Bringing Epigenetics to the Masses: Modbamtools for epigenetic analysis of nanopore data

Luke Morina[1], Roham Razaghi[1], Paul W. Hook[1], Ariel Gershman[2], Yuval Ebenstein, Jared Simpson[3], *Winston Timp*[1]

[1]Johns Hopkins University Department of Biomedical Engineering [2]Johns Hopkins University Department of Molecular Biology [3]Ontario Institute for Cancer Research

Abstract

Nanopore sequencing has enormous potential in epigenetic applications; unlike traditional sequencing-by-synthesis technologies, it can distinguish covalently modified nucleotides directly through their modulation of the electrolytic current. We can take advantage of the long read lengths (>100kb) generated by nanopore sequencing to precisely call methylation patterns and obtain phased methylation information across the genome. Using exogenous labeling methods, we can measure chromatin accessibility or even protein-DNA interactions.

We have begun to use long, single molecules as a proxy to interrogate the distribution of cellular states within a sample. We can even encode the cellular state into the DNA molecule itself via exogenous labeling, generating multiomic data. From ultralong sequencing reads, we can obtain methylation information across tens to hundreds of kilobases, identifying the cellular source of the read from distinct biomarker regions. From this point, we can deconvolve even complex mixtures of cells via their epigenetic state; we are now testing this approach by applying it to human blood samples.

To facilitate this, we have developed a toolset to work with ultralong methylation, modbamtools (https://rrazaghi.github.io/modbamtools/). Modbamtools allows us to manipulate and visualize DNA/RNA base modification data that have recently been added to the BAM file spec (MM and ML) so are stored directly. These tags have provided a better/efficient way for storing modification data inside alignment files.

modbamtools

Methods

Sample	GM12878	MCF10A	HCT116	Blood
Yield	163 Gb	196 Gb	137 Gb	136 Gb
N50	110 kb	110 kb	86 kb	106 kb
Avg. Depth	53X	65X	45X	45X

Table 1: Summary statistics from sequencing runs. Each column represents data from two PromethION flowcells.

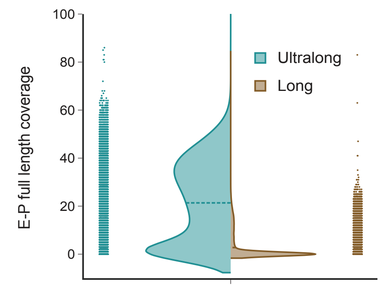


Figure 1: Distribution of sequencing coverage for reads covering both enhancers and promoters

