Simulating whole genome bisulfite sequencing data

Peter F Hickey*# and Terence P Speed*#
*Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia
#Department of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia

Introduction

It is very useful to be able to simulate realistic data when developing and assessing the performance of statistical methods. By using simulated data the truth is both known and controllable, whereas in real data it may be neither. It is of course vital that the simulated data is similar enough to the real data, where "similar enough" is defined with respect to the statistical questions under consideration.

Current software for simulating whole genome bisulfite sequencing data, such as Sherman, BSSim and DNemulator, focus on simulating data that are useful when comparing the performance of read mapping strategies. The stochastic models used by these programs, which generally make simplifying assumptions such as spatial-independence of methylation events at nearby CpG sites and independence of reads, do not generate appropriately realistic data for tasks such as comparing methods for identifying differential methylation.

We are developing software to simulate whole genome bisulfite sequencing data that captures the complex intra- and inter-sample heterogeneity found in real studies of DNA methylation, with a focus on CpG methylation. Some features of real data that we simulate include the strong spatial dependence of methylation at nearby CpGs, regions of hypo-, intermediate and hyper-methylation, and epipolymorphism.

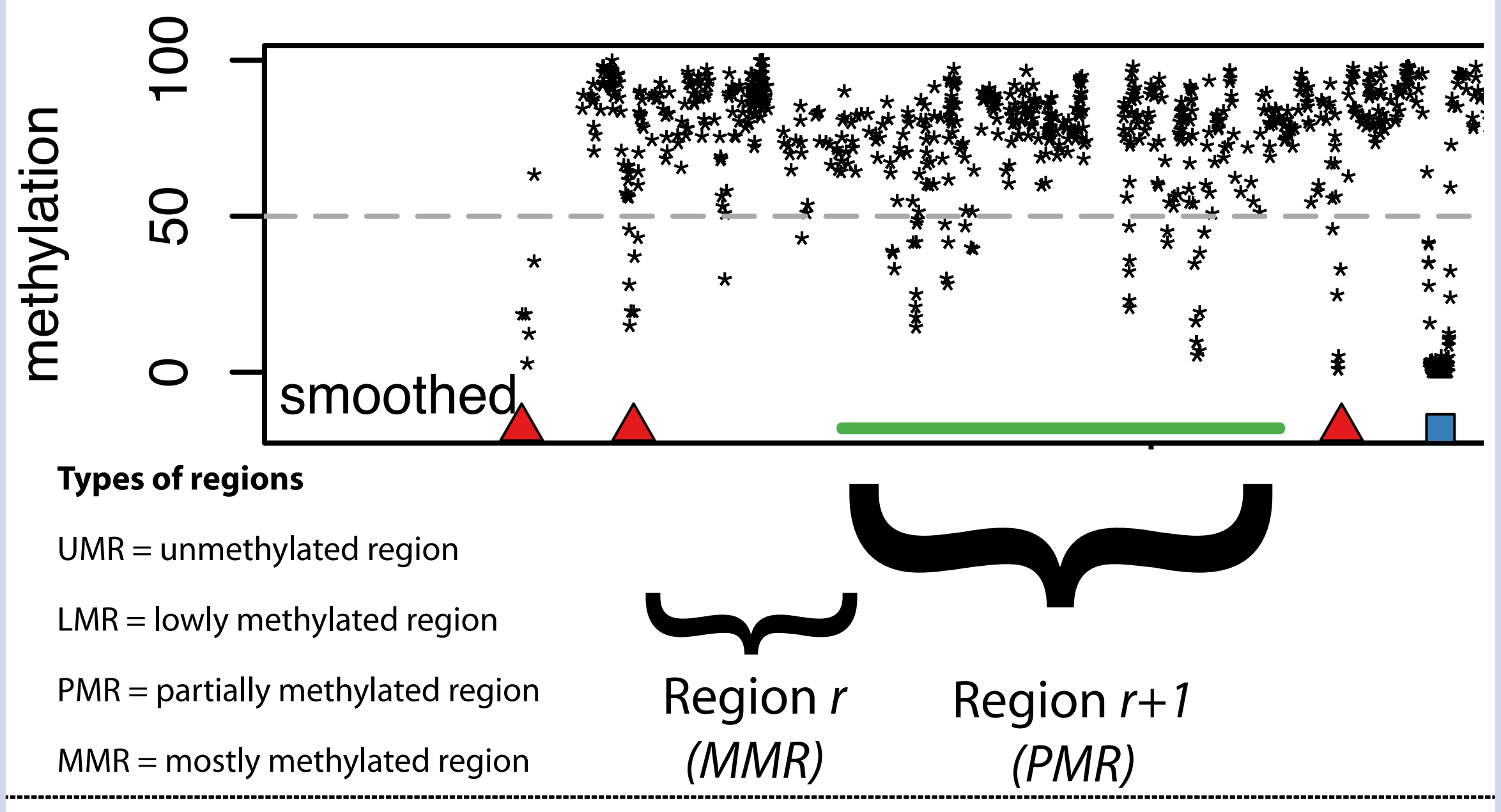
This work is motivated by our desire to resolve some questions that have arisen from our exploratory data analyses of tens of whole genome bisulfite sequencing experiments. This software will also be useful in developing and comparing statistical methods for analysing DNA methylation data.

