


Metblocks: An unsupervised method for the analysis of methylation data

Christopher Graham Fenton ^{1,2} and Ruth H Paulssen ^{1,2}

¹ Clinical Bioinformatics Research Group, Department of Clinical Medicine, UiT-The Arctic University of Norway, Tromsø, Norway. ² Genomics Support Centre Tromsø (GSCT), Department of Clinical Medicine, UiT-The Arctic University of Norway, Tromsø, Norway.  Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 12 September 2025

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/))

#Summary

Most methylation analysis methods contrast two or more sample groups. Metblocks is a R package that isolates variable methylated regions from methylation count data without sample group information. An unsupervised approach can be useful in identifying subgroups that can be overlooked in group contrasts providing complementary information important for precision medicine. Furthermore, in some cases obtaining samples from control or other groups is not feasible.

#Statement of need

DNA methylation is the addition of methyl groups to DNA, and occurs mostly at CpG sites, within CpG islands (Bernstein et al. (2007)). CpG islands are often found in the promoter regions of genes in mammalian genomes (Antequera & Bird (1999)). Methylation can have cis effects as methylation in promoter regions can lead to a decrease in transcript expression, or trans effects dependent on 3D chromosomal architecture (Qin et al. (2016)). Methylation status is dynamic and can vary significantly dependent on cell type composition, age, medication, and disease state among other factors (Hüls & Czamara (2019)). Methylation methods such as whole genome bisulfite sequencing (WGBS) analysis involves the comparison of millions of CpG sites. Treating all methylation sites as statistically independent may lead to overly harsh multiple-testing penalties. To reduce the number of observations, most algorithms aggregate signal from closely packed methylation sites into regions. Regions are compared between groups of samples, and those meeting the significance criteria are labelled as differentially regulated methylation regions (DMRs). Missing values are a common problem in analyzing large scale methylation data (Lena et al. (2020)). Although different methods are used to impute the value of missing values the methods rely on a priori group information. The goal of this approach was to address the above challenges of isolating variable methylation regions without group information.

#State of field

There are several software packages for the identification, visualization, and annotation of differentially methylated regions using sample group data. These include msPipe (Kim et al. (2022)), BAT (Hoffmann et al. (2017)), bicycle (Graña et al. (2018)), GemBS (Merkel et al. (2019)), Msuite (Sun et al. (2020)), methylseq (Nf-Core/Methylseq (n.d.)), PiGx (GitHub - BIMSBBioinfo/Pigx_bsseq (n.d.)), snake-Pipes (Bhardwaj et al. (2019)), wg-blimp (Wöste et al. (2020)), RnBeads2 (Müller et al. (n.d.)), Bismark (Krueger & Andrews (2011)) and coMET (Martin et al. (2015)) among others. Many of these software implementations share algorithms for imputing missing values and finding DMRs. Common algorithms for imputing missing values and region discovery include BSmooth (Hansen et al. (2012)), metilene (Jühling et al. (2016)), dmrseq (Korthauer et al. (2019)), or DSS algorithms (Park & Wu (2016)). Metblocks hopes to provide complementary information to these well-established algorithms.

#Test data

A WGBS dataset of 9 control and 17 treatment naïve ulcerative colitis (UC) mucosal colonic biopsies was used as a test dataset. UC is a complex heterogeneous disease with considerable variation between patients. Methylation data for chromosome 18 was produced using the Bismark software package. The metblocks results were compared to two widely used DMR finding software packages, metilene and dmrseq using the same dataset. Both dmrseq and metilene isolate DMRs by comparing control and UC groups.

