Detection of gene expression from DNA methylation of long reads

Samuel Terkper Ahuno (sta4008)

December 8, 2023

1 Abstract

DNA methylation, where methyl are attached to DNA, is an epigenetic mechanism with roles in health and disease. Our goal is to detect gene expression from DNA sequence and methylation using long reads from oxford nanopore sequencing technology, which provides DNA sequence and methylation calls. I proposed 3 models with increasing complexity to solve this problem. Final model is Bayesian Generalized linear model with priors on that are consistent with the type of dataset with Stochastic Variational Inference as inference algorithm. Posterior predictive showed fit to the data for out of samples data.

2 Introduction

DNA methylation is an important epigenetic modification process that involved the addition of methyl groups to DNA molecules. Various types of DNA modifications, including N6-methyladenine (6mA), N4-methylcytosine (4mC), 5-methylcytosine (5mC), and its oxidized derivatives like 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), are widespread in the genomes of life. These modifications have diverse distribution patterns across cells and tissues and serve essential functions in processes such as genomic imprinting, modulation of chromatin structure, inactivation of transposons, maintenance of stem cell pluripotency and differentiation, regulation of inflammation, and control of transcriptional repression.

3 Methods

4 Exploratory Data Analysis

For bench marking, Murine cell lines treated with Azacytidine (n=3) and DMSO(n=3) were sequenced with oxford nanopore long read sequences and bulk rna-seq. Azacytidine is a demethylation agent. DMSO is a control treatment and should not have a significant effect on methylation. DMSO and Azacytidine should be good samples to model the background methylation. DeSeq2 was used for Differential gene expression. Gene expression were normalized log transformed. For the ONT, base-calling was done with dorado software, reads were mapped to mouse reference genome(mm10) and methylation calling was done with Modkit. I limited analysis to 5mC in CpGs regions only. Each sample's CpG site was supported by minimum of 5 reads and should be presented in 75% all sample.

Previous study shows, correlation between gene expression and promoter methylation which is supported by the data1. Here, i took all the coding genes in the mice genome and computed hypergeometeric mean of cpg methylation at the promoter sites. i also computed other metrics like proportion of methylation and entropy and they all seem seem to be correlated with each other. PCA was done using promoter methylation to see how promoter methylation of genes clusters/explain variance in samples.2

4.1 Model 1: Simple linear regression with sci-kit learn

Here i created a one parameter model where normalized RNA expression was response variable (Y) and promoter methylation as explanatory variable (X). The dataset were divided into 80% training set; 20%

