Distribution Agreement

Signature:

______________________________ ______________

Jiahui Jiang Date

Evaluation of the impact of DNA Sequence Variations on in vivo Transcription Factor Binding Affinity

By

Jiahui Jiang

Master of Public Health

Biostatistics and Bioinformatics

____

Zhaohui (Steve) Qin, PhD

(Thesis Advisor)

____

Xiangqin Cui, PhD

(Reader)

Evaluation of the impact of DNA Sequence Variations on in vivo Transcription Factor Binding Affinity

By

Jiahui Jiang

B.A.

Zhejiang University

2017

Thesis Advisor: Zhaohui (Steve) Qin, PhD

Reader: Xiangqin Cui, 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 Public Health
in Biostatistics and Bioinformatics
2021

Abstract

Evaluation of the impact of DNA Sequence Variations on in vivo Transcription Factor Binding Affinity

By Jiahui Jiang

Genome-wide association studies (GWASs) identified huge amounts of single nucleotide variants (SNVs) and thousands of SNVs within non-coding regions have associations with complex diseases. However, how non-coding SNVs specifically affect diseases is not clear yet. Recently, the number of studies focusing on the impact of these SNVs are increasing rapidly. A possible mechanism is that some non-coding SNVs can alter regulatory elements such as disrupting transcription factor (TF) binding sites, leading to the change of gene expression which result in diseases. Traditionally, it is assumed that SNVs within TF binding sites will impact the TF binding. However, increasing studies show that not all SNVs contribute to the TF binding since most TF binding motifs are not well conserved. Therefore, more information is needed to annotate SNVs within TF binding sites. In this study, we conducted a comprehensive survey to quantify the impact of SNVs on TF binding affinity using a creative sequence-based machine learning method. We found that only 20% SNVs within putative TF binding sites would be possible to significantly impact the *in vivo* TF binding.

Evaluation of the impact of DNA Sequence Variations on in vivo Transcription Factor Binding Affinity

By

Jiahui Jiang

B.A.

Zhejiang University

2017

Thesis Advisor: Zhaohui (Steve) Qin, PhD

Reader: Xiangqin Cui, 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 Public Health
in Biostatistics and Bioinformatics
2021

Acknowledgments

I feel very honored to be a master student at Emory University. First of all, I want to thank my parents. Without their support and dedication, I would not have the opportunity to stand here and become an alumnus of Emory University with so many great people. Then I want to thank my thesis supervisor Zhaohui (Steve) Qin and the PhD candidate Yutong Jin who lead me to do the thesis study. Their spirits of hard work have inspired me a lot, and I decide to become a PhD student in the future and follow in their footsteps.

The life in the past year was very special. Although everything has changed from offline to online because of COVID pandemic, the connection between me and professors and friends has not diminished. They helped me a lot. I will take their encouragement and support to move on. Many people have helped me in the past two years, such as my family, my classmates, my roommates, my ADAP, my recommender, my isss advisor, my APE supervisor, my APE advisor, etc. I can't list their names one by one, some people I don't even know their names, but I really appreciate their help.

Finally, I want to thank my girlfriend Zhuyun Mao and thank you for your toleration and encouragement all the time. In the past year, my temper has become much worse than before. Although we have had quarrels and unpleasantness, you continue to encourage me and let me overcome a lot of difficulties. If without you, maybe I still don't know what I should do in the future. In addition, I also want to thank the Australian Shepherd Raven that we raised during COVID pandemic. Although it is very hard to raise it, it really brings us a lot of happiness.

Table of contents

1 Introduction

Single nucleotide variant (SNV), also known as single-nucleotide polymorphism (SNP), is a variant of a single nucleotide at a specific location in genome. It is pretty common in human genome. In the past 15 years, considerable SNVs are the main topics in thousands of Genome-wide association studies (GWAS), which is a powerful tool to identify the association between SNVs and the phenotypes. According to these studies, SNVs are considered to play an important role on a wide array of phenotypes (Welter, et al., 2014). With the development of technologies, SNVs are not only found in the protein-coding regions of the human genome, but also occur to the non-coding regions (Zhou, et al., 2020). Moreover, it has been demonstrated that SNVs, which are identified through GWAS, are enriched in non-coding regulatory regions to control and modulate gene expression (Williams, et al., 2019). It is expected that genetic variants have an association with the pathogenesis of diseases.

