Sequencing of DNA modifications
part of “High-Throughput Analyzes of Genome Sequenzes”

Jan Engelhardt

Bioinformatics
University of Leipzig

Leipzig, WS 2014/15
Chemical modifications
DNA modifications: 5-Methylcytosine (5mC)

- the most commonly known DNA modification
- DNA methyltransferases (DNMT1, DNMT3a/b (DNMT3L))
- predominantly occurs at 'CpG' motifs in (vertebrate) genomes
- genes can be silenced by 5mC in their promoter
- cell-type specific 5mC pattern
5mC - Easy maintenance over cell divisions

- DNMT1: Maintenance methyltransferase
- DNMT3a/b: De-novo methyltransferases

DNA modifications: 5-Hydroxymethylcytosine (5hmC)

- is processed from 5mC
- reported to be present in human and mouse in 2009
- Ten-eleven translocation methylcytosine dioxygenase (TET 1-3): catalyses oxidation of 5mC
- lower levels than 5mC but present in “most” cells
- controversial reports about the function
5-formylcytosine (5fC) and 5-carboxycytosine (5caC)
What is the problem?

What happens if we sequence DNA including 5mC?

- methylated cytosine (5mC) appear as normal cytosine (C) in standard sequencing
- we need to find a way to distinguish C and 5mC
Solution - Bisulphite treatment

What happens if we treat DNA with bisulphite?

- all unmethylated cytosines (C) are converted to uracil (U after PCR T)
- all methylated cytosines (5mC) stay the same

Conversions:

- C → T
- 5mC → C
Bisulphite treatment
Bisulphite sequencing

Cokus et al’s library protocol

- DNA fragments
- Ligated with adapters of DpnI restriction sites
- Bisulfite converted
  - Digested by DpnI restriction enzyme
  - 5-bp sequence tags formed
- Ligated with Solexa adapters
  - PCR I
  - PCR II
- BS reads

Lister et al’s library protocol

- DNA fragments
- Ligated with cytosine-methylated adapters
- Bisulfite converted
  - PCR
- BS reads
Problem?

Unfortunately we just messed up our whole genome...

Read: \textcolor{red}{GTTATTTCGATTTTGACGT}

Genome: \textcolor{blue}{GCTACTCGACCTGACGT}

- 4 introduced mismatches
- results in a pretty bad score despite of a “perfect” hit
- C/T mismatches are not only introduced by the conversion (sequencing errors, SNPs)
Standard segemehl

**Workflow:**
1) Use enhanced suffix array to find the position of a read in the genome
   *(Efficient data structure to handle big genomic data and find (almost) exact hits)*
2) Use the hits from step 1) as anchors and try to extend the regions according to given parameters (e.g. maximum number of mismatches)
   *(Efficient algorithm to build “good” semi-global alignments)*
Extended segemehl - Overview

- bisulfite reads
- genome

segemehl:
- seed search on collapsed alphabet
- E-value & maxocc filter
- bisulfite-sensitive semi-global alignment
- accuracy filter

 genome indices of collapsed alphabet

.sam file
Extended segemehl - Seed search on collapsed alphabet

Normal 4-letter nucleotide alphabet: $\sum_{DNA} = \{A, C, G, T\}$

We use two collapsed alphabets successively:

$$f_{C\rightarrow T}(x) = \begin{cases} T, & \text{falls } x = C \\ x, & \text{otherwise} \end{cases}$$

$$f_{G\rightarrow A}(x) = \begin{cases} A, & \text{falls } x = G \\ x, & \text{otherwise} \end{cases}$$

Workflow

1) Map $f_{C\rightarrow T}$ converted reads to $f_{C\rightarrow T}$ converted reference genome (+FW)
2) Map $f_{G\rightarrow A}$ converted reads to $f_{G\rightarrow A}$ converted reference genome (-FW)
Extended segemehl - Bisulphite-sensitive semi-global alignment

Normally, only equal character are regarded as matches (e.g. A= A, T=T).
Now possible bisulphite conversions are taken into account: IUPAC symbols were implemented Y = C or T, R = A or G; New comparison: C = Y (C or T), T = Y, A = R, G = R;
Methylation calling

For every cytosine in the genome the most frequent base (in the reads) is calculated. 
If the most frequent base is C the cytosine is called methylated. 
If the most frequent base is T the cytosine is called unmethylated. 
(Otherwise there is a problem with the reference)
Possible problems?
Possible problems?

- reduced complexity of the genome during mapping (longer reads are necessary)
- incomplete bisulphite conversion
- methylation is different in individual cells; normally a pool of cells is sequenced;
- degradation of DNA during bisulphite treatment (according to wikipedia)
- what happened to sequencing errors and SNPs?
- 5mC and 5hmC cannot be distinguished
Other modifications

Limitation of BS-seq
In 2009 it was reported that 5hmC can be found in human and mouse brains and in addition that BS-seq can not distinguish it from 5mC.
(Still makes people unhappy)

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>5mC</th>
<th>5hmC</th>
<th>5fC</th>
<th>5caC</th>
</tr>
</thead>
<tbody>
<tr>
<td>untreated</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>BS-seq</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>oxBs-seq</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>TAB-seq</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>redBS-seq</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>hydroxylamin</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>CAB-Seq</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
</tr>
</tbody>
</table>
oxBS-seq

Literature
