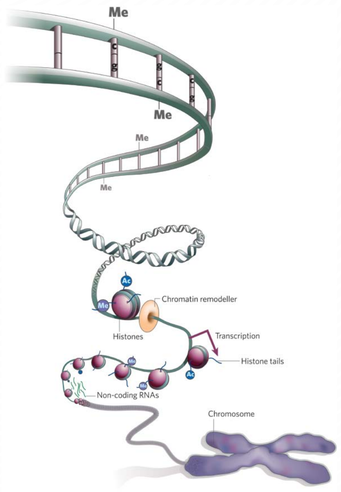
The Role of Histone Modification Patterns in Transcription Factor Binding and Gene Expression

Jason Young, Wei Wang

Epigenetics is the study of differences in the phenotype of a cell or organism that do not involve changes in the nucleotide sequence of DNA. Epigenetic mechanisms include histone modification, DNA methylation, and small and non-coding RNAs, all of which form elements of chromatin signatures that interact with transcription factors to regulate gene expression patterns, actively play a role in cell development, differentiation, and identity

The Epigenome: Functions of a Landscape



The Missing Link: Histone Modification Patterns and Transcription Factors

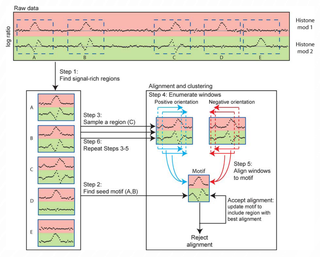
In this study, histones were the focus of our examination of epigenetic elements. Histones are proteins comprised of a globular domain and a tail of polypeptides, and they form octamer aggregates called nucleosomes which play an integral role in chromatin formation and DNA packing. Modifications of histone N-terminal tails include methylation, acetylation, and phosphorylation, and combinations of these modifications have been found to affect gene expression levels by affecting chromatin packing and chromatin DNA-interactions, forming the basis of what is known as the histone code hypothesis. Recent findings have demonstrated the relationship between chromatin modifications at enhancers as they relate to cell-type-specific gene expression, and recently, histone modification patterns have been used in conjunction with other methods to predict putative enhancer sites. However, the precise relationship between histone modification patterns and the protein elements that transcribe genes remains unclear, and a systematic and widespread search for the relation between such elements and histone modification patterns has yet to be attempted.
One such protein element is the transcription factor (TF). Transcription factors are proteins that bind to RNA Polymerases and functional elements in the DNA to facilitate RNA transcription and ultimately help to dictate gene expression. They often interact with each other and the DNA to form complexes around transcription start sites (TSSs), and other functional elements in the genome. Indeed, transcriptional programs comprised of networks of TFs have been correlated to differential gene expression, for by binding to transcription factor binding sites (TFBSs) in varying combinations of complexes, TFs have proven to be critical to regulating gene expression in cells and carrying out cell-specific transcription. Unfortunately, the mechanisms by which TF networks and other DNA associating proteins regulate gene expression are not well studied, partly due to a lack of information regarding TFBSs, prompting the ongoing search for motifs. One such protein element is the transcription factor (TF). Transcription factors are proteins that bind to RNA Polymerases and functional elements in the DNA to facilitate RNA transcription and ultimately help to dictate gene expression. They often interact with each other and the DNA to form complexes around transcription start sites (TSSs), and other functional elements in the genome. Indeed, transcriptional programs comprised of networks of TFs have been correlated to differential gene expression, for by binding to transcription factor binding sites (TFBSs) in varying combinations of complexes, TFs have proven to be critical to regulating gene expression in cells and carrying out cell-specific transcription. Unfortunately, the mechanisms by which TF networks and other DNA associating proteins regulate gene expression are not well studied, partly due to a lack of information regarding TFBSs, prompting the ongoing search for motifs.

