# methylcircleplot: a tool for visualizing CpG and GpC methylation status across multiple samples and loci

Dai-Ying Wu [1,*]

[1] Department of Biochemistry and Molecular Biology, USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California 90089-9176

**ABSTRACT**

**Motivation:** Analyzing bisulfite converted sequence data is often laborious for researchers lacking a computational background. With the advent of Nucleosome Occupancy Methylome Sequencing (NOMe-Seq), bisulfite sequencing becomes even more informative by identifying both CpG methylation and nucleosome occupancy in a single molecule. We present a flexible software tool to simultaneously visualize DNA methylation and nucleosome occupancy using sequencing data from bisulfite treated samples. This tool is a vast improvement over the current process of using PowerPoint or Illustrator to manually create a circle for all CpG and GpC sites from every cloned amplicon of each region of interest. Additionally, this software tool is straightforward to use and requires little knowledge of R-programming to generate publication quality figures while including user-friendly options to customize the layout of the figure.

**Availability**: This source code is freely available under GPL license from Github https://github.com/ying-w/bioinformatics-figures/tree/master/methylcircleplot

**Contact**: daiyingw@usc.edu

## 1 INTRODUCTION

CpG methylation and nucleosome occupancy are two important epigenetic marks that are commonly studied for their roles in modulating gene expression. Aberrant DNA methylation has been found in many cancers and diseases (Taberlay and Jones, 2011). Nucleosome occupancy is a strong indicator of gene silencing and nucleosome depleted regions allow access for transcription factors and DNA binding proteins to interact and activate gene expression. A recent advancement using bacterial methyl-transferases to methylate nucleosome-depleted GpCs coupled with standard bisulfite treatment allows for the detection of both endogenous CpG methylation and nucleosome occupancy at different genomic regions. (Kelly et al., 2012).

While much progress has been made on the technology side of bisulfite sequencing, such as higher sequencing throughput and corresponding drop in price, there has been a delay in tools developed for data visualization. One of the challenges is creating a tool that is both simple to use but also flexible enough to suit the requirements of the community.

We introduce a tool that generates a graphical map that visualizes the methylation status of CpGs and optionally GpCs with proper scaling. One of the aims of the tool is to be accessible to users with no computational background. To achieve this, we incorporate flexible input formats, rigorous input validation, and include detailed examples. Our tool is written in R, a programming language that has attracted a large audience of biologists, to allow for easy installation and integration with existing R workflows.

## 2 METHOD OVERVIEW

Our software tool requires two sources of data input: a reference sequence and bisulfite converted sequencing data. Bisulfite converted samples are expected to contain multiple sequences (one for each sample sequenced) that overlap the reference sequence. An example experimental setup of this would be cell populations treated with a drug. After performing NOMe-seq on the cells before and after treatment, this tool could be used to generate two figures visualizing the methylation and nucleosome changes in the treated cells versus untreated cells.

The reference sequence is used to identify potential CpG and GpC sites while excluding ambiguous GCG sites. GpC site detection is disabled by default and should only be enabled for NOMe-seq experiments. After identifying sites of potential methylation in the reference sequence, the same sites in bisulfite converted samples will be checked for a cytosine base pair (methylated) or thymine base pair (unmethylated) and a percentage methylation will be calculated based on the number of sites methylated in all samples (see Figure 1). If the site contains a base that is not cytosine or thymine, that site is skipped and assumed to be a sequencing error (Fig 1c third amplicon).

In non-genome wide experiments, it is common to know the primers used to amplify the bisulfite treated sequence of interest. By specifying both forward and reverse primers in the 5' to 3' orientation, our tool can extract the bisulfite-converted sequence of interest from the sequencing input and compare only the sequence of interest to the reference sequence. In the case of an incorrectly specified primer or if the primer cannot be found, the tool will do a local alignment between reference and bisulfite sequence. Pairwise sequence alignments are computed using Biostrings library (Pages et al., 2009) in Bioconductor (Gentleman et al., 2004). For whole genome bisulfite sequencing, the sequences are usually aligned so it is possible to output only the reads that overlap the region of interest. Similarly, these outputted reads will be aligned and the same figure can be generated displaying only the sites with overlapping reads.

This tool requires R/bioconductor and Biostrings to be installed. These requirements are readily available for all major operating systems. The run time of this tool is usually instantaneous but can take up to several minutes on slower computers and with large regions of interest and numerous amplicons. This tool can be used on headless servers to generate figures since user interaction is not required

**a.**

Start R and load script

```
> source(”methylcircleplot.R”)
```

Specify the reference

```
> ref.seq = “atatctaggactctaggcgggtgggtaa
gcaagaactgaggagtggccccagaaataattggcacacga
acattcaatggatgttttaggctctccagaggatggctgag
tgggctgtaaggacaggccgaga”
```

Specify the file that holds clone sequence:

```
> bis.seq = “clone.fasta”
```

Specify primers (if any)

```
> fwd.primer = “tgggctgaaatactgggttcaccc”
> rev.primer = “atggctggacacctggcttcag”
```

Run script:

```
> methylcircleplot(ref.seq, bis.seq, fwd.primer,
  rev.primer, reference=TRUE, NOME=2)
```