Methods

DNA methylation is a non-stationary process; both the average methylation level and the variation in the methylation level vary as a function of position in the genome. However, within smaller regions, say within a CpG island or within a partially methylated domain, DNA methylation is more "stable" or "similar" and may be approximated by a locally stationary stochastic process.

Our basic idea is to segment the genome into "regions of similarity" and then simulate data using region-specific parameters. The algorithm is sketched in Figure 1.

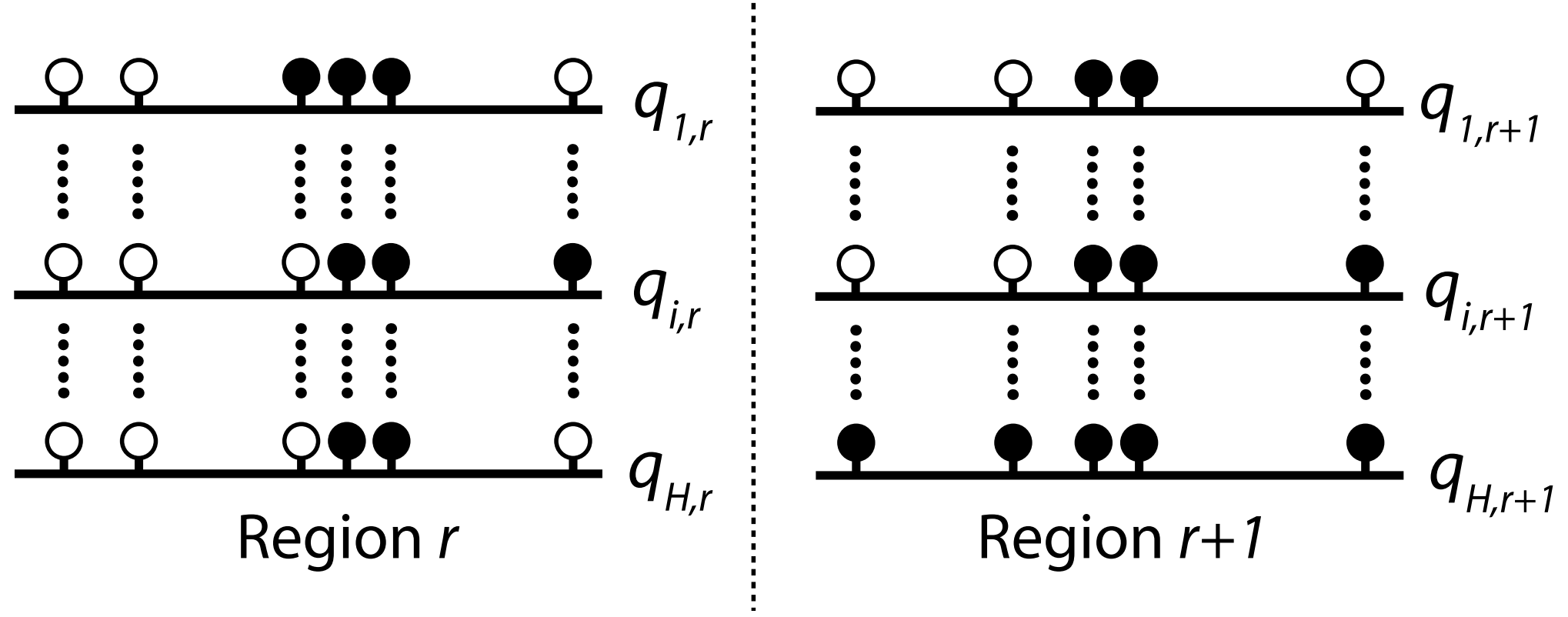
(1) Segment genome into "regions of similarity" (MethylSeekR)