Interplay between Histone Modification Patterns and Transcription Factor Binding

In order to elucidate the relation between histone modifications and DNA associating proteins like TFs, in this study, histone modification clusters were analyzed around known promoter sites in 7 different cell lines across eight different histone modifications. We hypothesized that functional elements and genes possessing similar roles in the cell would have similar histone marks, and that these clusters of similar genomic elements would be bound by the same TFs and DNA-binding proteins. An analysis of these clusters of promoters and enhancers and DNA-interacting protein binding was performed to systematically determine whether histone patterns or clusters of genes possessing similar patterns could elucidate functions of DNA-binding proteins or corroborate known relationships. Ever since the discovery of the epigenome, studying the interactions between epigenomic elements and underlying DNA has proven to be critical to understanding cell differentiation and specificity, and by combining histone modification mapping, transcription factor motif data, and computational tools, we hoped to shed light upon the link between epigenetics and protein binding in determining cell identity cell-specific gene expression.

ChromaSig: Histone Modification Pattern Extraction

Recent studies have characterized unique histone modification patterns at functional genomic regions such as promoters and enhancers, and because DNA interacting proteins often bind to these genomic elements, this study focuses on the relationship between known motifs at these elements in humans and histone modifications. A list of known promoter regions was downloaded from the UCSC table browser, an open source database that retrieves data and DNA sequences associated with specific genes or functional elements in the DNA. The assembly used was from the March 2006 hg18 selection, and the list of promoters was garnered from the track of UCSC genes. The promoters were rounded to the nearest 100 base pair in order to accommodate the ChIP-seq data of histone modifications we downloaded later, which was of a 100bp resolution. ChIP-Seq is a chromatin immunoprecipitation technique that can find the enrichment of histone modifications on a genome-wide scale.



ChIP-seq functions by precipitating out the proteins of study that bind to DNA using highly specific antibodies (ChIP). A library of the resulting bound DNA nucleotides is complied and sequenced, providing data for enriched bound locations of the proteins (seq). Because ChIP-seq maps of histone modifications are becoming widely available, ChIP-seq data was thus the natural choice for this study to generate modification cluster maps around promoter locations. In order to systematically search for the interplay between histone modification patterns and DNA associated protein binding motifs, and because these patterns are often cell specific, we studied seven different cell lines, Gm12878, Hmec, Hsmm, Huvec, K562, Nhek, Nhlf, and eight modification patterns (H3K27me3, H3K27ac, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H4K20me1) to gain a broader view of the role of chromatin in motif discovery in various environments by extracting ChIP-seq histone modification patterns from the ENCODE project's ChIP-seq data (freely available at http://hgdownload.cse.ucsc.edu/ goldenPath/hg18/ encodeDCC/wgEncodeBroadChipSeq/) downloaded from the UCSC Genome Bioinformatics Browser. In particular, data from the Broad lab was used in this study in the .tagAlign format of the files. Next, the files were input into a program called ChromaSig, an unsupervised learning method that locates commonly occurring chromatin signatures in the data. The ChromaSig package (available from the STAR Pipeline http://wanglab.ucsd.edu/star/ pipeline.php) includes a program called PreChromaSig, a Perl language program that performs data normalization on the ChIP-Seq files, a necessary step because the absolute tag counts for the ChIP-Seq data varies widely, which may present an inaccurate picture of the histone modification landscape. The ChromaSig program scans the genome-wide histone maps to locate enriched regions of chromatin patterns then creates a seed pattern based on the maps, then finally enumerates each locus to the seed and aligns it with the model to either reject the locus if poorly aligned or update the motif if the locus has a similar alignment The output gave 114 unique clusters of similar histone modification reads across 13914 unique promoter locations. We next used Treeview, a freely available software tool that we used to generate heatmaps of our histone clusters for visualization. The generated heatmaps provide a visual model of the histone modification patterns in a 10 kilo base pair region around each promoter location.

Histone Modification Heatmap of ChromaSig Clusters Generated from Treeview

Here, 114 histone modification patterns are shown across 13914 Unique promoter regions in the human genome. Every eight columns represents a different cell line (Gm12878, Hmec, Hsmm, Huvec, K562, Nhek, Nhlf) and each of the eight columns represents a histone modification (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H4K20me1). Every column is a 10kb region around the TSS of each promoter region. Similar histone modification patterns across all studied cell lines suggests a similarity in the epigenetic landscape around promoters.

