A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data

Xiwei Sun et al.

Supplementary Data
**Supplementary Data**

**Supplementary Tables**

**Table S1.** Sequencing quality of RRBS libraries of 18 lung tumor tissues and matched normal tissues.

**Supplementary Figure Legends**

**Figure S1.** Comparison of methylation results from different software tools in real RRBS data. (A) Distribution of concordant and discordant CpG sites under different sequencing depths. (B) Distribution of concordant CpG sites with different levels of DNA methylation. (C) Distribution of concordant and discordant CpG sites on different genomic regions (CGI, CGI shore, gene body, TSS, and promoter). In concordant CpG sites, the differences in the methylation levels of the same CpG sites estimated from any two of these aligners are less than 0.1, and in discordant CpG sites, the differences are more than 0.1.

**Figure S2.** Distribution of methylation levels among three types of sequencing depth (A) and distribution of sequencing depth among three types of methylation levels (B).

**Figure S3.** Distribution of read depth (A) and methylation level (B) on CpG sites in a real RRBS dataset.

**Figure S4.** FastQC metrics for reads derived from real and simulated RRBS data. (A) Distribution of phred quality scores across positions on sequencing reads. (B) Distribution of averaged phred quality scores of each read.
Figure S5. Recall and precision of different mapping algorithms for simulated RRBS datasets under different sequencing depth and genomic features (threshold of 0.1 is used to define concordant and discordant CpG sites).

Figure S6. Recall and precision of different mapping algorithms for simulated RRBS datasets under different levels of methylation and genomic features (threshold of 0.1 is used to define concordant and discordant CpG sites).

Figure S7. Comparison of alignment and methylation calling times (seconds).

Figure S8. Distribution of overestimated, unbiased, and underestimated CpG sites at different methylation levels for different mapping algorithms. Trend of distribution of unbiased CpG sites (green) was U-shaped.

Figure S9. Distribution of the recall of detecting CpG sites with different methylation levels for different mapping algorithms. Its curve was also U-shaped.

Figure S10. An illustrative example of distinct known methylation levels showing a diverse performance on their methylation estimation, if discarded reads occur. For easier view of performance differences, the sequencing reads are six bases in length and their coverage is four. Some of reads covering the same CpGs may be discarded (as error rate in the chart) by bisulfite aligners because of bisulfite-converted error, single nucleotide variation, insertion and deletion, sequencing error and multiple positions alignment after bisulfite-conversion, etc. The red color base represents CpGs and the blue color base is mutated base.

Figure S11. The probability of correctly estimating methylation levels is inversely
related to error rate on CpGs with intermediate methylation levels (as 0.5 here) in 30x sequencing depth. Error rate represents the fraction of discarded reads in all reads.

**Figure S12.** Methylation levels of five genomic regions (CGI shore, CGI, gene body, promoter, and TSS) in real RRBS data.
Supplementary Method

All Linux commands were used in this study, including sequence trimming, generating simulated RRBS data, and executing software tools for methylation data analysis. Users can repeat our simulations using these codes. The computing environment is high-performance computing cluster platform with Intel Xeon E5-2690v2 and Intel Xeon E7-8837 CPU.

Installation Prerequisite

1. Python (version=2.7)
   https://www.python.org/

2. Perl (version=5.10)
   https://www.perl.org/

3. Trim_galore (version=0.4.0)
   https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

4. Bismark (version=0.14.1)
   https://www.bioinformatics.babraham.ac.uk/projects/bismark/

5. BS-Seeker2 (version=2.0.9)
   http://pellegrini.mcdb.ucla.edu/BS_Seeker2/

6. BSMAP (version=2.89)
   https://github.com/genome-vendor/bsmap/

7. bwa-meth (version=0.10)
   https://github.com/brentp/bwa-meth/

8. GSNAP (version=2016-11-07)
http://research-pub.gene.com/gmap/

9. bowtie2 (version=2.2.4)

10. bowtie1 (version=1.1.1)