Allele-specific epigenome interaction

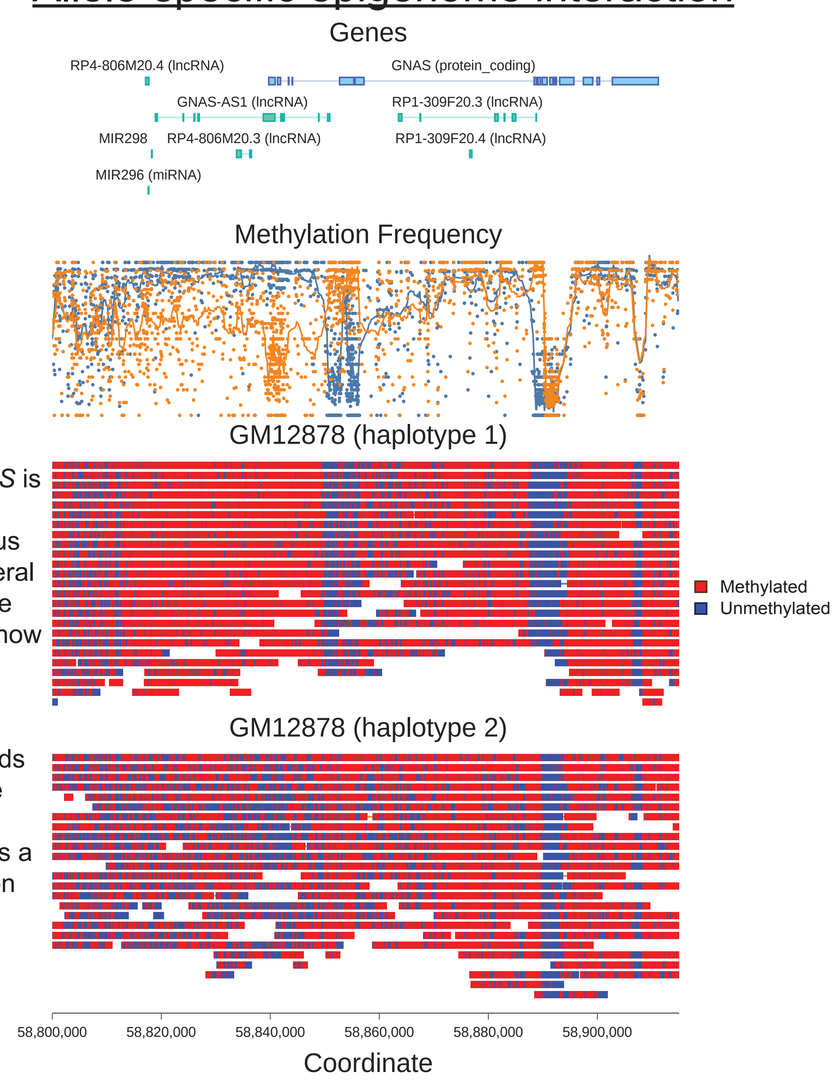
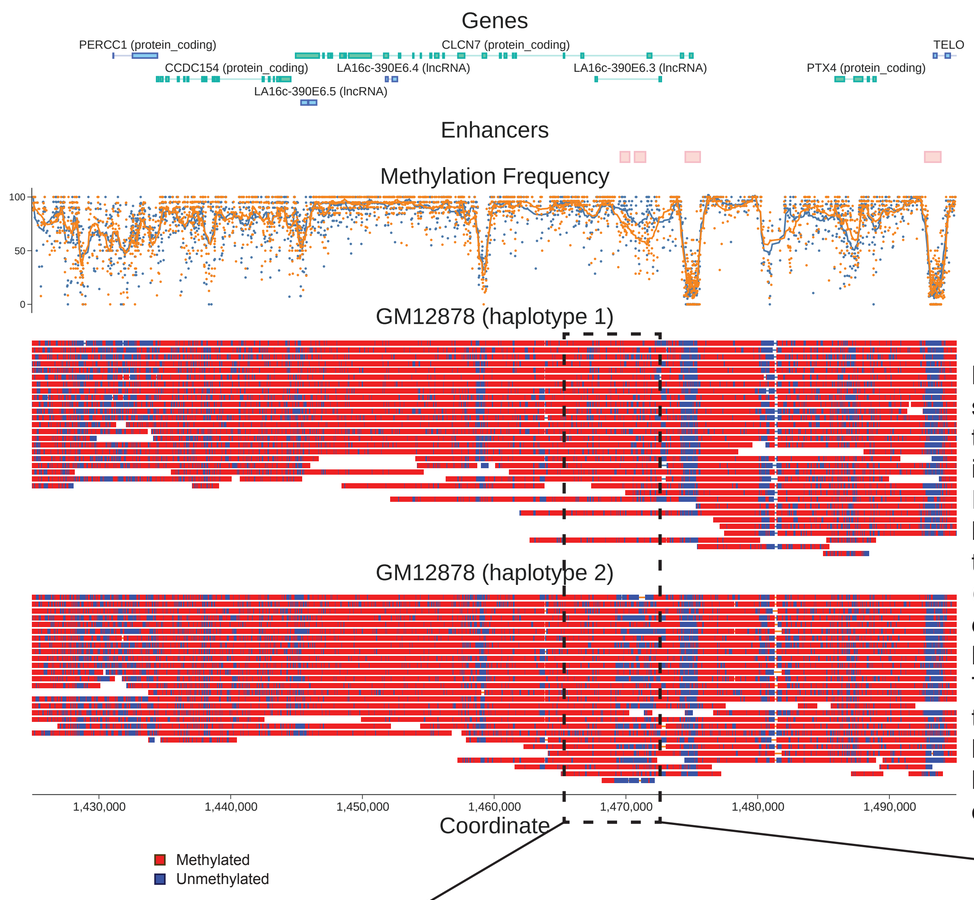


