Canadian Bioinformatics Workshops

www.bioinformatics.ca
Module 3
Introduction to WGBS and analysis

Guillaume Bourque
Epigenomic Data Analysis
June 20-21, 2016

Learning Objectives of Module

• Know the different technologies used to measure DNA methylation
• Understand their strengths and weaknesses
• Be familiar with the bisulfite-sequencing (BS-seq) data analysis workflow
• Understand the underlying principles and challenges
• Be able to extract methylation levels from BS-seq data
• Know how to visualize BS-seq data
• Be able to identify differentially methylated regions (DMRs)
What is DNA methylation?

• Most common form of DNA methylation is 5-methylcytosine (5mC)
• Affects 70 to 80% of CpGs in the human genome
• High levels of 5-methylcytosine (5mC) in CpG-rich promoters is strongly associated with repression
• In CpG poor regions, the relationship between DNA methylation and transcription is more complex

Bock, Nat Rev Genet, 2012

What is DNA methylation?

Day and Sweatt, Nat Neuro, 2010
Why study methylation?

- Methylation is the only epigenetic mark with demonstrated mitotic inheritance
- DNA methylation has been shown to be important for:
  - Genomic imprinting
  - Transposable element silencing
  - Stem cell differentiation
  - Embryonic development
  - Inflammation

Bock, Nat Rev Genet, 2012

Abnormal methylation in Cancer

Rodríguez-Paredes & Esteller
Nat Med, 2011
Assays for measuring DNA methylation

- **Bisulphite microarrays:**
  bisulphite mutations are mapped using genotyping microarrays that cover a selection of Cs

- **Enrichment-based methods:**
  methylated (alternatively unmethylated) DNA fragments are enriched and then measured using NGS

- **Whole-genome bisulphite sequencing (WGBS):**
  bisulphite mutations are mapped using NGS across the whole-genome
Bisulphite microarrays

- DNA preparation
- Bisulfite conversion
- Hybridization onto microarray (e.g. Illumina 450K)
- Data normalization and analysis

Methylation profile with microarray

Bock et al., Nat Biotech, 2010
Processing bisulphite microarray data

<table>
<thead>
<tr>
<th>Processing bisulphite microarray data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ComillaT</td>
<td>R script for correcting known or suspected batch effects using an empirical Bayes method <a href="http://www.bu.edu/lab/wp-assets/ComillaT">link</a></td>
</tr>
<tr>
<td>Illumina BeadScan</td>
<td>Machine control and image processing software for Illumina Infinium microarray scanners <a href="http://www.illumina.com/support/array/array_instruments/beadarray_readertest">link</a></td>
</tr>
<tr>
<td>Illumina GenomeStudio</td>
<td>Graphical tool for data normalization, analysis and visualization of Illumina Infinium microarrays (and other genomic data types) <a href="http://www.illumina.com/software/genomestudio">link</a></td>
</tr>
<tr>
<td>lsva</td>
<td>R package for batch effect correction using an algorithm that is based on singular value decomposition <a href="https://cran.r-project.org/web/packages/lsva">link</a></td>
</tr>
<tr>
<td>methylumi</td>
<td>R/Bioconductor package for Infinium data normalization and general data handling <a href="http://www.bioconductor.org/packages/release/bioc/html/methylumi.html">link</a></td>
</tr>
<tr>
<td>minfi</td>
<td>R/Bioconductor package for Infinium data normalization, analysis and visualization <a href="http://www.bioconductor.org/packages/release/bioc/html/minfi.html">link</a></td>
</tr>
<tr>
<td>RnBeads</td>
<td>R package providing a software pipeline for Infinium data normalization, quality control, exploratory visualization and differentially methylated region (DMR) identification <a href="http://rnbeads.computational-epigenetics.org">link</a></td>
</tr>
<tr>
<td>SVA</td>
<td>R/Bioconductor package for correcting batch effects that are directly inferred from the data using surrogate variable estimation <a href="http://www.bioconductor.org/packages/release/bioc/html/SVA.html">link</a></td>
</tr>
</tbody>
</table>

