Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yutong Jin Date

Systematic Evaluation of The Impact of DNA Sequence Variants on The in vivo Binding Affinity at Transcription Factor Binding Sites

By

Yutong Jin

Master of Science in Public Health

Biostatistics and Bioinformatics

Zhaohui (Steve) Qin, PhD, MS

(Thesis Advisor)

Tianwei Yu, PhD

(Reader)

Systematic Evaluation of The Impact of DNA Sequence Variants on The in vivo Binding Affinity at Transcription Factor Binding Sites

By

Yutong Jin

Bachelor of Medicine

Fudan University

2016

Thesis Committee Chair: Zhaohui (Steve) Qin, PhD, MS

Reader: Tianwei Yu, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics and Bioinformatics

2018

Abstract

Systematic Evaluation of The Impact of DNA Sequence Variants on The in vivo Binding Affinity at Transcription Factor Binding Sites

By Yutong Jin

The majority of the single nucleotide variants (SNVs) identified by genome-wide association studies (GWASs) fall outside of the protein-coding regions. Elucidating the functional implications of these variants has been a major challenge due to the lack of functional annotation in the non-coding part of the genome. A possible mechanism for some of the functional non-coding variants is that they disrupted the canonical transcription factor (TF) binding sites which affect the *in vivo* binding affinity of the TF. However, not all variants located within TF binding sites will impact TF binding since a substantial proportion of most TF binding motifs is not well conserved. Therefore, simply annotate all variants located in putative TF binding sites is not ideal. In this project, we conducted a comprehensive survey to study the effect of SNVs on the TF binding affinity. Using CTCF as an example, we found that mutations occur at about 30% of positions inside a putative CTCF binding motif sites will likely to have significant effect on the TF-DNA binding. Our results provide key guidance on annotating variants in terms of the impact of TF binding.

Table of Contents

1. Introduction

Thousands of genome-wide association studies (GWAS) have been conducted in the past decades and resulted in tens of thousands of single nucleotide variants (SNVs) being discovered as robustly associated with a wide array of phenotypes (Welter et al., 2013). The vast majority of disease- and trait-associated variants detected by these studies lie within the noncoding part of the human genome (Maurano et al., 2012) and are hypothesized to play a regulatory role to control gene expression of genes important for disease pathogenesis. In particular, whereas it has been demonstrated that GWAS-identified variants are enriched in regulatory regions (Cookson et al., 2009).

A possible mechanism for non-coding GWAS variants is the alteration of short DNA motif influence the *in vivo* binding affinity of transcription factors (TFs) to the regulatory elements (promoters, enhancers, and boundary elements) (Pasquali et al., 2012). A common practice to study TF binding motif patterns (Hertz & Stormo, 1999; Lawrence et al., 1993) is to use the position weight matrices (PWMs) to model the aligned short DNA sequences in terms of the frequencies of the four types of nucleotides at each position. By scanning the genome, such matrices assign a matching score to each position of the candidate sequences, and only sequences with a score exceeding a predefined threshold are considered as a putative binding site. Subsequently, all mutations found within such binding sites are considered detrimental and have strong functional impacts (Boyle et al., 2012; Ward & Kellis, 2011). However, it is well known that many positions within TF binding motifs are not well conserved and the impact of a mutation at these positions are not well understood. It is likely that mutations found within motifs may have various levels of functional impact (Vaquerizas et al., 2009). Moreover, using the overall PWM probability difference for a motif instance with or without a mutation may not be a good way to quantify the

effect of the mutation since the PWMs assume independence among all positions. Hence, we need more information beyond the PWM level to quantify the impact of variants such as SNVs within motifs.

In this work, we conducted a comprehensive survey on the TF binding impact of SNVs. We employ established methodology of gapped k-mer support vector machine (gkm-SVM) to determine the cell type-dependent binding strength, then quantify the effect of sequence changes between every pair of motifs by the deltaSVM method (Lee et al., 2015). We believe the affinity scores of each position in all putative motif sites in the genome can serve as a useful resource to study and quantify the position-specified impact of a mutation inside a motif site.

Using CTCF chromatin immunoprecipitation and sequencing (ChIP-Seq) (Johnson et al., 2007) data on the Encyclopedia of DNA Elements project (ENCODE Project Consortium, 2012), we measure the preference among all the 10-mers for CTCF in GM12878, H1-hESC and K562 cells. We then conduct a comprehensive survey of the impact of all mutations occurring within all putative CTCF binding sites across the genome. Our results show that the majority of these mutations will have limited impact on the binding affinity and suggest more detailed elucidation of the functional impact of SNVs is needed to more accurately annotate these SNVs. Our findings are consistent across all three representative cell lines.