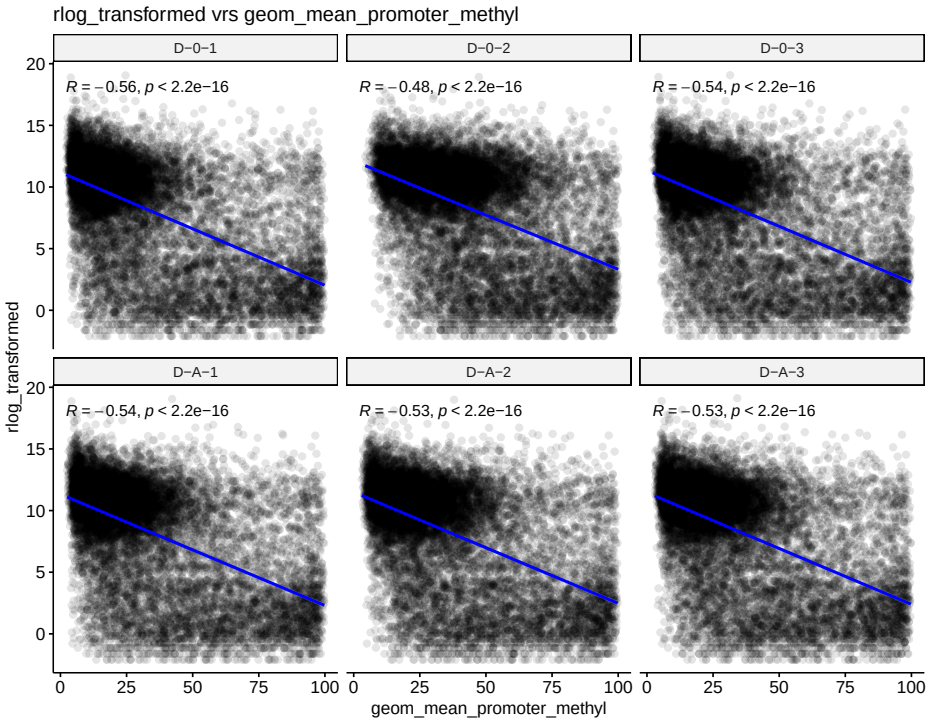


Figure 1: Correlation between gene expression and promoter methylation

test set. Each individual sample was modelled separately given the small sample size. An average R^2 0.3 was achieved for all but one control sample (D-0-2) 3. The was unexpected so as confirmatory analysis i inspected the pca2 which also show that this sample is not consistent with control sample.

4.2 Model 2: BMAP with SGD

models for the project has evolved from simple linear regression to bayesian linear regressions with Gausian priors where the BMAP was generated with stochastic gradient descent. For the stochastic gradient descent. Number of batches was 10, number of iterations was 50000, tolerance was 1e-6, alpha to robsinson and monroe was $1e-8$ and prior beta variance (lambda) hyperparameter was set between 10 to 800. i demonstrated performance of this model on simulated data4.

4.3 Final model

$$
\begin{aligned}
Y^{GE} &\sim \beta_1 X^{'M} + \beta_2 X^C + \epsilon \\
Y &\sim N(\beta^T \cdot X, \quad \sigma^2) \\
\beta_1 &\sim Uniform(1, \quad Y_{max})
\end{aligned}
$$

$$
\begin{aligned}
&Y^{GE} > 0 \\
&X^{'M} \in [0,1] \\
&X^{'M} = (1 - \frac{X^M}{100}) \in [0,1] \\
&X^C \text{ is a one-hot encoding if } \begin{cases} x_i \in \{0,1\} & \forall i = 1,2,\ldots,n \\ \sum_{i=1}^{n} x_i = 1 \end{cases}
\end{aligned}
$$

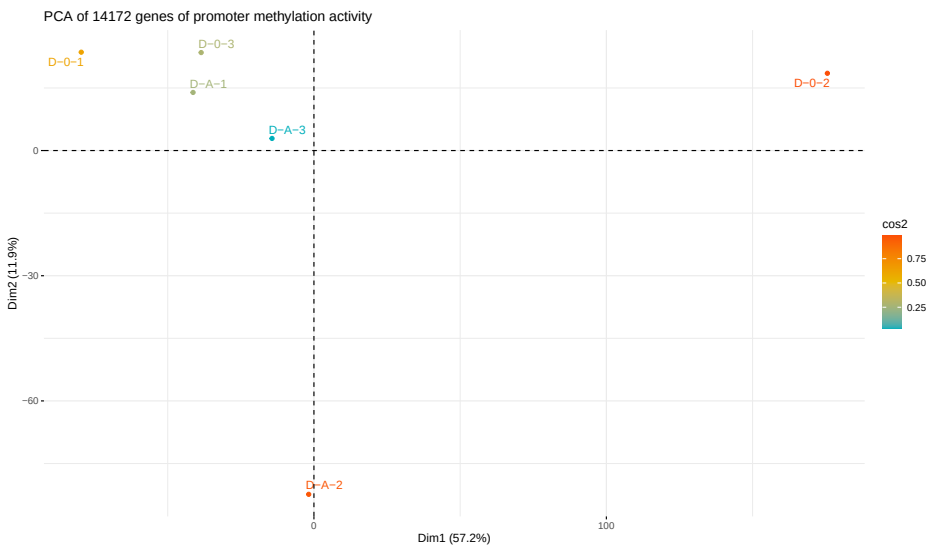


Figure 2: Principal Components (PCA) of samples

4.3.1 Model Assumptions

The hypergeometric mean of cpg methylation at promoter sites X^M was transformed from $[0-100]$ to $[0-1]$ and $1-X^M$. To generate consistent beta values whose linear combination with the X will approximate the mean/expectation of a distribution from which the responses are drawn from, i drew the β_1 from a uniform distribution from 1 to Ymax. This was informed by the distribution of the data(response distribution)6.Given that individuals or samples were modelled separately, the X^C becomes negligible. β_2 would have come from Bernoulli with parameter=0.5 or Dirichlet distribution with uniform priors. Furthermore, for simplicity the model assumes the ϵ term in the regression is negligible. In addition, the plate diagram for the model is shown below. 5