Bock, Nat Rev Genet, 2012

Module 3: Introduction to WGBS and analysis

Enrichment-based: MeDIP-seq

- Sonification of DNA
- Library preparation
- Denaturation
- Enrichment with antibody for 5-methylcytosine
- Library amplification
- High-throughput sequencing
**Enrichment-based: MethylCap**

- Sonification of DNA
- Enrichment with methyl-binding domain protein
- Washing and elution
- Library preparation and amplification
- High-throughput sequencing

**Enrichment-based: RRBS**

- Digestion with MspI
- Library preparation
- Gel-based size selection
- Bisulfite treatment
- Library amplification
- High-throughput sequencing
Enrichment-based methylation profiles

Bock et al., Nat Biotech, 2010

Processing enrichment-based data

Bock, Nat Rev Genet, 2012
Module 3: Introduction to WGBS and analysis

WGBS

- Genomic DNA Isolation
- Bisulfite treatment
- Library preparation
- High-throughput sequencing ( $$$)
Enrichment-based: MCC-Seq

- Genomic DNA Isolation
- Library preparation
- Bisulfite treatment
- Capture of targeted DNA-fragments
- High-throughput sequencing


Tools for processing WGBS (and other BS-seq)

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bismark</td>
<td>Probably the most widely used three-letter bisulphite aligner; supports both Bowtie (fast, gap-free alignment) and Bowtie 2.0 (sensitive, gapped alignment)</td>
<td><a href="http://www.bioinformatics.ualberta.ca/bismark">http://www.bioinformatics.ualberta.ca/bismark</a></td>
</tr>
<tr>
<td>Bis-SNP</td>
<td>Variant caller for inferring DNA methylation levels and genomic variants from bisulphite sequencing reads that have been aligned by other tools</td>
<td><a href="http://epigenome.wustl.edu/publicationdata/bisrank2011">http://epigenome.wustl.edu/publicationdata/ bisrank2011</a></td>
</tr>
<tr>
<td>BRAT</td>
<td>Highly configurable and well-documented three-letter bisulphite aligner</td>
<td><a href="http://compbio.cs.ucsd.edu/brat">http://compbio.cs.ucsd.edu/brat</a></td>
</tr>
<tr>
<td>BS-Seeker</td>
<td>Basic three-letter bisulphite aligner based on Bowtie</td>
<td>[<a href="http://galluppr.rockhuffuck.ee/BS">http://galluppr.rockhuffuck.ee/BS</a> Seeker/ BS Seeker.html](<a href="http://galluppr.rockhuffuck.ee/BS">http://galluppr.rockhuffuck.ee/BS</a> Seeker/BS Seeker.html)</td>
</tr>
<tr>
<td>BSMAF</td>
<td>Probably the most widely used wild-card bisulphite aligner</td>
<td><a href="http://code.google.com/p/bsmap">http://code.google.com/p/bsmap</a></td>
</tr>
<tr>
<td>CSMAP</td>
<td>Wild-card bisulphite aligner included in a widely used general-purpose alignment tool</td>
<td><a href="http://shese.genes.com/csnap">http://shese.genes.com/csnap</a></td>
</tr>
<tr>
<td>Last</td>
<td>Recent and well-validated wild-card bisulphite aligner included in a general-purpose alignment tool</td>
<td><a href="http://last.chrc">http://last.chrc</a></td>
</tr>
<tr>
<td>MethylCoder</td>
<td>Three-letter bisulphite aligner that can be used with either Bowtie (high-speed) or CSMAP (high sensitivity)</td>
<td><a href="https://github.com/brents/methylcode">https://github.com/brents/methylcode</a></td>
</tr>
<tr>
<td>Pash</td>
<td>Wild-card bisulphite aligner included in a general-purpose alignment tool</td>
<td><a href="http://bklouk.fmr.c.ee/pash">http://bklouk.fmr.c.ee/pash</a></td>
</tr>
<tr>
<td>RNAP</td>
<td>Wild-card bisulphite aligner included in a general-purpose alignment tool</td>
<td><a href="http://www.cshl.edu/people/andrewk/cmap">http://www.cshl.edu/people/ andrewk/cmap</a></td>
</tr>
<tr>
<td>RRBSMAF</td>
<td>Variant of BSMAF that is specialized on reduced representation bisulphite sequencing WGBS data</td>
<td><a href="http://rrbsnap.computational-epigenetics.org">http://rrbsnap.computational-epigenetics.org</a></td>
</tr>
<tr>
<td>segemehl</td>
<td>Wild-card bisulphite aligner included in a general-purpose alignment tool</td>
<td><a href="http://www.biostat.uni-leipzig.de/Software/segemehl">http://www.biostat.uni-leipzig.de/Software/ segemehl</a></td>
</tr>
</tbody>
</table>

