wgbsGSEA: Evaluating Gene Set Enrichment Analysis Methods (methylGSA & ebGSEA) for Whole Genome Bisulfite Sequencing Data

Maya Arvanitis

Department of Computer Science, McGill University
Supervisor: Dr. Claudia Kleinman, Department of Human Genetics

Abstract

Whole Genome Bisulfite Sequencing (WGBS) provides detailed levels of methylation data by investigating patterns at single-base resolution and providing valuable insights into epigenetic regulation. Traditional Gene Set Enrichment Analysis (GSEA) methods are robust in terms of RNA-seq data. However, these methods present challenges for WGBS data due to its sparse CpG coverage in addition to inherent uneven CpG distribution across genes. This study aims to address these challenges by analyzing the feasibility and performance of two GSEA methods, **methylGSA** and **ebGSEA**. Through adaptation of the two methods, including empirical Bayes and rank-based approaches, sensitivity, specificity and the detection of enriched pathways was analyzed. Findings underscore the importance in further developing robust workflows and designing statistical frameworks in order to address the specific challenges WGBS data poses for Gene Set Enrichment Analysis.

GitHub Repository for this project: https://github.com/mayaarvanitis/wgbsGSEA

1 Introduction

DNA methylation is a pivotal modification in epigenetics that has significant influences on gene expression, genomic stability and development and disease (Moore et al., 2013). Whole Genome Bisulfite Sequencing (WGBS) provides single base resolution on the DNA methylation patterns across the entire genome. This is an invaluable tool that involves treating genomic DNA with sodium bisulfite. Consequently, the unmethylated cytosines are transformed into uracils, whilst the methylated cytosines remain unchanged. Further sequencing allows the differentiation between cytosines that are methylated and those that are unmethylated, representing a formidable tool for viewing detailed methylation profiles (Clark et al., 2021).

Whilst WGBS provides unparalleled amounts of data, current Gene Set Enrichment (GSEA) methodologies are traditionally developed for RNA-sequencing data. Several challenges can be encountered when applying GSEA methods to WGBS data. For instance, the sparse CpG coverage per gene. Additionally, the alignment of CpGs to genes is challenging due to the unequal number of CpGs per associated gene. For this reason, in traditional GSEA methods genes with more CpGs tend to have higher significance simply due to this inequality in mapped genes (Hansen et al., 2021).

This project aims to evaluate two GSEA specialized tools, **ebGSEA** (Dong et al., 2019) and **methylGSA** (Ren et al., 2018) for WGBS data. Although these tools are not inherently designed to evaluate WGBS data, this project aims to evaluate both their effectiveness and performance on WGBS datasets.

methylGSA, developed by Ren and Kuan (2018) provides a complementary approach by implementing two methods. 1) *methylRRA*, a rank aggregation method that will adjust for the number of CpGs per gene, and 2) *methylglm*, a method that utilizes logistic regression in order to account for the density biases in CpG.

ebGSEA, developed by Dong et al. (2019), is an empirical Bayes-based method that will rank genes by their overall differential methylation by using CpGs mapped to a gene. Overall, this approach will reduce bias that is caused by the variability in CpG representation and enhance the sensitivity in order to further detect biologically relevant pathways.

Outcomes of this work have potential to advance the availability of effective tools designed for WGBS data. The development of a robust framework for GSEA on WGBS data would significantly contribute to our understanding of biological processes in addition to their consequential implication for disease.

2 Background

DNA methylation is a modification that involves the addition of a methyl group to the 5-carbon position of a cytosine. In terms of epigenetics, this modification plays a significant role in gene regulation as well as developmental processes. Hypermethylation, in tumor suppressor genes for example, have been implicated in diseases such as cancer and neurodevelopmental disorders (Baylin & Jones, 2011).

2.1 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is an invaluable tool for interpreting high-throughput data. GSEA will evaluate whether a set of genes show statistically significant enrichment. The traditional GSEA algorithm will evaluate enrichment through a calculation of a running-sum statistic, by taking a ranked list of genes and then computing the observed enrichment score (ES) and generating a comparison to the null

distribution that is obtained through permutation testing. When computing the normalized enrichment score (NES), we take into account the varying gene set sizes as well as a p-value adjustment measure in order to determine significance. GSEA is well developed and established for RNA-seq data, as it leverages the well-defined gene expression profiles where there is a direct link between each gene's transcription abundance and its activity level in cells (Subramanian et al., 2005).

