Hidden Markov Models Characterize Consistency Between Methylation Signals

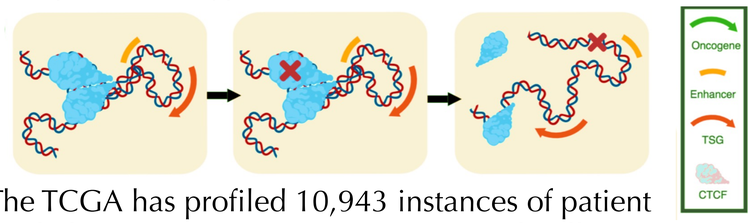
Judy Du[1] and **Mona Singh**[1,2]
[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University
jtdu@princeton.edu

Motivation

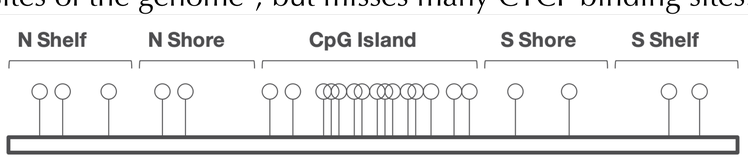
Human Methylation 450 (HM450) arrays have been extensively used to study population methylomics by many large consortia, such as The Cancer Genome Atlas (TCGA) and ENCODE. These arrays miss many CpG base steps of the genome, often at biologically and clinically relevant sites such as CTCF binding sites. Inference on these sites can leverage existing data to explore population trends. The goal of this project is to annotate methylation states and characterize the correspondence between nearest neighbor signals, leveraging the assumption that nearby CpG sites are more likely to have correlated signals.

Background and Data

Alongside the sequence of DNA a person is born with, epigenomic signals across the genome define the normal function of the genome, regulating which genes are turned on ("expressed") and in what proportion.
In particular, the methylation of CpG base steps is a chemical signal that sterically hinders the association of regulatory molecules to DNA. Recently, methylation at CTCF binding sites were shown to disrupt TADs[5], leading to cancer phenotypes[3]. Cases of oncogenic TAD disruptions have been experimentally verified in ovarian cancer[1], T-cell acute lymphoblastic leukemia[2], gliomas[3], and acute myeloid leukemia[4], but whole-genome effects multiple cancer types have not been well characterized.



The TCGA has profiled 10,943 instances of patient methylomes across 33 cancer types via the Illumina Human Methylation 450 Array. This experiment estimates the methylation fraction of CpG base steps across 485,577 sites of the genome[6], but misses many CTCF binding sites.



Hidden Markov Model

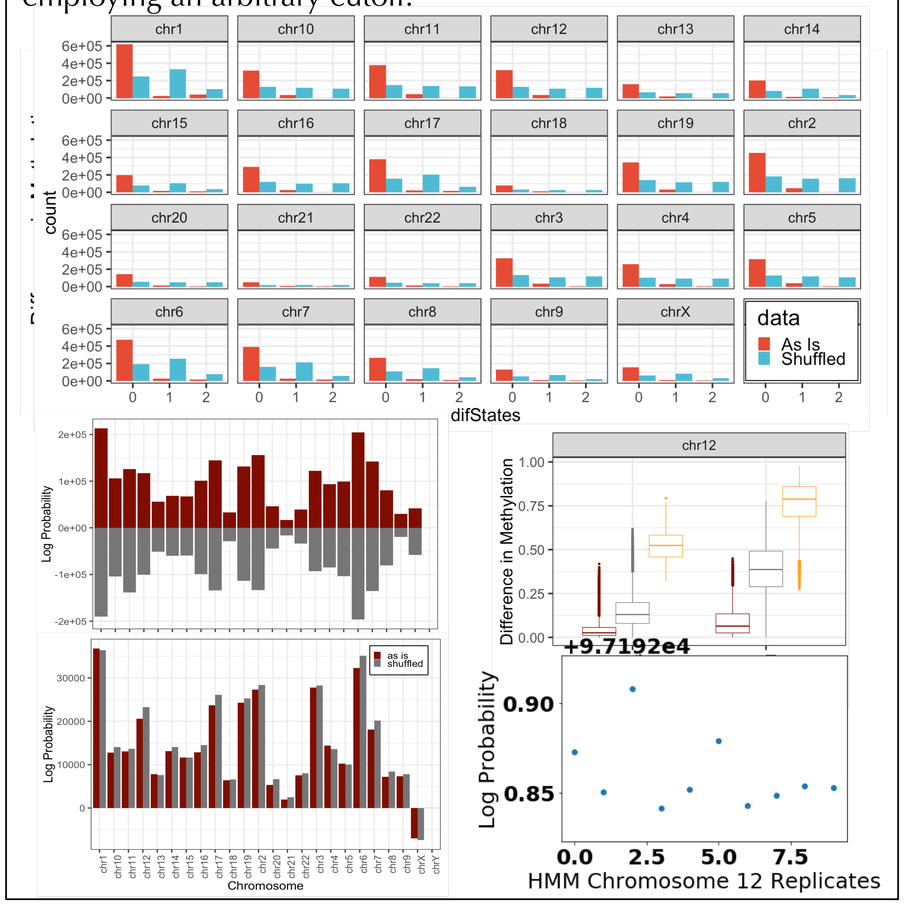
Here, we model colon adenocarcinoma (TCGA-COAD) methylation signals with a Hidden Markov Model with a generative Gaussian process. For the i-th chromosome locus where $i \in [L]$, let $x_i \in [0,1]$ denote the observed methylation fraction at that site and $z_i \in \{0,1,2\}$ denote the hidden methylation state. The joint probability is given by