SNVs fall within protein-coding regions have been verified to be associated with human diseases, but it is still not clear about the mechanism of non-coding variants. However, plenty of evidence suggests that genetic variants in the non-coding region may cause missing heritability, which results in human diseases (Zhang, et al., 2015). One possible mechanism currently considered is that the variation of short DNA motif may disrupt the binding affinity between transcription factor and enhancers (Pasquali, et al., 2014). Generally, using the position weight matrices (PWMs) to score a DNA motif is a common way to study the TF binding specificity (Stormo, 2000). By scanning sequences in the possible binding sites, PWM would assign a matching score to those sequences. The positions of those sequences whose matching score above the arbitrary threshold would be considered as candidate TF binding sites (Hertz, et al., 1999). Then all SNVs occur to the binding sites would be marked for their potential functional impacts. Nevertheless, much information cannot be well conserved within TF binding sites, resulting in low functional impacts of SNVs at these locations. Although highly informative and experimentally determined PWMs are accessible in open databases such as FlyFactorSurvey (Zhu, et al., 2011) and JASPAR (Portales-Casamar, et al., 2010), they poorly consider the influence of mutations in a motif on probability because they assume all positions are independent. In order to accurately express the information contained in SNVs, more research besides PWM is needed.

More researchers believe that an accurate computational model is helpful and necessary to identify and predict functional variation in specific binding sites. Recently, Ghandi's group introduced an excellent method, kmer support vector machine (kmer-SVM), to predict regulatory DNA sequence using combinations of short (6 – 8 bp) k-mer frequencies. Compared to PWM approach, it is not necessary to have large amounts of data to determine the scoring threshold for kmer-SVM method. The key is whether k-mers are present or not. However, this method would be inaccurate when k becomes large, especially when Transcription Factor Binding Sites (TFBS) are over 8 bp. Based on the kmer-SVM, Ghandi's group developed alternative method using gapped k-mers, gapped k-mers support vector machine (gkm-SVM) (Ghandi, et al., 2014), which can be applied on longer and more general sequences. The authors use sequencing-based assays to define gkm-SVM weights for k-mers so that they can quantify the functional importance of k-mers at TFBS. Moreover, the authors defined deltaSVM score which is the gkm-SVM weight difference for a k-mer with or without SNVs. Therefore, they successfully quantify the effect of SNVs.

In this study, we evaluated the impact of SNVs on TF binding using gkm-SVM method. First, we used gkm-SVM to train all 10-mers based on the TF's ChIP-seq data to evaluate the TF binding potential. Then we used the deltaSVM method to quantify the effect of a SNV on the TF binding sites. It is essential to study and quantify the position-specified impact of SNVs throughout the genome, and we believe the deltaSVM scores based on ChIP-seq data can be a powerful resource for relative studies.

In this work, we used ChIP-seq data from the Encyclopedia of DNA Elements (ENCODE) project (Consortium, 2012). Considering well-defined PWMs and availability from ENCODE, we selected 18 TFs in GM12878, which are BCL11A, CTCF, EGR1, GABPA, JUN, JUND, MAX, NANOG, POU5F1, RAD21, RFX5, SIX5, SRF, STAT1, TCF12, USF1, USF2 and YY1. For each TF, we measured SVM weights of all the 10-mers by counting the number of their occurrences in a ChIP-seq dataset. Then we surveyed the impact of all mutations within those 10-mers with top SVM weights. Our results show that the most SNVs have little impact on the binding affinity, so more information of the functional SNVs is needed.