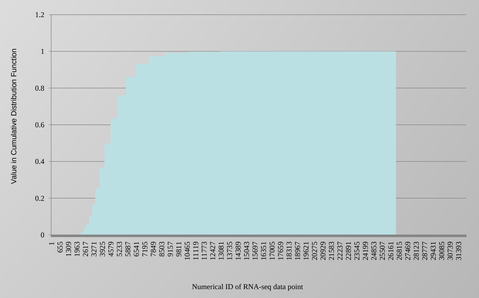
Motif Analysis and Gene Ontology

In order to determine whether specific epigenetic patterns are related to DNA-binding proteins and motifs, the approach was two-fold. First, determine significantly enriched binding motifs at promoters by performing motif analysis, and combining it with RNA-seq data to find transcription factors that are ubiquitous and cell specific in each cluster. Second, perform gene ontology analysis to computationally determine the function of those proteins in each histone cluster. In the former case, a list of histone modification patterns sites which form the chromatin clusters determined from ChromaSig was used. Position specific scoring matrices (PSSMs or PWMs) were obtained for a list of 1169 known motifs gathered off of 5 motif databases, bulyk, hPDI, Jaspar, Uniprot, and TRANSFAC enriched in each of the promoter and enhancer regions. The PSSM is a commonly used method to measure motifs in the DNA, and in this study, each promoter region determined from ChromaSig was scanned *in silico* for motifs that would suggest the presence of that TF or other protein. The PSSM data was then complied and input into R, freely available software which was used for statistical analysis and graphing. R was employed in this study due to its immediate availability and ease of use. The program was used to obtain the hypergeometric distribution result of each motif from the PSSM data in each cluster. The group of study for each calculation was the sum of all the PSSM scores representing each motif for all the promoters of each cluster, and the background was designated as the sum of motifs in all other histone cluster locations. A log-odd ratio value between 1 and 0 for each motif in each cluster, with numbers closer to zero suggesting greater significance of that particular motif in that cluster. An arbitrary and strict cutoff p-Value of 10^-5 was determined to obtain the highest prevalent motifs in each cluster which would correlate to the statistically most significant binding of the associated protein in the cluster.

RNA-seq data: RPKM frequencies

Next, enriched DNA-interacting proteins in the studied cells determined from the pValues were checked back to the motif counts determined in the PSSM data. Because simply determining the pValue will return false-positives (since if a single count appears in the foreground while there is none in the background, the read will be deemed significant), we calculated the number of appearances of each PSSM for each promoter region of each cluster and determined to reject any motifs with high pValues that did not appear in at least 30% of the promoter regions, an arbitrarily determined cutoff. The remaining motifs with significant pValues were compared with RNA-seq data and filtered to determine the actual associated proteins binding in each histone modification cluster. Comparing the pValues of the proteins with actual RNA-seq data is crucial to determining whether a TF or other DNA-binding protein is ubiquitously expressed or whether it is cell-specific. RNA-seq is a method that sequences cDNA that reflects mRNA transcribed in the cell, and hence, the actively transcribed gene. RNA-seq data was obtained from the UCSC ENCODE project using data specifically from this URL (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeCaltechRnaSeq/). This study used RPKM data from the first repetition of each experiment as data input. Unfortunately, due to the unavailability of data for all of the cell types being studied, only 4 of the 7 cell types had RNA-seq data. We thus localized our study of protein-DNA interaction with histone modification patterns to those 4 cell types. RNA-seq enrichment is related to both the length of the transcript and concentration of enrichment. RPKM is a method of quantifying the transcription level to "reads per kilobase of exon model per million mapped reads" (RPKM). This is a measure of read density that allows a simple comparison of transcription levels relative to other RNA-seq data and other reads within the same file. A quick plot of the frequency of enriched expression based on the RPKM values demonstrated that the data followed a rough Poisson Distribution. Thus, using R, we determined p-Values for each transcribed gene using the Poisson distribution function to determine the probability that any other gene would be more highly expressed that the gene in question. A strict pVal cut of 10^-3 was used to take only the most highly expressed genes. After cross-comparing the filtered RNA-seq genes with the TFs and other proteins determined from motif analysis a list was found for the proteins that appeared in both data sets for each histone modification cluster to be the set of DNA-interacting proteins at the cell's promoters in the cluster.

