Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet

Coby Viner[1, 2], James Johnson[3], Nicolas Walker[4], Hui Shi[4], Marcela Sjöberg[5], David J. Adams[5], Anne C. Ferguson-Smith[4], Timothy L. Bailey[3], and Michael M. Hoffman[1, 2, 6, *]

[1]*Department of Computer Science, University of Toronto, Toronto, ON, Canada*
[2]*Princess Margaret Cancer Centre, Toronto, ON, Canada*
[3]*Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia*
[4]*Department of Genetics, University of Cambridge, Cambridge, England*
[5]*Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, England*
[6]*Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada*
[*]Correspondence: michael.hoffman@utoronto.ca

Motivation. Many transcription factors (TFs) initiate transcription only in specific sequence contexts, providing the means for sequence specificity of transcriptional control. A four-letter DNA alphabet only partially describes the possible diversity of nucleobases a TF might encounter. Cytosine is often present in the modified forms: 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC). TFs have been shown to distinguish unmodified from modified bases. Modification-sensitive TFs provide a mechanism by which widespread changes in DNA methylation and hydroxymethylation can dramatically shift active gene expression programs.

Methods. To understand the effect of modified nucleobases on gene regulation, we developed methods to discover motifs and identify TF binding sites (TFBSs) in DNA with covalent modifications. Our models expand the standard `A/C/G/T` alphabet, adding `m` (5mC) and `h` (5hmC). We additionally add symbols to encode guanine complementary to these modified cytosine nucleobases and represent states of ambiguous modification. We adapted the position weight matrix model of TFBS affinity to an expanded alphabet. We developed a program, Cytomod, to create a modified sequence. We also enhanced the MEME Suite to be able to handle custom alphabets. We created an expanded-alphabet sequence using whole-genome maps of 5mC and 5hmC in naive ex vivo mouse T cells.

Results. Using this sequence and ChIP-seq data from Mouse ENCODE and others, we identified modification-sensitive cis-regulatory modules. We elucidated various known methylation binding preferences, including the preference of ZFP57 and C/EBPβ for methylated motifs and the preference of c-Myc for unmethylated E-box motifs. We demonstrated that our method is robust to parameter perturbations, with TF sensitivities for methylated and hydroxymethylated DNA broadly conserved across a range of modified base calling thresholds. Hypothesis testing across different threshold values was used to determine cutoffs most suitable for further analyses. Using these known binding preferences to tune model parameters enables discovery of novel modified motifs.

Discussion. Hypothesis testing of motif central enrichment provides a natural means of differentially assessing modified versus unmodified binding affinity. This approach can be readily extended to other DNA modifications. As more high-resolution epigenomic data becomes available, we expect this method to continue to yield insights into altered TFBS affinities across a variety of modifications.