However, developing GSEA methods to WGBS data is not as straightforward, as not all CpGs are uniform across the genome as well as the mapping challenges as CpGs must be assigned to genes.

2.2 Statistical Basis for ebGSEA and methylGSEA methods

2.21 methylGSA

methylGSA implements two approaches for gene set enrichment analysis. The first is done through the methylRRA (Robust Rank Aggregation) method. This method aggregates the p-values of CpGs within each gene and then adjusts for the number of CpGs per gene.

In terms of computation, we have given $P_1, P_2, \ldots, P_n$ p-values, with n CpGs in a gene g, the robust rank aggregation method will compute the following score:

$$RRA_g = min\{P_1, P_2, \ldots, P_n\}$$

Where p-values are adjusted using the Bonferroni correction due to multiple hypothesis testing.

Next, methylGSA implements a logistic regression model in order to evaluate the probability of the gene g being enriched in a given pathway. Additionally, this method adjusts for CpG density.

The model is specified in methylglm as the following:

$$logit(p_i) = \beta_0 + \beta_1 X_i + \beta_2 C_i$$

With p_i being the probability of enrichment for a gene i , C_i being the number of CpGs in a gene i and X_i being the number of CpGs in gene i.

2.22 Empirical Bayes GSEA

ebGSEA utilizes the empirical bayes framework. ebGSEA will rank genes based on their overall differential methylation levels by using all the CpGs that are mapped to a gene. The method first applies a global test.

$$S_g = \sum_{j \varepsilon g} T_j$$

Computing the test statistic T_j for the jth CpG in a given gene g.

3 Methods

3.1 Data

The data used in this project consisted of Whole Genome Bisulfite Sequencing datasets generated in order to study the effects of histone H3.3 G34R mutations on DNA methylation in the cortex of 10-week-old-mice. The samples were aligned to the mm10/GRCm38 genome. The analysis included two experimental groups (Khazaei, 2023):

- Wild-Type (WT): N=2 samples (`SK059_2427_WT_CRX`, `SK060_2616_WT_CRX`)
- G34R Mutant: N=2 samples (`SK061_2617_G34R_CRX`, `SK062_2428_G34R_CRX`)

Two key data files were utilized to prepared inputs for analysis methods:

1. WTvsG34R_CRX_10W.percentage_meth.tsv.gz
    a. This file contains the percentage methylation values at each CpG site across the genome. It provides the proportion of methylation cytosines for each position, providing a direct measure of methylation at the single-base level.
2. WTvsG34R_CRX_10W.diff_meth.tsv.gz
    a. This file contains the differential methylation statistics for each CpG site, comparing the WT and G34R groups. It provides the p-values indicating the significance of methylation differences between the groups.

3.2 methylGSA

To evaluate the feasibility of applying the methylGSA package for gene set analysis on Whole Genome Bisulfite Sequencing (WGBS) data, a series of tests were conducted using differential methylation data (Figure 1).

chr	start	end	strand	p-value	q-value	meth.diff
1	3000827	3000827	+	7.890358e-01	0.93115340	-0.94696970
1	3000827	3000827	-	2.69319e-01	0.70704502	2.85714286

***Figure 1**: Sample View of Methylation Data.*

CpGs were filtered based on statistical significance threshold (p-value < 0.05) and categorized into either hypermethylated (meth.diff > 0) or hypomethylated (meth.diff < 0) subsets.

Next steps involved mapping CpGs to genes using the **biomaRt** package to query the Ensembl database (Durinck et al., 2009). Overlaps between CpG coordinates and gene regions were identified with the GRanges framework, and the resulting mappings were further processed to eliminate the possibility of duplicates. These mappings were then formatted into compatible methylGSA annotation files using the `prepareAnnot` function.

Gene sets relevant to the mouse genome were retrieved from the Molecular Signatures Database (MSigDB), focusing on hallmark gene sets. The analysis tested three of the methods provided by methylGSA: logistic regression (methylglm), over-representation analysis (ORA), and gene set enrichment analysis (GSEA). Logistic regression assessed the enrichment of gene sets whilst adjusting for CpG density biases. ORA identified the pathways which were over-represented and GSEA evaluated whether gene sets were significantly enriched, determined by their ranked p-values. Parameters were then adjusted to reflect the data, limiting gene set sizes to between 2 and 800 genes.

3.3 ebGSEA

Further analysis was conducted with a customized implementation of the ebGSEA method to evaluate gene set enrichment in WGBS data.