Types of regions

UMR = unmethylated region
LMR = lowly methylated region
PMR = partially methylated region
MMR = mostly methylated region

(2) For each region: simulate each haplotype from a Markov model
Transition matrices depend on distance between CpGs and the type of region
Assign haplotype *i* in region *r* frequency $q_{i,r}$
Keep track of how haplotypes link across regions



(3) Simulate read positions
Simulate reads for region *r* by sampling from i^{th} haplotype with probability $q_{i,r}$
Simulate sequencing error

***Figure 1**: Schematic of the simulation procedure (MySim). Simulatation parameters were estimated from the ADS methylC-seq data from Lister et al. Nature, 2011. Paired-end reads were simulated to an average of 20x coverage.*

Results

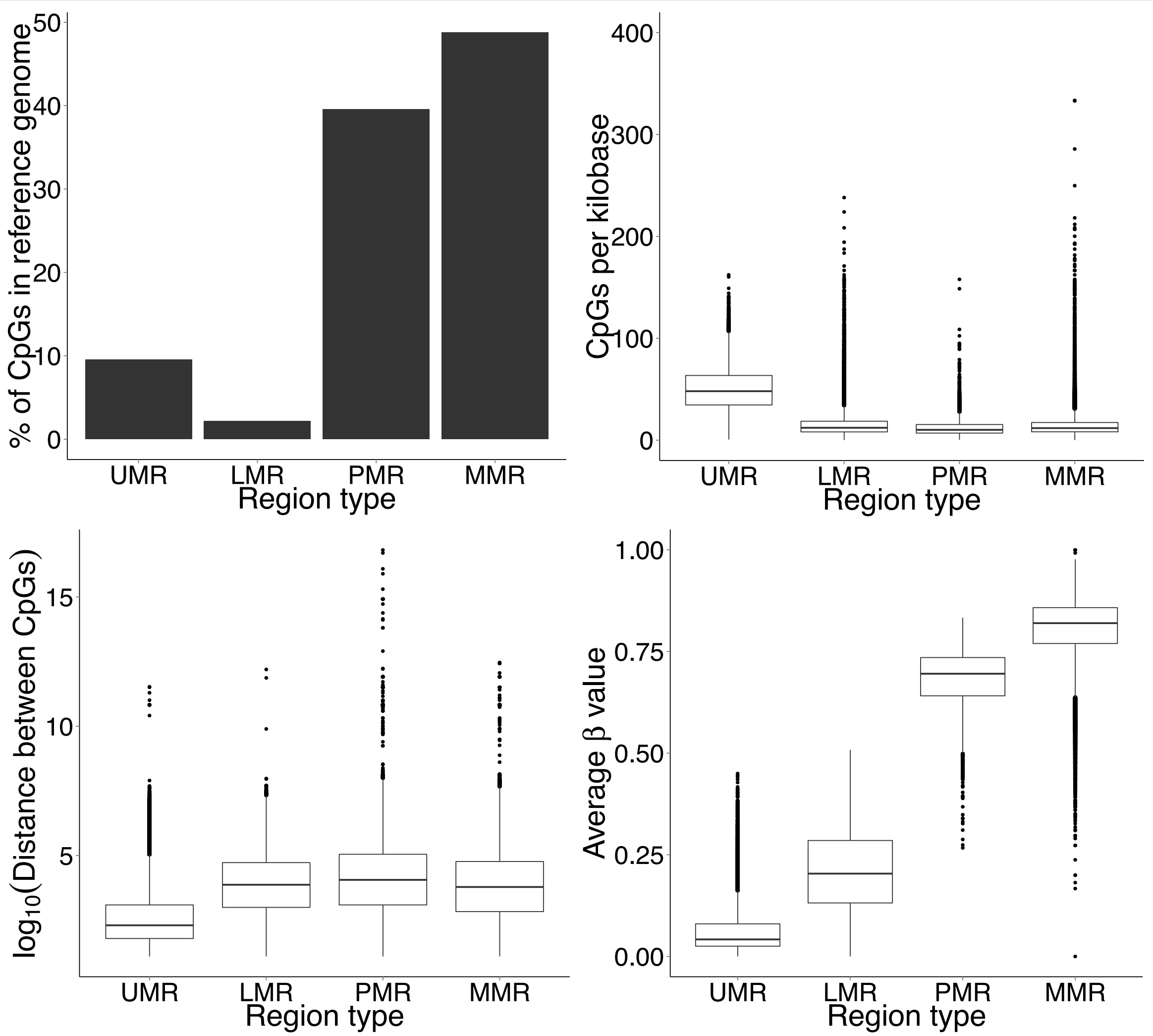


***Figure 2**: Distributions of summary statistics from the segmentation process. The segmentation is based entirely on beta values and is done using MethylSeekR.*

Figures 3, 4, 5 and 6 show summary statistics computed on real methylC-seq data (ADS), data simulated from an independence model that is similar to that implemented in Sherman (Sherman) and data simulated from the model described in the Methods (MySim).

