Interplay between DNE sequence motifs and the human epigenome

John William Whitaker[1], Zhao Chen[2], Wei Wang[2]

[1]UCSD, San Diego, United States
[2]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, United States

Motivation and objectives

Different mammalian cell-types display distinct phenotypes but posses the same genome. It is well established that epigenomic modification is important in establishing cell-type specific patterns of gene expression. Epigenomic modifications include the covalent modification of histone tails and the methylation of DNA. Epigenomic modifications function to mark regions of the genome as being active or repressed and their correct establishment is a critical aspect of mammalian development. Furthermore, correct recapitulation of the epigenome is key during cellular reprogramming, such as induced pluripotent stem cells (Lister *et al.* 2011; Won *et al.* 2012). Moreover, alterations in the epigenome are associated with disease such as cancers (Hon *et al.* 2012) and autoimmune diseases (Nakano *et al.* 2012).

The establishment and maintenance of the epigenome is regulated by many factors including: modifying enzymes, DNA binding proteins, non-coding RNAs, signaling molecules and three dimensional genomic organization. Herein, we investigate the involvement of DNA sequence motifs in the regulation of the epigenome. We demonstrate the involvement of DNA sequence motifs by constructing a series of predictive models that can predict the presence of six histone modifications in five different developmental cell-types, including human embryonic stem cells. Epigenomic modifications can span long variably length regions that are associated with GC content biases. Thus, great care was taken to avoid biases from influencing predictive performance.

Methods

We analyzed a comprehensive dataset of six core histone modifications (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3) in five primitive cell-types (human embryonic stem cells (H1), trophoblast-like, neural progenitor cell, mesendoderm and mesenchymal cells). To identify a broad set of DNA motifs that are associated with epigenome we used two *de novo* motif discovery programs: Homer (Heinz *et al.* 2010), and our own, Epigram. Then a LASSO logistic regression was use to identify the subset of motifs that had the greatest prediction performance (Friedman *et al.* 2010). Then a Random forest was trained to distinguish genomic regions that posses a modification from regions that do not.

Results and discussion

An integrative analysis of over 70 separate ChIP-Seq experiments shall be presented. The analysis pipeline first used to distinguish genomic regions that possess a modification from regions that do no possess any modifications. In H1 the average prediction performance across the six modifications was AUC = 0.85 (Figure 1). Further comparisons, identified that prediction performance was constituent in other cell-types and that models could be trained to distinguish a specified modifications from other modifications.

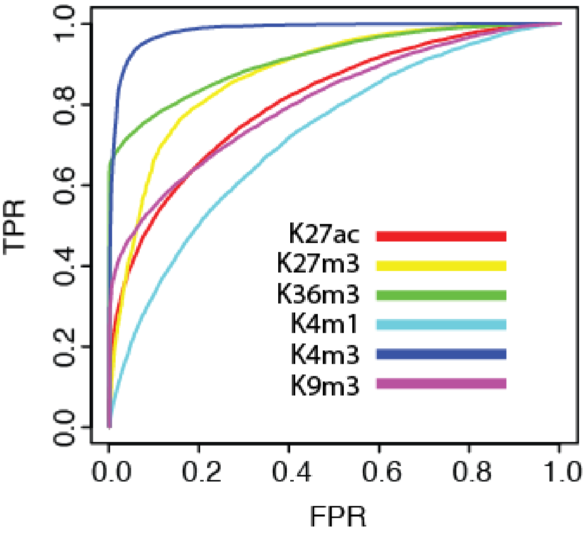

Figure 1. The prediction performance in of six histone modifications in human embryonic stem cells

Comparison of the identified motifs to known motifs identified interplay between known factors and the epigenome. For example, cell-type specific marker genes are identified as being associated with H3K27ac, which marks active regions of the genome. Motif location preference analysis revealed that motifs occur at the edge of modification peaks and suggests they may function by establishing barriers.

The identified DNA motif and epigenome associations demonstrate interplay between sequence and epigenome. This work demonstrates the importance of large-scale integrative genomic analysis to gain complex biological insight. Identification of factors that interplay with epigenome in stem cells should improve the efficacy of cellular reprogramming strategies.

Acknowledgements

This work was partially supported by NIH.

References

Friedman, J., T. Hastie and R. Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33**(1): 1-22.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, *et al.* (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**(4): 576-589.

Hon, G. C., R. D. Hawkins, O. L. Caballero, C. Lo, R. Lister, *et al.* (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome research* **22**(2): 246-258.

Lister, R., M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, *et al.* (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**(7336): 68-73.

Nakano, K., J. W. Whitaker, D. L. Boyle, W. Wang and G. S. Firestein (2012). DNA methylome signature in rheumatoid arthritis. *Ann Rheum Dis.* **72**(1): 110-117

Won, K. J., Z. Xu, X. Zhang, J. W. Whitaker, R. Shoemaker, *et al.* (2012). Global identification of transcriptional regulators of pluripotency and differentiation in embryonic stem cells. Nucleic Acids Res. **40**(17): 8199-8209