Methplotter: a python package for the visualization of DNA methylation data

Karisa Ke[1], Zhenhui Zhong[2], Steve Jacobsen[2]

[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Jacobsen Lab, Department of Molecular, Cell, and Developmental Biology, UCLA

Abstract

DNA methylation is an inheritable epigenetic mark that locks genes or transposable elements (TE) in the “off position” , which serves as an important component in various cellular processes such as genomic imprinting, embryonic development, differentiation, and maintenance of cellular identity through the epigenetic regulation of gene expression. High-throughput sequencing is widely used to profile genome-wide DNA methylation in a single-base resolution. Here, using the Python, we build a package, methplotter, for the visualization of high-throughput DNA methylation data. Functions of methplotter include converting methylation report from different methylation pipelines into a consistent DNA methylation format, generating wiggle report, graphing methylation data on a bar plot, line plot, and box plot over chromosomes, genes, TEs, or specific regions. In summary, our work provides a new tool for the downstream visualization of DNA methylation data.

Introduction

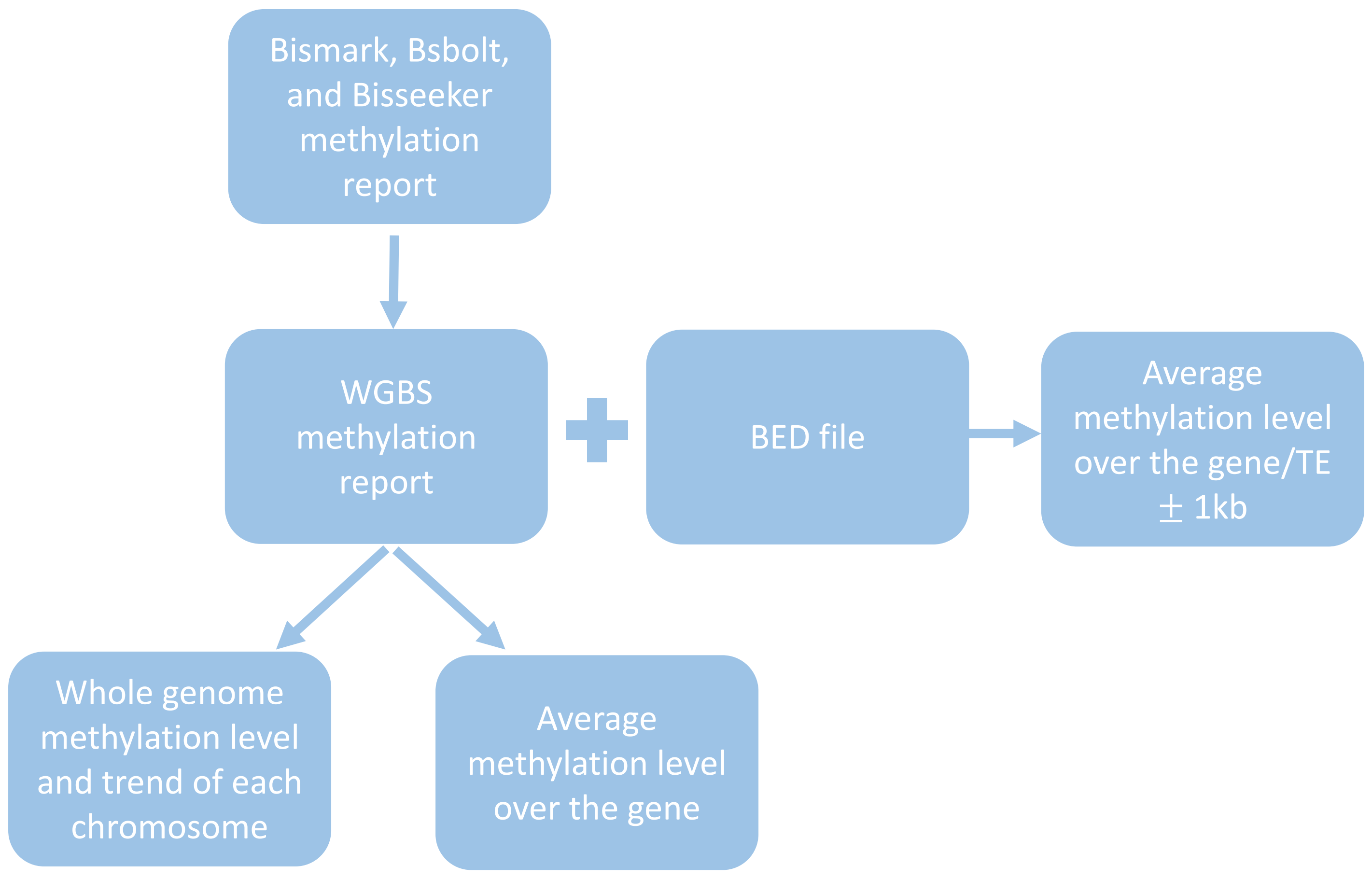
Methplotter python package contains functions that are commonly used in analyzing DNA methylation data. The basic steps to use this package is as follows:

1. Receive methylation report. If the report format is not in WGBS format, use the converter function to convert it into WGBS methylation report.

2. Receive BED file. Using WGBS methylation report and BED file as inputs to generate a plot showing the average methylation level over the gene/TE ± 1kb.

3. Using WGBS methylation report as the input to generate a line plot and a bar plot showing the whole genome methylation level and trend of each chromosome.

4. Using WGBS methylation report as the input to generate a box plot showing the average methylation level over the gene.



Bismark, Bsbolt, and Bisseeker methylation report

WGBS methylation report

BED file

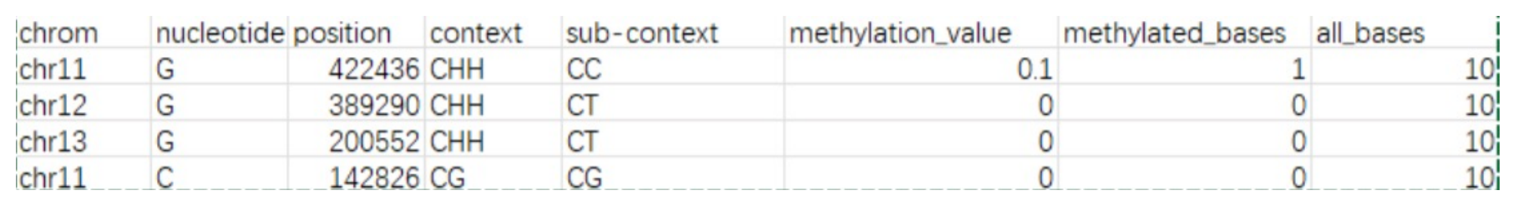
Average methylation level over the gene/TE ± 1kb

Whole genome methylation level and trend of each chromosome

Average methylation level over the gene

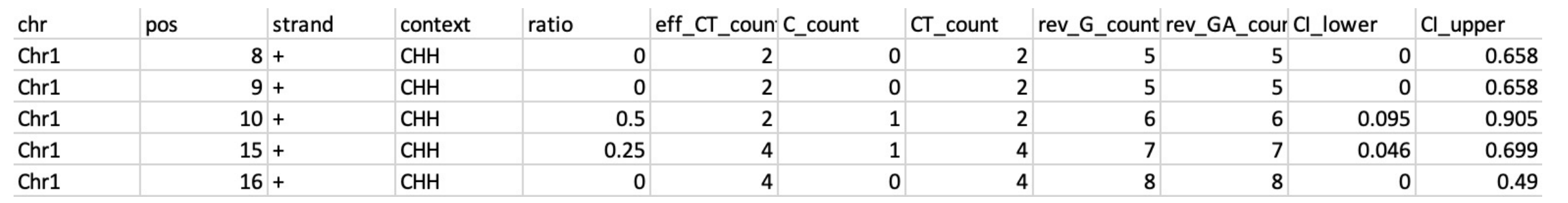
Results

1 Some common methylation data report formats include Bismark, Bsbolt, Bsseeker, and WGBS format. In this project, we used WGBS format consistently for the later steps. So, when given a methylation data report with a format other than WGBS, the first thing we need to do is to convert it into WGBS format. Here we wrote a script that is used to convert other methylation data report formats into WGBS format.



chrom	nucleotide	position	context	sub-context	methylation_value	methylated_bases	all_bases
chr11	G	422436	CHH	CC	0.1	1	10
chr12	G	389290	CHH	CT	0	0	10
chr13	G	200552	CHH	CT	0	0	10
chr11	C	142826	CG	CG	0	0	10