4.3.2 Inference and Model Evaluation

For Inference i used stochastic Variation inference in pyro. AutoDiagonalNormal was to generate the corresponding guide for my model. Adam was used as the optimizer and Trace_ELBO() was used to calculate the ELBO. Number of iterations=10000 and learning rate was 0.03 was used. The model did fairly well given that, most of the data points in the test set were within the posterior prediction distribution with 90% CI. As shown in Figure 7a, the posterior predictive distribution illustrates significant variability. Meanwhile, Figure 7b demonstrates the regression line with a 90% confidence interval. Also the distribution of the slopes/weights is positive centered around 8.804.

5 Further step

Subsequent steps would be to 1. Explain the variability between groups. for example what can we lean by pooling samples in a model. Further, I'll look into matrix factorization where we can learn about sample-genes interaction to be able to glean information from disease phenotypes like cancer. 2. Fit models for groups of genes rather than whole gene sets. Exploring mixed membership models or mixture models 3. Account for tissue/cell heterogeneity given that this is bulk long read sequencing data and not single cell. it is possible that the observation/methylation is being contributed by mixture of cells with different methylation status at each cpg site. Approaches such as rate of read disconcordance of methylations at cpgs sites

References

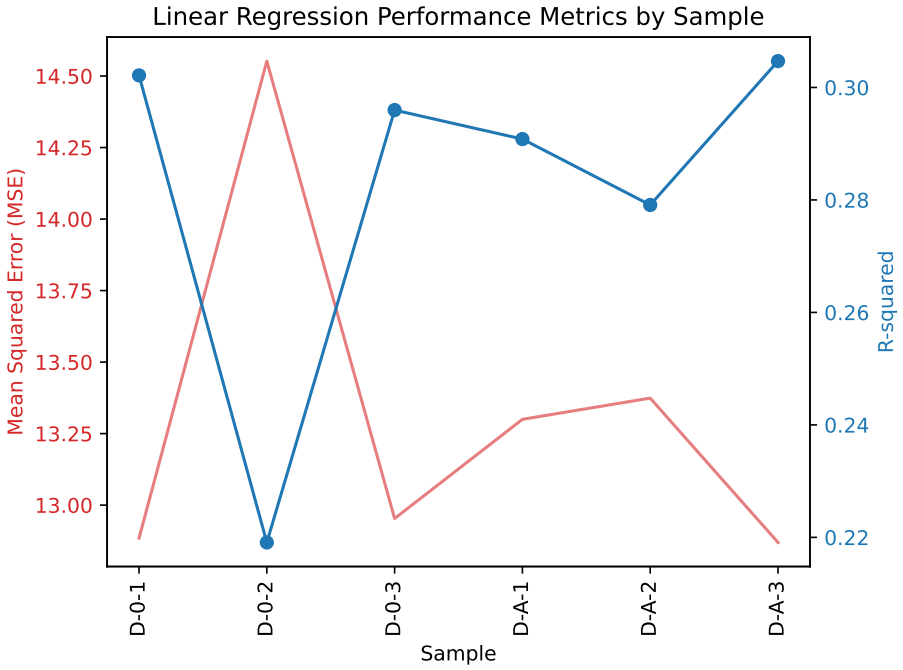


Figure 3: linear model performance.