11. samtools (version=1.3.1)
  http://samtools.sourceforge.net/

12. RRBSsim (version=0.1.0)
  https://github.com/xwBio/RRBSsim
  https://github.com/xwBio/Docker-RRBSsim

**Commands for RRBSsim simulation**

```
python RRBSsim.py -f $region.fa -d $coverage --seed $i --min $min --max $max --CG_rate $rate --mCG_level $level
```

**Command for removing adapters and trimming sequences**

```
trim_galore -a GATCGGAAGAGCA -a2 GATCGGAAGAGCA --rrbs --paired end1.fq end2.fq
```

**Command for software tools for reads mapping**

**Bismark:**
```
bismark $path_to_reference_genome_folder -1 end.1_val_1.fq -2 end.2_val_2.fq --
bowtie2 -L 15 -N 1 -D 50 --score_min L,-0.6,-0.6 -X 600 -I 0 -p 2
```

**BS-Seeker2-bowtie:**
```
python bs_seeker2-align.py -1 end.1_val_1.fq -2 end.2_val_2.fq -t Y -r -m 1 --aligner=bowtie -f bam -g $path_to_reference_genome -d
```
Commands for software tools for calling CpG sites

**Bismark:**
```
bismark_methylation_extractor -- --genome_folder $path_to_genome_folder
```

**Output from bismark.bam --multicore 4**

**BS-Seeker2-bowtie:**
```
python bs_seeker2-call_methylation.py -i $output_from_align.bam --db $path_to_reference_genome_library
```

**BS-Seeker2-e2e:**
```
python bs_seeker2-call_methylation.py -i $output_from_align.bam --db $path_to_reference_genome_library
```
**BS-Seeker2-local:** python bs_seeker2-call_methylation.py -i$output_from_align.bam -db $path_to_reference_genome_library