2. Materials and Methods

2.1 Overview

Gkm-SVM classifier (Ghandi et al., 2014) requires a positive training set of putative regulatory sequences and a negative training set of non-regulatory sequences to produce a scoring function characterized by weights on a list of all possible 10-mers. Given training sets, gkm-SVM computes the contribution of each DNA sequence (10-mers) feature to the prediction of regulatory function. Accordingly, each 10-mer will receive a SVM weight that quantifies its contribution. After training, the SVM weights can be used to measure the predicted impacts of any possible SNV on the regularity activity in a particular cell type via the delta-SVM method, which sums the difference in weights between alleles for each of the TF binding sites-associated 10-mers encompassing the SNV. (Lee et al., 2015). Here, the delta-SVM score obtained by the delta-SVM method is the difference between the sum of all 19 10-mers weights in reference alleles and that in alternative alleles across the affected bases. A significant difference indicates high impact of the SNV in terms of the binding affinity.

2.2 Statistical Analysis Methods

2.2.1 Comparing correlation between difference in PWM score and weights

We trained a gkm-SVM, proposed by Lee, et al. (2015), by applying their method on ENCODE ChIP-Seq data. The regions within called peaks were treated as the positive training set and the regions outside called peaks as the negative training set to estimate 10-mers weights. Different sets of 10-mers weights were constructed for each human cell line.

We ranked all 10-mers based on their weights and selected 1000 unique 10-mers that have the highest weights as our study sequences. Since the motif length was 15 base pairs (bps), there were six possible alignments of the 10-mers inside a motif, we first assessed the probabilities according

to the PWM and assigned the largest one among all six potential matching alignments as the probabilistic value of the specific 10-mer, named as PWM scores. Then we used Pearson correlation to measure the correlation between PWM scores and weights among all 10-mers. The results were visualized by scatter plots.

We also measured the differences between log-transformed probabilistic values among 10-mers with the other three variants on each particular position and their corresponding reference 10-mer sequences, named delta-PWM. To further demonstrate the correlation between probability difference calculated from the PWM and weights difference, we employed Pearson correlation between differences in probabilities and differences in weights for the 1000 top ranked 10mers. The results were visualized by scatter plots.

We visualized the top 100 and 1000 10-mers (based on weight) averaged change in weights among 30 variants for each sequence and their corresponding confidence interval with one standard deviation.

2.2.2 Using delta-SVM score to determine the impact of variations on each particular position

Using the CTCF motif PWM (Xu et al., 2015), we scanned the human reference genome GRCh37 (UCSC version: hg19) to determine all potential positions of binding motifs. There were 139,084 sites in the genome and 48,804 unique motif sequences were identified.

In the set of all potential 15-bp motifs, we ranked and retrieved the first 20 most frequent motif sequences for cell line GM12878 and the first ten most frequent motif sequences for cell line H1-hESC and K562, respectively. To detect the difference between motif sites and their adjacent regions, we extended each 15-bp motifs by 10-bp on both sides.

By using the set of all 10-mer weights, the delta-SVM was estimated by summing ten 10-mer weights across 15-bp motif region with one mutated position in the middle. Based on this method for calculating delta-SVM score, we could then quantify the changes in binding affinity of the motif sequence due to the variants. Furthermore, we assessed delta-PWM and assigned the mean value of them as the probabilistic value of each position successively within a particular motif.

2.2.3 Determine the threshold value to identify significant SNPs

In order to determine the threshold value to identify significant SNPs, we generated 10,000 random sequences and excluded the sequence overlapping with motif regions to be our control set. Repeat the work using random control sequences to validate the difference above. Since negative delta-SVM value associated with large impact of binding ability, we used the 5th percentile of the empirical distribution from the left tail as the threshold to identify significant SNPs. To demonstrate the difference between delta-SVM and delta-PWM for motif and adjacent region, the results were visualized at the same scale.

3. Result

We performed data analysis on three human cell lines: GM12878, H1-hESC and K562, and the results for GM12878 were shown in this thesis. Results from the other two cell lines were presented in Supplemental materials.

The correlations between 10-mer PWM scores and weights are 0.620, 0.605 and 0.614 for GM12878, H1-hESC and K562 respectively. The correlations between probability change and weights difference are 0.507, 0.505 and 0.512 for GM12878, H1-hESC and K562 respectively. There is a pronounced positive relationship between weights difference and probability difference. However, the weak correlation indicates that the purely probabilistic calculation based on PWM is not well-performed in presenting the importance of 10-mers (Figure 1).

