# EPIGENOMICS

A. J. HARTEMINK

*Department of Computer Science, Box 90129*
*Duke University*
*Durham, NC 27708-0129, USA*
*Email: amink@cs.duke.edu*

M. KELLIS

*Computer Science and Artificial Intelligence Laboratory*
*Broad Institute of MIT and Harvard*
*Stata Center - 32D.524 Cambridge, MA 02142, USA*
*E-mail: manoli@mit.edu*

W. S. NOBLE

*Department of Genome Sciences, Box 355065*
*University of Washington*
*Seattle, WA 98109, USA*
*E-mail: william-noble@uw.edu*

Z. WENG

*Biochemistry & Molecular Pharmacology*
*364 Plantation Street, LRB*
*University of Massachusetts Medical School*
*Worcester, MA 01605, USA*
*E-mail: Zhiping.Weng@umassmed.edu*

Epigenomics involves the global study of mechanisms, such as histone modifications or DNA methylation, that have an impact on development or phenotype, are heritable, but are not directly encoded in the DNA sequence. The recent availability of large epigenomic data sets, coupled with the increasing recognition of the importance of epigenetic phenomena, has spurred a growing interest in computational methods for interpreting the epigenome.

*Keywords*: Epigenomics, histone modifications, chromatin, DNA methylation

Scientists have known for a long time that the sequence of nucleotides that comprise the genome is not sufficient to explain the heritability of traits from one generation to the next, nor is that sequence sufficient to drive the myriad functions of a living cell. Recently, however, catalyzed by the rapid acquisition of a wide variety of genome-scale data sets from projects such as ENCODE,[1] modENCODE,[2] and Roadmap Epigenomics,[?] scientists have begun to characterize just how much information is encoded beyond the primary DNA sequence. Accordingly, many of the central questions facing biology today concern the interpretation and integration of epigenomic data with our existing knowledge of the molecular pathways within the cell, including DNA, RNA, proteins, and metabolites. This session includes three papers, each of which describes a novel computational method for the analysis and interpretation of one or more types of epigenomic data.

The first paper analyzes a single type of data, derived from a DNase 1 sensitivity assay. The endonuclease DNase 1 has long been known to preferentially cleave in short regions of open chromatin, known as DNase 1 hypersensitive sites.[3] Such regions are of great interest because they correspond to various types of regulatory elements, including promoters, enhancers, insulators and boundary elements. Recently, a series of DNase 1-based assays have been described for ascertaining the cleavage profile of DNase 1 across the entire genome. Originally based on quantitative PCR[4] and microarrays,[5,6] these assays were quickly adapted for next-generation sequencing platforms.[7,8] Importantly, in addition to recognizing classical hypersensitive sites, which have a typical size of 225–250 bp, subsequent work demonstrated that a detailed DNase 1 cleavage profile could localize protein-binding events at basepair resolution.[9,10]

Given the importance of transcription factor binding for gene regulation, and given the increasing availability of DNase 1 data for a wide variety of human cell types, a variety of computational methods have been developed to interpret DNase 1 sensitivity data. Luo and Hartemink contribute to this literature by introducing a method, called Millipede, that aims to identify transcription factor binding events on the basis of DNase 1 sensitivity data as well as analysis of the primary sequence. Millipede improves upon the previously described Centipede algorithm[11] by reducing the number of parameters and switching from unsupervised to supervised learning. Luo and Hartemink benchmark Millipede using data from human and yeast.

The second paper, by Sahu et al., proposes a machine learning approach to enhancer detection. An enhancer is a gene regulatory element that is responsible for upregulating one or more genes. Enhancers are notoriously difficult to detect because they often do not occur proximal to their target gene, relying instead upon DNA looping or other complex chromatin structures to carry out their regulatory effect. No single high-throughput assay can be used to identify the "enhancerome" because different types of enhancers presumably rely upon different regulatory mechanisms. The gold standard method for identifying an enhancer involves knocking it out and observing the resulting downregulation of the target gene. This approach, obviously, does not scale to whole-genome analysis. Currently, closest proxy we have for genome-wide enhancer detection is ChIP-seq for the DNA-binding protein p300. Although almost all p300 binding sites are enhancers, many known enhancers are not bound by p300.

This lack of a high-quality and high-throughput enhancer assay has led to the development of a series of computational methods that aim to identify putative enhancers.[12–15] Sahu et al. contribute to this ongoing project by introducing a support vector machine classifier that learns to identify enhancers on the basis of ChIP-seq histone modification and DNase 1 sensitivity data. They demonstrate that, not only does their classifier perform well in cross-validation, but it also can be used to identify putative enhancers associated with SNPs from genome-wide association studies of cardiac phenotypes.