#Methodology

Sites with greater than 30% missing values were removed. Each chromosome was then split into segments dependent on the distance between consecutive CpG sites. A new segment was created if the distance between consecutive CpG sites was greater than 3000 bp. Relative methylation values for the segments were then calculated by dividing the number of methylated sites by total coverage. Any missing values in segment relative methylation were imputed using the R impute package impute (version 1.76.0) impute.knn function as nearest neighbor works well with less than 30% missing values(Jadhav et al. (2019)). The imputed segment relative methylation values were used in metblocks region detection. Blocks were detected in metblocks from each segment using clustering based on Pearson correlation penalized by the genomic distance between CpG sites. An initial distance matrix was computed based on the correlation distance. Correlations were penalized by chromosomal distance using a gaussian decaying function with a bandwidth parameter. Distance decay is an important consideration in region selection as methylation levels of proximal sites are often more related than distal sites(Affinito et al. (2020)). The distance penalized correlations were clustered using single linkage hierarchical clustering (hclust) to isolate blocks. Initial analysis of metblocks results revealed that only one site or sample was often responsible for most of the variation in methylation levels. To exclude these regions an IQR (interquartile range) filter was applied. Metblocks excluded blocks where the methylation level range (maximum-minimum) divided by the interquartile (Q3-Q1) range was greater than or equal to 10. User parameters include the minimum number of CpG sites required to keep segment, bandwidth for distance decay function, minimum number of CpG sites required to keep block, the hclust threshold, iqr filtering cutoff, number of neighbors for KNN comparison, and number of cores. For example, lowering the hclust threshold gives fewer and smaller blocks or increasing the distance decay parameter penalizes CpG sites that are farther apart.

#Results

The average size of the 585 segments meeting the criteria was 2690 bp with an average of 81 CpG sites. Metblocks found more blocks (404) as shown in Table 1 retaining a total of 0.6% of total methylation sites roughly three times more than metilene and dmrseq DMRs. Metblocks results were both smaller and contained more densely packed with CpG sites than DMRs found by the dmrseq and metilene supervised methods.

Table 1 Comparison of regions

	metblocks	dmrseq	metilene
number of regions	404	162	245
pct of total sites**	0.6	0.23	0.26
mean width of region	154.42	293.13	186.05
mean number of CpGs per region	18.95	18.41	13.38

The density distribution of regions found by all three methods is shown in Figure 1A. There is considerable genomic positional overlap between the results of all three methods as shown in Figure 1B. Metblocks isolated 245 regions that did not overlap with either of the supervised methods. These are regions that were not significantly (q.value < 0.1) differentially methylated between UC and control samples. Figure 1C shows a principal component analysis done on a matrix of relative methylation levels of all CpG sites located within metblock results. Figure1C