2 Method

The traditional way to evaluate the binding affinity of a TF is to calculate the probability of the TF based on the PWM. The impact of a SNV can be quantified through the probability difference a TF with or without a mutation. However, PWM method poorly express the probability difference while the alternative method, gkm-SVM, can accurately convey the information using deltaSVM scores. In this study, we compared the difference between these two methods in assessing the TF binding affinity.

2.1 Using Phenotype–Genotype Integrator (PheGenI) and UCSC Table Browser to find SNVs

There are thousands of SNVs identified through GWAS, we selected those SNVs associated with certain diseases. We chose 11 common diseases including Alzheimer Disease, Asthma, Breast neoplasms, Cardiovascular diseases, Child Development Disorders, Colorectal Neoplasms, Crohn diseases, Lung neoplasms, Obesity, Psoriasis and Type 2 diabetes. We set p-value 10^{-6} as the threshold for each disease on Phenotype–Genotype Integrator (PheGenI) website. Then we got thousands of disease-related GWAS SNVs and their specific position in the genome. Subsequently, we extended 10kb region centered on each GWAS SNV and this 10kb region is regarded as case region. We input this 10kb region to UCSC Table Browser to locate all single SNV fall within the 10kb region. We believe all these SNVs may be associated with corresponding diseases.

2.2 Using PWM method to evaluate motif

PWM is one of the most popular bioinformatic methods for investigating motifs (Xia, 2012). In this part, we used 19 TF motif PWMs to scan the human reference genome GRCh37 (UCSC version: hg19). Take CTCF as an example: we identified 139,084 15-mer CTCF motif sites in the whole genome using PWM method with the 80% minimum score. There are lots of duplicated 15-mer motif sequences and there are 48,804 unique 15-mer motif sequences. Then we screened out the motif sequences that contain the positions of the previously selected disease-related SNVs. Here, we regarded the hg19 genome as reference genome. Those selected motifs with the reference allele at the SNV position is the reference motif, while motifs with the alternative allele at the SNV position the alternative motif. A motif PWM score is the probability of a DNA motif, which is the product of relevant probabilities for each position based on the PWM matrix. The log-transformed probability difference between the reference motif and the alternative motif is defined as the motif delta-PWM score. The difference is regarded as the impact of a SNV. Since a 15-bp motif contains six different consecutive 10-mers, we set the one with largest PWM score as the probabilistic value of the specific 10-mer, which is defined as 10-mer PWM score.

2.3 Using gkm-SVM method to evaluate motif

Gkm-SVM is a new sequence-based computational method to predict the effect of regulatory variation (Lee, et al., 2015). It requires positive training set and negative training set. We obtained ENCODE TF ChIP-Seq datasets from the GM12878 cell line. According to the narrow peaks provided by ENCODE, we treated peak regions and non-peak region as positive training set and negative training set respectively in order to apply gkm-SVM method. As a result, gkm-SVM estimated the SVM weights for all possible 10-mers. For each SNV, in its flanking area, there are ten 10mers containing the SNV. The deltaSVM score of the SNV is defined as the sum of weights difference between the ten 10-mers containing the reference allele or the alternative allele.

However, there is no standard threshold to determine the significance of the weight difference for SNVs. Here, we selected 10,000 random non-motif sequences which are excluded by PWM scan probability threshold. These non-motif sequences have the same length with the motif sequences and are treated as the control set. For each base within the control set, we calculated the averaged deltaSVM scores. After we obtained large dataset of deltaSVM scores, we set the value of 2.5 percentile and 97.5 percentile of the deltaSVM scores as the empirical significant threshold.