RPKM Cumulative Distribution Function
Demonstrates Poisson Distribution



Gene Ontology

These corresponding TFs were analyzed through Gene Ontology, a bioinformatics database that seeks to annotate gene data and provide tools to analyze that data. One such tool, known as the Ontologizer, was employed to find significant GO terms for each of the promoter clusters and the corresponding proteins in each cluster. GO terms are properties of gene products that are tagged onto specific search terms like TFs and genes, and this study hypothesized that clusters of promoters or enhancers represent similarly acting or functioning genes, thus possessing similar GO terms. The Ontologizer is a tool that performs enrichment analysis of input genes and outputs significant GO terms associated with the input genes by using a model-based gene set analysis (MGSA) to search all the associated GO terms and categories at once for the input in a Bayesian network. This approach thus addresses the categorical overlap inherent in many of the GO terms to reduce excessive returns of correlated GO terms. Next, using Perl, scripts were created that processed the data to generate tables for heatmap input. Three heatmap tables were created. One heatmap demonstrates the TF binding occurrences in each cluster as a fraction of all the promoters it bound to over the total number of promoters in the cluster. A second heatmap displays the significance of the binding enrichment by plotting the pValues determined for each TF expressed in each of the four RNA-seq cells in the cluster. A third heatmap demonstrates expression level by plotting the normalized RPKM value for the expressed TF in each of the cell types per cluster.

Results

Our results validated the hypothesis that each cluster genes possessing similar histone marks also possessed similar functions. While a large portion of clusters had similar functions relating to transcription and gene expression, within each cluster, the functions were largely the same. For instance, in cluster 16 (below), the GO terms revealed that besides the broad terms like "cell" and "biological process", the cluster was mainly involved in cell adhesion, binding, and reproduction to a small extent. Indeed, the GO terms also discovered that the DNA-binding proteins in each cluster also had largely similar functions. Thus, by comparing the functions of the cluster's genes, histone patterns, and protein functions, both novel insights and known relationships could confidently be drawn. During our analysis, to discover the specific functions of individual DNA-binding proteins, the Uniprot database was used, an open database that provides annotated and curated protein sequence and functional information.
Eighty unique DNA-binding proteins to associate across the 114 clusters of promoters in differing combinations for each cluster (Insert supplemental table).These combinations represent the most statistically significant and highly expressed DNA-binding proteins in each cluster. These clusters, the associated genes, and the associated DNA-binding proteins were compared to find any significant relationships between the data. From the analysis, four distinct categories of interactions were discovered. (1) Proteins that bound ubiquitously across all cell types for multiple clusters. (2) Proteins that showed cell specificity but were expressed ubiquitously across histone patterns. (3) Cell specific histone patterns that were demonstrated for proteins that bound over several clusters. (4) Cell specific histone modification patterns that correlated with cell specific proteins and transcription factors.

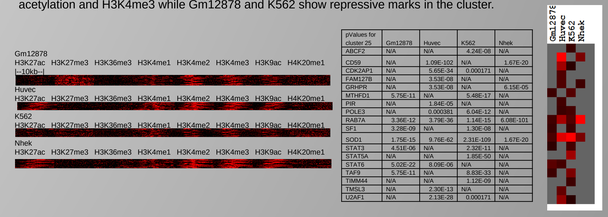
For example, we examined our data to search for cell-specific proteins that would be expressed in a single cluster of genes. Although cell specific proteins binding to promoters would be rare, we did indeed find a few DNA-binding proteins that were involved with unique clusters. In general, our results demonstrated a strong corroboration between the protein and cluster functions, demonstrating a strong correlation between histone patterns and the associated genes and protein function.