3.31 Input Data Preparation

Input data was prepared consisting of:

1. A binary phenotype vector (`pheno.v`) indicating the experimental groups (Wild Type [WT] and G34R Mutant) structured as [0, 0, 1, 1] where 0 = WT and 1 = G34R
2. A methylation data matrix (`data.m`) containing CpG IDs as row names and sample beta values (methylation levels ranging from 0 to 1) as columns, CpG IDs were formatted as chr_start_end format for consistency. The structure of

the matrix ensured compatibility with ebGSEA requirements.

3.32 Custom Gene Mapping

New functions (described below) were developed as R scripts, and a custom gene mapping list (`custom_map`) was created to associate CpG IDs with genes using their ENTREZ IDs. The list was structured with gene IDs as keys and vectors of corresponding CpG IDs as values. This format allowed to incorporate genome-wide methylation data, diverging from the array-specific formats used in the original ebGSEA.

3.33 Custom functions doGT_custom and doGSEAwt_custom

The analysis incorporated two custom functions, **`doGT_custom`** and **`doGSEAwt_custom`** from their original implementations in ebGSEA. These functions were designed to incorporate properties lacking in the original doGT doGSEAwt ebGSEA implementations.

The `doGT_custom` function builds upon the original doGT source code from ebGSEA by introducing support for custom arrays and CpG-to-gene mappings. It then performs a global test as implemented in the original doGT function in order to assess whether specific gene sets are enriched for differential methylation signals. An additional array input was set to "`custom`" to enable the use of user-defined CpG-to-gene mappings. Moreover, to further allow for custom array support beyond the predefined arrays such as Illumina 450k/850k in the original function, a `custom_map` was imputed as a named list mapping ENTREZ IDs to their associated CpG IDs. This mapping allowed for genome-wide analysis through its linking of CpG sites to specific genes.

The `doGSEAwt_custom` function extends the functionality of the original source code of doGSEAwt in ebGSEA. It performs GSEA using the Wilcoxon Rank Sum Test (WT) and the Known-Population Median Test (KPMT). The main additional functionality in this implementation is the cross-species support allowing for the analysis of datasets from multiple organisms by querying the appropriate annotation database and allowing users to input an "`org.db`" annotation database object for the selected analysis species. In this study, the `org.Mm.eg.db` mouse genome was used.

Input validation was conducted in order to ensure consistency between `pheno.v` and `data.m` as well as verifying that the CpG IDs in `data.m` aligned with those in `custom_map`.

Subsequently, the global test was executed through the `doGT_custom` function applying the Wilcoxon Rank Sum Test (WT) and Known-Population Median Test (KPMT) to assess the enrichment levels across the gene sets. Pathways were further filtered to include those with a minimum of valid genes (`minN = 10`) and adjusted p-values (`adjPVth = 0.05`) were used in order to identify the pathways deemed significant.

4 Results

4.1 methylglm

Results from the application of the methylglm function from the methylGSA method using hypermethylated CpGs and hallmark genes identified several pathways with notable raw p-values. That being said, none of these pathways seemed to have been statistically significant (p-value < 0.05) following adjustment. The pathway with the lowest raw p-value was found to be Il2 Stat5 Signaling Hallmark Pathway with p = 0.0086, and second, The Bile Acid and Metabolism Pathway with p = 0.0373. However, once adjusted, their p-values were 0.43 and 0.93,

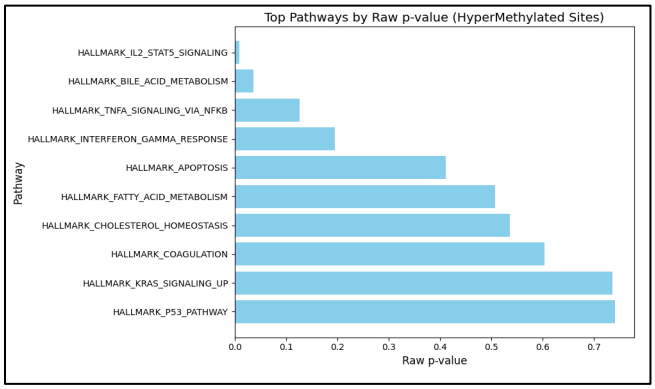

Figure 2: *Bar plot of top pathways for hypermethylated CpGs sorted by raw p-value.*