We next surveyed the weights of top-ranked 10-mers and their corresponding one-standard-deviation confidence intervals. (Figure 2-3) As expected, the mean change in weights among 30 variants of top 100 sequences exceeds our threshold value. The trend of GM12878 shows that SNVs within high-ranked 10-mers would also lead to a large weight difference, which means the change within CTCF motif were more likely to indicate a significant impact of its binding ability.

To understand whether SNVs at each particular position affect the binding affinity of the TF differently, we retrieved the 20 most frequent unique motifs (Figure 4) and extended to flanking region by 10 bps on each side. We also provided a graphical assessment demonstrating the distribution of the mean delta-SVM of 35-bp sequences and mean delta-PWM of 15-bp motif regions (Figure 5). Combining three variants together, mutations at different position lead to different scores. We find the observed distribution of the mean delta-SVMs are not always consistent with the distribution of delta-PWMs, hence our comprehensive annotation method will

provide more accurate assessment of binding affinity of putative motif sites in the genome. For example, the 4th most frequent motif sequence "GCACCACCTAGTGGA" shows a highly conserved performance among all positions within regulatory region. To further shed light to the prediction on different motifs, we employed "logo plots" (Schneider & Stephens, 1990) to demonstrate the significant SNVs and found about 30% of positions inside a putative binding motif site had a large effect on the TF-DNA binding affinity based on our pre-determined threshold. (Figure 6) For cell line GM12878, the positive prediction rate among top 20 motifs with 300 positions is 29.0%.

4. Discussion

A significant challenge in molecular genetics is to predict the impact of variants on TF binding and to understand how genetic variation alters genetic regulation. While changes are frequently observed in binding affinity of the transcription factor for altered variants within the canonical transcription factor, it is always ignored by adopting classic PWM-based methods to capture the position-specified traits. Assuming independence for PWM, the simply impact determination based on purely probability calculation may cause inaccuracy. Our results suggest that positions inside binding motif sites should be considered when determine the impact on TF-DNA binding. By embracing flanking region in quantifying the effect of SNV, our method incorporates all alterations for particular sequence to capture the impact of each positon.

Admittedly, the predictive power of our method is inherently limited. Although the machine learning approaches learns directly from the determined sequences, any training data sets constructed on PWM collections will probably be skewed and with many false negatives since the current incomplete understanding of most TF binding motifs. Future advances in motif

identification will enhance the liability of the prediction on the TF binding affinity. Besides, we identify the averaged delta-SVM among all three potential variants to represent the position-specified impact. However, we find that for some particular positions, the delta-SVM score for three variants vary significantly. It indicates the impact of different variants for the same position within the motif may also have disparate impact. In the future, we could expand current simplified quantification method to a more comprehensive way, such as quantifying variant-specified impact of all positions within cell-specific motifs.

5. Reference

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & Parkinson, H. (2013). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(D1), D1001-D1006.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., & Shafer, A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 1222794.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3), 184.

Pasquali, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, İ., Tena, J.J., Morán, I., Gómez-Marín, C., Van De Bunt, M., & Ponsa-Cobas, J. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature genetics*, 46(2), 136.

Hertz, G. Z., & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, 15(7), 563-577.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *science*, 262(5131), 208-214.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., & Cherry, J. M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, 22(9), 1790-1797.

Ward, L. D., & Kellis, M. (2011). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, 40(D1), D930-D934.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4), 252.

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature genetics*, 47(8), 955.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830), 1497-1502.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57.

Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology*, 10(7), e1003711.

Xu, T., Li, B., Zhao, M., Szulwach, K.E., Street, R.C., Lin, L., Yao, B., Zhang, F., Jin, P., Wu, H., & Qin, Z. S. (2015). Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic acids research*, 43(5), 2757-2766.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20), 6097-6100.