Bock, Nat Rev Genet, 2012
Microarray/enrichment-based/WGBS

- All provide accurate DNA methylation measurements
- Microarrays typically have lower-costs and provide accurate measurements across a large number of CpGs targeting important regions (promoters, CpG islands)
- Enrichment-based methods (e.g. MeDIP-seq, MethylCap) have relatively low-resolution and can be a challenge to analyze/ normalize
- Bisulfite-based methods (e.g. RRBS, MCC-Seq and WGBS) provide absolute DNA measurements at a base-pair resolution
- WGBS is very expensive

Workflow for analyzing BS-data

- Processing of bisulfite-sequencing data:
  - Quality control and pre-processing
  - Bisulfite sequence alignment
  - Quantification of absolute DNA methylation
- Data visualization and statistical analysis
  - Visual inspection in a genome browser of selected regions
  - Visualization of global distribution of methylation values
  - Clustering of samples based on similarity
- Downstream analysis
  - Identification of Differentially Methylated Regions (DMRs)
  - Global analysis of DMRs
Workflow for analyzing BS-data

• Processing of bisulfite-sequencing data:
  – Quality control and pre-processing
  – Bisulfite sequence alignment
  – Quantification of absolute DNA methylation
• Data visualization and statistical analysis
  – Visual inspection in a genome browser of selected regions
  – Visualization of global distribution of methylation values
  – Clustering of samples based on similarity
• Downstream analysis
  – Identification of Differentially Methylated Regions (DMRs)
  – Global analysis of DMRs

Importance of quality control

• Before you start an analysis, it’s very important to look at your raw data!
• Are all of your samples sequenced using the same protocol and instruments?
• Are there any technical issues affecting some of the samples?
• This is especially important if you plan to compare different samples or different conditions
Module 3: Introduction to WGBS and analysis

Running FastQC on read 1

Very good!

QC: Base Composition

WGBS

RRBS

Felix Krueger, Babraham Bioinformatics
Read trimming

- Simple trimming: The method scans from the 5’ end of each read. As soon as it detects a position with quality scores below a preset threshold, it discards this position and the remaining positions at the 3’ end of the read.
- Dynamic trimming (window based): The method searches for the longest stretch of positions (window) in each read such that the quality scores of each position in the window exceed a preset threshold.
- Mott trimming (running sum): The method starts from the 3’ end of each read, subtracts a preset cutoff quality score from the quality score at each position and adds the remainder to a cumulative score at the position. The 3’ portion of the read starting from the position with the minimum cumulative score is trimmed.

Removing poor quality basecalls

[Graphs showing Phred score before and after trimming]

Felix Krueger, Babraham Bioinformatics
Sequence duplication

Complex/diverse library:

Duplicated library:

Felix Krueger, Babraham Bioinformatics

Quality metrics

- Read quality
- Presence of adapter sequencers
- Duplicate rates
- Conversion rate
Workflow for analyzing BS-data

- Processing of bisulfite-sequencing data:
  - Quality control and pre-processing
  - Bisulfite sequence alignment
  - Quantification of absolute DNA methylation
- Data visualization and statistical analysis
  - Visual inspection in a genome browser of selected regions
  - Visualization of global distribution of methylation values
  - Clustering of samples based on similarity
- Downstream analysis
  - Identification of Differentially Methylated Regions (DMRs)
  - Global analysis of DMRs