Finally, the paper by Ahn and Wang describes a statistical testing methodology for identifying genomic regions in which patterns of variability in DNA methylation across individuals may be indicative of disease. DNA methylation involves the addition of a methyl group either to an adenine or (most commonly in animals) a cytosine. Methylation is used extensively by

the cell to shut off expression of individual genes or large chromosomal regions, and plays a critical role in regulating cellular processes such as embryonic development, X chromosome inactivation, genomic imprinting and chromosome stability.[16] Methylated cytosines can be identified by first subjecting the DNA to bisulfite conversion, which changes cytosine residues to uracil unless the cytosines are methylated, and then sequencing the converted DNA. The result, by comparison to a reference genome, is a map of the frequency of methylation at each cytosine residue. Methylation is associated with a set of heritable syndromes—imprinting disorders—that result from asymmetric expression of the alleles of one or more genes, as well as with a variety of repeat-instability diseases.[16] More recently, aberrant methylation has been increasingly implicated in various types of cancer.[17]

The primary goals of Ahn and Wang's work is to improve our ability to detect patterns of aberrant methylation that are potentially associated with a given disease. Their proposed statistical framework draws upon the observation that such loci differ not only in the mean level of methylation but also its variance. Accordingly, Ahn and Wang propose a regression-based testing framework that captures more features of the methylation profile of a given locus and, in so doing, boosts statistical power relative to approaches based only on the mean.

The topics covered by these three papers are quite diverse, reflecting the wide range of challenging computational and statistical problems posed by epigenomic data.

## References


1. ENCODE Project Consortium, *Nature* **489**, 57 (2012).
2. T. modENCODE Consortium, *Science* **330**, 1775 (2010).
3. C. Wu, *Nature* **286**, 854 (1980).
4. P. J. Sabo, R. Humbert, M. Hawrylycz, J. C. Wallace, M. O. Dorschner, M. McArthur and J. A. Stamatoyannopoulos, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4537 (2004).
5. P. J. Sabo, M. S. Kuehn, R. Thurman, C. Grant, B. Johnson, S. Johnson, H. Kao, M. Yu, J. Goldy, M. Weaver, M. A. Singer, T. Richmond, M. Dorschner, P. Navas, R. Green, W. S. Noble and J. A. Stamatoyannopoulos., *Nature Methods* **3**, 511 (2006).
6. G. E. Crawford, S. David, P. C. Scacheri, G. Renaud, M. J. Halawi, M. R. Erdos, R. Green, P. S. Meltzer, T. G. Wolfsberg and F. S. Collins, *Nature Methods* **3**, 503 (2006).
7. P. J. Sabo, M. Hawrylycz, J. C. Wallace, R. Humbert, M. Yu, A. Shafer, J. Kawamoto, R. Hall, J. Mack, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 16837 (2004).
8. A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey and G. E. Crawford, *Cell* **132**, 311 (Jan 2008).
9. J. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields and J. A. Stamatoyannopoulos, *Nature Methods* **6**, 283 (2009).
10. A. P. Boyle, L. Song, B. K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford and T. S. Furey, *Genome Research* **21**, 456 (2011).
11. R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad and J. K. Pritchard, *Genome Research* **21**, 447 (2011).
12. D. Lee, R. Karchin and M. A. Beer, *Genome Research* **21**, 2167 (2011).
13. M. Fernandez and D. Miranda-Saavedra, *Nucleic Acids Research* **40**, p. e77 (2012).
14. D. May, M. J. Blow, T. Kaplan, D. J. McCulley, B. C. Jense, J. A. Akiyama, A. Holt, I. Plajzer-

Frick, M. Shoukry, C. Wright, V. Afzal, P. C. Simpson, E. M. Rubin, B. L. Black, J. Bristow, L. E. Pennacchio and A. Visel, *Nature Genetics* **44**, 89 (2011).
15. L. Narlikar, N. J. Sakabe, A. A. Blanski, F. E. Arimura, J. M. Westlund, M. A. Nobrega and I. Ovcharenko, *Genome Research* **20**, 381 (2010).
16. K. D. Robertson, *Nature Reviews Genetics* **6**, 597 (2005).
17. M. A. Dawson and T. Kouzarides, *Cell* **150**, 12 (2012).