Exploring cis-regulatory Models of the Genome to Predict Epigenetic State and Variation

By Daniel Kang

Submitted to the
Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Masters of Engineering in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

August 2015 [September 2015]

Author: Signature redacted
Department of Electrical Engineering and Computer Science
September 8, 2015

Certified by: Signature redacted
David Gifford, Thesis supervisor
September 8, 2015

Accepted by: Signature redacted
Dr. Christopher Terman, Chairman, Masters of Engineering Thesis Committee

Abstract

We introduce a new model called CCM+ that rectifies the inability of previous sequence-based methods to generalize across cell-types. We permit generalization by introducing arbitrary base-pair resolution covariates. We show that the addition of base-pair resolution chromatin accessibility covariate greatly aids in the prediction of cis-regulatory marks. Additionally, we show that by using cell-type specific covariates, CCM+ can generalize across cell-types. Finally, we show CCM+ can be used for downstream analysis that matches state-of-the-art methods when chromatin accessibility is used as a covariate.

1. Introduction

A fundamental task in regulatory genomics is to understand the relationship between genome sequence and its consequence on molecular phenotypes. Molecular phenotypes include high-resolution chromatin accessibility [1], gene expression, and the protein occupancy of the genome. These phenotypes play a critical role in transcription factor (TF) binding, gene regulation, cellular identity, among other cellular processes [2–4]. The ability to predict molecular phenotypes directly from genome sequence would permit us to evaluate the consequences of DNA sequence variants that have been implicated in diseases [5].

Computational approaches that model the relationship between DNA sequence and molecular phenotypes include the gapped k-mer SVM (gkm-SVM) method [6] and the Cooperative Chromatin Model (CCM) [1]. gkm-SVM uses gapped k-mers as features in a support vector machine (SVM) for prediction [6]. gkm-SVM performs well at predicting a variety of regulatory elements in in different cell conditions [6]. However, gkm-SVM requires examples to be labeled as positive or negative, such as labels defined by ENCODE's uniform processing pipeline [6]. This approach, and similar ones [7,8], often suffer from the ad-hoc interpretation of genome measurements. For example, an arbitrary cutoff is often chosen to pick "interesting regions" of accessible chromatin. CCM has focused on rectifying this issue in the context of chromatin accessibility [1]. CCM predicts chromatin accessibility over the genome, as measured by assays such as DNase-seq and ATAC-seq, from sequence alone [1]. From a modeling perspective, CCM is a genome-wide Poisson regression. Its main technical novelty is in the state-of-the-art methods to fit L_1 regularized linear models over the entire genome. The CCM framework can predict arbitrary quantitative, non-negative integer-valued traits measured per

base on the genome. For example, in addition to DNase-seq, the framework can be used to predict other sequencing assays such as ChIP-seq.

While CCM does remarkably well on simple assays, such as DNase-seq, ATAC-seq, and Nrf1 ChIP-seq, it does not perform nearly as well on certain types of ChIP-seq assays. As we show, CCM and gkm-SVM are limited because no auxiliary data can be used to aid their models, yet it is often the case that more information than pure sequence is available during data analysis. Consider the case of predicting a transcription factor (TF) ChIP-seq. In a simple model of TF binding, a given TF binds at its motif wherever chromatin is open. Under this simple model, having chromatin accessibility information in the form of DNase-seq would help predict TF binding. Additionally, training on pure sequence offers no way generalize across conditions, as sequence is identical between different cell-types. Thus, gkm-SVM and CCM must be retrained to predict binding in different cell-types.

Here we introduce a new model called CCM+ that rectifies the inability of previous methods to use high-resolution auxiliary information by allowing for arbitrary base-pair resolution covariates. We find that this enables CCM+ models to generalize across cell-types. We also show that CCM+ significantly outperforms CCM for predicting p300 binding and histone marks. Finally, we show that CCM+ can be used for downstream analysis that matches even state-of-the-art methods when chromatin accessibility is used as a covariate. Thus, accessibility information provides substantial information on genome regulatory elements.

2. Summary of methods

CCM+ is a fully generative model of arbitrary non-negative signals over the genome. In this work we focus on the readout of ChIP-seq experiments. The genome is treated as a single regulatory sequence with k-mers as elements that have invariant, proximal spatial effects. Additionally, a covariate may be provided as a functional prior at base-pair resolution. Here we use chromatin accessibly as measured by DNase-seq as the covariate. The read counts of a ChIP-seq experiment at a given base are generated by the log-linear combination of nearby sequence effects and DNase-seq signal. We note that this model is inherently proximal and therefore cannot model trans-effects that are not mediated by cis regulatory sequences.

For the majority of this work, we compare CCM+ against CCM for three major reasons: 1) this comparison is the most controlled to see if accessibility can help in predicting regulatory elements, 2) further post-processing is required to compare against other methods, and 3) adding covariates to linear models has been well studied. To make the appropriate comparison with gkm-SVM, we would need to add chromatin accessibility to the gkm-SVM model, which is beyond the scope of this work. Many other models, including gkm-SVM, address a binary prediction task, whereas CCM+ outputs continuous signals over the genome at base-pair resolution. Thus we process the output of CCM+ to make binary predictions. Finally, we note that adding covariates to linear models has been well studied and characterized [9].

3. CCM+ is the best predictor of p300 binding signals

p300 is an important transcription factor that has been shown to interact with hundreds of proteins [10]. It also plays an important role in cellular tasks ranging from transcription [11]

to tumor suppression [12]. p300 has been shown to often bind at promoters and enhancers and is likely to be related to enhancer function [13,14]. Because p300 often binds at enhancers, p300 binding sites have been used as a proxy for enhancer locations [15,16]. Thus, p300 binding is important in studies of gene regulation.

We find that CCM+ outperforms CCM for predicting p300 binding when DNase-seq is used as a covariate. We first trained CCM and CCM+ on chromosomes 1-12 of a human GM12878 p300 ChIP-seq binding dataset using a matched DNase-seq track as a covariate for CCM+; the special effect of the chromatin accessibility is positive, with a peak offset to the left of center (Figure 9). We tested the resulting CCM and CMM+ models on chromosome 14, which was not used in training, and measured the correlation between the predicted rates and the observed reads. CCM achieves a correlation of 0.31 and CCM+ achieves a correlation of 0.49, a significant improvement.

