Assignment 5 Solutions

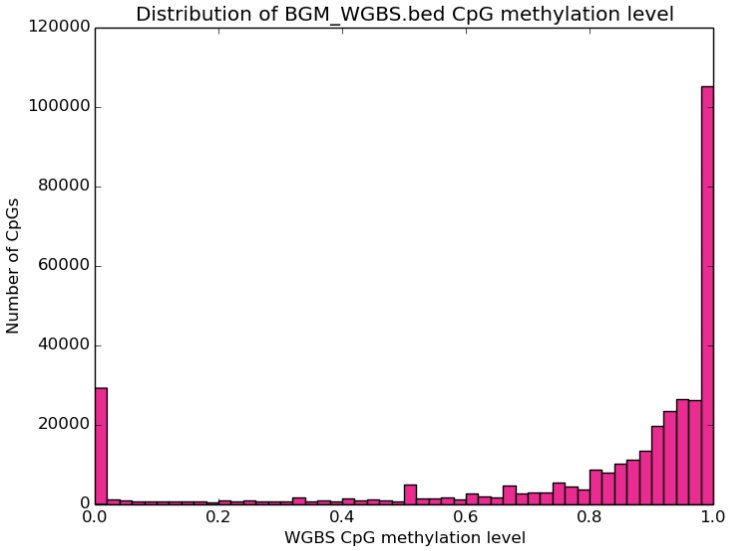
Part 1 — Analyzing DNA methylation data

Part 1.0

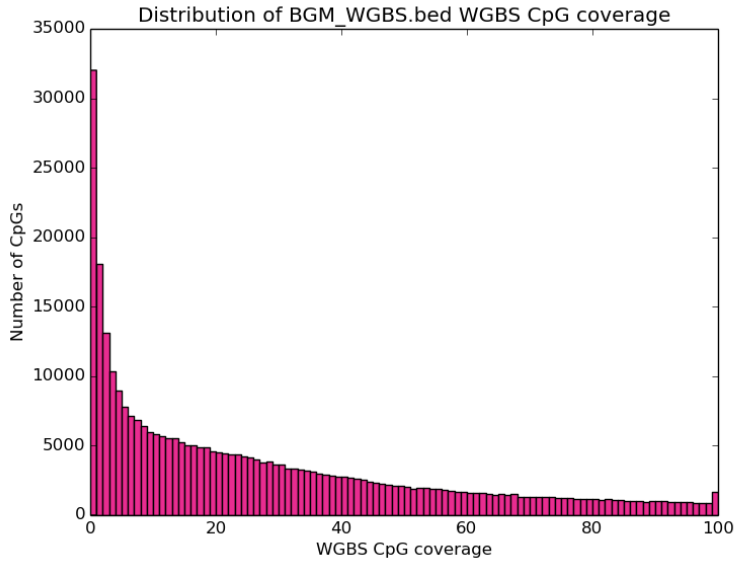
Command for running analyze_WGBS_methylation.py:

```
python3 analyze_WGBS_methylation.py BGM_WGBS.bed
```

BGM_WGBS_methylation_distribution.png looks like this:



BGM_WGBS_CpG_coverage_distribution.png looks like this:



Question 1
What does DNA methylation look like across chromosome 21?

Biomodal distribution with most CpGs hypermethylated.

Question 2
What does the CpG coverage look like across chromosome 21?

Even with such a deeply sequenced sample, there is still a large fraction of CpGs with little to no coverage.

Question 2.1
What fraction of the CpGs have 0X coverage?

Fraction 0X CpG Coverage: 0.08414904690309218 (32014 / 380444)

Part 1.1

Command for creating a bed file with the average CpG methylation level in each CGI.

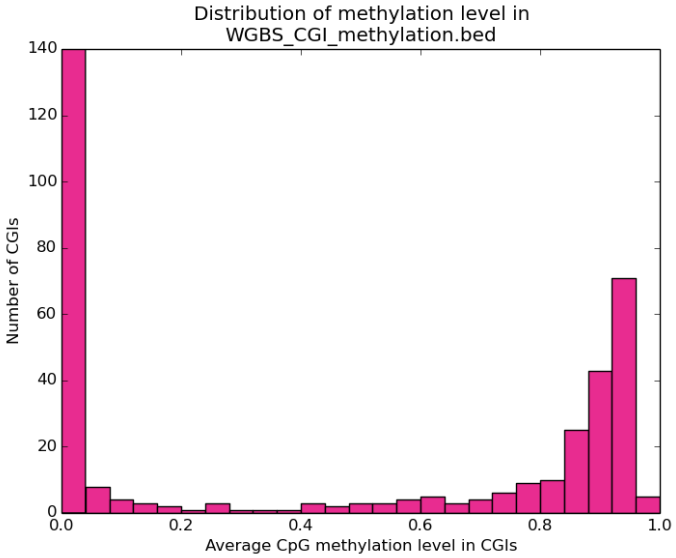
```
bedtools intersect -a CGI.bed -b BGM_WGBS_CpG_methylation.bed -wa -wb |
bedtools groupby -i - -g 1-4 -c 9 -o mean > WGBS_CGI_methylation.bed
```