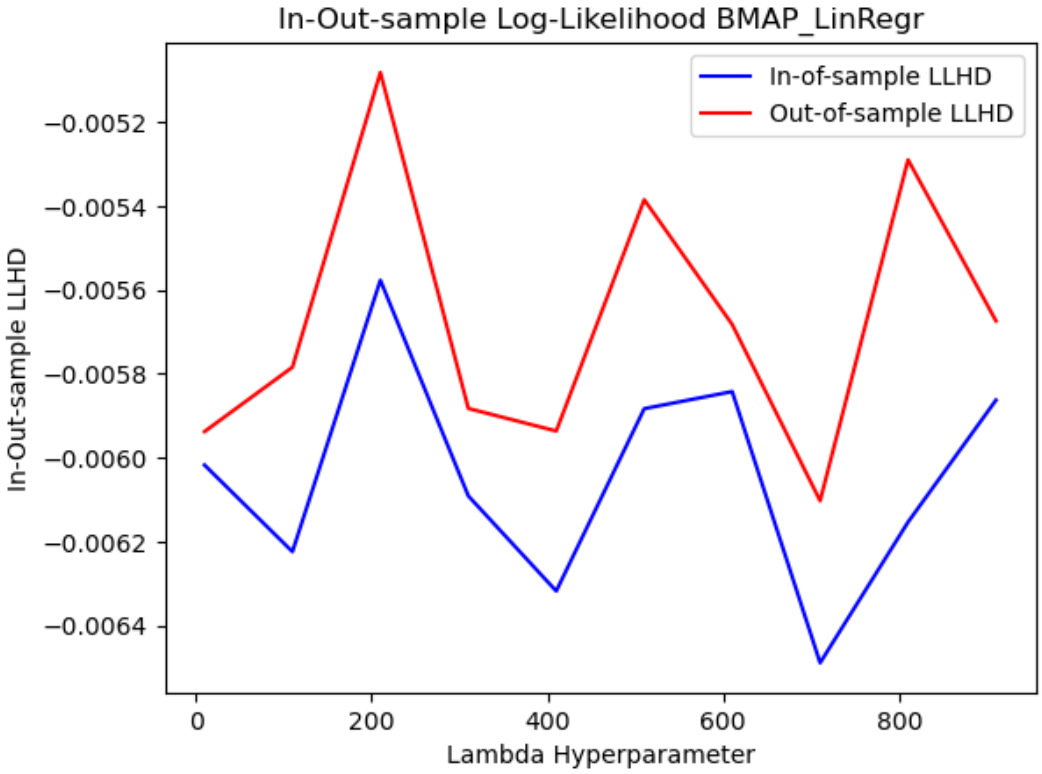


Figure 4: Comparison of out-of-samples log likelihood

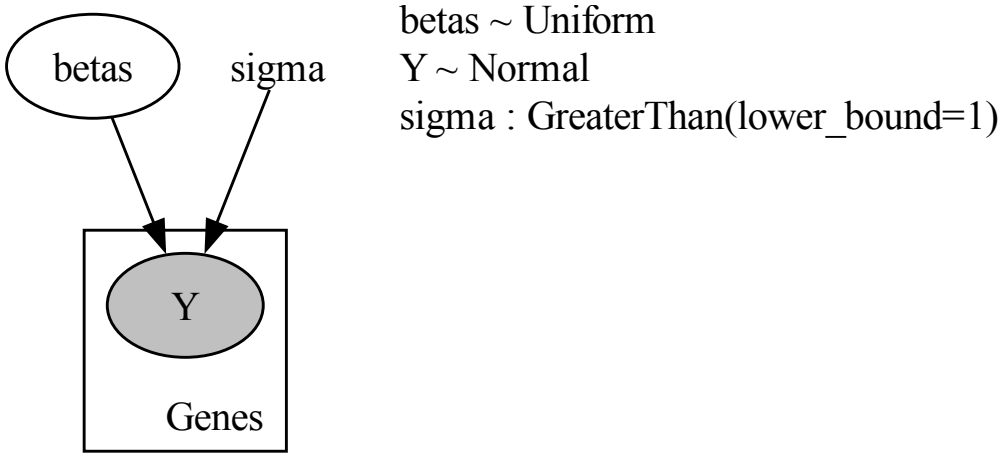