We also find that when CCM+ output is run through an event caller the resulting binding events are more faithful than those produced by CCM. We used the resulting output Poisson rates from the CCM and CCM+ p300 models to generate predicted read counts at every base position. Using the predicted reads, we used GPS [17] to call binding events over the genome for 1) the real p300 ChIP-seq dataset, 2) the CCM predicted reads, 3) the CCM+ predicted reads. Using the calls from the observed p300 ChIP-seq dataset as ground truth, we labeled the calls from CCM and CCM+ as correct if they fell within 300 bp of a ground truth call. We plotted the precision-recall curves using calls from only chromosomes 14-22 to avoid overfitting errors (Figure 1). CCM+ strictly dominates CCM in precision at all recall levels and in the metric of area under the precision-recall curve (PR curve) (CCM: 0.14, CCM+: 0.39). Moreover, CCM+ achieves

higher precision at low recall, where CCM suffers. Thus, CCM+ clearly outperforms CCM on this task.

Finally, we show that CCM+ can be used to classify p300 peaks better than BinDNase [18], the current state-of-the-art TF binding predictor. BinDNase takes sequence and DNase-seq to produce TF binding calls, which is similar in spirit to our method [18]. We used the recommended training procedure to train a BinDNase model for p300 binding. The ground truth calls were generated from p300 ChIP-seq data with GEM with the recommended settings [17]. For training, only calls from chromosomes 1-13 were used. We note that BinDNase recommends using only 6000 total examples. Using the same call locations as BinDNase, we trained a logistic classifier using a +/- 70 bp window around each call site, with the predicted Poisson rates as features. To compare the performance, we used the motif generated from GEM to find all the motif matches with a p-threshold below 1e-5 using FIMO [19]. Motif hits within 300 bp of a GEM call were considered a real binding event. Using the scores from BinDNase and our classifier, we plotted the ROC and PR curves (Figure 2). Our classifier slightly outperforms BinDNase in the metric of area under the ROC curve (BinDNase 0.87 vs CCM+ 0.89) and performs the same in the metric of area under the PR curve (BinDNase 0.64 vs CCM+ 0.64). We note our classifier is much simpler than BinDNase, but performs slightly better.

4. Histone mark analysis

In addition to affecting chromatin accessibility, histones and their covalent marks are known or suspected to play a role in the regulation of transcription, DNA replication, and splicing [22,23]. Specific histone marks are associated with enhancers (e.g. H3K4me2) and

promoters (e.g. H3K4me3, H3K9ac) [24]. Thus, to further understand gene regulation we were interested in predicting the location of histone marks. While there has been work in predicting histone marks from sequence features [20,21], these methods similar shortcomings to other methods. For example, Epigram [21] requires histone peaks to be called and, as it uses only sequence, has no way to generalize across cell-types. We note that the GM12878 and Dermal Fibroblast H3K4me3 observed ChIP-seq datasets are only minimally correlated (R^2 of 0.59), so pure sequence cannot be the only determinant of epigenetic marks (Figure 8).

Since CCM+ predicts proximal effects, we studied histone marks that are designated as "peaks" by ENCODE [24]. Specifically, we studied the histone H2A.Z and histone marks H3K4me2, H3K4me3, H3K9ac, and H3K27ac. We found that while CCM+ and CCM have roughly equal performance on predicting histone data based on the correlation of their predictions with observed data (Table 1, Figure 6), CCM+ dominates CCM in the task of peak finding. Additionally, we found that CCM+ can use the parameters from one cell-type to predict the histone ChIP-seq from another cell-type, a task that CCM and gkm-SVM cannot do.

We explored the ability of CCM+ and CCM to predict histone data peaks by using MACS [25] to call histone peaks on the observed human GM12878 ChIP-seq dataset for every histone mark listed above. The resulting peaks were considered ground truth. Predicted reads were generated from CCM and CCM+ models using the same procedure as with the p300 analysis. The predicted reads were used as input to MACS to produce peak locations. Using theses peaks, we plotted the PR curves, using the calls from chromosomes 14-22 to avoid overfitting issues. In every case, CCM+ dominates CCM in the metric of area under the PR curve (Table 2).

Additionally, we see that the sequence parameters for CCM and CCM+ are sufficiently different (Figure 7).

5. Transfer learning

When CCM+ uses DNase-seq as a covariate we posit that CCM+ can learn cell-type invariant parameters that can "transfer" a model to a new cell-type with high accuracy when supplied with DNase-seq data for that cell type. This type of analysis is known as "transfer learning." Transfer learning is not possible with sequence-only based methods such as CCM since sequence remains constant between cell-types.

We find that CCM+ can find differential histone peaks with high accuracy using transfer learning. We first used the parameters from GM12878 CCM+ models and DNase-seq from dermal fibroblast to generate predicted rates for all the histone marks we studied above. Using these rates, we used the procedure above to generate predicted reads over the genome, and we used these predicted reads to generate histone peak calls from MACS. We also generated ground truth peaks for dermal fibroblast with MACS. Then, we removed peaks that were in GM12878 from both the observed and predicted dermal fibroblast data. Using the dermal fibroblast unique peaks we plotted the PR curves by comparing observed and predicted peaks. Remarkably, the area under the PR curve is high for CCM+ (above 0.53) for all the curves (Figure 4).

As the sequence features we used for prediction did not change between cell-types, it is reasonable to assume that the chromatin accessibility covariate is the sole source of the performance across cell types. To show this is not the case we created an alternative model

that left out sequence features. We used the modified model to predict rates, and used the predicted rates to generate predicted reads and histone peak predictions, as above. Using these predictions, we plotted the PR curves (Figure 4). In terms of overall performance, the full CCM+ model dominates the DNase-seq covariate only model for every histone mark, although for some histones the DNase-seq only model has higher precision at low recall. However, it is visually clear that the covariate is not dominating the performance of CCM+.