Bisulfite treatment (again)

1) Denaturation
Watson >> ACGGGCGTTGAG >> Crick << TGGCAGCGAATCG <<
C\(^\text{m}\) methylated
C Un-methylated

2) Bisulfite Treatment
BSW >> AC\(^\text{m}\)GGGCGTTGAG >> BSC << TGG\(^\text{m}\)CAGCGAATCG <<

3) PCR Amplification
BSW >> AC\(^\text{m}\)GGGCGTTGAG >> BSC << TGG\(^\text{m}\)CAGCGAATCG <<
BSWR << TG\(^\text{m}\)CAGCGAATCG >> BSCR >> ACG\(^\text{m}\)CGTTACTTAAA >

Xi and Li, *BMC Bioinformatics*, 2009
3 main strategies for processing WGBS reads

- Wild-card alignment
- Three-letter alignment
- Reference-free processing

Wild-card aligners

- Replace Cs in the genomic DNA sequence by the wild-card letter Y, which matches both Cs and Ts in the read sequence
- Or they modify the alignment scoring matrix in such a way that mismatches between Cs in the genomic DNA sequence and Ts in the read sequence are not penalized.
- Software: BSMAP, GSNAP, Last/bisulfighter, Pash, RMAP, RRBSMAP and segemehl

Three-base aligner

• Simplify bisulphite alignment by converting all Cs into Ts in the reads and for both strands of the genomic DNA sequence
• Software: Bismark, BRAT, BS-Seeker and MethylCoder

Bisulphite sequence alignment

Strengths and weaknesses

• Three-letter aligners have lower coverage in highly methylated regions because they purge the remaining Cs from the bisulphite-sequencing reads and thereby decrease their sequence complexity and they become ambiguous.

• Wild-card aligners typically have higher genomic coverage but at the cost of introducing some bias towards increased DNA methylation levels because the extra Cs in a methylated sequencing read can raise the sequence complexity.

• These problems are more prevalent in repetitive regions of the genome and are reduced with longer reads.
3 main strategies for processing WGBS reads

- Wild-card alignment
- Three-letter alignment
- Reference-free processing

Reference-based variant detection

Michael Stromberg
Reference-free variant detection

Reference-free methylation analysis

Moncunill et al. Nature Biotech 2014

Klughammer et al. Cell Reports 2015
Workflow for analyzing BS-data

- Processing of bisulfite-sequencing data:
  - Quality control and pre-processing
  - Bisulfite sequence alignment
  - Quantification of absolute DNA methylation

- Data visualization and statistical analysis
  - Visual inspection in a genome browser of selected regions
  - Visualization of global distribution of methylation values
  - Clustering of samples based on similarity

- Downstream analysis
  - Identification of Differentially Methylated Regions (DMRs)
  - Global analysis of DMRs

Goal: measure absolute methylation levels

Impact of SNPs

Figure 1: Detecting single nucleotide polymorphisms from Bisulfite-seq data. Hypothetical bisulfite-sequencing data is shown, with reference genome at top. Genome of the individual sequenced (unobserved) in the middle, and bisulfite sequencing reads bottom. [a] shows three reference cytosine positions, with the first being a match to the reference genome and the second two being homozygous single nucleotide polymorphisms. The first case shows a true C/G genotype, and all reads on the same strand as the C/G (C strand) are read as T, indicating an unmethylated state (shown as blue). Because the Illumina Bisulfite-seq protocol is directional, reads on the opposite strand (the G strand) are read as the true genotype. [b] (genotype reads on the G strand are boxed in this figure). The second case illustrates a true C>T SNP, which can be distinguished by the A reads present on the G strand. In this case, the reads on the C strand are inferred to be from a true T and should not be used for methylation calling (crossed out here). The third case shows a T>C SNP, which again can be identified based on G-strand reads. [c] A cytosine position with 50% unmethylated (S) and 50% methylated (C) reads can be associated with a heterozygous SNP on the same sequencing reads. In this case, the unmethylated reads are those on the "A" allele chromosome (here shown as maternal) and the methylated reads are on the "T" allele chromosome.