**BSMAP:** python methratio.py -d $path_to_reference_genome_file $output_from_align.sam

**Bwa-meth:** PileOMeth extract $path_to_reference_genome_file $output_from_align.bam

**GSNAP:** sam_sort -D $path_to_genome_directory -d $genome_database

$output_from_align.sam > $output_from_align-sorted.sam

samtools view -bS $output_from_align-sorted.sam > $output_from_align-sorted.bam

PileOMeth extract $path_to_reference_genome_file $output_from_align-sorted.bam
Table S1  Sequencing quality of RRBS libraries of 18 lung tumor tissues and matched normal tissues.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Yield (Gb)</th>
<th># Reads</th>
<th>% of &gt;= Q30 Bases</th>
<th>Mean Quality Score</th>
<th>% of mapping rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRBS01</td>
<td>2.2</td>
<td>31372024</td>
<td>96.7</td>
<td>35.8</td>
<td>79.3</td>
</tr>
<tr>
<td>RRBS02</td>
<td>2.1</td>
<td>28805676</td>
<td>95.4</td>
<td>35.5</td>
<td>74.6</td>
</tr>
<tr>
<td>RRBS03</td>
<td>3.1</td>
<td>44274236</td>
<td>95.1</td>
<td>35.2</td>
<td>75.7</td>
</tr>
<tr>
<td>RRBS04</td>
<td>2.2</td>
<td>29283696</td>
<td>95.1</td>
<td>35.3</td>
<td>79.5</td>
</tr>
<tr>
<td>RRBS05</td>
<td>2.4</td>
<td>32224840</td>
<td>96.5</td>
<td>35.8</td>
<td>81.4</td>
</tr>
<tr>
<td>RRBS06</td>
<td>1.7</td>
<td>22676546</td>
<td>94.3</td>
<td>35.0</td>
<td>76.8</td>
</tr>
<tr>
<td>RRBS07</td>
<td>2.2</td>
<td>31255550</td>
<td>93.3</td>
<td>34.7</td>
<td>67.8</td>
</tr>
<tr>
<td>RRBS08</td>
<td>2.0</td>
<td>27124986</td>
<td>92.9</td>
<td>34.7</td>
<td>68.4</td>
</tr>
<tr>
<td>RRBS09</td>
<td>2.0</td>
<td>28654870</td>
<td>95.1</td>
<td>35.3</td>
<td>77.4</td>
</tr>
<tr>
<td>RRBS10</td>
<td>2.3</td>
<td>31739056</td>
<td>93.8</td>
<td>34.9</td>
<td>71.5</td>
</tr>
<tr>
<td>RRBS11</td>
<td>2.3</td>
<td>36771622</td>
<td>92.8</td>
<td>34.7</td>
<td>64.7</td>
</tr>
<tr>
<td>RRBS12</td>
<td>3.3</td>
<td>42067804</td>
<td>95.7</td>
<td>35.5</td>
<td>79.4</td>
</tr>
<tr>
<td>RRBS13</td>
<td>3.2</td>
<td>45171848</td>
<td>96.5</td>
<td>35.8</td>
<td>79.7</td>
</tr>
<tr>
<td>RRBS14</td>
<td>2.1</td>
<td>27890198</td>
<td>95.1</td>
<td>35.4</td>
<td>74.6</td>
</tr>
<tr>
<td>RRBS15</td>
<td>2.4</td>
<td>33806060</td>
<td>93.8</td>
<td>35.1</td>
<td>68.9</td>
</tr>
<tr>
<td>RRBS16</td>
<td>2.3</td>
<td>33747438</td>
<td>93.4</td>
<td>34.8</td>
<td>70.0</td>
</tr>
<tr>
<td>RRBS17</td>
<td>1.1</td>
<td>16661076</td>
<td>93.0</td>
<td>34.8</td>
<td>66.0</td>
</tr>
<tr>
<td>RRBS18</td>
<td>2.2</td>
<td>29504070</td>
<td>94.6</td>
<td>35.3</td>
<td>73.1</td>
</tr>
<tr>
<td>RRBS19</td>
<td>1.9</td>
<td>26269456</td>
<td>96.4</td>
<td>35.8</td>
<td>80.8</td>
</tr>
<tr>
<td>RRBS20</td>
<td>1.1</td>
<td>16951500</td>
<td>97.0</td>
<td>36.0</td>
<td>80.0</td>
</tr>
<tr>
<td>RRBS21</td>
<td>2.3</td>
<td>36715718</td>
<td>91.6</td>
<td>34.5</td>
<td>59.6</td>
</tr>
<tr>
<td>RRBS22</td>
<td>2.0</td>
<td>27893308</td>
<td>96.5</td>
<td>35.8</td>
<td>79.0</td>
</tr>
<tr>
<td>RRBS23</td>
<td>1.9</td>
<td>28090374</td>
<td>93.4</td>
<td>34.9</td>
<td>71.0</td>
</tr>
<tr>
<td>RRBS24</td>
<td>2.7</td>
<td>40024296</td>
<td>94.0</td>
<td>35.0</td>
<td>73.7</td>
</tr>
<tr>
<td>RRBS25</td>
<td>1.9</td>
<td>27633944</td>
<td>93.9</td>
<td>35.0</td>
<td>72.1</td>
</tr>
<tr>
<td>RRBS26</td>
<td>2.0</td>
<td>30276240</td>
<td>93.9</td>
<td>35.0</td>
<td>71.3</td>
</tr>
<tr>
<td>RRBS27</td>
<td>1.9</td>
<td>29337700</td>
<td>94.5</td>
<td>35.2</td>
<td>74.5</td>
</tr>
<tr>
<td>RRBS28</td>
<td>2.0</td>
<td>30374740</td>
<td>94.6</td>
<td>35.2</td>
<td>75.5</td>
</tr>
<tr>
<td>RRBS29</td>
<td>2.0</td>
<td>26364426</td>
<td>95.9</td>
<td>35.6</td>
<td>80.3</td>
</tr>
<tr>
<td>RRBS30</td>
<td>2.2</td>
<td>30224484</td>
<td>93.5</td>
<td>34.8</td>
<td>71.4</td>
</tr>
<tr>
<td>RRBS31</td>
<td>1.1</td>
<td>19534716</td>
<td>97.3</td>
<td>36.1</td>
<td>77.6</td>
</tr>
<tr>
<td>RRBS32</td>
<td>1.7</td>
<td>23164446</td>
<td>96.2</td>
<td>35.7</td>
<td>78.8</td>
</tr>
<tr>
<td>RRBS33</td>
<td>2.3</td>
<td>31883906</td>
<td>96.6</td>
<td>35.8</td>
<td>79.7</td>
</tr>
<tr>
<td>RRBS34</td>
<td>2.0</td>
<td>27384542</td>
<td>95.5</td>
<td>35.5</td>
<td>76.4</td>
</tr>
<tr>
<td>RRBS35</td>
<td>2.1</td>
<td>28542478</td>
<td>94.7</td>
<td>35.3</td>
<td>72.6</td>
</tr>
<tr>
<td>RRBS36</td>
<td>2.3</td>
<td>30019170</td>
<td>95.7</td>
<td>35.5</td>
<td>77.8</td>
</tr>
</tbody>
</table>
Figure S1