3 Results

Currently, annotating non-coding variants is generally based on PWM method. All SNVs obtained from PWM method which are within the TF binding sites are considered to be associated with TF binding affinity. However, the impacts of those SNVs are not well-quantified and there is no evidence about the indeed impact of every single SNVs within the TF binding sites. In this study, we applied a new method, gkm-SVM method to tell if all SNVs within the TF binding sites affect the TF binding affinity or only some of SNVs affect the TF binding affinity. We applied gkm-SVM method to 18 TFs in GM12878 cell line, including BCL11A, CTCF, EGR1, GABPA, JUN, JUND, MAX, NANOG, POU5F1, RAD21, RFX5, SIX5, SRF, STAT1, TCF12, USF1, USF2 and YY1. For each TF, according to the number of times of occurrence, we obtained every 10-mers' weight from peak region. Then we selected top 1000 10-mers with highest PWM scores to conduct a completed survey on the impact of all

SNVs appearing within these 10-mers and compare the results obtained from two results.

3.1 Correlation between PWM scores and gkm-SVM weights

We first applied PWM method to get PWM scores for all 10-mers and screened out top 1000 10-mers with highest PWM scores. Then we applied gkm-SVM method to calculate gkm-SVM weights for these 1000 10-mers. We found that the correlations between PWM scores and gkm-SVM weights vary in terms of TFs and the correlations range from -0.081 to 0.787. Take CTCF, USF1, SIX5 and BCL11A for example (Figure 1), the correlations are 0.620, 0.787, 0.773, and 0.021 respectively, indicating these two methods are not consistent to some extent. The complete results for all 18 TFs are in Supplementary Materials. For CTCF, USF1, and SIX5, two methods have a relatively strong and positive correlation relationship, while such correlation is not obvious in BCL11A. This indicates the PWM method may not well assess the TF binding potential considering PWM method assume mutually statistical independence between positions. Moreover, we selected top 20 10-mers with highest gkm-SVM weights and compare their corresponding PWM scores for the four TFs mentioned above (Figure 2). As we can see, some 10-mers with high gkm-SVM scores do not have high PWM scores. This phenomenon is particularly obvious for USF1 and SIX5, once again confirming the inconsistency between PWM method and gkm-SVM method.

3.2 Potential association between TFs and complex diseases

GWAS have shown that plenty of disease-related SNVs fall within the non-coding regions of the genome (Zhu, et al., 2017). These SNVs may be associated with complex diseases. A possible mechanism is that some of these SNVs are located in the regulatory regions of the genome and they can weaken or disrupt the binding of TF, causing changes in gene expression to manifest diseases. In order to see whether the impacts of SNVs on a specific TF have a relationship with any complex diseases, we conducted a comprehensive survey on 11 diseases including Alzheimer Disease, Asthma, Breast neoplasms, Cardiovascular diseases, Child Development Disorders, Colorectal Neoplasms, Crohn diseases, Lung neoplasms, Obesity, Psoriasis and Type 2 diabetes. For each disease, we set p-values 10^{-6} as the threshold on PheGenI website to obtain all disease-related SNVs that meet the p-value requirement. We believe these SNVs are significantly associated with the specified disease. However, PheGenI only provide index SNVs while ignore other nearby SNVs which may be the most important factor in the disease. Therefore, we extended 5kb region up and down with those GWAS index SNVs as the center to increase the probability to capture the causal SNVs. We used UCSC table browser tool to obtain all SNVs located in the 10kb region.

Next, we screened out SNVs in the 10kb region overlapped with any putative TF binding sites which is identified with PWM method. Then we calculated how many selected SNVs have significant deltaSVM scores exceeding the empirical thresholds. The false positive was defined as the proportion of SNVs with insignificant deltaSVM scores among selected SNVs. Traditionally, all SNVs within putative TF binding sites identified by PWM method are considered to be significantly associated with designated diseases. However, we found many SNVs within putative TF binding sites do not have significant deltaSVM scores, indicating inconsistency between two methods. The significance level is 0.05. Figure 3 shows the heatmap of the false positive rates. Taking POU5F1 (TF) and AD (disease) as an example, there are 693 SNVs found in the POU5F1 binding site while only 18 of them have significant deltaSVM scores, indicating not all SNVs inside the binding site have significant impact on POU5F1 *in vivo* binding. The false positive rates vary in terms of TFs and diseases, but the average rate is around 80% across 18 TFs for 11 diseases. This is a good evidence that in addition to PWM method, more information is needed to study the impact of SNVs on complex diseases.