6. Tables and Figures

Figure 1 The Correlation Between Probability Difference and Weights Difference – GM12878

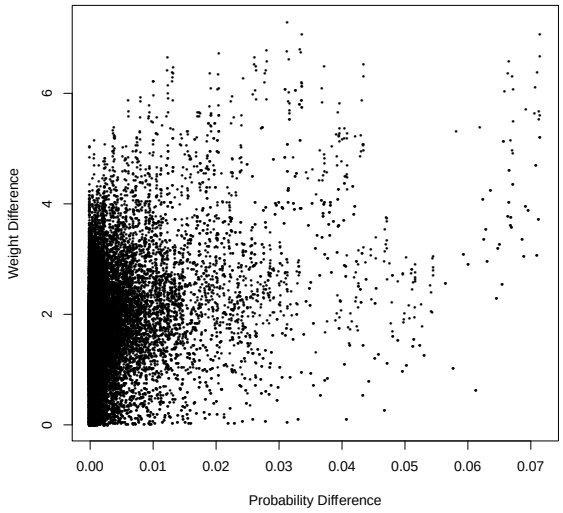

Figure 2 Average Delta-SVM Scores and One-standard-deviation Confidence Intervals for Topped 100 10-mers – GM12878

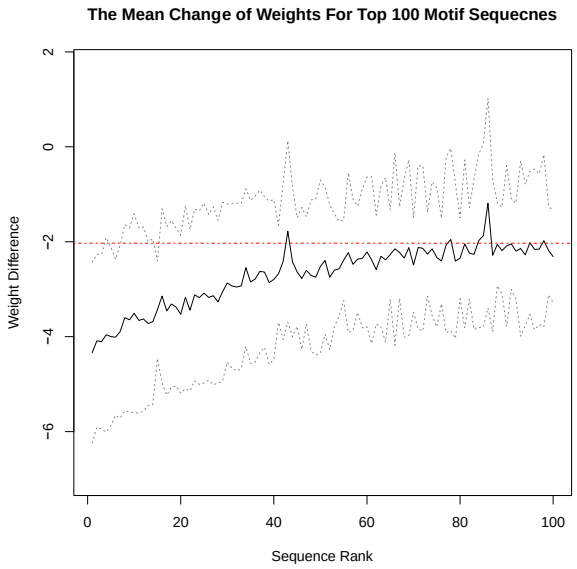

Figure 3 Average Delta-SVM Scores and One-standard-deviation Confidence Intervals for Topped 1000 10-mers – GM12878

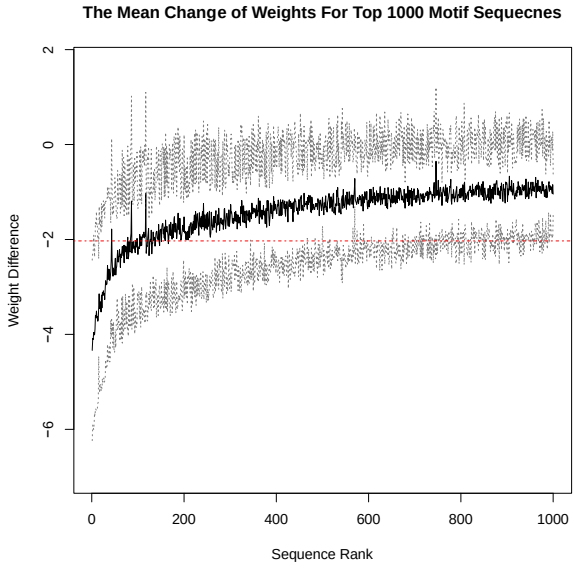

Figure 4 The 20 Most Frequent Unique Motifs

Log-sum	Sequence	Frequency
-11.806	CCTCCCCCTCCTGGG	561
-10.657	CCTCCACCTCCTGGC	493
-10.904	CCTCCACCTGCTGGG	387
-10.323	GCACCACCTAGTGGA	377
-11.523	CCGCCCCCTGGGGTT	316
-8.738	GTGCCACCTCCTGGC	288
-10.632	GTGTCTCCTGGTGGC	287
-12.207	GCTCCCTCTGCTTGC	284
-12.796	GTTCCTCCTGGTGGG	255
-10.445	GCACCACCTCCTGGC	250
-12.451	TCGACCCCTGCTGGG	237
-12.796	GCTCCCTCTGAAGGC	221
-8.423	GAGCCCTCTGGTGGC	197
-12.389	CCGCCGCCTCCTGGG	139
-11.635	CAGCCCCCTGCTGCC	138
-13.739	TCGCCCTCTGTGGGC	131
-7.991	GCGCCACCTCCTGGC	127
-10.65	CCGCCACCTCCTGGG	119
-11.921	TTGCCTTCTGCTGGC	118
-11.214	CCGCCCCCAGCTGGG	117