Part 1.2

Command for plotting the distribution of average CGI methylation levels

```
python3 analyze_CGI_methylation.py average_CGI_methylation.bed
```

WGBS_CGI_methylation_distribution.png looks like this:



Question 3
What does DNA methylation look like for CpGs in CGIs? How does it compare to all the CpGs on chromosome 21?

Bimodal distribution. There is a much larger fraction of unmethylated CpG.

Part 1.3

Command for generating the promoter bed file
```
python3 generate_promoters.py refGene.bed
```

Justification for promoter definition
See relevant publications. Common promoter definitions are:

- -1kb to TSS
- -3kb to +1 kb of TSS

Commands for generating promoter-CGI and non-promoter-CGI bed files
```
bedtools intersect -a promoter_CGI.bed -b WGBS_CpG_methylation.bed -wa -wb >
promoter_CGI_and_WGBS_methylation.bed
bedtools intersect -a non_promoter_CGI.bed -b WGBS_CpG_methylation.bed -wa
-wb > non_promoter_CGI_and_WGBS_methylation.bed
```

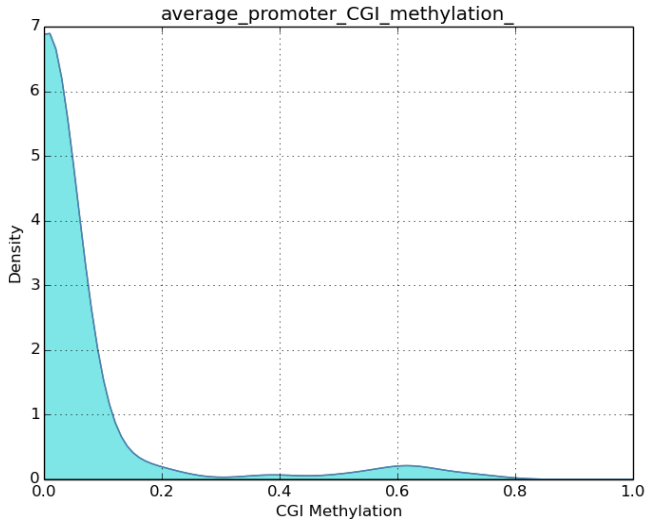
Justification for overlapping criteria
See relevant publications. Common overlap criteria are:

- 1 bp of overlap

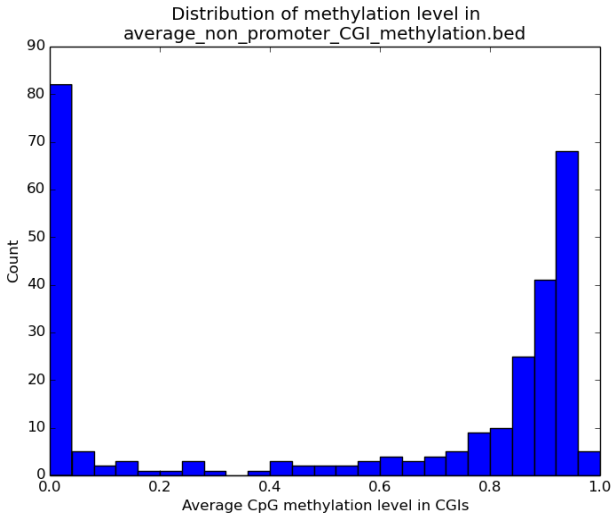
Commands for calculating the average CpG methylation for each promoter-CGI and non-promoter-CGI
```
bedtools groupby -g 1,2,3,4 -c 9 -o mean -i
promoter_CGI_and_WGBS_methylation.bed > average_promoter_CGI_methylation.bed
bedtools groupby -g 1,2,3,4 -c 9 -o mean -i
non_promoter_CGI_and_WGBS_methylation.bed >
average_non_promoter_CGI_methylation.bed
```

Commands for running analyze_CGI_methylation.py on average_promoter_CGI_methylation.bed and average_non_promoter_CGI_methylation.bed
```
python3 analyze_CGI_methylation.py average_promoter_CGI_methylation.bed
python3 analyze_CGI_methylation.py average_non_promoter_CGI_methylation.bed
```

average_promoter_CGI_methylation.png looks like this:



average_non_promoter_CGI_methylation.png looks like this:



Question 4

How do the DNA methylation profiles of promoter-CGIs and non-promoter-CGIs differ?