A

Proportion (%)  
75  50  25  0

<10x  10x-30x  >30x

Sequencing depth

B

Proportion (%)  
75  50  25  0

<1/3  1/3-2/3  >2/3

Methylation level

Discordant CpG sites
Concordant CpG sites

C

CGI Shore  CGI  Gene Body

28.8  71.2

16.1  83.9

31.7  68.3

Promoter  TSS

15.3  84.7

12.7  87.3
Figure S2

A

- Methylation level:
  - >2/3
  - 1/3~2/3
  - <1/3

- Proportion of concordant (con.) and discordant (dis.) CpGs across different methylation levels:
  - <10x
  - 10x~30x
  - >30x

B

- Sequencing depth:
  - >30x
  - 10x~30x
  - <10x

Concordant CpGs: conc.
Discordant CpGs: dis.
Figure S3
Figure S4

A

Real data

Simulated data

Quality scores

Position in read (bp)

B

Read count (x10^6)

Mean sequence quality (Phred Score)

Read count (x10^6)

Mean sequence quality (Phred Score)
Figure S5

- CGI Shore
- CGI
- Gene Body
- Promoter
- TSS

Sequencing depth:
- 10x
- 30x
- 50x

- Bismark
- BS-Seeker2-bowtie
- BS-Seeker2-ce
- BS-Seeker2-local
- BSMAP
- bow-meth
- GSNAP
Figure S6

The figure shows a comparison of precision and recall across different regions of the genome (CGI Shore, CGI, Gene Body, Promoter, TSS) for various sequencing tools and libraries. The tools and libraries are represented by different symbols and colors, and the methylation levels are indicated by markers of different shapes:

- Bismark
- BSMAP
- BS-Seeker2-bowtie
- BS-Seeker2
- BS-Seeker2-local
- bwa-meth
- GSNAP

Methylation levels are represented by:
- High
- Low
- Medium

The data points are plotted on a grid with precision (%) on the x-axis and recall (%) on the y-axis, allowing for a visual comparison of performance across different tools and conditions.
Figure S7

The figure shows a bar chart comparing the runtime (in seconds) for various tools and methods, specifically focusing on Methylation calling and Alignment. The tools include:

- BS-Seeker 2
- bwa-meth
- BS-Seeker 2-local
- Bismark
- GSNAP

Each bar represents the cumulative runtime, with the top portion indicating Methylation calling and the bottom representing Alignment.
Figure S8
Figure S9
Figure S10

<table>
<thead>
<tr>
<th>Genomic sequence</th>
<th>DNA Methylation Level</th>
<th>Bisulfite-seq reads</th>
<th>Probability</th>
<th>Alignment</th>
<th>DNA Methylation Level</th>
<th>Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGGT\text{CG}CTGA</td>
<td>0%</td>
<td>GT\text{TG}CT</td>
<td>1</td>
<td>1/6</td>
<td>0%</td>
<td>0</td>
</tr>
<tr>
<td>AGGT\text{CG}CTGA</td>
<td>50%</td>
<td>GT\text{TG}CT</td>
<td>1/6</td>
<td>4/6</td>
<td>50%</td>
<td>0.083</td>
</tr>
<tr>
<td>AGGT\text{CG}CTGA</td>
<td>100%</td>
<td>GT\text{CG}CT</td>
<td>1/6</td>
<td>1/6</td>
<td>100%</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure S11

Sequencing depth: 30X

Accuracy probability

Error rate (%)