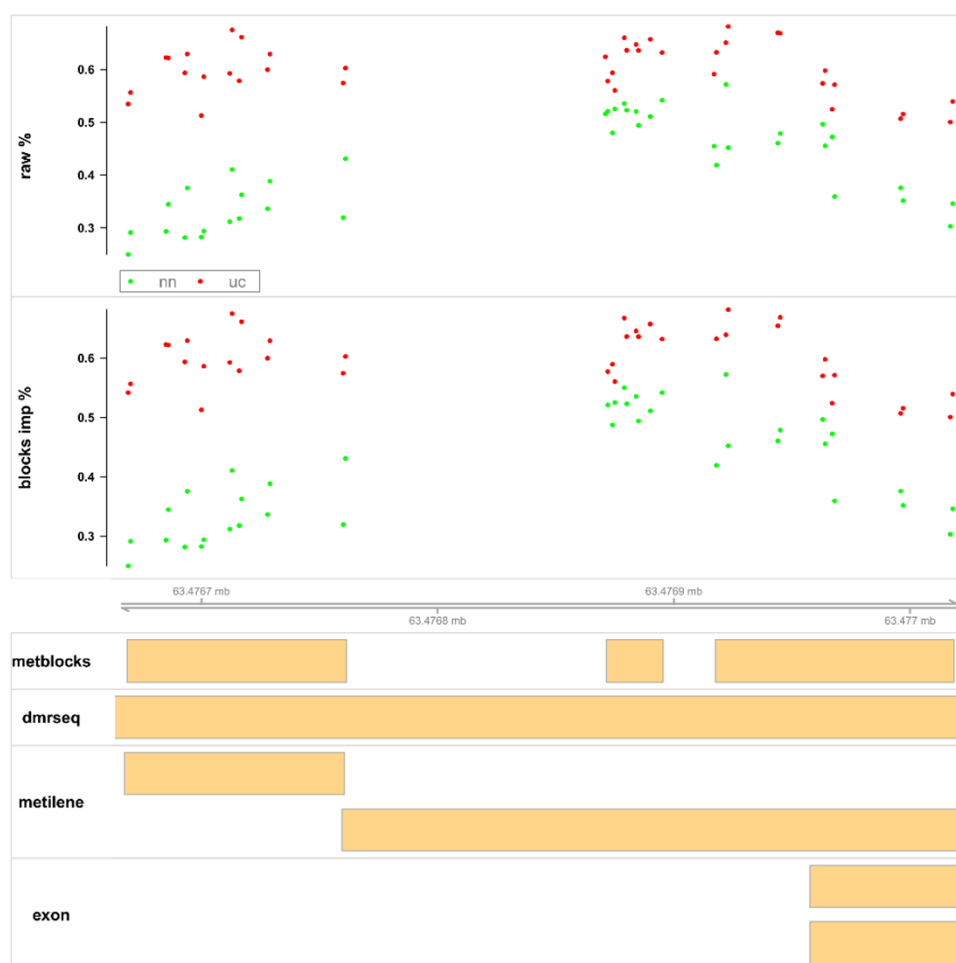


Figure 2: Example overlaps of regions .

Figure2, Shows a visualization of methylation events from a region on chromosome 18. The top panel shows the raw relative methylation data for all samples. The second panel shows the imputed methylation level for each CpG site for each metblocks result for all samples. In these panels red is ulcerative colitis (uc) and green is control (nn). The yellow boxes show the identified regions for all three methods along with exon location at the bottom.

#Conclusion

The metblocks approach was designed with flexible parameters to help reduce methylation data to variable methylated regions. Usage includes exploration analyses in datasets with single or unclear grouping information. The methodology is complementary to traditional supervised approaches. The identification of sub-groups or additional regions may aid in the discovery of methylated regions that are biologically significant at a group or individual level.

#Data availability

Metblocks is provided as an R package at <https://github.com/christopher047/metblocks>. All the information needed to reproduce these results can be found on the website in <https://github.com/christopher047/metblocks/tree/main/misc>.

Affinito, O., Palumbo, D., Fierro, A., Cuomo, M., Riso, G. D., Monticelli, A., Miele, G., Chiariotti, L., & Coccozza, S. (2020). Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics*, 112, 144–150. <https://doi.org/10.1016/j.ygeno.2019.05.007>