6. Discussion

We have shown that CCM+ significantly outperforms CCM on a variety of tasks and regulatory marks. Additionally, we show that CCM+ can be used to classify p300 binding sites slightly better than the state-of-the-art. Finally, we show that CCM+ can be used for transfer learning to new cell types and that the performance is not dominated by the DNase-seq covariate. Thus, we see that chromatin accessibility can aid greatly in the prediction of regulatory marks, but chromatin accessibility does not dominate the prediction.

While CCM+ performs remarkably well, it has deficiencies. First, the model is inherently proximal. As a result, CCM+ can only predict regulation that is a consequence of cis-regulatory sequences. For example, CCM+ cannot predict the histone mark H3K27me3 [24] or the effects of heterochromatin spreading [26] (Figure 5). Additionally, while CCM+ can predict histone binding as defined by MACS, the correlation for several histone marks remains low. The low correlation is likely due to non cis-regulatory effects. Thus, there remains potential for improvement by incorporating distal effects into the model.

We have found that CCM+ is a useful tool in studying proximal regulation in the genome. As CCM+ requires no parameter tuning, aside from setting parameters to the maximum that memory and time constraints allow, and since CCM+ takes arbitrary covariates, it can also be applied to a wide range of base-pair resolution measurements.

7. Model

As CCM+ extends CCM, we make the same six core assumptions regarding sequence. First, the entire genome is treated as one continuous regulatory sequence. This is achieved in practice by concatenating chromosomes. Second, we view k-mers as "code words" which induce invariant spatial effects to the genomic area surrounding the k-mer. Each k-mer has a +/- 1 kb spatial effect. Third, only a small number of k-mers are necessary for prediction. This assumption is achieved in practice through L_1 regularization. Fourth, a given k-mer has the same spatial effect regardless of its location in the genome. Fifth, the spatial profiles from the k-mers non-specifically synergize in a log-linear fashion. Sixth, only k-mers up to 8bp long are needed for prediction.

Aside from the sequence parameters, CCM+ also includes a covariate track. The covariate track is the same size as the genome, and in this work is a binary indicator of a DNase-seq read at each base. Similarly to the sequence features, each base in the covariate track has a location-invariant spatial effect of +/- 200 bp. The effect from the covariate track synergizes with the sequence effects in the same log-linear fashion.

We note the maximum k-mer size and the spatial effect window sizes are parameters, but they are set to be as large as possible given memory and time constraints.

These assumptions naturally lead to a regularized Poisson regression. We first introduce the notation and representation for the model. We treat the genome as a single sequence with coordinates from 0 to N. The k-mer effect parameters are denoted as θ^k, which has size $4^k \times 2M$ for $k = 1 \ldots 8$ and $M = 1000$ (the effect window size). For a given k-mer of length k starting at base i in the reference genome, we denote g_i^k as the row index in θ^k. Thus, $\theta^k_{(g_i^k, j)}$ is the effect of that k-mer at offset j. There is also a special parameter θ_0 which is used to set the average read rate globally. The covariate is denoted as κ and is the same length as the genome. The regression coefficients are denoted as β and we denote L as the covariate effect window size.

The regularized Poisson regression to solve has the objective function

$$\max_{\theta,\beta} \left(\sum_i c_i \log \lambda_i - \lambda_i \right) - \eta \sum_k |\theta^k|_1$$

with the intermediate variables

$$\lambda_i = \exp\left(\left(\sum_{k \in [1,8]} \sum_{j \in [-M.M-1]} \theta^k_{\left(g^k_{i+j}, -j\right)} \right) + \left(\sum_{l \in [0,2L]} \beta_l \cdot \kappa_{i+l-L} \right) - \theta_0 \right)$$

as the per-base Poisson rates. The generative model is that the read counts are generated per-base from a Poisson distribution with rates λ_i.

8. Inference

CCM+ has many parameters and the training data is often half the genome; thus exact inference is infeasible. However, we can use approximate inference techniques to learn the

parameters. At the heart of our inference algorithm is a batch proximal gradient descent, which is as follows.

1. Given the current iterate θ and β, calculate λ_i for each base $i \in [0, N]$ as above.
2. Given λ, calculate the per-base gradient vector: $d \log \lambda_j = err_j = c_i - \lambda_i$
3. Propagate the errors back to the parameters. Let s be the index corresponding to a given k-mer and j be the offset. Let l be the covariate offset. Then, the gradients are:

$$\mathrm{d}\theta^k_{\mathrm{s,j}} = \sum_{\{\mathrm{i}:\mathrm{g}^k_\mathrm{i}=s\}} err_{\{i+j\}}$$

$$\mathrm{d}\beta_l = \sum_{i=1}^{N} err_{i+l-L} \cdot \kappa_i$$

and

$$d\theta_0 = \sum_{i=1}^{N} err_i$$

4. Update the current parameters with stepsize α:

$$\theta^k = \theta^\mathrm{k} + \alpha\, \mathrm{d}\theta^\mathrm{k}$$

$$\beta_l = \beta_l + \alpha\, \mathrm{d}\beta_l$$

5. Update the constant offset:

$$\theta_0 = \theta_0 - \alpha\, d\theta_0$$

6. Apply the proximal operator for L_1 regularization

$$\theta^k_{\{s,j\}} = \begin{cases} \theta^k_{\{s,j\}} - \alpha\eta & if\ \left|\theta^k_{\{s,j\}}\right| > \alpha\eta \\ 0 & otherwise \end{cases}$$

7. Repeat until convergence.

The naïve proximal gradient descent suffers from three major problems: 1) computing λ is extremely slow, 2) propagating back the errors is slow, and 3) many iterations are required for convergence. We solve the first two problems with prefix compression and decompression. The third problem is solved with Nesterov's method and step-size selection heuristics.

a. Prefix compression

We can represent a 3-mer as the sum of the 4 possible 4-mers that prefix match the 3-mer. Using this fact, we can compress the parameter matrix to speed up the algorithm and substantially increase cache coherence, as there are fewer effective parameters in memory. In the regime where $N \gg 4^K \times 2M$, prefix compression will speed up the computation, as it needs to only be performed once per step.

To use prefix compression, we maintain a matrix ϕ of size $4^K \times 2M$, where $K = 8$, that represents only the longest k-mers. The first through fourth steps are modified to use ϕ instead of θ. Since every k-mer has a unique prefix match, the reduction maintains the correctness of the algorithm.