Figure 5 Mean Delta-SVM of Each Position for topped 4th motifs – GM12878

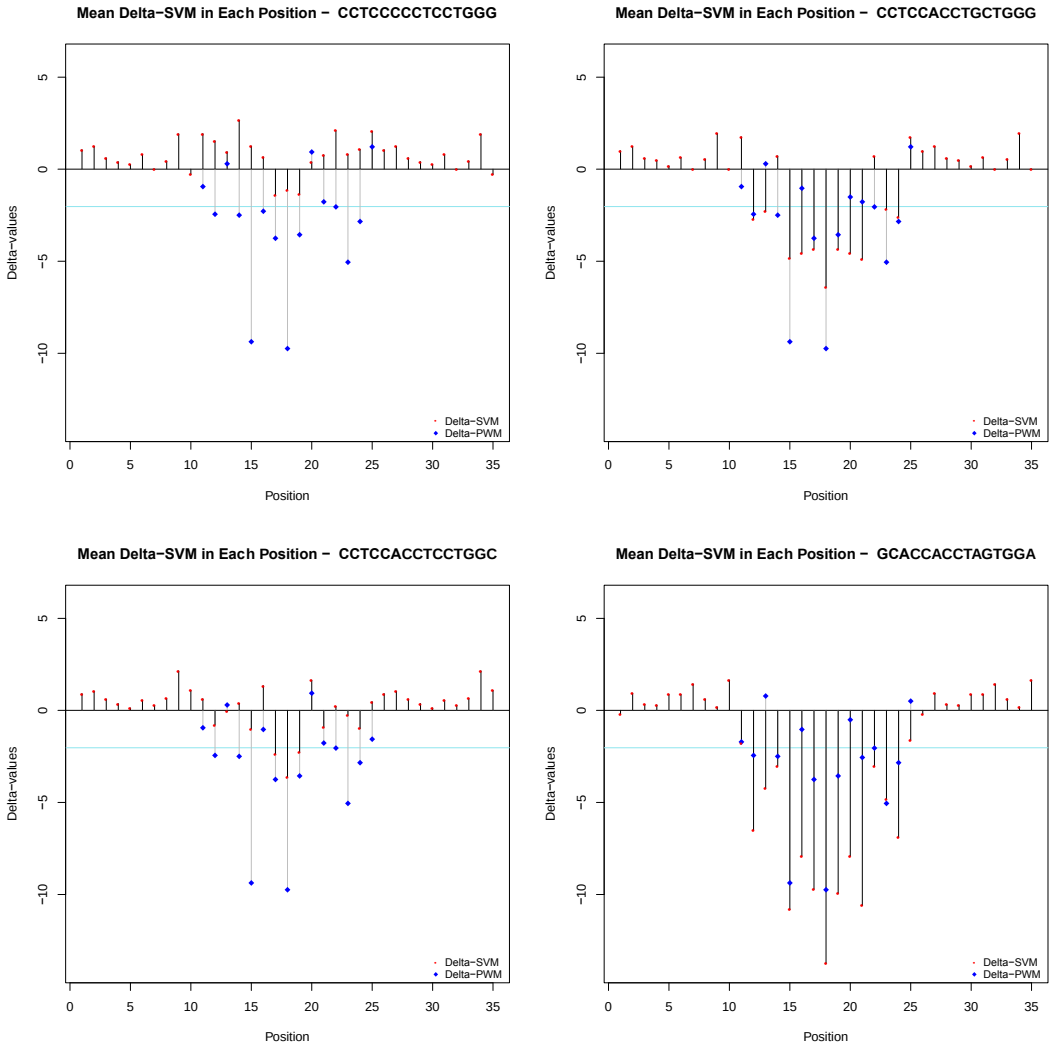

Figure 6 Prediction on Top 20 15-bp Motifs Using Means of Delta-SVM – GM12878

7.Supplementary Materials

Figure 7 The Correlation Between Probability Difference and Weights Difference – H1-hESC

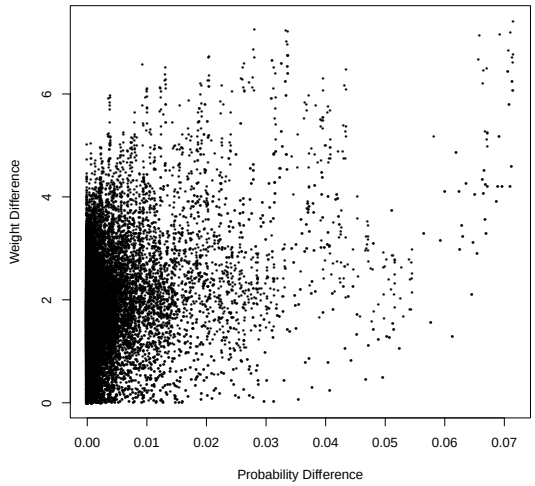

Figure 8 Average Delta-SVM Scores and One-standard-deviation Confidence Intervals for Topped 100 10-mers – H1-hESC