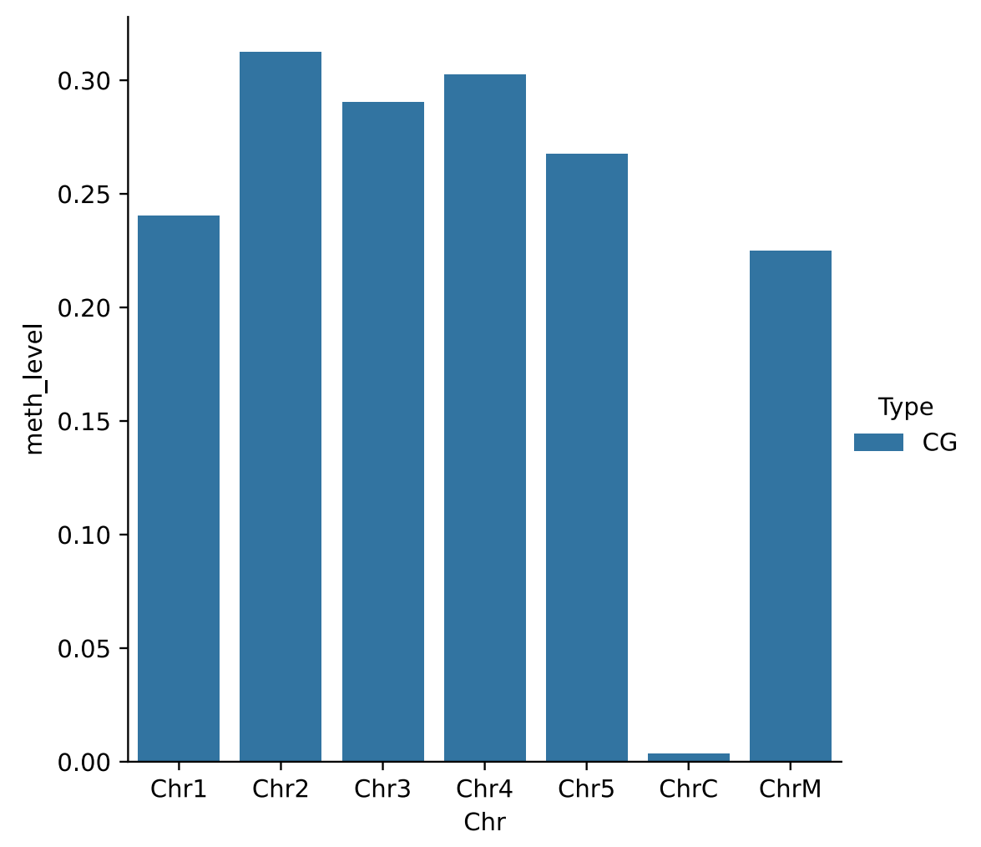
Bsbolt File Format



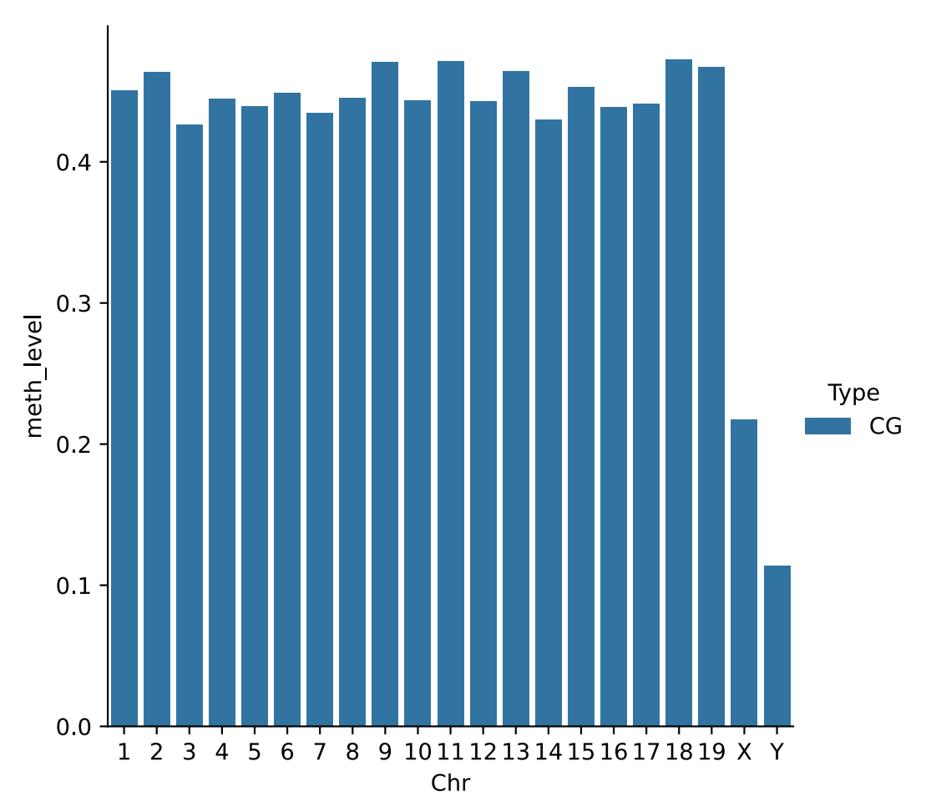
chr	pos	strand	context	ratio	eff_CT_coun	C_count	CT_count	rev_G_count	rev_GA_cour	CI_lower	CI_upper
Chr1	8	+	CHH	0	2	0	2	5	5	0	0.658
Chr1	9	+	CHH	0	2	0	2	5	5	0	0.658
Chr1	10	+	CHH	0.5	2	1	2	6	6	0.095	0.905
Chr1	15	+	CHH	0.25	4	1	4	7	7	0.046	0.699
Chr1	16	+	CHH	0	4	0	4	8	8	0	0.49

WGBS File Format