4 Discussion

It is a big challenge to understand the functional impact of non-coding variants considering there is no golden standard to quantify the impact currently. Fortunately, studies on SNVs increase dramatically (Rojano, et al., 2019). More and more genomics and epigenomics data is provided to increase the chance to better understand the internal mechanism. With more comprehensive data, we aim to develop a better metric to quantify the impact of SNVs on *in vivo* TF binding affinity. Also, we believe the better metric could help us understand how non-coding SNVs are associated with complex diseases.

PWM method is the one of the most used method to evaluate the impact of SNVs on TF binding affinity. However, this method was developed for small amount of motif-enriched sequences, it would not be effective when the dataset is large (Hu, et al., 2010). Moreover, PWM method assumes mutually statistical independence between positions within a motif site (Zhou and Liu, 2004), which limits our complete understanding of the binding of TF. Therefore, we applied an alternative metrics to assess the impact of SNVs using ENCODE ChIP-seq data. More information would be offered in ChIP-seq data and this would help us better understanding the impact of SNVs besides traditional PWM method.

In our results, for most TFs, the correlation between gkm-SVM method and PWM method is positive and consistent. However, for some TFs such as BCL11A and POU5F1, the correlations are pretty small and the correlation for POU5F1 is negative, indicating inconsistency between two methods. We believe gkm-SVM method is the better way to study the impact of SNVs to *in vivo* TF binding. Gkm-SVM method not only consider the mutual-dependence between positions, but also can apply on large dataset. Nevertheless, there are some limitations in our method. The ChIP-seq data we used to train is cell-type specific and the quality of the data varies over time. Having said that, these are both a disadvantage and an advantage. Cell-type specific and latest data can bring new and comprehensive information on TF binding, helping us understand the mechanism more deeply.

5 References

[1] Welter, Danielle, et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." Nucleic acids research 42.D1 (2014): D1001-D1006.

[2] Zou, Hecun, et al. "Significance of single-nucleotide variants in long intergenic non-protein coding RNAs." Frontiers in Cell and Developmental Biology 8 (2020).

[3] Williams, Sarah M., et al. "An integrative analysis of non-coding regulatory DNA variations associated with autism spectrum disorder." Molecular psychiatry 24.11 (2019): 1707-1719.

[4] Zhang, Feng, and James R. Lupski. "Non-coding genetic variants in human disease." Human molecular genetics 24.R1 (2015): R102-R110.

[5] Pasquali, Lorenzo, et al. "Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants." Nature genetics 46.2 (2014): 136-143.

[6] Stormo, Gary D. "DNA binding sites: representation and discovery." Bioinformatics 16.1 (2000): 16-23.

[7] Hertz, Gerald Z., and Gary D. Stormo. "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics (Oxford, England) 15.7 (1999): 563-577.

[8] Zhu, Lihua Julie, et al. "FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system." Nucleic acids research 39.suppl_1 (2011): D111-D117.

[9] Portales-Casamar, Elodie, et al. "JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles." Nucleic acids research 38.suppl_1 (2010): D105-D110.

[10] Ghandi, Mahmoud, et al. "Enhanced regulatory sequence prediction using gapped k-mer features." PLoS Comput Biol 10.7 (2014): e1003711.

[11] Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57.

[12 Hu, M.*, et al*. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Research* 2010;38(7):2154-2167.

[13] Lee, D.*, et al*. A method to predict the impact of regulatory variants from DNA sequence. *Nature genetics* 2015;47(8):955.

[14] Rojano, E.*, et al*. Regulatory variants: from detection to predicting impact. *Briefings in bioinformatics* 2019;20(5):1639-1654.

[15] Xia, X. Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica* 2012;2012.