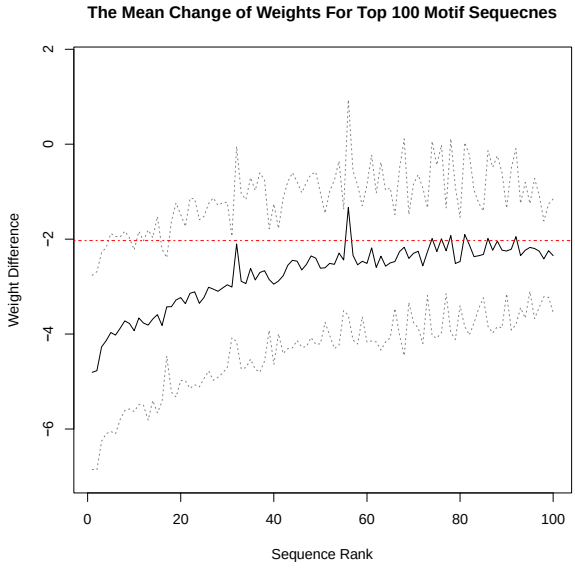

Figure 9 Average Delta-SVM Scores and One-standard-deviation Confidence Intervals for Topped 1000 10-mers – H1-hESC

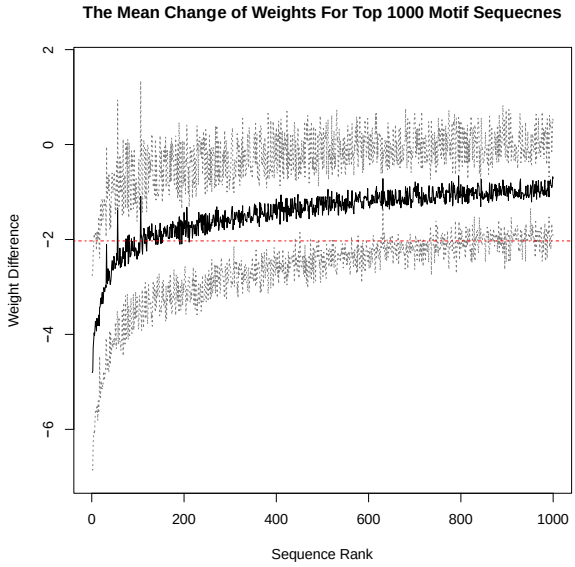

Figure 10 Mean Delta-SVM of Each Position for topped 4th motifs – H1-hESC

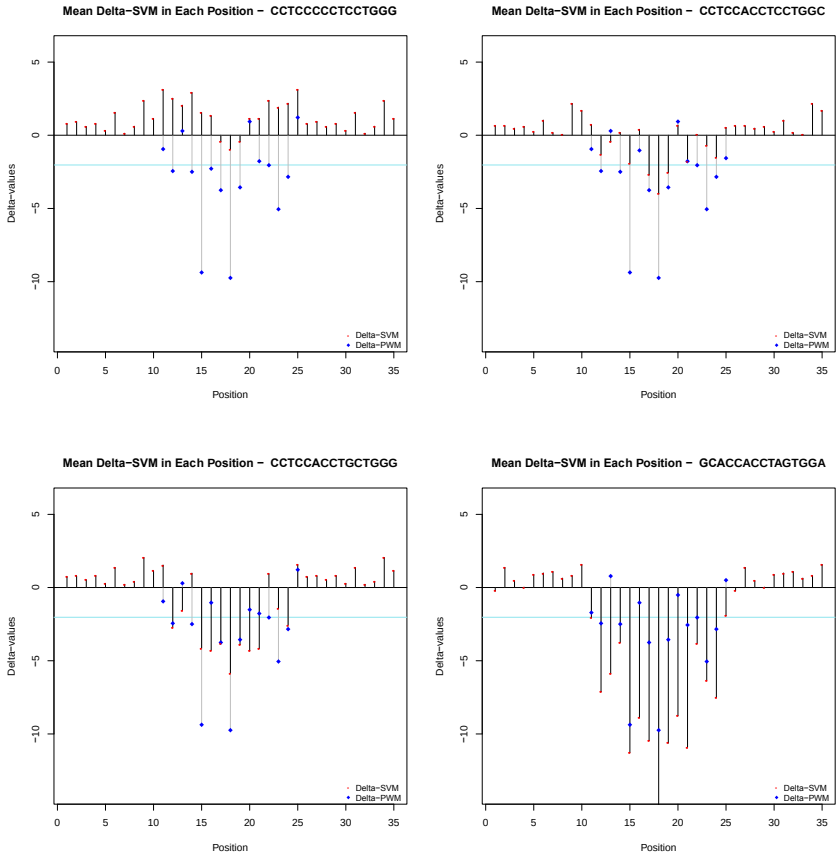