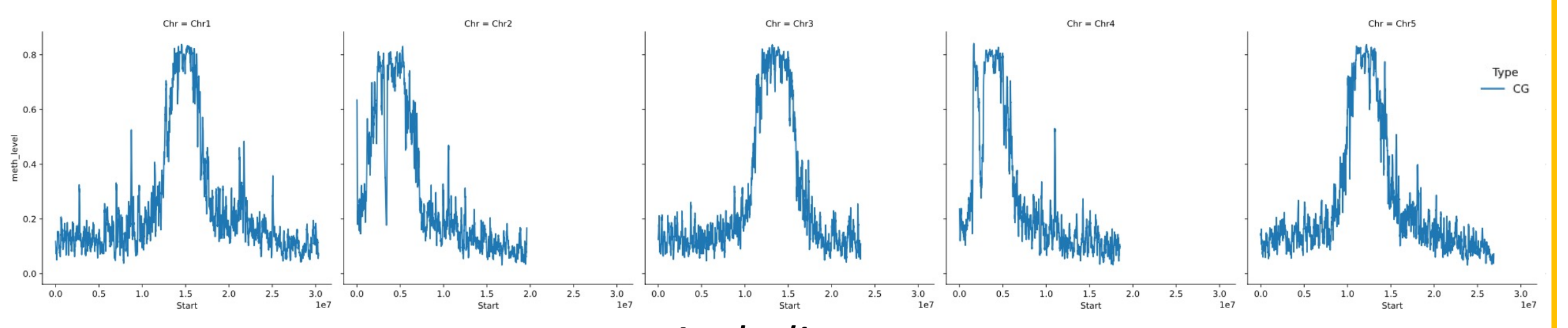
2 The visualization of the whole genome methylation data is important. In our second script, the final goal we want to achieve is to generate graphs that show the whole genome methylation. Here we have a bar plot and a line plot. The bar plot shows the average methylation level of each chromosome, and the line plot contains panels that shows the methylation level trend of each chromosome.