CpGs in promoters-CGIs are more unmethylated in comparison to the CpGs in non-promoter-CGIs. CpGs in promoter-CGIs are almost always unmethylated. In contrast, CpGs in non-promoter CGIs form a big peak at 0 methylation, and a second big peak at ~<75% methylated.

Part 1.3.1

Commands for calculating CpG frequency for each promoter type

```
bedtools getfasta -fi data/hg19_chr21.fa -bed promoter_CGI.bed -fo
promoter_CGI.fa
```

```
python3 data/nuc_count_multisequence_fasta.py promoter_CGI.fa

bedtools getfasta -fi data/hg19_chr21.fa -bed non_promoter_CGI.bed -fo
non_promoter_CGI.fa
python3 data/nuc_count_multisequence_fasta.py non_promoter_CGI.fa
```

CpG frequencies for each promoter type
Using [TSS - 1kb, TSS]:
promoter-CGI: 0.10628583639018484
non-promoter-CGI: CG:0.0909164155534428

Question 5
What is a possible biological explanation for the difference in CpG frequencies? Interpret your results from part 1.4 (parts 1.4.0 and 1.4.1): what are the "simple rules" for describing regulation by DNA methylation in promoters?

The answer is open ended. Promoter CGIs are almost always unmethylated (hence kept on and not regulated by DNA methylation). Non-promoter-CGIs are more likely to subject to regulation by DNA methylation. There may evolutionary pressure to decrease the number of CpGs in non-promoter-CGIs since they are vulnerable to deamination.

Part 2 — Comparing CGI MeDIP-seq, MRE-seq, and WGBS methylation level.

Commands to calculate CGI RPKM methylation scores

```
perl bed_reads_RPKM.pl CGI.bed BGM_MeDIP.bed > MeDIP_CGI_RPKM.bed
perl bed_reads_RPKM.pl CGI.bed BGM_MRE.bed > MRE_CGI_RPKM.bed
```

Command to generate the correlation plots

```
python3 compare_methylome_technologies.py MeDIP_CGI_RPKM.bed MRE_CGI_RPKM.bed
WGBS_CGI_methylation.bed
```

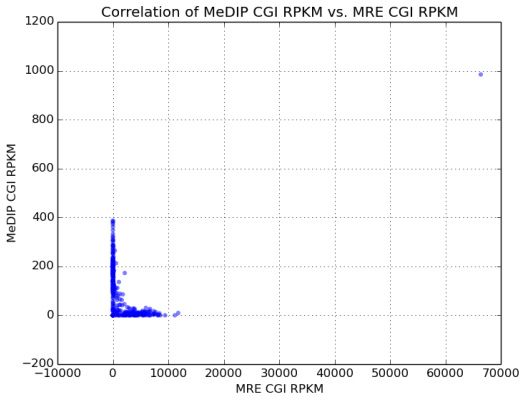
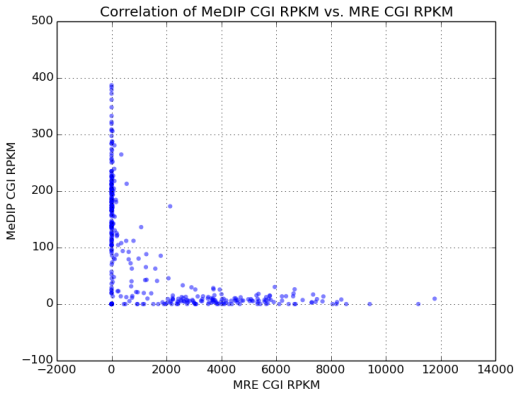
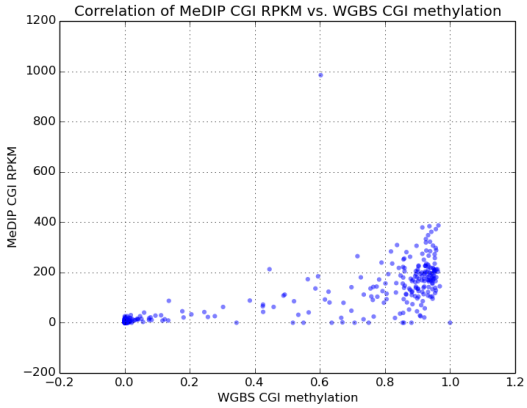
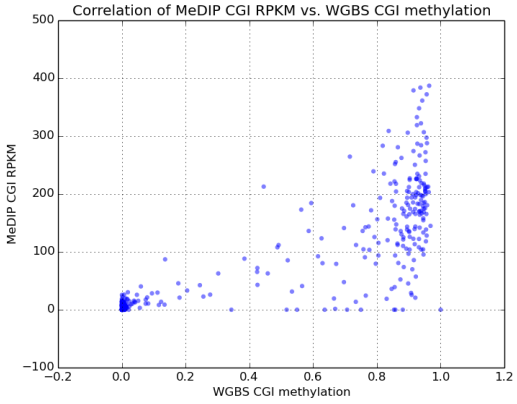
Correlations for each comparison:

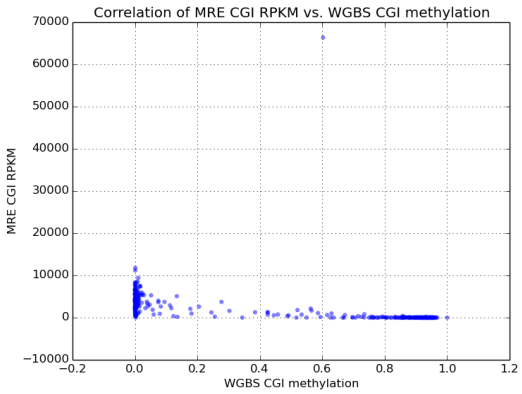
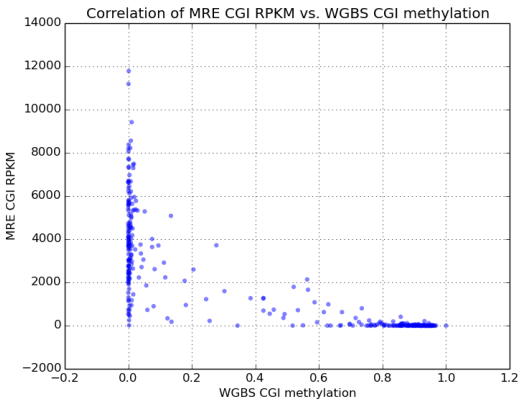
Comparison	Spearman's correlation (All data points included)	Spearman's correlation (With outlier removed)
MeDIP_CGI_RPKM vs. MRE_CGI_RPKM	-0.6428	-0.6567
MeDIP_CGI_RPKM vs. WGBS_CGI_methylation	0.8015	0.8045
MRE_CGI_RPKM vs. WGBS_CGI_methylation	-0.8405	-0.8436

Justification for chosen correlation metric

Spearman's ρ is a nonparametric measure of the dependence between two variables. Since it is a nonparametric measure, fewer assumptions about the input data have to be assumed. The method is also less sensitive to outliers than say for example Pearson's r.

The comparison plots look like this:

	All data points included	**With outlier removed**
MeDIP vs. MRE		
MeDIP vs. WGBS		

MRE vs. WGBS		

Question 6

How do MeDIP-seq and methylation correlate? How do MRE-seq and methylation correlate? How do MeDIP-seq and MRE-seq correlate?

There is a general correlation between MeDIP with methylation level, anti-correlation between MRE and methylation level, and anti-correlation between MeDIP and MRE.

There is at least one outlier. In your README, list their locations and explain the potential cause(s) for the outlier(s). (Hint: look at the CGIs in genome browser.) Explain why (or why not) the outlier(s) should be removed. If you removed them, recreate the scatter plots and recalculate correlations.

List of outlier locations

- chr21:9825442-9826296

Explain the potential cause(s) for the outlier(s).

This data point is most likely an artifact in mapping repeats.
The outlier is a CGI that spans two MIR elements and a simple repeat. This CGI has both very high MeDIP and very high MRE.

Explain why (or why not) the outlier(s) should be removed.

This outlier skews the graph/analysis. In practice we remove this outlier because it is the result of an artifact in mapping. Therefore we MUST remove it. Once the outlier is removed, the plots look much better.

Correlations for each comparison with outlier removed

See table above.

Comparison plots with outlier removed

See table above.