Figure 2: *GNAS* is a complex imprinted locus leading to several different gene products that show exclusive monoallelic expression. Ultra-long reads deconvolute methylation patterns across a ~140kb region

Clustering methylation heterogeneity in medically relevant genes



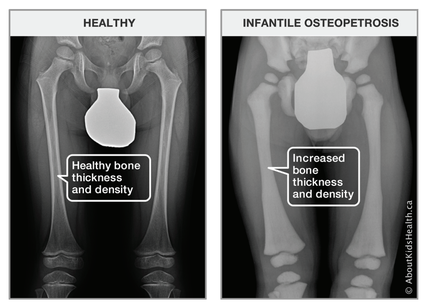
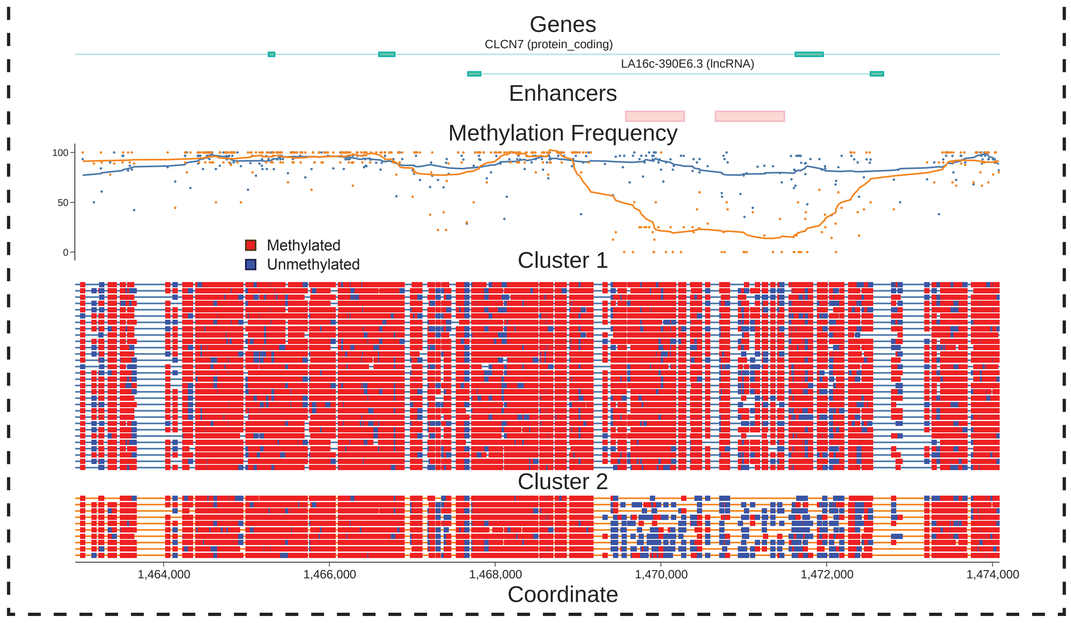


Figure 4: Chloride channel 7 alpha subunit also known as H+/Cl− exchange transporter 7 is a protein that in humans is encoded by the *CLCN7* gene.
In melanocytic cells this gene is regulated by the Microphthalmia-associated transcription factor. Mutations in the *CLCN7* are associated with autosomal dominant osteopetrosis type II, a rare bone disorder.
There is an intergenic enhancer region for this gene. While we observe hypermethylation on one of the haplotypes, haplotype 2 reveals a mixture of unmethylated/methylated reads.