Before step 6, we perform a decompression step. Let $h(s, k)$ be the set-valued function consisting of all the k-mers whose first $k - 1$ characters match s. Then,

$$d\theta_s^k = \sum_{s' \in h(s,k)} d\phi_s'$$

and we can use dynamic programming to generate $d\theta^k$ for each k. The total runtime for the decompression is $O(4^K \times 2M)$.

After step 6, we re-compress the parameter matrix. Given a k-mer s, let $f(s,k)$ be set set-valued function that returns the k-character prefix of s. Then,

$$\phi_s = \sum_{k\in[1,K]} \theta^k_{f(s,k)}$$

which has a runtime of $O(4^K \times 2M \times K)$.

In addition to the prefix compression, we can represent k-mers as bit-strings where every two bits represents a base. This allows for the above operations to be done via bit-shifts and cache-coherence additions, which allows for fast encoding and decoding.

b. Improving batch gradient descent

Naïve gradient descent requires too many iterations for convergence in a reasonable timeframe. Thus, we improve upon the naïve gradient descent in two main ways: 1) Nesterov's method and 2) step-size selection heuristics. Nesterov's method has been shown to improve convergence in theory and in practice. Additionally, we pre-optimize the covariate parameters.

We use two heuristics to select the step size. The first heuristic is to prevent against selecting step sizes too large. If the function value increased by 30% in the first five iterations or 0.5% past the first five iterations, we backtrack to find a step size which decrease the function value. The second heuristic is to intelligently increase the step size. After five iterations of decreasing function value, the step size is increased.

Thus, in-between steps 3 and 4 of the previous algorithm, the following is added (in addition to step 4 being the Nesterov update instead of the normal gradient update):

1. If the function value increases by more than 30% in the first five iterations or 0.5% pas the first five iterations, backtrack.
2. If the function value has been monotonically decreasing for 5 iterations, increase the step size.

9. Data sources

All datasets used in this work were for human. The accession codes and short descriptions of each dataset are provided below.

Accession code	Description
GSM736567	Dermal Fibroblast DNase-seq
GSM736620	GM12878 DNase-seq
GSM935559	P300 GM12878 ChIP-seq
GSM935562	P300 GM12878 ChIP-seq
GSM935294	P300 GM12878 ChIP-seq
GSM1003505	Dermal Fibroblast H2AFZ ChIP-seq
GSM733753	Dermal Fibroblast H3K4me2 ChIP-seq
GSM733650	Dermal Fibroblast H3K4me3 ChIP-seq
GSM733709	Dermal Fibroblast H3K9ac ChIP-seq
GSM733662	Dermal Fibroblast H3K27ac ChIP-seq
GSM733767	GM12878 H2AFZ ChIP-seq
GSM733769	GM12878 H3K4me2 ChIP-seq
GSM733708	GM12878 H3K4me3 ChIP-seq
GSM733677	GM12878 H3K9ac ChIP-seq
GSM733771	GM12878 H3K27ac ChIP-seq
GSM945196	GM12878 H3K27me3 ChIP-seq
GSM945212	GM12878 H3K36me3 ChIP-seq

10. Parameters

For all CCM and CCM+ models, the maximum k-mer size was 8 and the k-mer effect window size was +/- 1000 bp. For all CCM+ models, the covariate window size was +/- 200 bp.

These parameters were chosen primarily due to memory constraints. For the covariate only models, the same +/- 200 bp window was used.

For p300 binding site discovery, GPS was run with 8 threads and the options "--v 2 --a 3", in addition to the recommended settings for human, for the predicted reads to ensure the first pass of the algorithm had a sufficient number of examples to update the read distribution. The recommended settings were used for the observed reads.

For histone peak discovery, MACS was used with default broad peak settings for the observed reads. As the predicted reads did not have strand information, the additional parameters "--extsize 147 --nomodel" were used to ensure MACS could bypass the shifting model building stage.

11. Acknowledgements

I would like to thank my supervisor, David Gifford, for his continued support through the majority of my time at MIT. Also in the Gifford lab, special thanks to Tatsunori Hashimoto for guiding me through my research and answering all my questions, no matter how trite, Matt Edwards for his insights as a senior graduate student, and Jeanne Darling and Patrice Macaluso for helping me navigate through the administrative parts of MIT. Special thanks to Richard Sherwood for his helpful insights on the biological aspects of my work, where I would otherwise have been lost.

Works cited

1. Hashimoto T, Sherwood RI, Kang DD, Rajagopal N, Barkal AA, Zeng H, et al. A Cooperative DNA Logic Predicts Genome-wide Chromatin Accessibility.

2. Weintraub H, Groudine M. Chromosomal subunits in active genes have an altered conformation. Science. 1976;193: 848–56. Available: http://www.ncbi.nlm.nih.gov/pubmed/948749

3. Soufi A, Donahue G, Zaret KS. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell. Elsevier; 2012;151: 994–1004. doi:10.1016/j.cell.2012.09.045

4. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol. 2014;32: 171–178. doi:10.1038/nbt.2798

5. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–5. doi:10.1126/science.1222794

6. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput Biol. Public Library of Science; 2014;10: e1003711. doi:10.1371/journal.pcbi.1003711

7. Megraw M, Pereira F, Jensen ST, Ohler U, Hatzigeorgiou AG. A transcription factor affinity-based code for mammalian transcription initiation. Genome Res. 2009;19: 644–56. doi:10.1101/gr.085449.108

8. Schultheiss SJ, Busch W, Lohmann JU, Kohlbacher O, Rätsch G. KIRMES: kernel-based identification of regulatory modules in euchromatic sequences. Bioinformatics. 2009;25: 2126–33. doi:10.1093/bioinformatics/btp278

9. McCullagh P, Nelder JA. Generalized Linear Models, Second Edition [Internet]. CRC Press; 1989. Available: https://books.google.com/books?hl=en&lr=&id=h9kFH2_FfBkC&pgis=1

10. Kasper LH, Fukuyama T, Biesen MA, Boussouar F, Tong C, de Pauw A, et al. Conditional knockout mice reveal distinct functions for the global transcriptional coactivators CBP and p300 in T-cell development. Mol Cell Biol. 2006;26: 789–809. doi:10.1128/MCB.26.3.789-809.2006