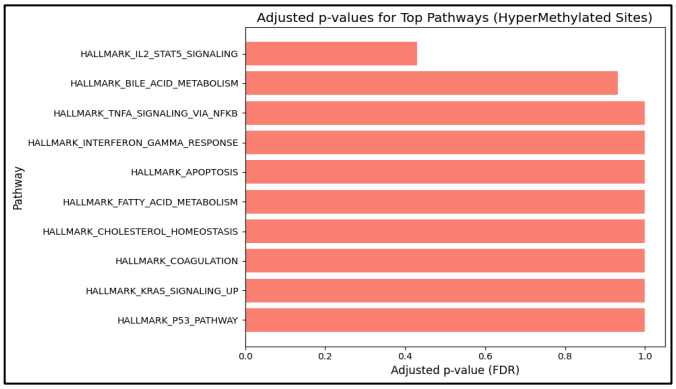

Figure 3*: Bar plot of top pathways for hypermethylated CpGs sorted by adjusted p-value.*

deemed to not be significant as seen in the comparison done between Figure 2 and Figure 3.

For hypomethylated sites, no hallmark pathways were found to be significantly enriched after adjustment for multiple testing. However, pathways such as the Oxidative Phosphorylation pathway with 197 genes showed a raw p-value of 0.0071 before being adjusted to 0.356, perhaps warranting further investigation.

4.2 methylRRA

4.21 GSEA

The application of methylRRA for Gene Set Enrichment Analysis on hypermethylated CpGs identified several hallmark pathways with significant enrichment scores. First, the pathway with the strongest enrichment was UV Response pathway ($NES = 1.94,\ p = 1 \times 10^{-10}, padj = 1.67 \times 10^{-9}$) and second the Mitotic Spindle pathway ($NES = 1.83,\ p = 1 \times 10^{-10}, padj = 1.67 \times 10^{-9}$) (See Figure 4).

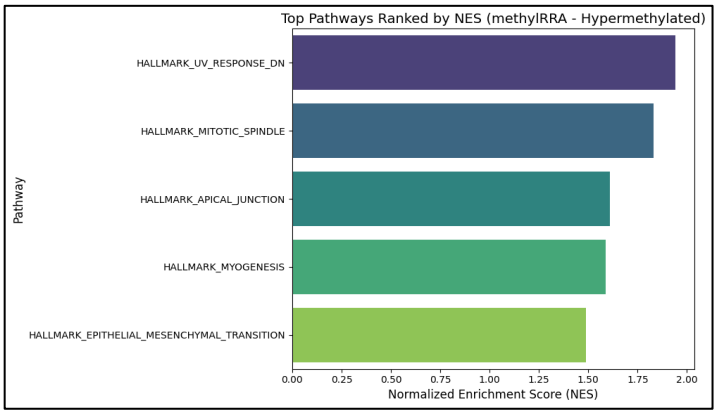

Figure 4: *Top Pathways ranked by NES for Hypermethylated CpGs. Found top pathways UV Response with NES = 1.94, Mitotic Spindle (NES = 1.83), Apical Junction (NES = 1.59), Myogenesis (NES = 1.59) and Epithelial Mesenchymal Transition (NES = 1.50)*

GSEA for hypomethylated CpG sites revealed enriched pathways such as UV Response with a normalized enrichment score of 1.99 and an adjusted p-value of $padj = 1.25 \times 10^{-9}$. Other enriched pathways included Mitotic Spindle, Apical Junction, and Myogenesis (Figure 5)

ID	Size	Enrichment Score	NES	p-value	Adjusted p-value (padj)	Leading Edge	Core Enrichment
HALLMARK_UV_RESPONSE_DN	244	0.6124	1.9862	1×10^{-10}	1.25×10^{-9}	tags=60%, list=22%, signal=47%	Atrnl1/Erbb4/...
HALLMARK_MITOTIC_SPINDLE	287	0.5357	1.7491	1×10^{-10}	1.25×10^{-9}	tags=56%, list=25%, signal=42%	Rabgap1l/Dock2/...
HALLMARK_APICAL_JUNCTION	272	0.4971	1.6199	1×10^{-10}	1.25×10^{-9}	tags=48%, list=26%, signal=36%	Cntn4/Cntn5/...
HALLMARK_MYOGENESIS	256	0.4963	1.6143	1×10^{-10}	1.25×10^{-9}	tags=40%, list=19%, signal=33%	Erbb4/Sgcd/...
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	235	0.4559	1.4767	5.66×10^{-7}	5.66×10^{-6}	tags=45%, list=28%, signal=32%	Lrp1b/Opcml/...
HALLMARK_IL2_STAT5_SIGNALING	234	0.4443	1.4388	7.57×10^{-6}	6.31×10^{-5}	tags=57%, list=36%, signal=37%	Prkce/Rabgap1l/...
HALLMARK_APICAL_SURFACE	81	0.5291	1.6189	1.04×10^{-5}	7.39×10^{-5}	tags=43%, list=18%, signal=35%	Epha6/Pkhd1/...