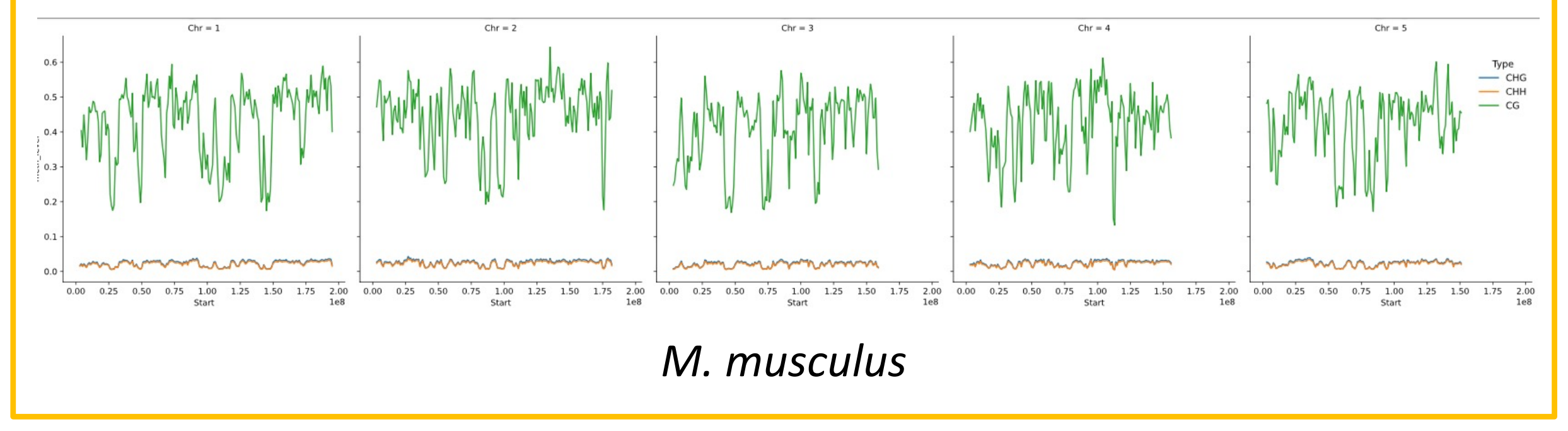
A. thaliana



M. musculus

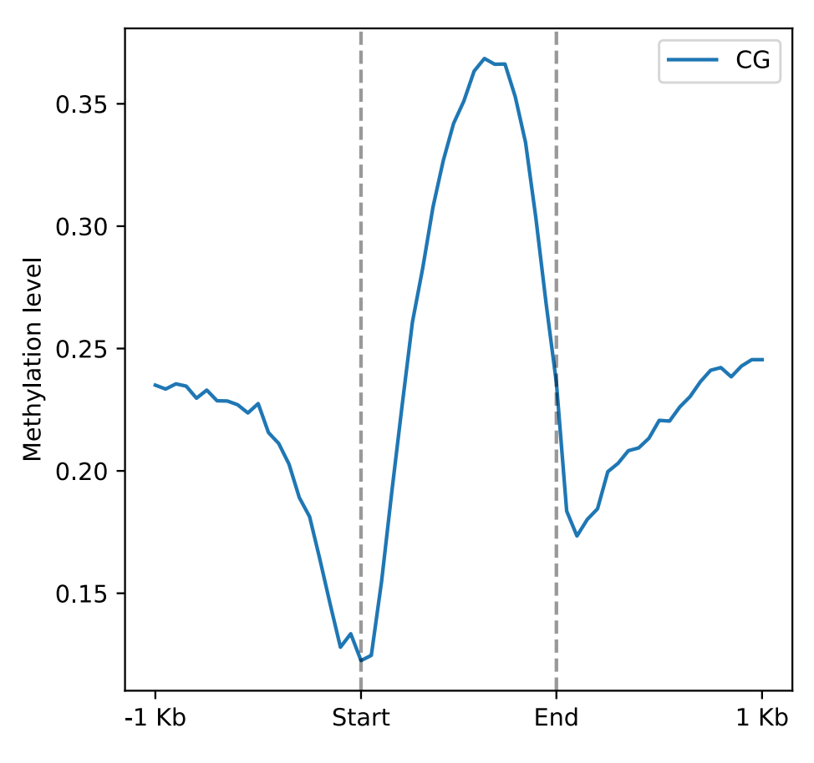


A. thaliana

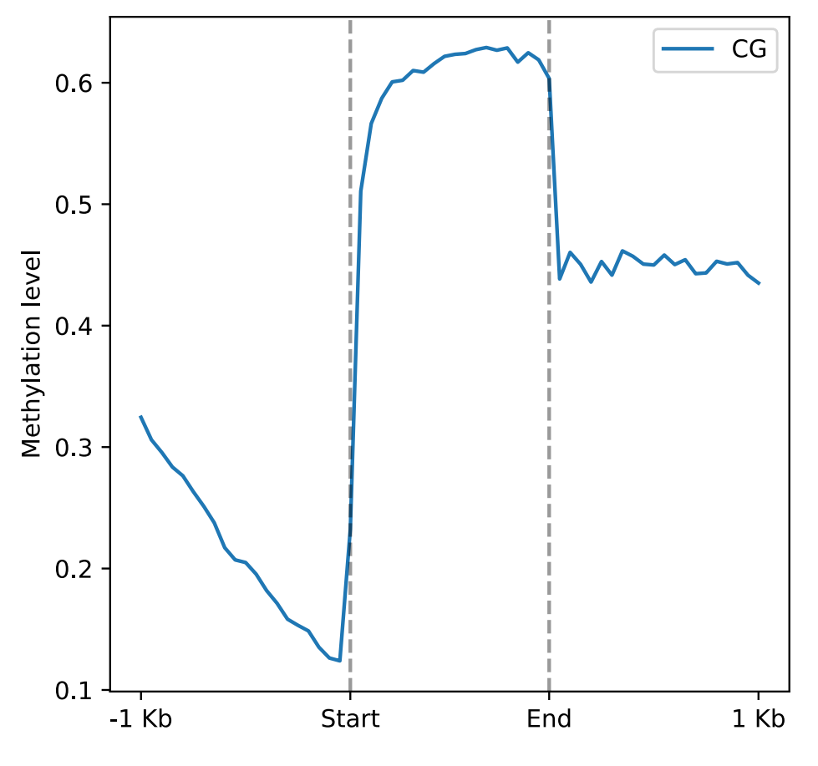


M. musculus

3 Going down to a smaller scale, we want to see the methylation level of each gene, including the 1kb region of its upstream or downstream. The final goal of this script is to generate a metaplot that shows the average methylation level from the 1kb point upstream of the gene, along the gene, and finally to the 1kb point downstream of the gene.