11. Vo N, Goodman RH. CREB-binding protein and p300 in transcriptional regulation. J Biol Chem. 2001;276: 13505–8. doi:10.1074/jbc.R000025200

12. Goodman RH, Smolik S. CBP/p300 in cell growth, transformation, and development. Genes & Dev. 2000;14: 1553–1577. doi:10.1101/gad.14.13.1553

13. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. Nature Publishing Group; 2007;39: 311–8. doi:10.1038/ng1966

14. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447: 799–816. doi:10.1038/nature05874

15. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 2011;21: 2167–80. doi:10.1101/gr.121905.111

16. Won K-J, Ren B, Wang W. Genome-wide prediction of transcription factor binding sites using an integrated model. Genome Biol. 2010;11: R7. doi:10.1186/gb-2010-11-1-r7

17. Guo Y, Mahony S, Gifford DK. High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. PLoS Comput Biol. 2012;8: e1002638. doi:10.1371/journal.pcbi.1002638

18. Kähärä J, Lähdesmäki H. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. Bioinformatics. 2015; doi:10.1093/bioinformatics/btv294

19. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27: 1017–8. doi:10.1093/bioinformatics/btr064

20. Ha M, Hong S, Li W-H. Predicting the probability of H3K4me3 occupation at a base pair from the genome sequence context. Bioinformatics. 2013;29: 1199–205. doi:10.1093/bioinformatics/btt126

21. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nat Methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;12: 265–272. doi:10.1038/nmeth.3065

22. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;12: 7–18. doi:10.1038/nrg2905

23. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007;128: 707–19. doi:10.1016/j.cell.2007.01.015

24. Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

25. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137. doi:10.1186/gb-2008-9-9-r137

26. Lee JT, Bartolomei MS. X-inactivation, imprinting, and long noncoding RNAs in health and disease. Cell. Elsevier; 2013;152: 1308–23. doi:10.1016/j.cell.2013.02.016

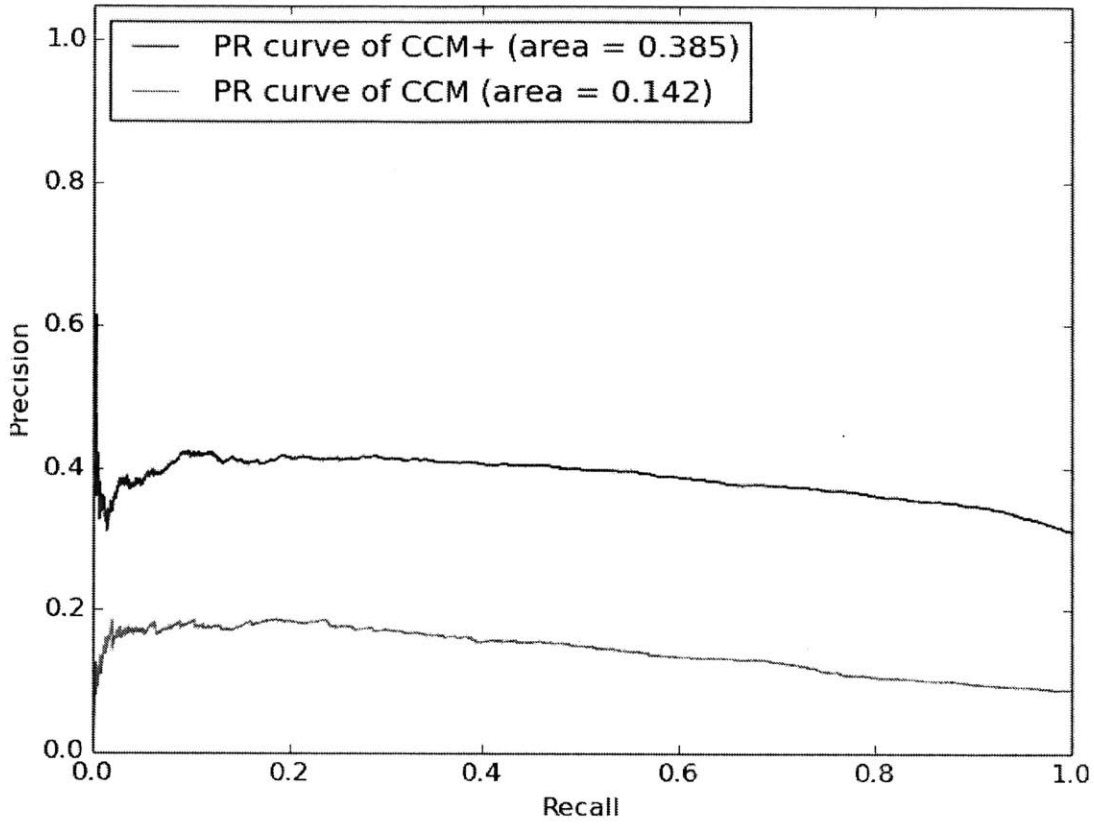

Figure 1. Precision-recall curves for the predicted p300 calls for CCM and CCM+. The CCM and CCM+ models for p300 were used to predict p300 binding sites with GPS. Predicted binding sites within 300 bp of the observed binding sites were called positive. CCM+ has a higher precision for every recall level. Only calls from chromosomes 14-22 were used to prevent overfitting errors.

Histone	**Correlation (CCM)**	**Correlation (CCM+)**
H2A.Z	0.46	0.47
H3K4me2	0.08	0.09
H3K4me3	0.56	0.58
H3K9ac	0.42	0.38
H3K27ac	0.16	0.18

Table 1. Correlation values for CCM and CCM+ predicted rates against the observed histone mark reads. The CCM and CCM+ histone mark models were used to generate predicted rates over chromosome 14. The correlation was measured over chromosome 14 with a smoothing window size of 2 kb.