Figure 5: Final Model by Pyro

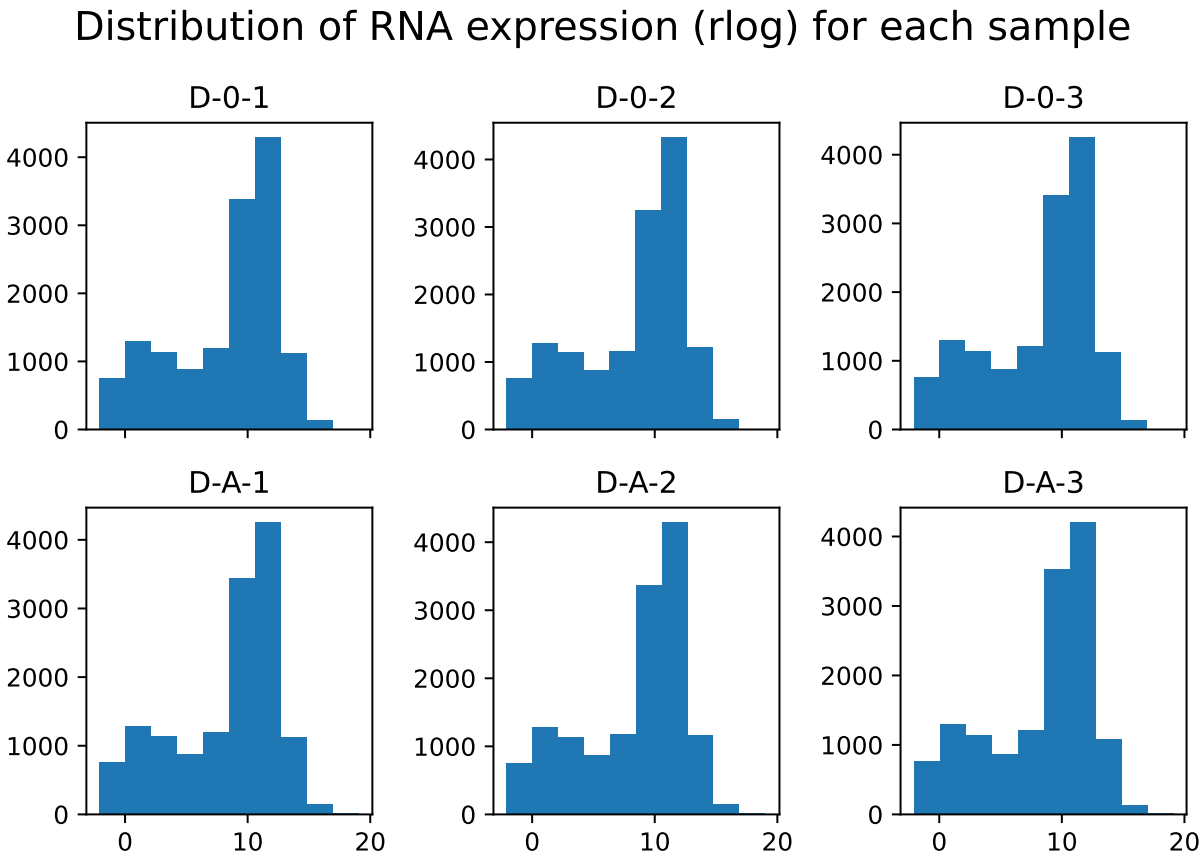
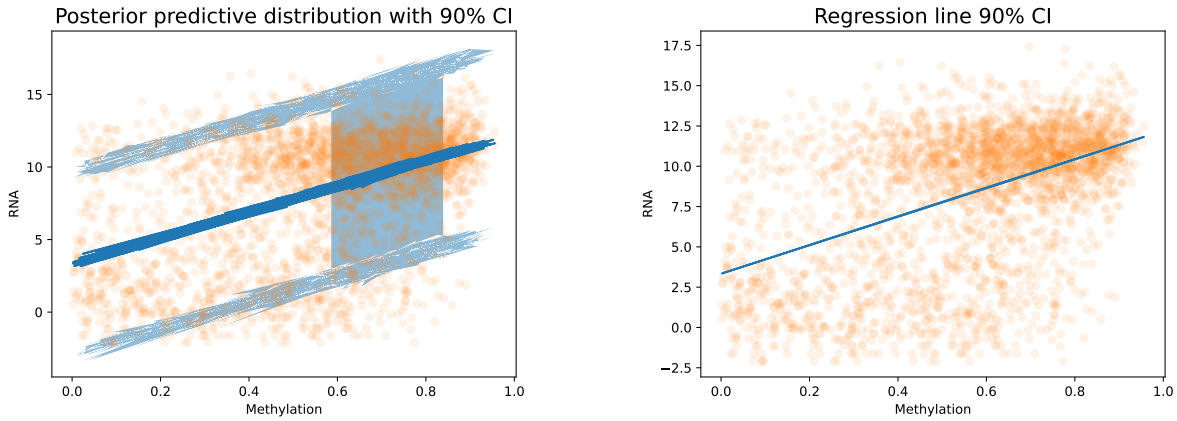