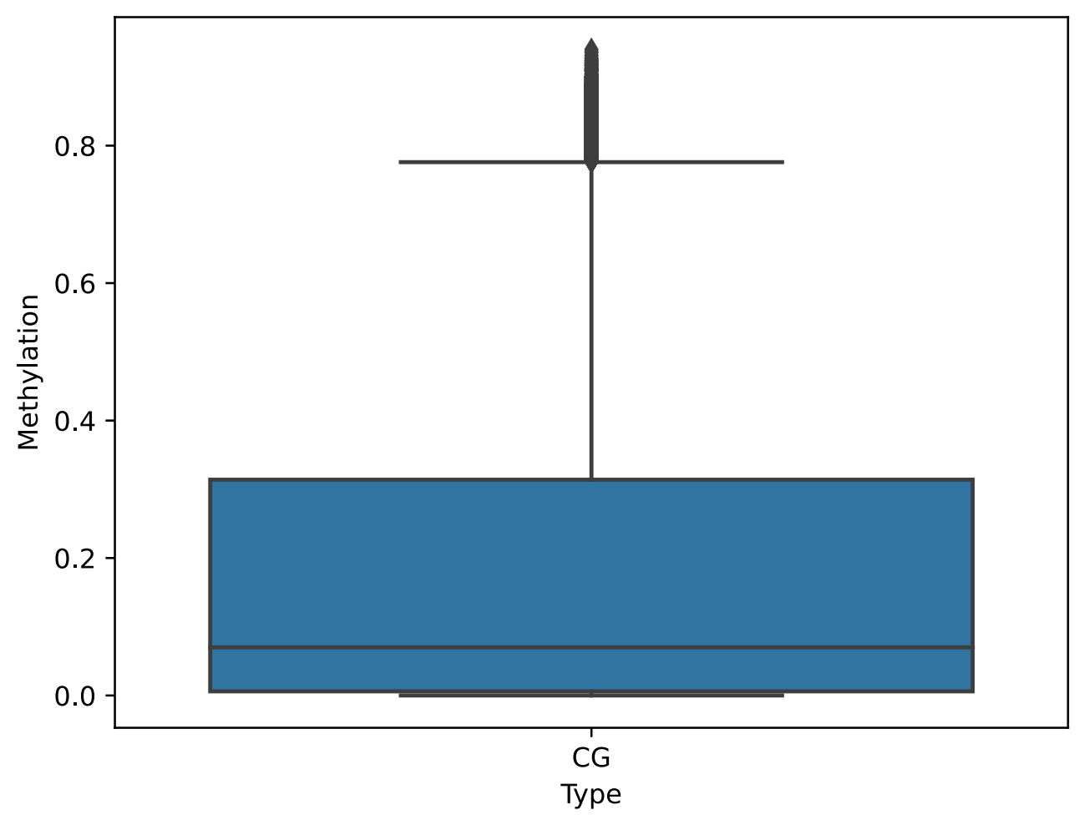


A. thaliana

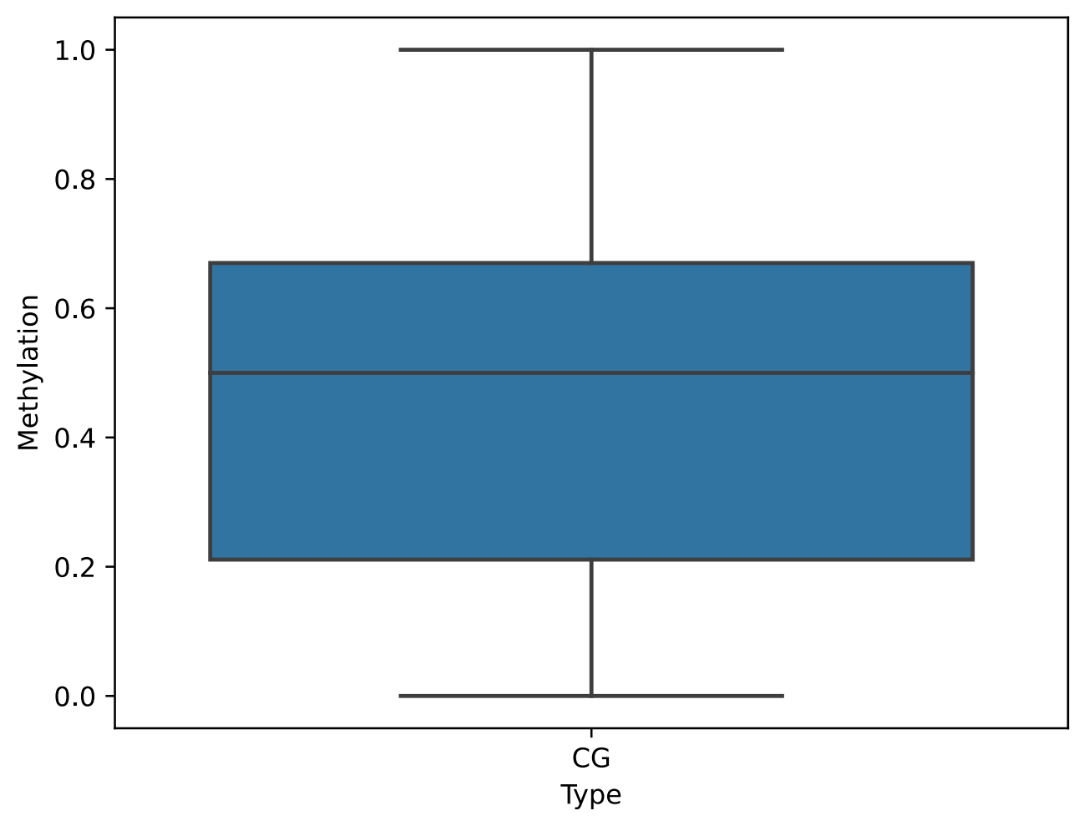


M. musculus

4 To have a better visualization of DNA methylation data, we also designed to generate a box plot for the average methylation level over the genes.



A. thaliana



M. musculus

Future Directions

1. As Methplotter was a newly developed package, we need to find more users to test this package. With more people using it, we could find potential bugs that we didn’t encounter during the development process.

2. Statistics is also important in the biology. Basic statistical concepts help biologists correctly prepare experiments, verify conclusions and properly interpret results. We could seek for people specialized in statistics and incorporate some functions involving the statistics into our package, such as calling DMR/DMC.

3. Our current version package is used to generate downstream plots from the methylation reports and BED files that are already prepared by others. In the future, we could develop scripts that can also achieve the goal of mapping and counting, such as mC total and CT total. By doing this, we’ll have a complete tool for analyzing DNA methylation data, from reading to plotting.

Acknowledgements

We would like to thank you the Institute for Quantitative and Computational Biosciences at UCLA for holding this program. I would like to thank you Zhenhui Zhong and Steve Jacobsen for supporting me and being my mentors to direct me on this research project.