Figure 5: *Enriched Pathways for hypomethylated CpG sites.*

4.3 ebGSEA

Results from the empirical Bayes-based Gene Set Enrichment Analysis using the global test (Sg) found several gene sets with significant enrichment scores.

Multiple gene sets (IDS 20220, 66218, 18424, 27390) were found to have strong significance with p-values near zero. Gene sets with strong significance reported high test statistics compared to their expected value of 33.33 and standard deviation of 38.49.

Moreover, several notable significant enrichments were found with p-values within the range between 0.001 and 0.002. Gene sets such as 17168 (p-value = 0.0009170) and 22195 (p-value = 0.0009170) were identified. Additional gene sets were identified with enrichment signals where the test statistic remained significant however close to the null expectation perhaps warranting further investigation (Figure 6).

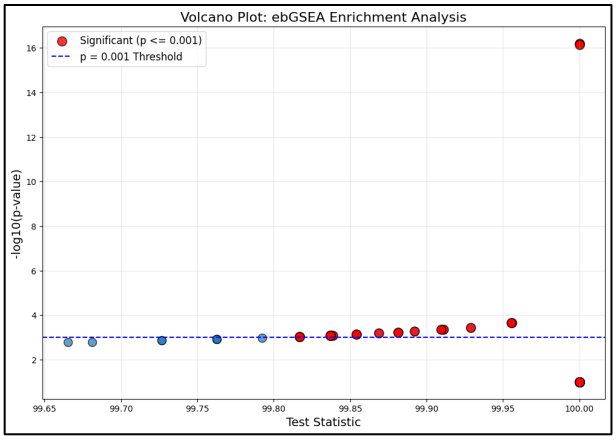

Figure 6: *Volcano plot of ebGSEA results. Test statistic plotted against -log10 of pvalues. Strong statistical significance found with p-values in red. Significance threshold set to 0.001*

Gene sets with coverage higher than 1, such as IDs 53331 and 13688 displayed statistical strength.

5 Discussion

This study aimed to assess the feasibility and performance of the two GSEA methods methylGSA and ebGSEA for Whole Genome Bisulfite Data through adaptation of these two tools to WGBS datasets.

5.1 Performance of methylGSA

The two approaches involved in methylGSA, methylglm and methylRRA showed varying levels of effectiveness. First, the logistic regression (methylglm) method did not identify significantly enriched pathways after p-value adjustment for multiple pathways. Although raw p-values were low and methylglm seems to detect these initial signals of enrichment, the adjustment for CpG density biases may reduce its sensitivity to the signals in WGBS data. While this adjustment for CpG density is a critical part of GSEA methods, it is possible that this is reducing the sensitivity of the method to detect the true biological signals that are present in Whole Genome Bisulfite Sequencing data.

Since CpGs are unevenly distributed across the genome, which leads to variability in the number of CpGs that are associated with each gene, we can see significant differential methylation in higher CpG dense genes. MethylGSA adjusts for this by correcting for CpG densities. However, this correction does reduce the amount of false positives (pathways that are incorrectly identified

as significantly enriched due to CpG density), consequently increasing specificity, it may also be decreasing sensitivity. This is a significant trade-off particularly for WGBS data, as with WGBS data being sparse, this careful consideration of adjustment for distribution biases is a crucial point in its analysis.

The rank aggregation method, methylRRA identified key pathways in the analysis of hypermethylated CpGs even after multiple testing adjustments. This suggests the potential for robust rank aggregation as a feasible method for WGBS datasets. Hypomethylated CpGs also identified biologically relevant pathways with methylRRA.

For instance, it was found that the UV response and mitotic spindle pathways were significantly enriched after adjustment. methylRRA seems to effectively capture the methylation signatures and identify the pathways that are known to hold regulatory roles of DNA methylation in response to UV damage and mitotic processes (Ren & Kuan, 2018; Laird, 2010).

Similarly, it was found that in hypomethylated CpGs pathways such as oxidative phosphorylation were enriched, although less significantly after adjustment. This could align with the findings that hypomethylation often occurs in the regions of active transcription, which contributes to transcriptional regulation (Jones, 2012; Dong et a., 2019).

