Analysis of shotgun bisulfite sequencing of cancer samples

Kasper Daniel Hansen
<khansen@jhsph.edu>
Postdoc with Rafael Irizarry
Johns Hopkins Bloomberg School of Public Health
Brixen, July 1st, 2011
The basis of phenotypic variation: species
The basis of phenotypic variation: tissues
Heritable changes in phenotype that are not caused by changes in DNA.
DNA Methylation

In humans: methylation occurs at CpG dinucleotides (28.2M)

CpGs are depleted genomewide.

CpGs tend to cluster together (clusters are termed CpG Islands), these clusters are enriched in or near promoters.

Methylation is associated with “openness” of the DNA. Hypermethylation (high) is associated with gene silencing. Hypomethylation (low) is associated with active genes.

Methylation is inherited (at least in cell division).
Measuring DNA methylation

PCR does not preserve methylation information. Hybridization is not affected by methylation.

<table>
<thead>
<tr>
<th>Pretreatment</th>
<th>Analytical step</th>
<th>Locus-specific analysis</th>
<th>Gel-based analysis</th>
<th>Array-based analysis</th>
<th>NGS-based analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzyme digestion</td>
<td></td>
<td>• HpaII-PCR</td>
<td>• Southern blot</td>
<td>• DMH</td>
<td>• Methyl–seq</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• RLGSS</td>
<td>• MCAM</td>
<td>• MCA–seq</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• MS-AP-PCR</td>
<td>• HELP</td>
<td>• HELP–seq</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• AIMS</td>
<td>• MethylScope</td>
<td>• MIRA–seq</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• CHARM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• MMASS</td>
<td></td>
</tr>
<tr>
<td>Affinity enrichment</td>
<td>• MeDIP-PCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sodium bisulphite</td>
<td>• MethyLight</td>
<td>• MeDIP</td>
<td>• MeDIP</td>
<td>• RRBS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• EpiTYPER</td>
<td>• mDIP</td>
<td>• mDIP</td>
<td>• BC–seq</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Pyrosequencing</td>
<td>• mCIP</td>
<td>• mCIP</td>
<td>• BSPP</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• MIRA</td>
<td></td>
<td>• WGSBS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Illumina methylation arrays:
GoldenGate (early 2007, 1.5k CpGs), “27k” (late 2007), “450k” (2011)

Bisulfite treatment

The gold standard for measuring DNA methylation at single CpGs is bisulfite treatment followed by Sanger or Pyro sequencing.

Bisulfite treatment converts unmethylated Cs to Us (= T).

Can be used genome-wide, but requires the same sequencing effort as whole genome DNA sequencing (= expensive).
DNA methylation in cancer was the first epigenetic modification discovered in cancer (~25 years ago).

Focus (at least lately) in the literature have been on hyper methylation of CpG islands in promoters (tumor supprs) hypo methylation of select repeat elements although global hypomethylation hypo methylation of selected genes (typically oncogenes) have also been described.

Methylation terminology
  Hyper: goes up,   Hypo: goes down
  DMR: differentially methylated region
Many changes are not in CpG islands, but in regions bordering CpG islands; termed CpG Island shores.
Increased methylation variation in epigenetic domains across cancer types

Kasper Daniel Hansen1,2,10, Winston Timp2–4,10, Héctor Corrada Bravo2,5,10, Sarven Sabunciyani2,6,10, Benjamin Langmead1,2,10, Oliver G McDonald2,7, Bo Wen2,3, Hao Wu8, Yun Liu2,3, Dinh Diep9, Eirikur Briem2,3, Kun Zhang9, Rafael A Irizarry1,2 & Andrew P Feinberg2,3

Nature Genetics, Advance Online
Increased methylation variation across all cancers

Increased variation between normals and cancers, for the same regions across all 5 cancer types (lung, colon, breast, thyroid and Wilms).

151 regions in 290 samples.

The same regions that distinguish cancers from normals, distinguishes normal tissue types.
Design

3 colon cancers and their matched normal mucosa
2 adenomas
ABI SOLiD, 50bp reads
~5x coverage on CpGs

We traded coverage for biological replicates.
Mapping

Bisulfite conversion makes the genome into an (appr) 3 letter alphabet, making mapping hard.

We could not use existing tricks for unbiased alignment of bisulfite sequencing data: we wrote a custom aligner, *Merman*.

We can map ~20M CpGs uniquely.

```
 Genome
```

Coverage (for this CpG): 8
3 M’s and 5 U’s (Unmethylated)
Global levels of methylation

Bisulfite conversion rates estimated using λ phage spike-in to be 99.7-99.8%
One sample, small region ~ 14kb

Smoothing using a binomial model (local likelihood)
Adaptive bandwidth (← important)
Loss of methylation boundaries in cancer
Boundary Shifts (inwards and outwards)

Methylation

Islands
Genes

Methylation

Islands
Genes

0.2 0.5 0.8
1kb

no r mals
cancers

normals
cancers

1kb
Novel hypomethylation

Methylation

0.2 0.5 0.8

1kb

Islands

Genes
Capture bisulfite

~40,000 capture regions, ~400,000 CpGs
Red: Average of cancers
Blue: Average of normals
Green: Difference between cancers and normals
Capture bisulfite

~40,000 capture regions, ~400,000 CpGs
Red: Average of cancers
Blue: Average of normals
Green: Difference between cancers and normals
More capture

[Graph with multiple lines and markers indicating methylation and DMRs.]
Large blocks of hypomethylation

(Some) coincides with LADs, LOCKs and PMDs.

Related to structural conformation of the DNA in the nucleus.
What predicts hypomethylation in blocks?

- Methylation change
- Blocks: 
  - Repeats
  - Not repeats
- Not blocks: 
  - Repeats
  - Not repeats

- Density plots for normals and cancers
- Methylation density for:
  - Normals, in blocks
  - Cancers, in blocks
  - Normals, outside blocks
  - Cancers, outside blocks
Blocks are enriched for hyper-variables genes

Some of these genes are associated with tumor progression
Quality control: M-bias WGBS

Supplementary Figure 22: SOLiD 3+ Read position bias in evidence for methylation. The horizontal axis represents an offset into the nucleotide alignment from the 5' end. The vertical axis represents the fraction of filtered CpG methylation evidence from that offset that indicates that methylation is present. Only reads aligning uniquely to the GRCh37 human genome assembly are considered. In a perfect assay, the fraction should be independent of alignment offset and each line should be flat and horizontal. In practice, the lines are not flat due to sequencing error and other noise arising from sample preparation and alignment.
Based on this, we trim 15bp.
This improves the concordance
Conclusions

- Large blocks of hypomethylation in cancer
  Global hypomethylation, expression variability
  LOCKs/LADs

- Structural changes (boundaries) in small regions
  Unified framework for shore/islands hypo/hyper methylation

- With our smoothing technique, 4-5x is good enough
  Verified by high coverage padlock bs capture

- Biological replicates are very useful

- Quality assessment (M-bias plots)
Advantages of biological replicates

Met\text{ylation} 0.2 0.5 0.8
normal  cancer
1kb

p=0.00056

p=0.52
The effect of copy number variation (CNV)