Figure 6: Distribution of Responses



(a) Posterior Predictive Distribution with 90% CI

(b) Regression Line with 90% CI

Figure 7: Comparative Analysis

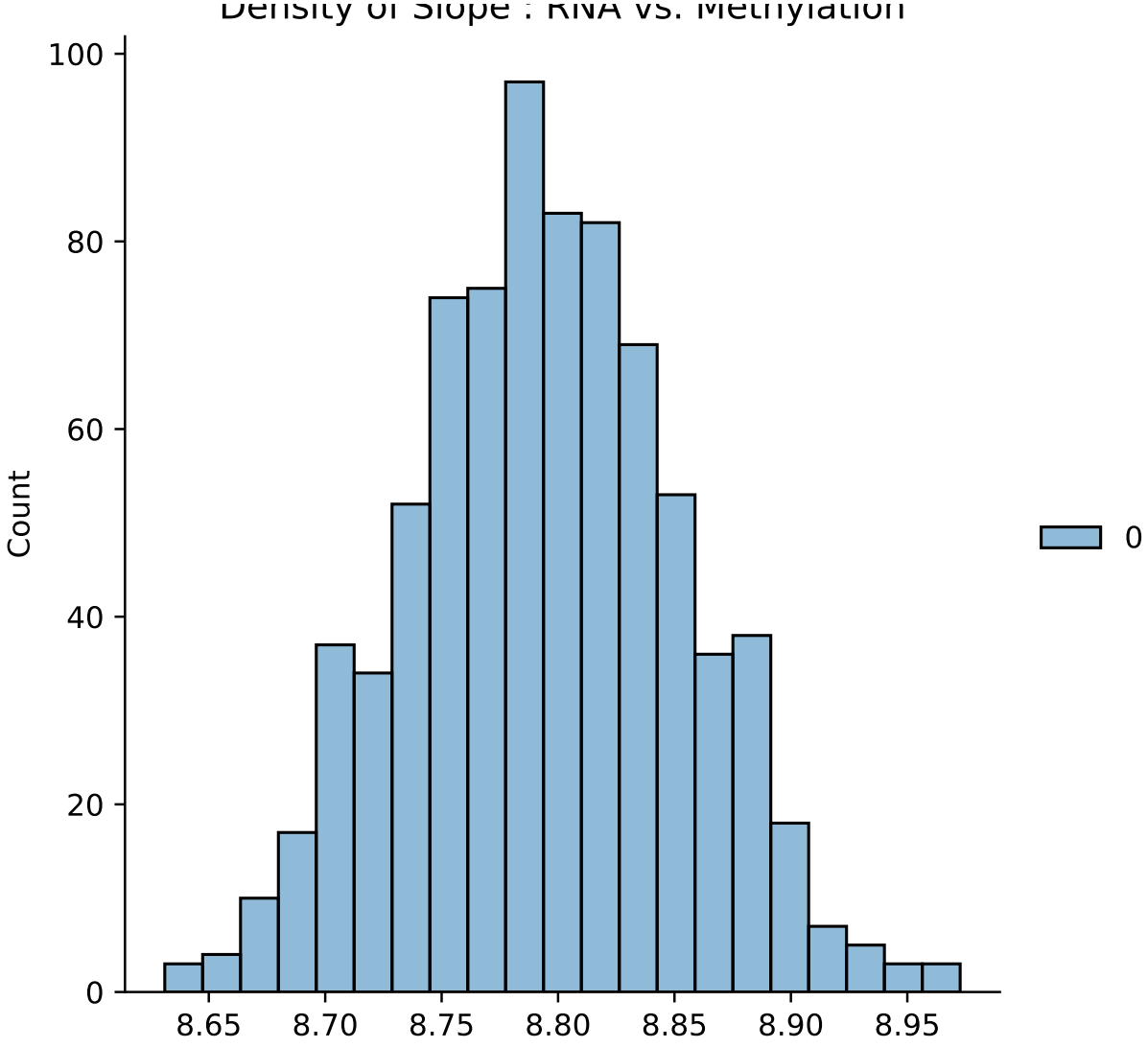


Figure 8: Distribution of Weights