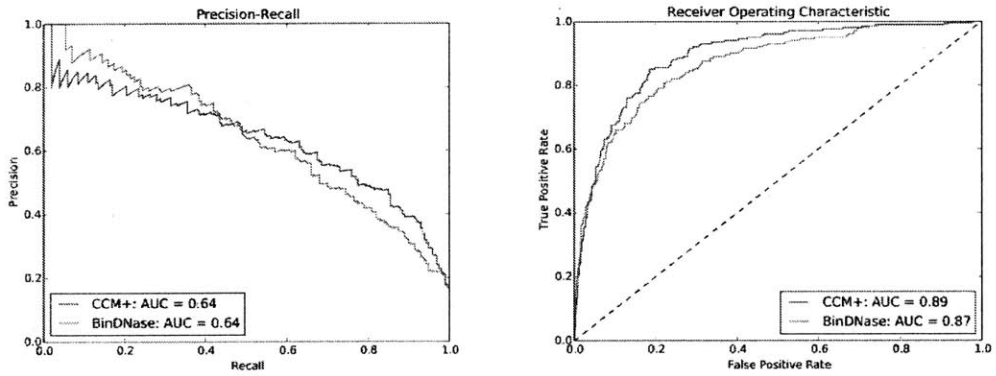

Figure 2. Precision-recall and receiver operating characteristic curves for BinDNase and CCM+ on predicting p300 binding. BinDNase and CCM+ were trained on the same number of sites from chromosomes 1-13 and all p300 motif hits from chromosomes 14-22 were used for testing. As seen, BinDNase and CCM+ perform approximately equally, while CCM+ has a much simpler model.

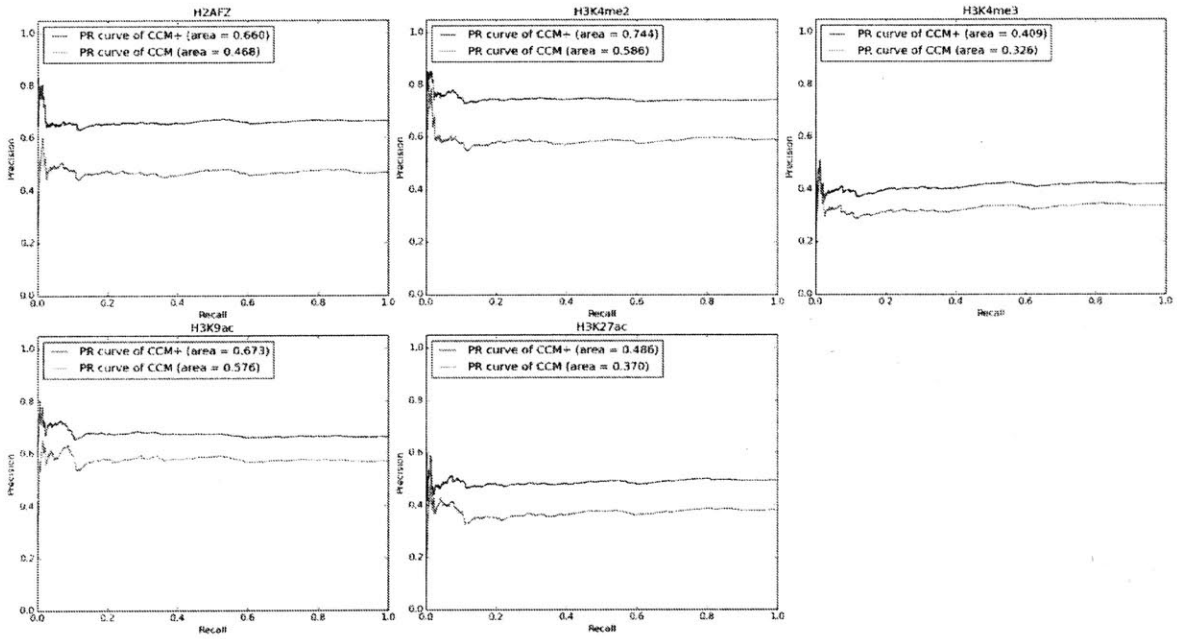

Figure 3. Precision-recall curves for the predicted histone mark calls for CCM and CCM+. The CCM and CCM+ models for histone marks were used to predict histone mark binding with MACS. Predicted binding sites that overlapped with an observed binding site were called positive. CCM+ has a higher precision for every recall level and histone mark. Only calls from chromosomes 14-22 were used to prevent overfitting errors.

Histone	aucPR (CCM)	aucPR (CCM+)
H2A.Z	0.47	0.66
H3K4me2	0.59	0.74
H3K4me3	0.33	0.41
H3K9ac	0.58	0.67
H3K27ac	0.37	0.49

Table 2. Area under the precision-recall curves (aucPR) for histone marks. The precision-recall curves were generated for chromosomes 1-14 as described and the aucPR was computed for each histone mark.

