>26.0 (1.8)</td>
</tr>
<tr>
<td>LINE-1, 5% mC (SD)</td>
<td>76.9 (2.0)</td>
<td>80.2 (3.2)</td>
<td>71.7 (5.9)</td>
<td>78.8 (1.0)</td>
<td>78.7 (3.2)</td>
<td>76.2 (6.2)</td>
</tr>
</tbody>
</table>

Zhu et al., Int J Epidemiology 2012
## Blood Count and Methylation

<table>
<thead>
<tr>
<th></th>
<th>Alu</th>
<th>LINE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beta *</td>
<td>P-value *</td>
</tr>
<tr>
<td>White blood cells, $10^3$cell/mm$^3$</td>
<td>0.002</td>
<td>0.938</td>
</tr>
<tr>
<td>Neutrophils, %</td>
<td>0.009</td>
<td>0.226</td>
</tr>
<tr>
<td>Lymphocytes, %</td>
<td>-0.009</td>
<td>0.246</td>
</tr>
<tr>
<td>Monocytes, %</td>
<td>-0.001</td>
<td>0.981</td>
</tr>
<tr>
<td>Eosinophils, %</td>
<td>-0.014</td>
<td>0.643</td>
</tr>
<tr>
<td>Basophils, %</td>
<td>0.005</td>
<td>0.968</td>
</tr>
</tbody>
</table>

* Adjusted for age, gender and study.

Zhu et al., Int J Epidemiology 2012
## Age, gender, alcohol

<table>
<thead>
<tr>
<th>Variable</th>
<th>Alu No.</th>
<th>%5mC</th>
<th>95% CI</th>
<th>Beta</th>
<th>P-value</th>
<th>No.</th>
<th>%5mC</th>
<th>95% CI</th>
<th>Beta</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;48</td>
<td>285</td>
<td>26.2</td>
<td>25.9, 26.5</td>
<td></td>
<td></td>
<td>264</td>
<td>77.4</td>
<td>76.8, 77.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48-63</td>
<td>279</td>
<td>26.0</td>
<td>25.8, 26.2</td>
<td></td>
<td></td>
<td>271</td>
<td>77.5</td>
<td>77.1, 77.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64-69</td>
<td>314</td>
<td>26.0</td>
<td>25.8, 26.2</td>
<td></td>
<td></td>
<td>307</td>
<td>77.2</td>
<td>76.8, 77.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70-74</td>
<td>267</td>
<td>25.9</td>
<td>25.7, 26.1</td>
<td>-0.012</td>
<td>0.019</td>
<td>263</td>
<td>77.1</td>
<td>76.7, 77.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥75</td>
<td>320</td>
<td>25.9</td>
<td>25.7, 26.1</td>
<td>-0.012</td>
<td>0.019</td>
<td>313</td>
<td>77.1</td>
<td>76.7, 77.5</td>
<td>-0.003</td>
<td>0.780</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>206</td>
<td>26.3</td>
<td>26.0, 26.6</td>
<td></td>
<td></td>
<td>202</td>
<td>76.5</td>
<td>76.0, 77.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1,238</td>
<td>25.9</td>
<td>25.8, 26.0</td>
<td>-0.361</td>
<td>0.012</td>
<td>1,216</td>
<td>77.4</td>
<td>77.2, 77.5</td>
<td>0.874</td>
<td>0.001</td>
</tr>
<tr>
<td>Alcohol drinking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nondrinker</td>
<td>609</td>
<td>26.1</td>
<td>26.0, 26.3</td>
<td></td>
<td></td>
<td>604</td>
<td>77.2</td>
<td>76.9, 77.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ever-drinker</td>
<td>717</td>
<td>25.9</td>
<td>25.8, 26.0</td>
<td>-0.208</td>
<td>0.043</td>
<td>700</td>
<td>77.1</td>
<td>76.9, 77.3</td>
<td>-0.049</td>
<td>0.798</td>
</tr>
</tbody>
</table>

^a Adjusted for age, gender and study.

^b Beta and P-value for age were obtained using continuous variables.

Zhu et al., Int J Epidemiology 2012
Summary of pooled analysis

• Alu and/or LINE-1 methylation associated with:
  – WBC differential
  – Age
  – Gender
  – Alcohol

• It may not apply to gene specific methylation

• Need to adjust for WBC differential
How stable is blood methylation?

Variability between Day 1 and 4

Byun HM et al, under review
Examples of variability

Byun HM et al, under review
Determinants of variability

Byun HM et al, under review
Mela Valley Asthma Study (Sicily, Italy)

- DNA from nasal swab in 38 third graders with asthma symptoms
- Lung function testing and exhaled nitric oxide measures
Airway inflammation and DNA methylation

• Interleukine-6 (IL-6) expression:
  – Associated with reduced DNA methylation of the gene promoter
  – Associated with reduced forced expiratory volume in 1 second (FEV1) in asthma patients

• Inducible Nitric Oxide Synthase (iNOS) expression:
  – Associated with reduced DNA methylation of the gene promoter
  – Increased NO production in the airway epithelium
  – NO production in asthma can be non-invasively measured as Fractional exhaled Nitric Oxide (FeNO).
Piko FEV-1 (obstruction)

FeNO (inflammation)
iNOS Methylation in Nasal Epithelial Cells & Exhaled Nitric Oxide

Exhaled Nitric Oxide (log scale) vs. iNOS promoter methylation (%5mC)

p=0.001

Baccarelli et al. under review
IL-6 Methylation in Nasal Epithelial Cells vs. Lung Function

\[ p=0.003 \]

Baccarelli et al. under review
Graphical Chain Modeling

1. iNOS methylation
   - LINE-1 methylation
   - Alu methylation

2. FeNO
   - IL-6 methylation
   - FEV1

3. wheezing

an IQR decrease in IL-6 methylation:
OR = 2.5 (90% CI 1.1-5.9, p=0.076) for wheezing
Conclusions

• Different methods for different questions
  – Gene-specific, global, genome-wide
  – Need to balance number of samples vs. coverage
    • Pyrosequencing and Illumina chips are a good fit for large human studies (Bisulfite treatment is easier and cheaper)
• Progress is fast
  – First methylation microarray chip released in 2006
  – Need for additional head-to-head comparisons between different methods
  – Bioinformatic/biostat methods to integrate data between different platforms
• Tissue specificity
  – Target tissues
  – Pure cell populations