

Revealing Genomewide Methylation Patterns from High-Throughput Sequencing by Statistical Inference

Meromit Singer¹, Dario Boffelli², Joseph Dhahbi², David I.K. Martin² and Lior Pachter^{1,3,†}

1 Introduction

The availability of high-throughput short-read sequencing coupled with fast read mapping to a reference genome allows for the rapid identification of fragments from a restriction digest. In epigenetics, this can be leveraged to obtain site-specific methylation information in a fast and cost effective manner, allowing us to adapt comparative genomic methods to epigenomic applications. However, data generated by this approach has multiple biases. When conducting comparison studies a notable bias involves the dependence of the signal intensity at a site on the local geometric structure of restriction sites and their methylation status. We present here MetMap, an inference method for site-specific methylation status and CpG Island regions prediction. MetMap integrates the experimental signal at each site with the local geometric structure of restriction sites and prior biological hypotheses to achieve an accurate mapping of genomewide methylation. We used high throughput sequencing of the ends of fragments generated by methylation-specific restriction enzymes together with MetMap to determine the genomewide methylation maps of four humans and four chimpanzees. We present results regarding the inter-individual and inter-species variation in whole-genome methylation, and suggest a method for such comparative studies.

2 MetMap Inference Method

Though there are great advantages in conducting high-throughput sequencing experiments with methylation-sensitive restriction enzymes[1], a major bias is introduced by the dependence of the signal at each restriction site on the methylation status and geometric structure of its neighboring sites. We present MetMap - a statistical framework, based on Conditional Random Fields [2], which interprets such high throughput methylation data and infers two important characteristics with respect to DNA methylation. MetMap estimates, at all restriction sites within the scope of the experiment, the proportion of alleles in which each site was methylated, as well as the probability that each site is located within a CpG island. As seen in Figure 1, MetMap exploits in the inference process both the short-read experimental data and the geometric structure of the CpG sites on the genome, and may be used on either paired-end or non paired-end data-sets. The potential functions of MetMap incorporate the distances between CpG sites in the genome, the probability for sites to be methylated/unmethylated inside and out of CpG islands, and the probability of observing a restriction fragment given the methylation configuration of the restriction sites it holds. Fortunately the inference is tractable, as the cliques of the graph are small due to the inherent upper bound on the length of restriction fragments that will be sequenced.

3 Results

We have used MetMap to determine the methylomes of a single, homogeneous and uncultured cell type, the neutrophil, in four humans and four chimpanzees. We determined the true methylation status of 60 different sites using direct bisulphite sequencing and show MetMaps inferences are better correlated with the true methylation states than raw experimental signals. Furthermore, we present specific examples for regions that were predicted correctly by MetMap and falsely by the raw signal with respect to their methylation states (Figure 2), as well as regions for which the raw signal implies false differences between the human and chimp methylomes, which are normalized by MetMap. Our results highlight the weaknesses in the direct inference of methylation status from

¹Department of Computer Science and ³Department of Mathematics, University of California, Berkeley, CA.

²Childrens Hospital Oakland Research Institute, Oakland, CA

[†]Corresponding author. email: lpachter.math.berkeley.edu

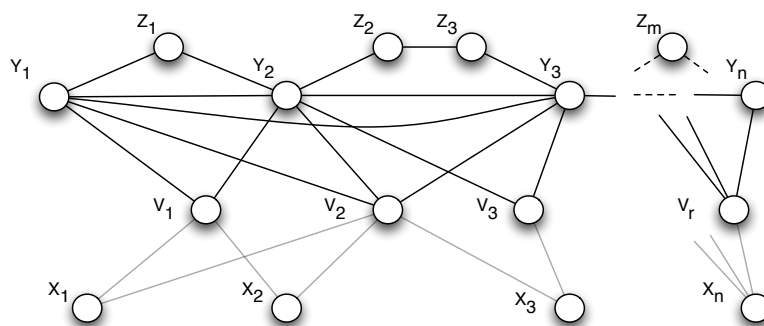


Figure 1: The graphical structure of MetMap. The hidden states of the model involve the Y nodes, representing restriction sites, and the Z nodes, representing the rest of the CpG sites. These nodes indicate the methylation status and the probability of the site to be in a CpG Island. The V nodes represent fragments that may appear in the data, and are observed in the case of paired-end sequencing. The X nodes represent read counts for specific restriction sites, and are observed in the case of non paired-end sequencing.

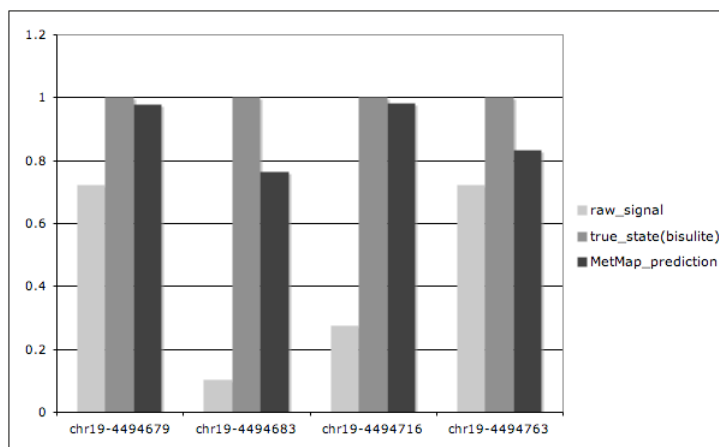


Figure 2: Predictions of the proportion of alleles that were unmethylated at four consecutive restriction sites on the human genome. MetMap's predictions (dark gray) are seen closer to the bisulfite results (gray) than the normalized read counts (light gray).

sequencing of fragment ends, and show that integration of the local geometry of the restriction sites along with the experimental sequencing data and prior biological hypotheses achieve an accurate mapping of genomewide methylation. We present results concerning the extent of variation in DNA methylation between the human and chimpanzee species, as well as within each species, and suggest efficient comparative methods for such studies.

References

- [1] Meissner, A., et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 2008. 454(7205): p. 766-771.
- [2] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of ICML, 2001*