Detection of gene expression from DNA methylation

Samuel Terkper Ahuno (sta4008)

November 13, 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 applied Stochastic gradient descent with maximum a posteriori probability (MAP) estimate to predict rna expression from DNA (methylation).

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 Description of dataset

For benchmarking, 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, basecalling was done with dorado software, reads were mapped to mouse reference genome(mm10) and methylation caling was done with modkit. I limited analysis to 5mC in CpGs regions only. Each sample's CpG site was supported by minimun 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

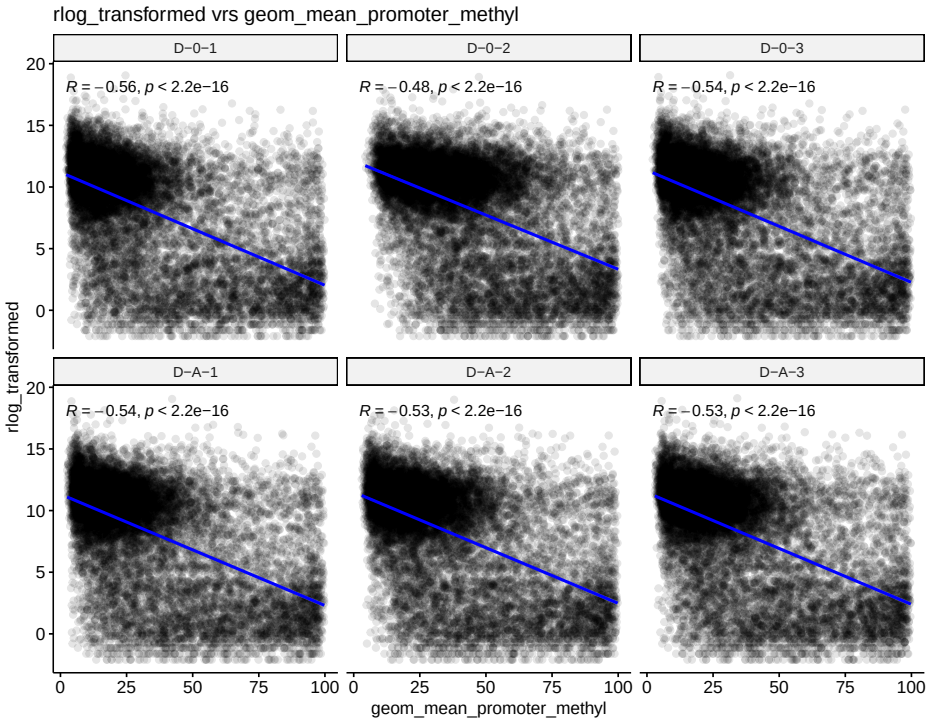

Figure 1: Correlation between gene expression and promoter methylation

5 Models

5.1 Variables and Split of dataset

Using normalized rna expression as response variable (Y) and promoter methylation as response variable (X), the dataset was split into train and test set 80% vrs 20%.

5.2 Model 1: linear Regression

Each individual sample was modelled separately to account for differences with samples gene expression and methylation using y - normalalized RNA expression and x-promoter methylation.

5.3 Model 2: BMAP with SGD

For the stochastic gradient descent. Number of bactches was 5, number of iterations was 100, tolerance was 1e-6, alpha to robsinson and monroe was $1e-3$ and prior beta variance (lambda) hyperparameter was set between 10 to 100.

6 Further step

Subsequent steps would be to 1. Explain the variability within groups. D-0-2 2. Fit models for groups of genes rather than whole gene sets. Exploring mixed membership models or mixture models 3. Account for sample heterogeneity (latent variable) given this is bulk data. Approaches such as rate of read disconcordance of methylations at cpgs sites.

References

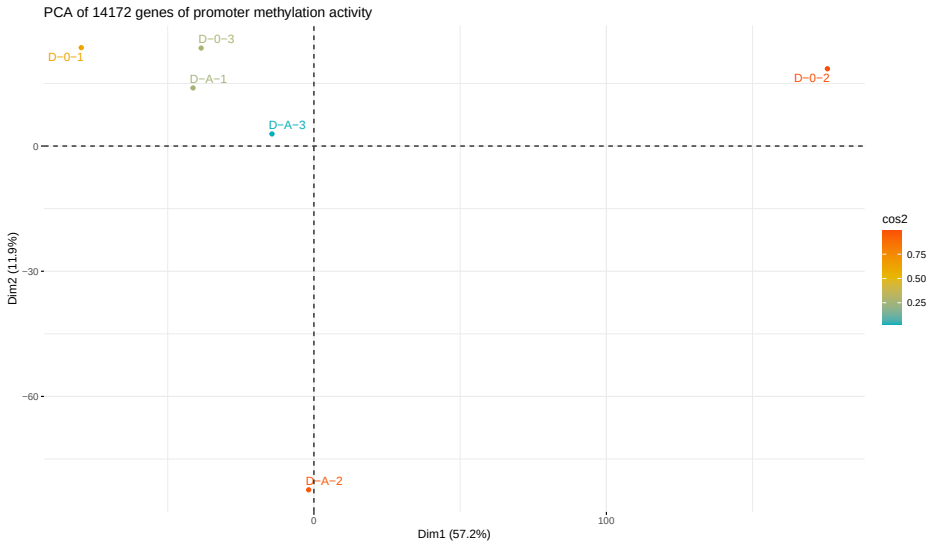

Figure 2: Principal Components (PCA) of samples

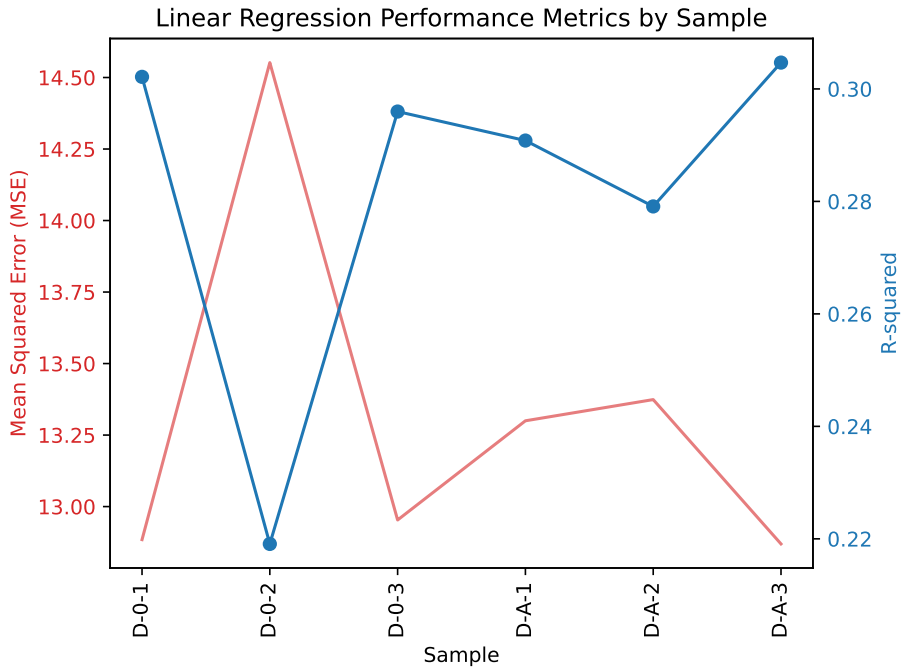

Figure 3: linear model performance.