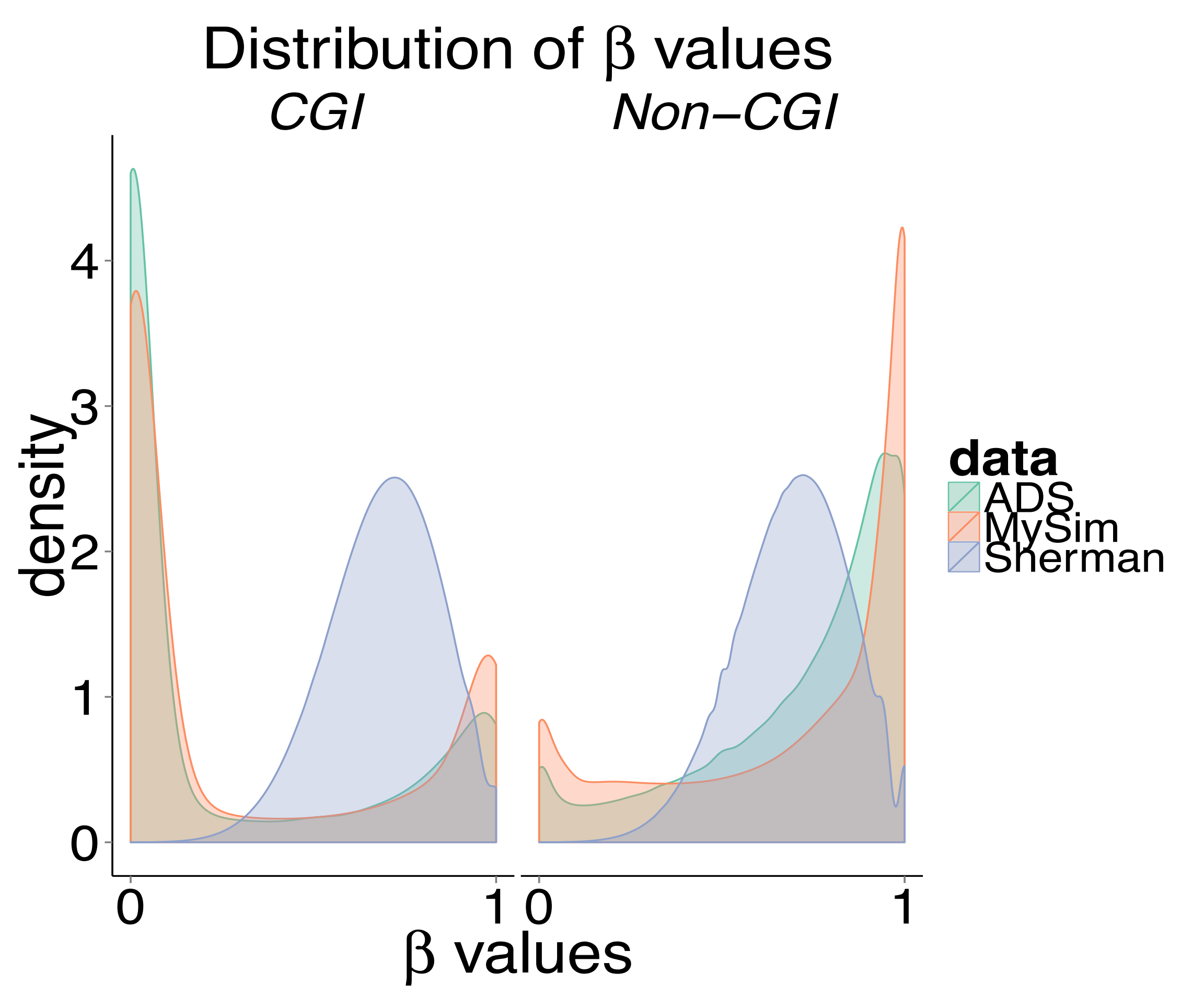


***Figure 3**: The distribution of beta values from MySim mimics that of real data. The Sherman data does not capture the lowly methylated CGIs. This can be improved by assigning different methylation levels to CGI and non-CGI regions.*