Save figure:

```
> dev.copy(png, “clone-lolipop.png”); dev.off()
```

**b.**

```
>methyl
tgggctgaaatactgggttcacccatatttaggattttaggcgggtgggcaagtaagaattga
ggagtggtttagaaataattggtatacgaatatttaatggatgttttaggtttttagagga
tggttgagtgggttgtaaggataggtcgagaatggctggacacctggcttcag
>unmethyl
tgggctgaaatactgggttcacccatatttaggattttaggcgggtgggtaagcaagaattga
ggagtggctttagaaataattggcatatgaatatttaatggatgttttaggctttttagagga
tggctgagtgggctgtaaggataggctgagaatggctggacacctggcttcag
>mix-error
tgggctgaaatactgggttcacccatatttaggattttaggcgggtgggtaagtaagaattga
ggagtggtttagaaataattggcatatgaatatttaatggatgttttaggctttttagagga
tggctgagtggggtgtaaggataggtcgagaatggctggacacctggcttcag
```

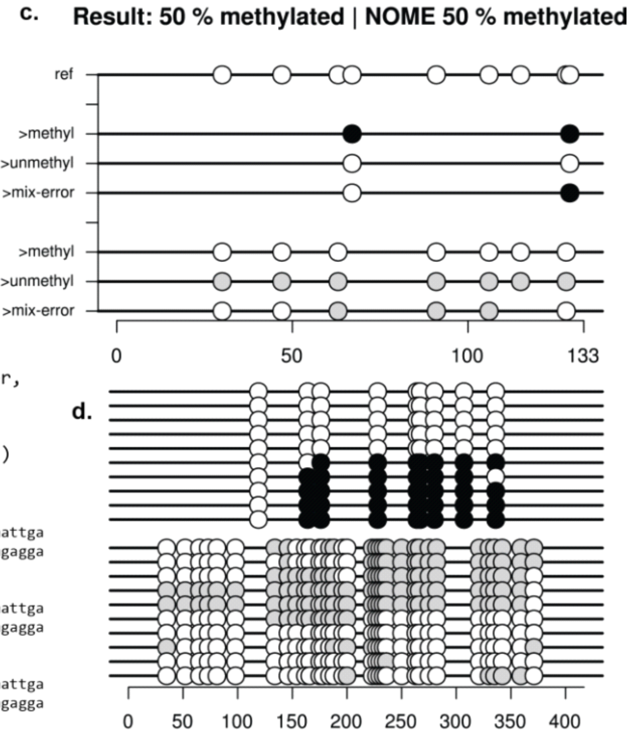

**Figure 1.** (A) Lists the R commands used to visualize methylation data. Black boxes around CG sites and grey boxes around GC sites have been added to help with interpretation. (B) Typical input file (in this case clone.fasta) that is required for this tool. In this example, the first clone amplicon is completely methylated and nucleosome occupied, the second clone is completely unmethylated and nucleosome depleted and the last clone is a mixture of both with a sequencing error in the second to last GC. The sequence in bold corresponds to the primers. Note that GCG is excluded. (C) The figure that is generated with unmethylated CpGs (white) methylated CpGs (black), nucleosome free regions (methylated GpCs, grey) and nucleosome occupied regions (unmethylated GpCs, white). (D) This is an example of a NOMe-seq experiment on the promoter of imprinted gene. The two alleles (first five and last five) are differentially methylated and this differential methylation is associated with differential nucleosome occupancy

## 3 CONCLUSION

The current process to visualize DNA methylation is time consuming and prone to errors due to the manual nature of creating a circle for every CpG and GpC methylation site. We present a novel software tool that can generate figures similar to the ones in existing publications and is accessible to researchers with no computation training. Our tool extends the functionality compared to previous tools (Block et al., 2005) by adding features such as GpC methylation detection while retaining quality control measures such as catching incomplete bisulfite conversion.

Our tool is also robust to common mistakes such as incorrectly specified primers or incorrect sequence orientation and will try to correct these errors while displaying helpful diagnostic messages to alert the user. This allows for less troubleshooting on the part of the end-user and thus greater ease of use. Finally, this tool is designed so it can be used with an automated R workflow since both input and output can be specified ahead of time so no user input is required.

## ACKNOWLEDGEMENTS

I would like to thank Danielle Bittencourt for the initial motivation for this project as well as Sheng-fang Su and Fides Lay for testing as well as Michael Stallcup and Kimberly Siegmund for proofreadng and publication advice.

*Funding*: This work was supported in part by a predoctoral traineeship from National Institute of Health grant T32 GM067587.

Conflict of interest: None declared

## REFERENCES

Block, C et al. (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics*., 21, 4067–4068.

Gentleman,R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., 5, R80.

Kelly TK et al. (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res*.,

Pages H, Aboyoun P, Gentleman R, DebRoy S. (2009). Biostrings: string objects representing biological sequences, and matching algorithms. R package version 2.24.1.

Taberlay PC, and Jones PA. (2011). DNA Methylation and Cancer. *Epigenetics and Disease*., 67, 1-23

## Notice

I learned later that this work mostly duplicates the functionality of MethylViewer software (http://pubmed.gov/20959287) which can be found here: http://dna.leeds.ac.uk/methylviewer/

I am going to leave this code up in case anyone wants to generate these plots on a Mac / linux machine or in a batch/automated way.