Liu et al. Genome Biol, 2012

BS-seq data SNP/methylation caller

- Bis-SNP
- MethylExtract
- BS-SNPer
- Etc.
Module 3: Introduction to WGBS and analysis

Bis-SNP

Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data

Liu et al. *Genome Biol.*, 2012

Bis-SNP accuracy

Liu et al. *Genome Biol.*, 2012
Workflow for analyzing BS-data

• Processing of bisulfite-sequencing data:
  – Quality control and pre-processing
  – Bisulfite sequence alignment
  – Quantification of absolute DNA methylation
• Data visualization and statistical analysis
  – Visual inspection in a genome browser of selected regions
  – Visualization of global distribution of methylation values
  – Clustering of samples based on similarity
• Downstream analysis
  – Identification of Differentially Methylated Regions (DMRs)
  – Global analysis of DMRs

Visualizing BS-seq data in IGV

• Reads are colored by DNA strand. For paired reads, this is the same as coloring by first-of-pair strand.
  – Gray for forward reads or forward read first pairs (F1 or F1R2)
  – Sage for reverse reads or reverse read first pairs (R1 or R1F2)
• The chosen mode is highlighted in reads with a red or blue nucleotide corresponding to the position of the cytosine in the reference genome.
  – For forward reads, a red C denotes a nonconverted cytosine, implying methyl or other protection, while a blue T denotes a bisulfite converted cytosine.
  – For reverse reads, a red G denotes a nonconverted cytosine, implying methyl or other protection, while a blue A denotes a bisulfite converted cytosine.
  – In zoomed-out views, colored nucleotides are represented by colored lines.
Visualizing BS-seq data in IGV (2)

https://www.broadinstitute.org/igv

Module 3: Introduction to WGBS and analysis  bioinformatics.ca

Visualizing BS-seq data in IGV (2)

https://www.broadinstitute.org/igv

Module 3: Introduction to WGBS and analysis  bioinformatics.ca
Workflow for analyzing BS-data

- Processing of bisulfite-sequencing data:
  - Quality control and pre-processing
  - Bisulfite sequence alignment
  - Quantification of absolute DNA methylation
- Data visualization and statistical analysis
  - Visual inspection in a genome browser of selected regions
  - Visualization of global distribution of methylation values
  - Clustering of samples based on similarity
- Downstream analysis
  - Identification of Differentially Methylated Regions (DMRs)
  - Global analysis of DMRs

Methylation values

![Histograms of %CpG methylation](a)
![Histograms of CpG coverage](b)

methylKit (Akalin et al. Genome Biol 2012)
Module 3: Introduction to WGBS and analysis

Pairwise correlations

Sample clustering

methylKit (Akalin et al. Genome Biol 2012)
**Workflow for analyzing BS-data**

- **Processing of bisulfite-sequencing data:**
  - Quality control and pre-processing
  - Bisulfite sequence alignment
  - Quantification of absolute DNA methylation

- **Data visualization and statistical analysis**
  - Visual inspection in a genome browser of selected regions
  - Visualization of global distribution of methylation values
  - Clustering of samples based on similarity

- **Downstream analysis**
  - Identification of Differentially Methylated Regions (DMRs)
  - Global analysis of DMRs

---

**Differentially methylated regions**
Need for biological replicates

Hansen et al. Genome Biol 2012

Advantages of smoothing

BSmooth (Hansen et al. Genome Biol 2012)
Module 3: Introduction to WGBS and analysis

DMRs example

Lee et al. Nat. Commun. 2014

Coverage requirements

Ziller et al. Nat Methods 2015
Conclusions

- BS-seq analysis is not easy
- Need to choose the appropriate DNA methylation technology
- Need to check quality and watch for biases
- Need to perform a multiple-step analysis workflow
- That’s what we’ll do in the lab!

Available WGBS datasets