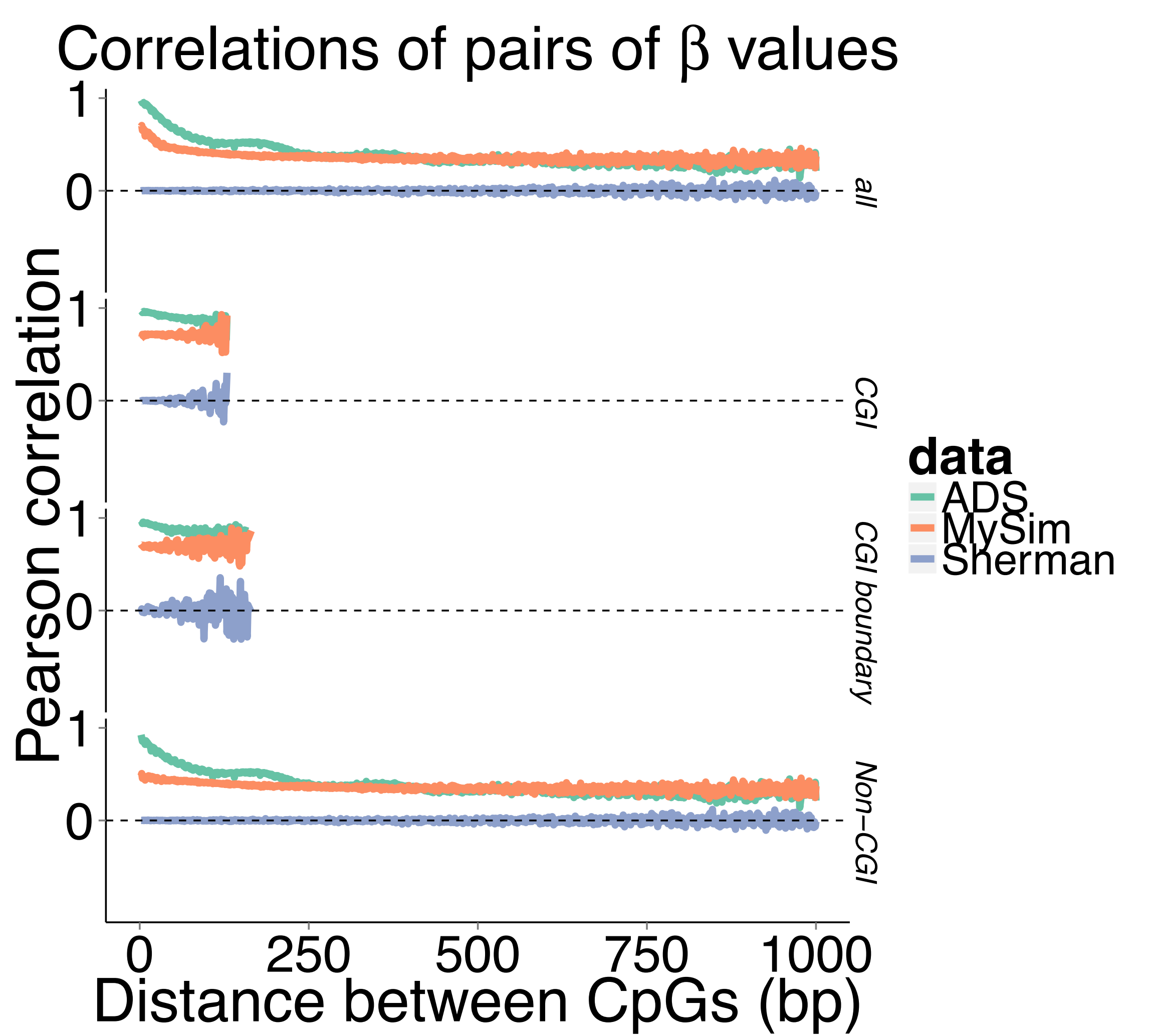


***Figure 4**: The MySim data includes the spatial correlation of beta values seen in real data. This is due to the Markovian spatial dependence model in the MySim method. The Sherman data does not capture the spatial correlations of beta values because each methylation event is simulated independently.*