Figure 11 Prediction on Top 10 15-bp Motifs Using Means of Delta-SVM – H1-hESC

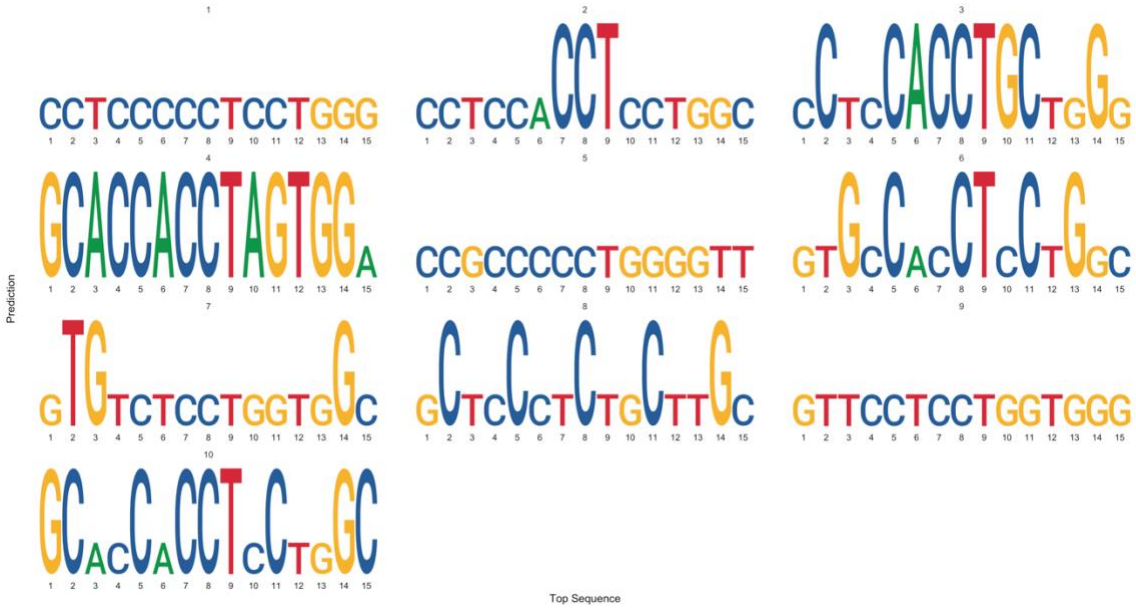

Figure 12 The Correlation Between Probability Difference and Weights Difference – K562

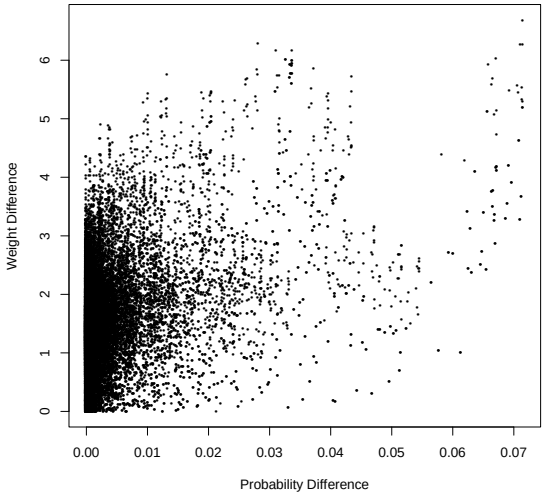

Figure 13 Average Delta-SVM Scores and One-standard-deviation Confidence Intervals for Topped 100 10-mers – K562