All in all, methylRRA seems to overcome some of the challenges inherently imposed when conducting GSEA with WGBS data. The robust rank aggregation method emphasizes the p-value aggregation in order to prioritize certain signals, leading to a reduction in the sparse and noisy coverage WGBS data holds. Therefore, in contrast to methylglm, which seems to compromise sensitivity in its adjustment, methylRRA is complementary in its approach, allowing for specificity and sensitivity without loss (Ren & Kuan, 2018).

5.2 Performance of ebGSEA

Results overall from the empirical Bayes-based Gene Set Enrichment Analysis (ebGSEA) show several gene sets with p-values that are near zero, suggestive of extreme statistical significance. While these results seem to indicate strong associations, they also raise concerns regarding the accuracy and reliability of this method when applied to WGBS data.

Since WGBS holds sparse CpG coverage where many genes may have few CpG sites represented this can lead to insufficient representation when examining the gene's overall methylation status. Therefore, it is possible that gene sets with a low number of CpGs will compute extreme test statistics as they will be amplified due to the minute differences in the sparse data. Consequently, this may cause the inflation of significance of enrichment scores and result in p-values that are near zero.

Particularly in empirical Bayes methods which assume a uniform distribution (Efron & Tibshirani, 2002), the imbalance of CpG sites across the genome may introduce bias with genes with high CpG densities being overrepresented in GSEA analysis, whereas, in contrast, genes with lower CpG densities may fall short in terms of statistical force.

6 Conclusions

This project assessed the performance of the two Gene Set Enrichment methods **methylGSA** and **ebGSEA** for Whole Genome Bisulfite Sequencing data. Both methods showed potential for analyzing methylation data although posing challenges and biases due to the particular features of WGBS data.

Whilst the methylglm approach in methylGSA identified several initial statistically significant pathways, the adjustment inherent to the method due to CpG density biases reduced the approaches' sensitivity.

The methylRRA method captured several significantly significant biological pathways such as UV response and oxidative phosphorylation. This was done by the aggregation of the p-values across the sparse data, seemingly overcoming the challenge of sparsity that WGBS imposes upon traditional GSEA methods.

In contrast, ebGSEA resulted in several extremely statistically significant pathways with p-values near zero. Implications of these results raise concerns regarding the effects of sparse CpG coverage as well as uneven CpG distribution which amplifies biases towards high CpG dense genes in empirical Bayes approaches.

The analysis of both of these methods addresses several inherent challenges in applying GSEA methods to WGBS data. Developing robust workflows to handle the sparsity of CpGs is critical and future direction of refining statistical models such as modification of the empirical Bayes framework could aid in the model's ability to reduce bias due to varying CpG densities.

References

[1] Baylin, S. B., & Jones, P. A. (2011). A decade of exploring the cancer epigenome – biological and translational implications. *Nature Reviews Cancer*, 11(10), 726–734. https://doi.org/10.1038/nrc3130

[2] Clark, S. J., Smallwood, S. A., & Krueger, F. (2021). Whole-genome bisulfite sequencing. In *Epigenetics Methods* (pp. 313–335). Springer Protocols. https://doi.org/10.1007/978-1-0716-4051-7_18

[3] Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A. Y., Wang, T., & Vijg, J. (2019). Accurate identification of single nucleotide variants in whole-genome-amplified single cells. *Nature Methods*, 14(5), 491–493. https://doi.org/10.1038/nmeth.4237

[4] Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

[5] Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A., & Feinberg, A. P. (2021). Increased methylation variation in epigenetic domains across cancer types. *Genome Biology*, 22(1), 55. https://doi.org/10.1186/s13059-021-02388-x

[6] Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies, and beyond. *Nature Reviews Genetics*, 13(7), 484–492. https://doi.org/10.1038/nrg3230

[7] Khazaei, S., et al. (2023). Single substitution in H3.3G34 alters DNMT3A recruitment to cause progressive neurodegeneration. *Cell*, 186(6), https://doi.org/10.1016/j.cell.2023.02.023

[8] Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3), 191–203. https://doi.org/10.1038/nrg2732

[9] Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1), 23–38. https://doi.org/10.1038/npp.2012.112

[10] Ren, X., & Kuan, P. F. (2018). methylGSA: A Bioconductor package and shiny app for differentially methylated region gene set analysis and visualization. *Bioinformatics*, 34(14), 2546–2548. https://doi.org/10.1093/bioinformatics/bty117

[11] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

[12] Efron, B., & Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology, 23(1),* 70–86. https://doi.org/10.1002/gepi.1124