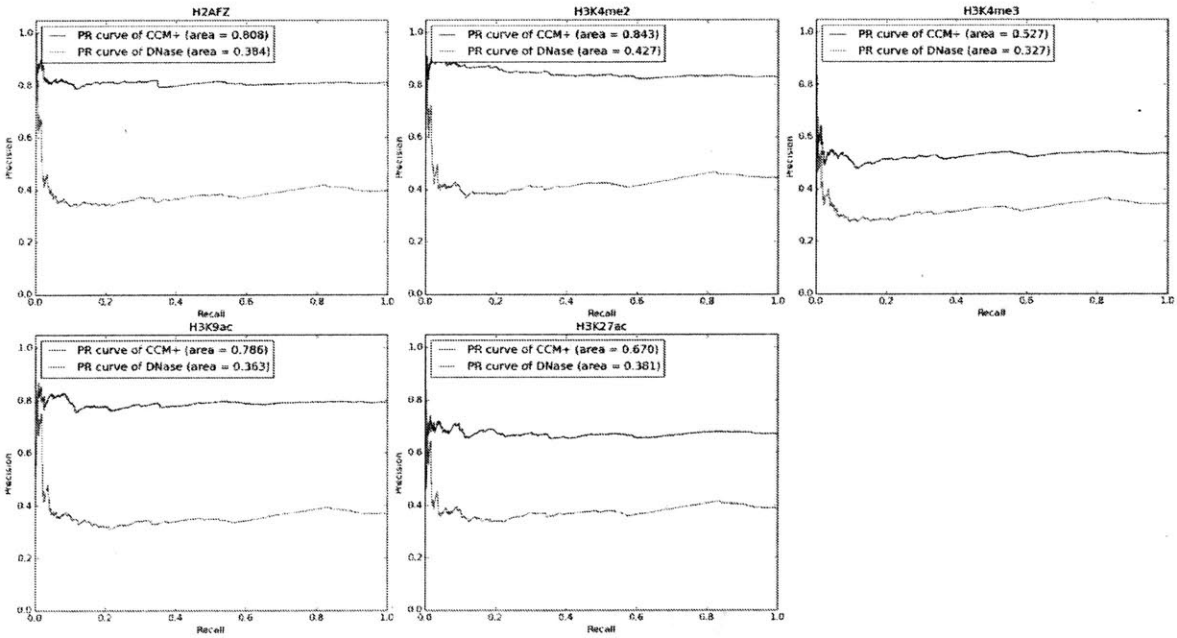

Figure 4. Precision-recall curves for transfer learning with histone marks. CCM+ GM12878 parameters were used with the dermal fibroblast DNase-seq covariate to generate dermal fibroblast predicted histone peaks. The differentials for the predicted and observed peaks on chromosomes 14-22 were used to generate the precision-recall curves.

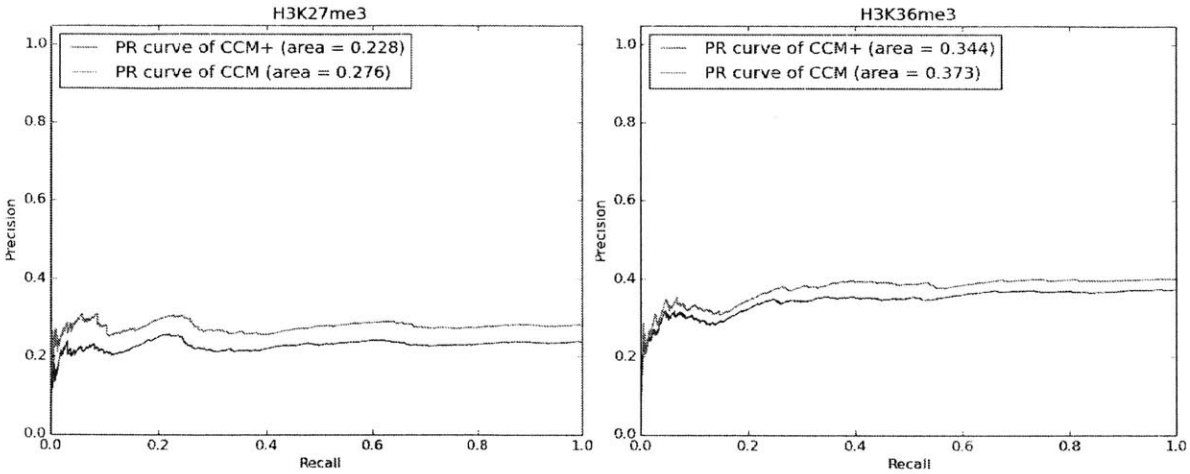

Figure 5. Precision-recall curves for histone marks H3K27me3 and H3K36me3. These marks are denoted "region" by ENCODE, which indicates they are trans-regulated. Both CCM and CCM+ cannot predict these marks as well as the "peak" marks, as denoted by ENCODE.