Figure 5: Utilizing Hierarchal Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), we can separate reads purely based on methylation status. This feature has been implemented in modbamtools and can be used to quantitatively measure clusters in a heterogenous sample. Ultra-long reads are especially beneficial as they offer uniform co interactions.



Enhancer-promoter interaction on active/inactive alleles in chromosome X

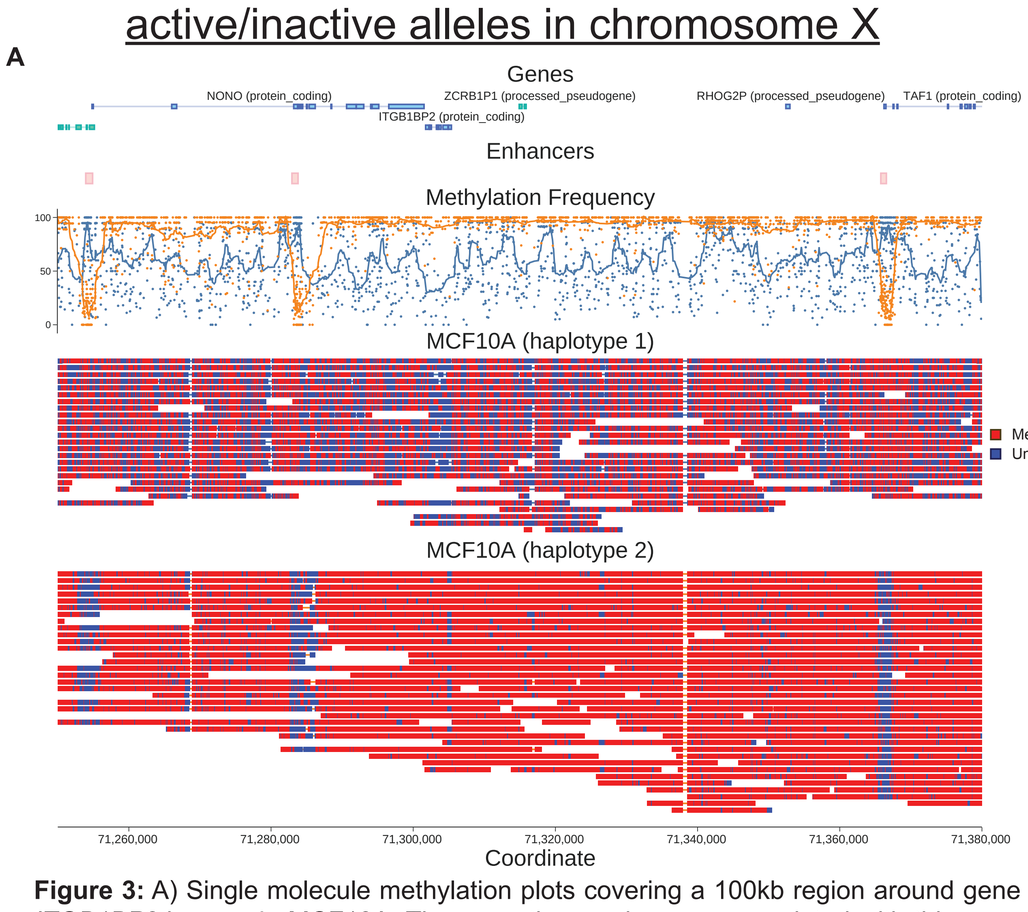


Figure 3: A) Single molecule methylation plots covering a 100kb region around gene *ITGB1BP2* in sample MCF10A. There are three enhancers associated with this gene (shown in pink recantgles). All enhancers are hypomethylated in haplotype 2 and hyper methylated in haplotype 1 suggesting haplotype 2 is the active allele B) Schematic of the proposed DNA looping in active/inactive alleles visualizing the enhancer-promoter interaction at this locus

Acknowledgements

This study was supported by grants from the NIH 5R01HG009190
.
We thank Miten Jain for helpful comments during the development of modbamtools.

We also thank Chris Wright for developing a fast C API compatible with HTSLIB MM and ML modification tag specifications.

NIH National Institutes of Health

TIMP LAB