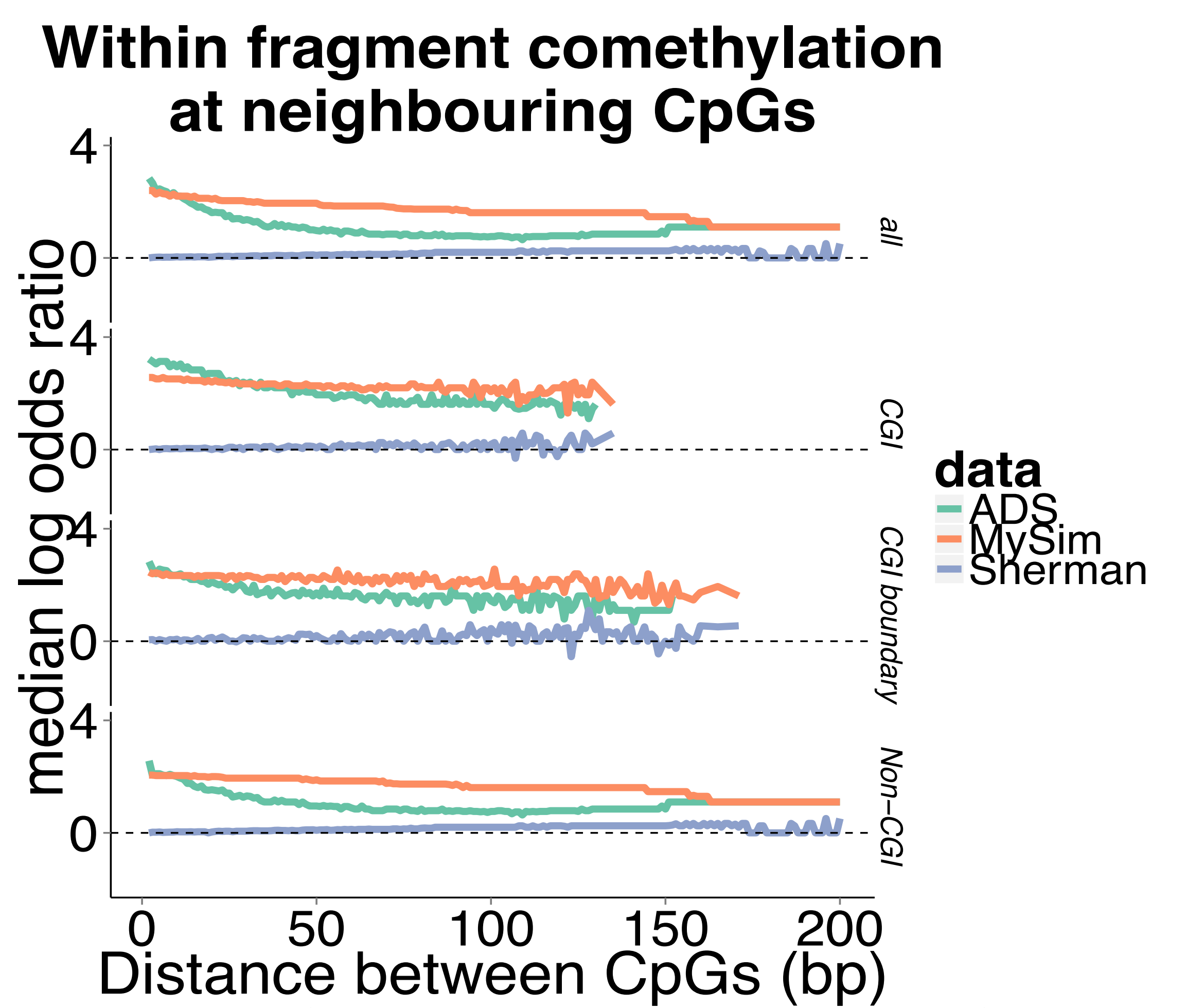


***Figure 5**: Within fragment comethylation is a measure of dependence of methylation states within individual DNA fragments. A value of zero corresponds to independence. The comethylation in the MySim data is generally too high compared to the real data. The Sherman data has more-or-less zero comethylation because each methylation event is simulated independently.*