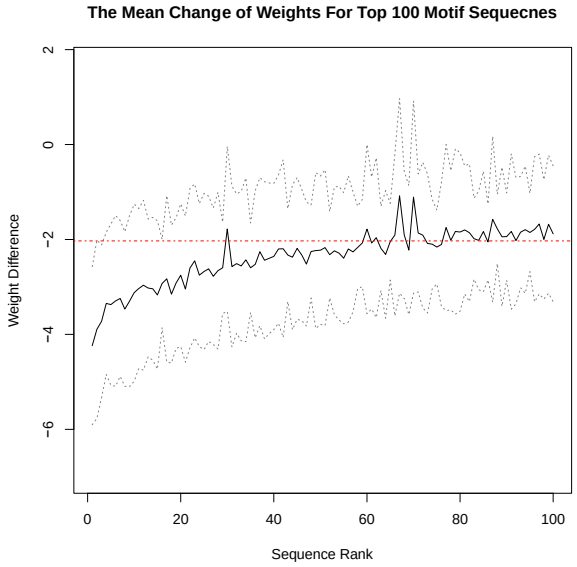

Figure 14 Average Delta-SVM Scores and One-standard-deviation Confidence Intervals for Topped 1000 10-mers – K562

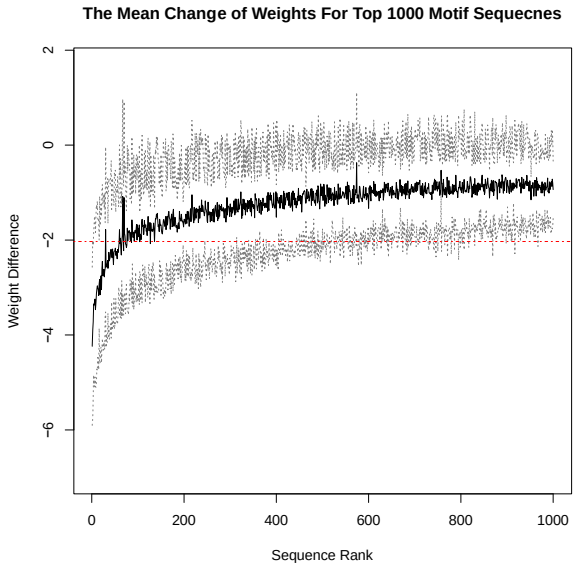

Figure 15 Mean Delta-SVM of Each Position for topped 4th motifs – K562

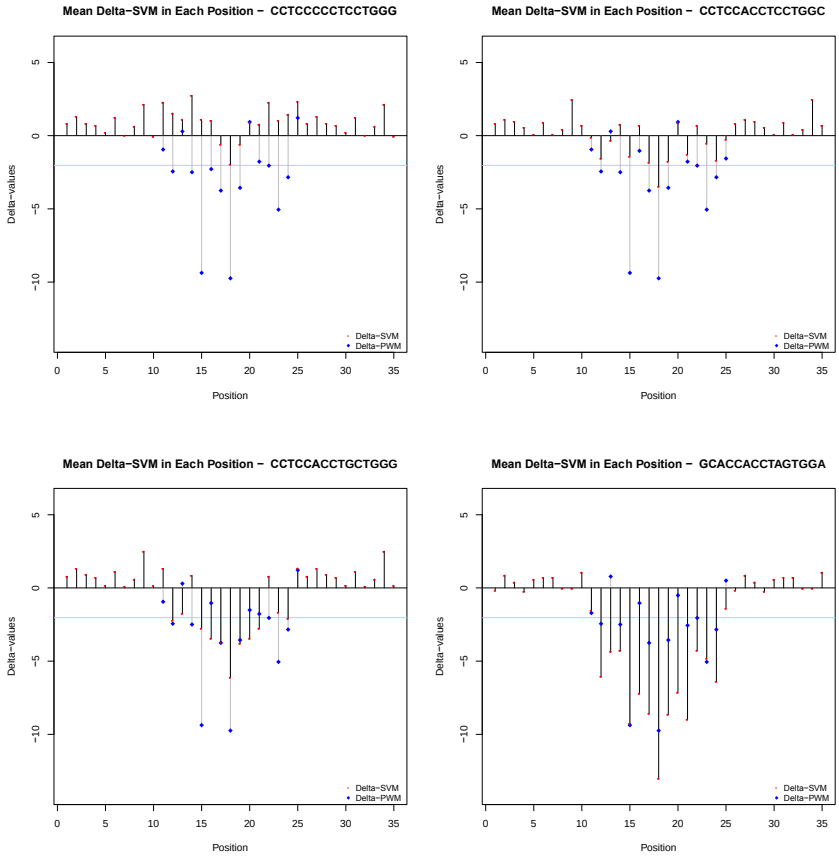

Figure 16 Prediction on Top 10 15-bp Motifs Using Means of Delta-SVM – H1-hESC