Genome-scale methyltyping by statistical inference for determining methylation states in populations

Meromit Singer[1], Dario Boffelli [2], Joseph Dhahbi[2], Alexander Schoenhuth[3,4], Gary P. Schroth[5], David I.K. Martin[2] and Lior Pachter[3,4]

Departments of [1] Computer Science [3] Mathematics and [4] Molecular and Cell Biology University of California, Berkeley, CA. [2] Children's Hospital Oakland Research Institute, Oakland, CA. [5] Illumina Inc., Hayward, CA
Email: meromit@cs.berkeley.edu

The ability to assay genome-scale methylation patterns using high-throughput sequencing makes it possible to carry out association studies to determine the relationship between epigenetic variation and phenotype. While whole-genome bisulfite sequencing can determine a methylome at high resolution, cost inhibits its use in comparative and population studies. MethylSeq, based on sequencing of fragment ends produced by a methylation-sensitive restriction enzyme (HpaII), is a method for methyltyping (survey of methylation states) and is a site-specific and cost-effective alternative to whole-genome bisulfite sequencing. Despite its advantages, data generated by MethylSeq has multiple biases. A notable bias involves the dependence of the signal intensity at a site on the local geometric structure of the other restriction sites and their methylation status.

We introduce a statistical method, MetMap, that produces corrected site-specific methylation states from MethylSeq experiments and annotates unmethylated islands across the genome, along with scores determining the islands' strength. MetMap integrates the experimental signal with the local geometric structure of restriction sites and prior biological hypotheses in a statistically sound and cohesive Bayesian Network, and achieves an accurate mapping of genomewide methylation that serves as a framework for comparative methylation analysis. We validated MetMap's inferences with direct bisulfite sequencing, showing that the methylation status of both single sites and annotated unmethylated islands is accurately inferred.

We demonstrate the use of MethylSeq with MetMap by methyltyping neutrophil samples from four male human individuals, and annotating their unmethylated islands. We show that such analysis gives significant insight into the methylome of each specimen, inside and outside of CpG islands, at site specific resolution. MetMap identified 3,767 novel unmethylated islands that are invisible to various sequence-based annotation strategies, and are associated with other features indicative of transcriptional function, such as open chromatin regions and transcription start sites. We conclude that MetMap leverages the cost-efficiency and practical ease of MethylSeq to produce informative genome-scale methylation annotations (methyltypes) that are suitable for both region- and site-specific comparative and case-control studies.