Summary Table

Proteins that bind in cell specific patterns in a single cluster	cluster number	cluster function	protein function
CAT*	23	Lipid and small molecule metabolic processes	Calcium Transport Protein
CD59*	25	Cell mediated signal response and reception	Glycoprotein
DUSP22	71	Sensory Perception	Signaling Pathway Protein
HMEX	18	Proteoglycan Synthesis	Homeobox Protein
HNRPA1	101	Transcription and signaling pathways and stimulus reception	Ribonucleoprotein
IRF6	23	Lipid and small molecule metabolic processes	DNA-binding Transcription Factor
POLE3	25	Cell mediated signal response and reception	DNA Polymerase Subunit
RARG	6	Transcriptional Regulation	Retinoic Acid Receptor
SUCLG1*	68	Cell adhesion, binding, and reproduction	Protein involved in Aerobic Respiration
TRMT1	68	Cell adhesion, binding, and reproduction	tRNA processing
TSN	68	Cell adhesion, binding, and reproduction	DNA Recombination
UTP18	38	Transcription and Transcriptional Regulation	Small Nucleolar RNA-associated protein

*Note: The known location for these proteins is outside the nucleus, suggesting they may be false positives, or that there may be an unknown function for these proteins that may potentially exist with these DNA-binding proteins.

One such example was the protein CD59. The study both corroborated known associations and suggested new functions for this protein. The data was indeed accurate in reflecting the presence and function of CD59, a cell surface receptor glycoprotein and membrane attack complex inhibition factor, for GO term analysis reveals that the cluster (cluster 25) is indeed involved in cell mediated signal response and reception, relating to the function of CD59. Additionally, the histone modification heatmap demonstrated how the histone landscape also demonstrated the increased binding density of CD59, for both RPKM data and the histone marks reflected its cell specificity in binding only to the genomes of HUVEC and NHEK cells. This is clearly visible in the histone heatmaps, which demonstrate how only HUVEC and NHEK exhibit the active acetylation and H3K4me3 while Gm12878 and K562 show repressive marks in the cluster.



pValues for cluster 25	Gm12878	Huvec	K562	Nhek
ABCF2	N/A	N/A	4.24E-08	N/A
CD59	N/A	1.09E-102	N/A	1.67E-20
COK2AP1	N/A	5.65E-34	0.000171	N/A
FAM127B	N/A	3.53E-08	N/A	N/A
GRHPR	N/A	3.53E-08	N/A	6.15E-05
MTHFD1	5.75E-11	N/A	5.48E-17	N/A
PIR	N/A	1.84E-05	N/A	N/A
POLE3	N/A	0.000381	8.04E-12	N/A
RAB7A	3.36E-12	3.79E-36	1.14E-15	6.08E-101
SF1	3.28E-09	N/A	1.30E-08	N/A
SOD1	1.75E-15	9.76E-62	2.31E-109	1.67E-20
STAT3	4.51E-06	N/A	2.32E-11	N/A
STAT5A	N/A	N/A	1.85E-50	N/A
STAT6	5.02E-22	8.09E-06	N/A	N/A
TAF9	5.75E-11	N/A	8.83E-33	N/A
TIMM44	N/A	N/A	1.12E-09	N/A
TMSL3	N/A	2.30E-13	N/A	N/A
UZAF1	N/A	2.13E-28	0.000171	N/A

While the study accurately described the binding of CD59 in the cluster and the similar functions between CD59 and the associated genes, it is curious to note the presence of CD59 (and other proteins determined in this category) in the nucleus of the cell. The RPKM data clearly shows its high expression density in the cluster, yet the high significance of the binding motif given by the pValue coupled with the obvious histone pattern difference may suggest that CD59 has an unknown chromatin modification function, though this statement will require future research to corroborate. The other proteins that are typically associated with functions in the membrane or cytosol may also have unknown functions in the nucleus that need further investigation.
Our data thus supports the idea that histone clusters and the associated genes exhibit distinct and significant functions that are related to the DNA-binding proteins/. However, how the proteins relate to the extent of gene transcription remains unknown, and an analysis of the RNA-seq data of the underlying genes remains an area of future work.