$$p(x_0, \ldots x_L, z_0, \ldots, z_L) = p(z_0)p(x_0|z_0)\prod_{i=1}^{L} p(z_i|z_{i-1}))p(x_i|z_i)$$

With the observed methylation values $\boldsymbol{x}_i$ for a given chromosome, we use the Python hmmlearn libraries to infer the transition probabilities $\mathrm{p}(z_i|z_{i-1})$, Gaussian emission probabilities $p(x_i|z_i)$, and the most likely state at each locus i.

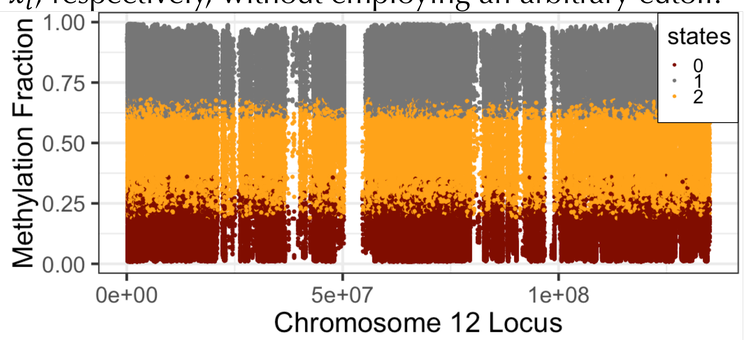
Model Performance and Consistency

Our model systematically annotates the three methylation states (methylated, hemimethylated, and unmethylated) without employing an arbitrary cutoff.



Hidden Markov Model Characteristics

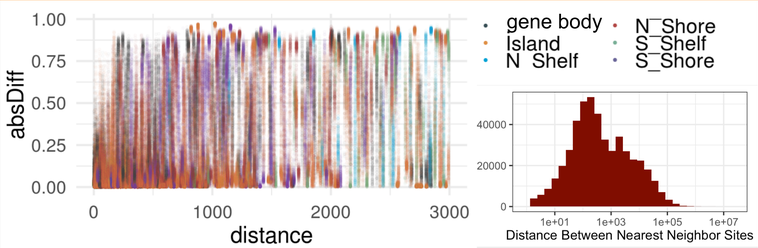
Our model systematically annotates the three methylation states (methylated, hemimethylated, and unmethylated) corresponding to high, medium and low x_i, respectively, without employing an arbitrary cutoff.



Moreover, the inferred transmission probabilities characterize the imputability of chromosome loci given the nearest neighbor methylation values, which can and have been used for imputation.

States	0	1	2
0	0.828 (1.1e-6)	0.073 (1.1e-6)	0.099 (8.5e-8)
1	0.154 (2.3e-6)	0.479 (6.2e-7)	0.367 (2.9e-6)
2	0.110 (1.5e-6)	0.178 (7.1e-7)	0.713 (2.1e-6)

Distance Between Probes



References

[1] Ji et al. 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell, 18(2):262–275, 2016
[2] Hnisz et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science, 351(6280):1454–1458, 2016.
[3] Flavahan et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nat Publ. Gr. 529(7584):110–114, 2016.
[4] Groschel et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. Cell, 157(2):369–381, 2014.
[5] Rao et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell,159(7):1665–1680, 2014.
[6] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using- tcga/technology/Illumina-HumanMethylation450-Data-Sheet

This work was funded in part by the NIH grant #5T32HG003284-13. The analyses shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. I'd like to thank my advisor, Mona Singh and the lab.