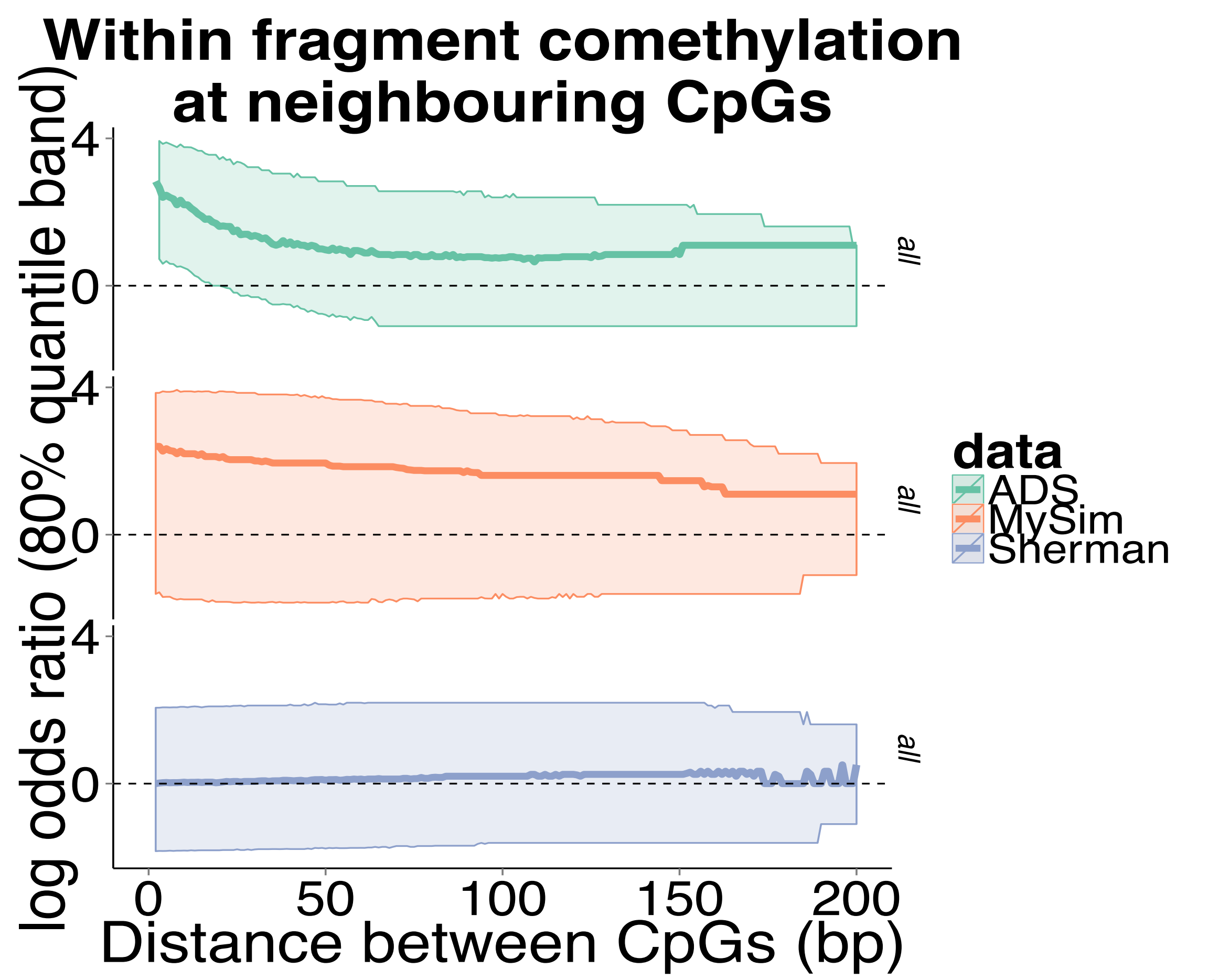


***Figure 6**: Similar to Figure 5 but with the addition of a band showing the 10^{th} and 90^{th} percentiles of the comethylation distributions. The comethylation of real data has less variability than the MySim data.*

Further work

These are very preliminary results and there is much work to be done. In particular, we wish to produce a more "locally stationary" segmentation by refining the segmentation process and to improve the distributions of within-fragment comethylation.

Further information

Poster: http://bit.ly/ECD13_PH

Software: www.github.com/PeteHaitch

Email: hickey@wehi.edu.au

References

1. Burger, L., Gaidatzis, D., Schübeler, D. & Stadler, M. B. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res (2013) (http://www.bioconductor.org/packages/release/bioc/html/MethylSeekR.html)

2. Lister, R. et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471, 68–73 (2011).

3. Sherman - bisulfite-treated Read FastQ Simulator (http://www.bioinformatics.babraham.ac.uk/projects/sherman/)

4. Frith, M. C., Mori, R. & Asai, K. A mostly traditional approach improves alignment of bisulfite-converted DNA. Nucleic Acids Res (2012) (http://www.cbrc.jp/dnemulator/)

5. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (http://www.R-project.org/)

6. Dirk Eddelbuettel and Romain Francois (2011). Rcpp: Seamless R and C++ Integration. Journal of Statistical Software, 40(8), 1-18. (http://www.jstatsoft.org/v40/i08/). Eddelbuettel, Dirk (2013) Seamless R and C++ Integration with Rcpp. Springer, New York. ISBN 978-1-4614-6867-7.