Citations

- Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, & Barbara Wold. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods, 5, 621. doi:10.1038/nmeth.1226
- Barrera, L. O., & Ren, B. (2006). The transcriptional regulatory code of eukaryotic cells – insights from genome-wide analysis of chromatin organization and transcription factor binding. Current Opinion in Cell Biology, 18(3), 291-298. doi:DOI: 10.1016/j.ceb.2006.04.002
- Barski, A., Cuddapah, S., Cui, K., Roh, T., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell, 129(4), 823-837. doi:DOI: 10.1016/j.cell.2007.05.009
- Bauer, S., Gagneur, J., & Robinson, P. N. (2010). GOing bayesian: Model-based gene set analysis of genome-scale data. Nucleic Acids Research, 38(11), 3523-3532. doi:10.1093/nar/gkq045
- Firpi, H. A., Ucar, D., & Tan, K. (2010). Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics, 26(13), 1579-1586. doi:10.1093/bioinformatics/btq248
- Gary Hon, Bing Ren, & Wei Wang. (2008). ChromaSig: A probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biology, 4(10) doi:10.1371/journal.pcbi.1000201.
- Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E., Kuan, S., Luu, Y., Klugman, S., Antosiewicz-Bourget, J., Ye, Z., Espinoza, C., Agarwahl, S., Shen, L., Ruotti, V., Wang, W., Stewart, R., Thomson, J. A., Ecker, J. R., & Ren, B. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell, 6(5), 479-491. doi:DOI: 10.1016/j.stem.2010.03.018
- Koch, C. M., Andrews, R. M., Flicek, P., Dillon, S. C., Karaöz, U., Clelland, G. K., Wilcox, S., Beare, D. M., Fowler, J. C., Couttet, P., James, K. D., Lefebvre, G. C., Bruce, A. W., Dovey, O. M., Ellis, P. D., Dhami, P., Langford, C. F., Weng, Z., Birney, E., Carter, N. P., Vetrie, D., & Dunham, I.The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Research, 17(6), 691-707. doi:10.1101/gr.5704207
- Kyoung-Jae Won, Bing Ren, & Wei Wang. (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. Genome Biology, 11(7) doi:10.1186/gb-2010-11-1-r7
- Lupien, M., Eeckhoute, J., Meyer, C. A., Wang, Q., Zhang, Y., Li, W., Carroll, J. S., Liu, X. S., & Brown, M. (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell, 132(6), 958-970. doi:DOI: 10.1016/j.cell.2008.01.018
- Nathaniel D. Heintzman, Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, Leonard K. Lee, Rhona K. Stuart, Christina W. Ching, Keith A. Ching, Jessica E. Antosiewicz-Bourget, Hui Liu, Ximmin Zhang, Roland D. Green, Victor V. Lobanenkov, Ron Stewart, James A. Thomson, Gregory E. Crawford, Manolis Kelli, & Bing Ren. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature, , 108. doi:10.1038/nature07829
- Peter A. Jones, Trevor K. Archer, Stephen B. Baylin, Stephan Beck, Shelley Berger, Bradley E. Bernstein, John D. Carpten, Susan J. Clark, Joseph F. Costello, Rebecca W. Doerge, Manel Esteller, Andrew P. Feinberg, Thomas R. Gingeras, John M. Greally, Steven Henikoff, James G. Herman, Laurie Jackson-Grusby, Thomas Jenuwein, Randy L. Jirtle, Young-Joon Kim, Peter W. Laird, Bing Lim, Robert Martienssen, Kornelia Polyak, Henk Stunnenberg, Thea Dorothy Tlsty, Benjamin Tycko, Toshikazu Ushijima, Jingde Zhu. (2008). Moving AHEAD with an international human epigenome project. Nature, , 711. doi:10.1038/454711a
- Whitington, T., Perkins, A. C., & Bailey, T. L. (2009). High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. Nucleic Acids Research, 37(1), 14-25. doi:10.1093/nar/gkn866
- Whitington, T., Perkins, A. C., & Bailey, T. L. (2009). High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. Nucleic Acids Research, 37(1), 14-25. doi:10.1093/nar/gkn866
- Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang, & Keji Zhao. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nature Genetics, , 897. doi:10.1038/ng.154

Acknowledgments

Thanks to Professor Wang for letting me work in his lab, and to all the graduate students Jie, John, Robert, and Kyoung-Jae for helping me along the way.