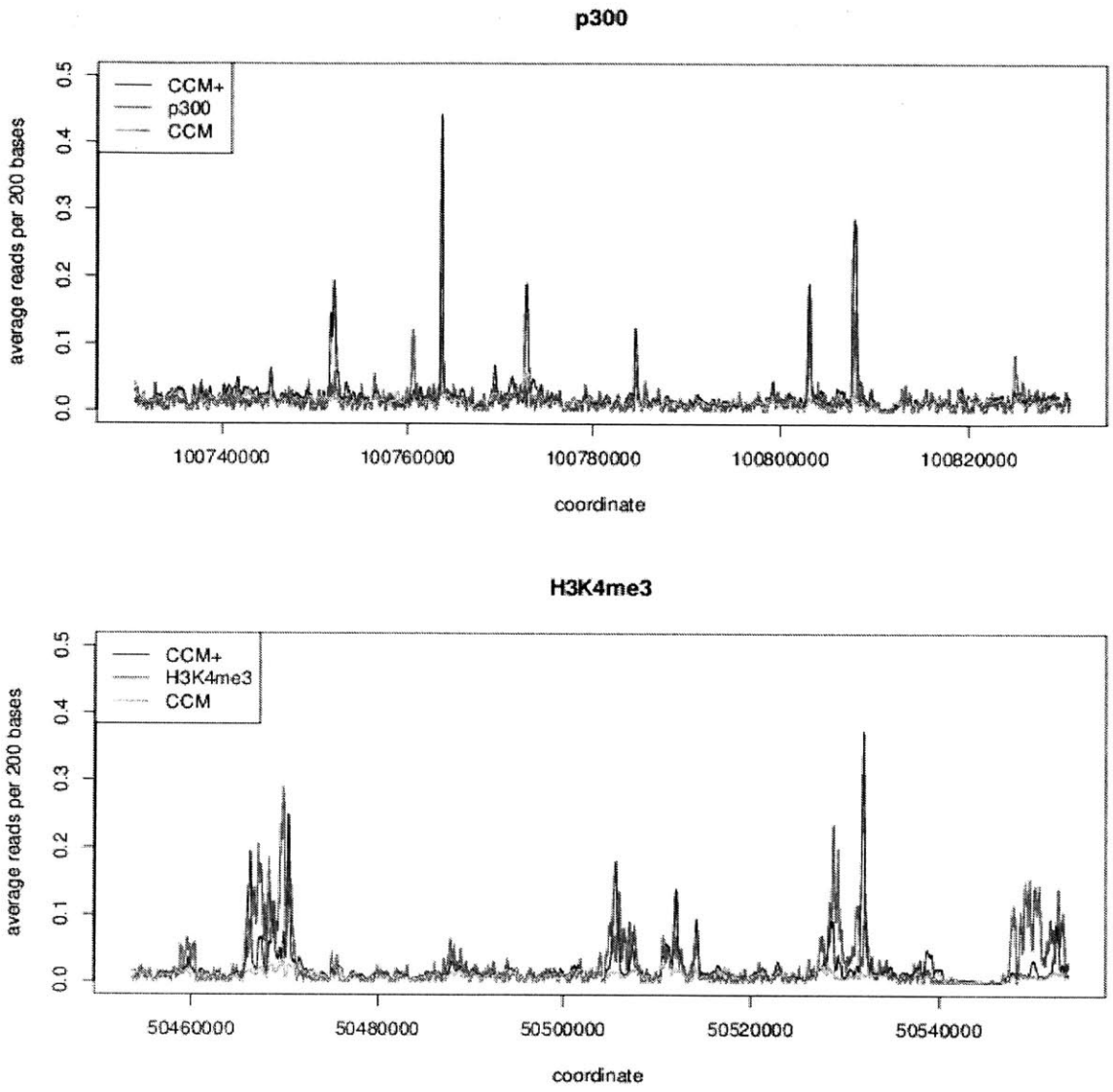

Figure 6. Example GM12878 held-out genomic region showing the observed reads (red), CCM+ predicted reads (black), and CCM predicted reads (green), all smoothed to 200 bp.

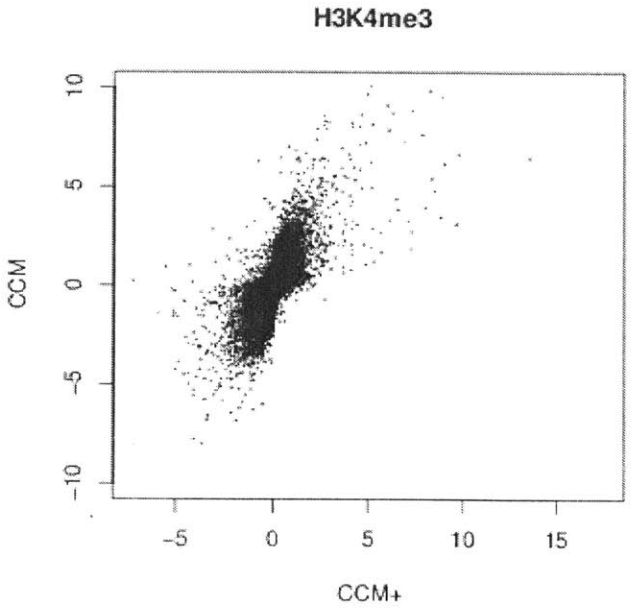

Figure 7. Summed k-mer effect sizes for each k-mer in CCM+ (x-axis) vs CCM (y-axis) for GM12878 H3K4me3. While there appears to be some correlation, many k-mers are different in the CCM and CCM+ models.

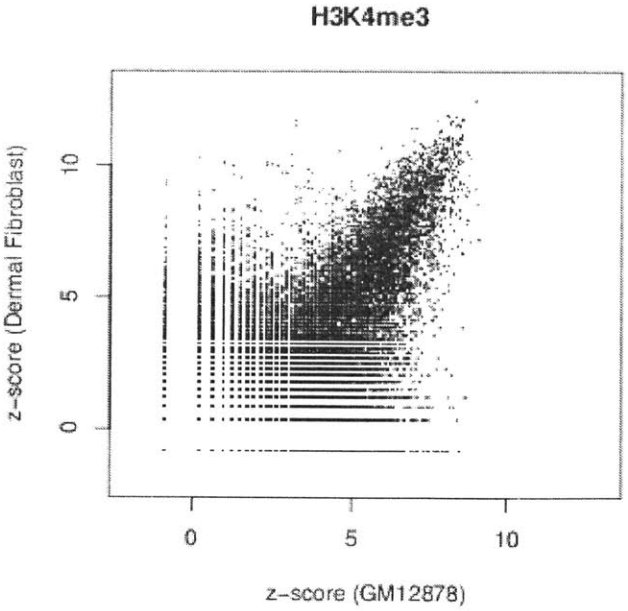

Figure 8. Comparison of observed GM12878 (x-axis) and observed Dermal Fibroblast (y-axis) reads in 2 kb binned regions of GM12878 H3K4me3 held-out chromosome 14. We note the R^2 is only 0.59.

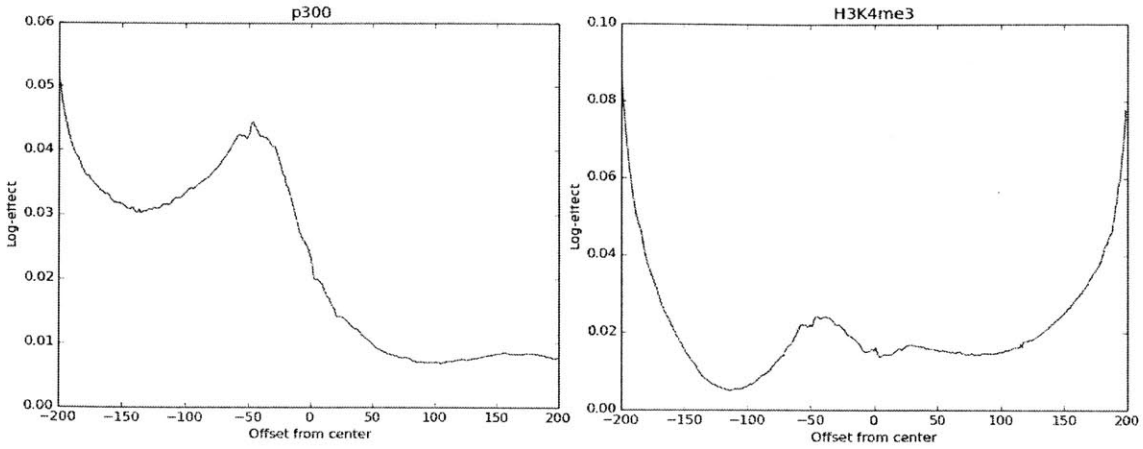

Figure 9. The spatial effect of the optimal covariate parameters for the GM12878 p300 and H3K4me3 models. The model is trained on the forward strand, so there is a lack of symmetry in the spatial effect.