- Antequera, F., & Bird, A. (1999). CpG islands as genomic footprints of promoters that are associated with replication origins. *Current Biology*, 9, R661–R667. [https://doi.org/10.1016/S0960-9822\(99\)80418-7](https://doi.org/10.1016/S0960-9822(99)80418-7)
- Bernstein, B. E., Meissner, A., & Lander, E. S. (2007). The mammalian epigenome. *Cell*, 128, 669–681. <https://doi.org/10.1016/J.CELL.2007.01.033>
- Bhardwaj, V., Heyne, S., Sikora, K., Rabbani, L., Rauer, M., Kilpert, F., Richter, A. S., Ryan, D. P., & Manke, T. (2019). snakePipes: Facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics*, 35, 4757–4759. <https://doi.org/10.1093/BIOINFORMATICS/BTZ436>
- GitHub - BIMSBBioinfo/pigx_bsseq: Bisulfite sequencing pipeline from fastq to methylation reports. (n.d.). https://github.com/BIMSBBioinfo/pigx_bsseq
- Graña, O., López-Fernández, H., Fdez-Riverola, F., Pisano, D. G., & Glez-Peña, D. (2018). Bicycle: A bioinformatics pipeline to analyze bisulfite sequencing data. *Bioinformatics*, 34, 1414–1415. <https://doi.org/10.1093/BIOINFORMATICS/BTX778>
- Hansen, K. D., Langmead, B., & Irizarry, R. A. (2012). BSmooth: From whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13, 1–10. <https://doi.org/10.1186/GB-2012-13-10-R83/FIGURES/4>
- Hoffmann, S., Kretzmer, H., & Otto, C. (2017). BAT: Bisulfite analysis toolkit. *F1000Research*, 6. <https://doi.org/10.12688/F1000RESEARCH.12302.1/DOI>
- Hüls, A., & Czamara, D. (2019). Methodological challenges in constructing DNA methylation risk scores. *Epigenetics*, 15, 1. <https://doi.org/10.1080/15592294.2019.1644879>
- Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33, 913–933. https://doi.org/10.1080/08839514.2019.1637138/ASSET/B09346F7-817F-47E3-B8BD-2117E2BEFFCD/ASSETS/GRAPHIC/UAAI_A_1637138_F0005_OC.JPG
- Jühling, F., Kretzmer, H., Bernhart, S. H., Otto, C., Stadler, P. F., & Hoffmann, S. (2016). Metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Research*, 26, 256. <https://doi.org/10.1101/GR.196394.115>
- Kim, H., Sim, M., Park, N., Kwon, K., Kim, J., & Kim, J. (2022). msPIPE: A pipeline for the analysis and visualization of whole-genome bisulfite sequencing data. *BMC Bioinformatics*, 23, 1–13. <https://doi.org/10.1186/S12859-022-04925-2/TABLES/2>
- Korthauer, K., Chakraborty, S., Benjamini, Y., & Irizarry, R. A. (2019). Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics*, 20, 367–383. <https://doi.org/10.1093/BIOSTATISTICS/KXY007>
- Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *BIOINFORMATICS APPLICATIONS NOTE*, 27, 1571–1572. <https://doi.org/10.1093/bioinformatics/btr167>
- Lena, P. D., Sala, C., Prodi, A., & Nardini, C. (2020). Methylation data imputation performances under different representations and missingness patterns. *BMC Bioinformatics*, 21, 1–22. <https://doi.org/10.1186/S12859-020-03592-5/TABLES/8>
- Martin, T. C., Yet, I., Tsai, P. C., & Bell, J. T. (2015). coMET: Visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. *BMC Bioinformatics*, 16, 1–5. <https://doi.org/10.1186/S12859-015-0568-2/FIGURES/1>
- Merkel, A., Fernández-Callejo, M., Casals, E., Marco-Sola, S., Schuyler, R., Gut, I. G., & Heath, S. C. (2019). gemBS: High throughput processing for DNA methylation data from

- 165 bisulfite sequencing. *Bioinformatics* (Oxford, England), 35, 737–742. <https://doi.org/10.1093/BIOINFORMATICS/BTY690>
166
- 167 Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., & Bock, C. (n.d.).
168 *RnBeads 2.0: Comprehensive analysis of DNA methylation data*. <https://doi.org/10.1186/s13059-019-1664-9>
169
- 170 *Nf-core/methylseq: Nf-core/methylseq version 1.3*. (n.d.). <https://doi.org/10.5281/ZENODO.2555454>
171
- 172 Park, Y., & Wu, H. (2016). Differential methylation analysis for BS-seq data under gen-
173 eral experimental design. *Bioinformatics*, 32, 1446–1453. <https://doi.org/10.1093/BIOINFORMATICS/BTW026>
174
- 175 Qin, Z., Li, B., Conneely, K. N., Wu, H., Hu, M., Ayyala, D., Park, Y., Jin, V. X., Zhang, F.,
176 Zhang, H., Li, L., & Lin, S. (2016). Statistical challenges in analyzing methylation and
177 long-range chromosomal interaction data. *Stat Biosci*, 8, 284–309. <https://doi.org/10.1007/s12561-016-9145-0>
178
- 179 Sun, K., Li, L., Ma, L., Zhao, Y., Deng, L., Wang, H., & Sun, H. (2020). Msuite: A
180 high-performance and versatile DNA methylation data-analysis toolkit. *Patterns* (New
181 York, N.Y.), 1. <https://doi.org/10.1016/J.PATTERN.2020.100127>
- 182 Wöste, M., Leitão, E., Laurentino, S., Horsthemke, B., Rahmann, S., & Schröder, C. (2020).
183 Wg-blimp: An end-to-end analysis pipeline for whole genome bisulfite sequencing data.
184 *BMC Bioinformatics*, 21, 1–8. <https://doi.org/10.1186/S12859-020-3470-5/FIGURES/3>

DRAFT