[16] Zhou, Q. and Liu, J.S. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* 2004;20(6):909-916.

[17] Zhu, Y., Tazearslan, C. and Suh, Y. Challenges and progress in interpretation of non-coding genetic variants associated with human disease. *Experimental Biology and Medicine* 2017;242(13):1325-1334.

6 Figures

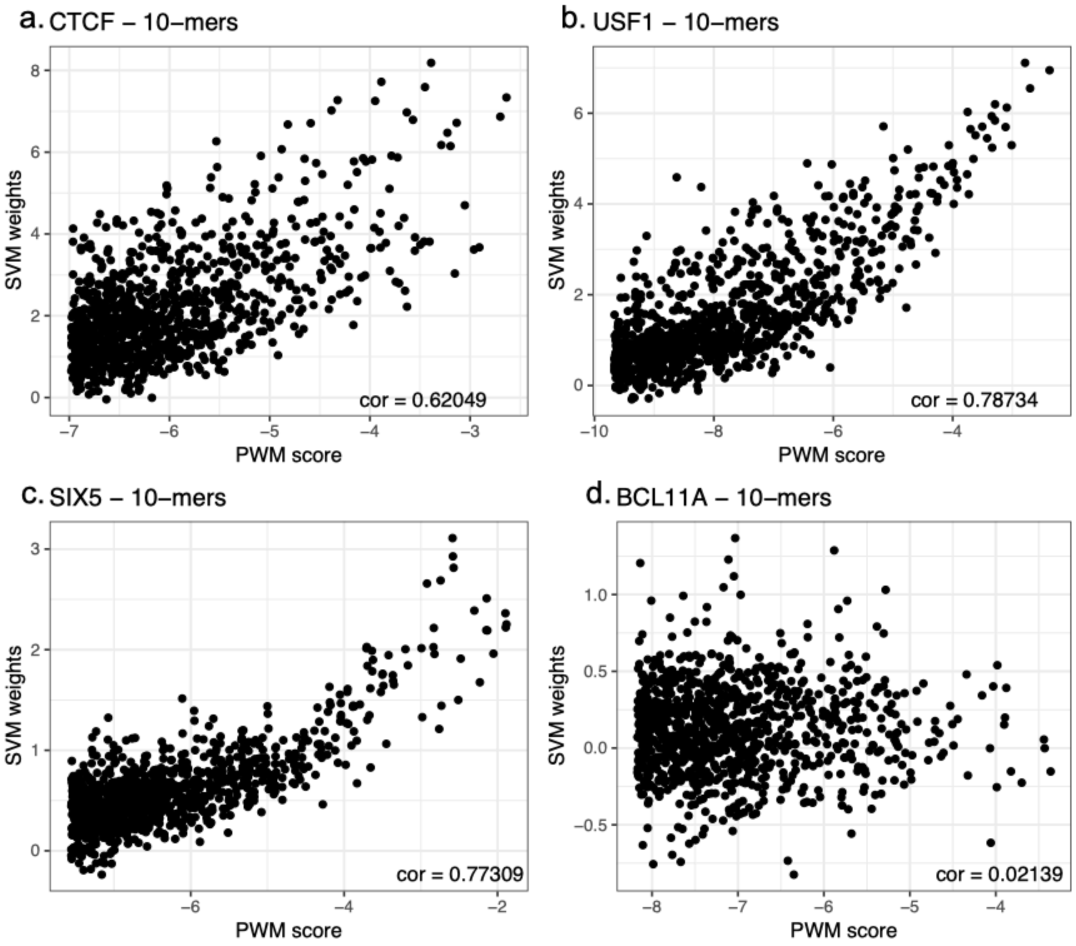

Fig1. Correlation between PWM scores and gkm-SVM weights for top 1000 10-mers

The correlation between PWM scores and gkm-SVM weights for (a) CTCF, (b) USF1, (c) SIX5, and (d) BCL11A are 0.620, 0.787, 0.773, and 0.021 respectively.

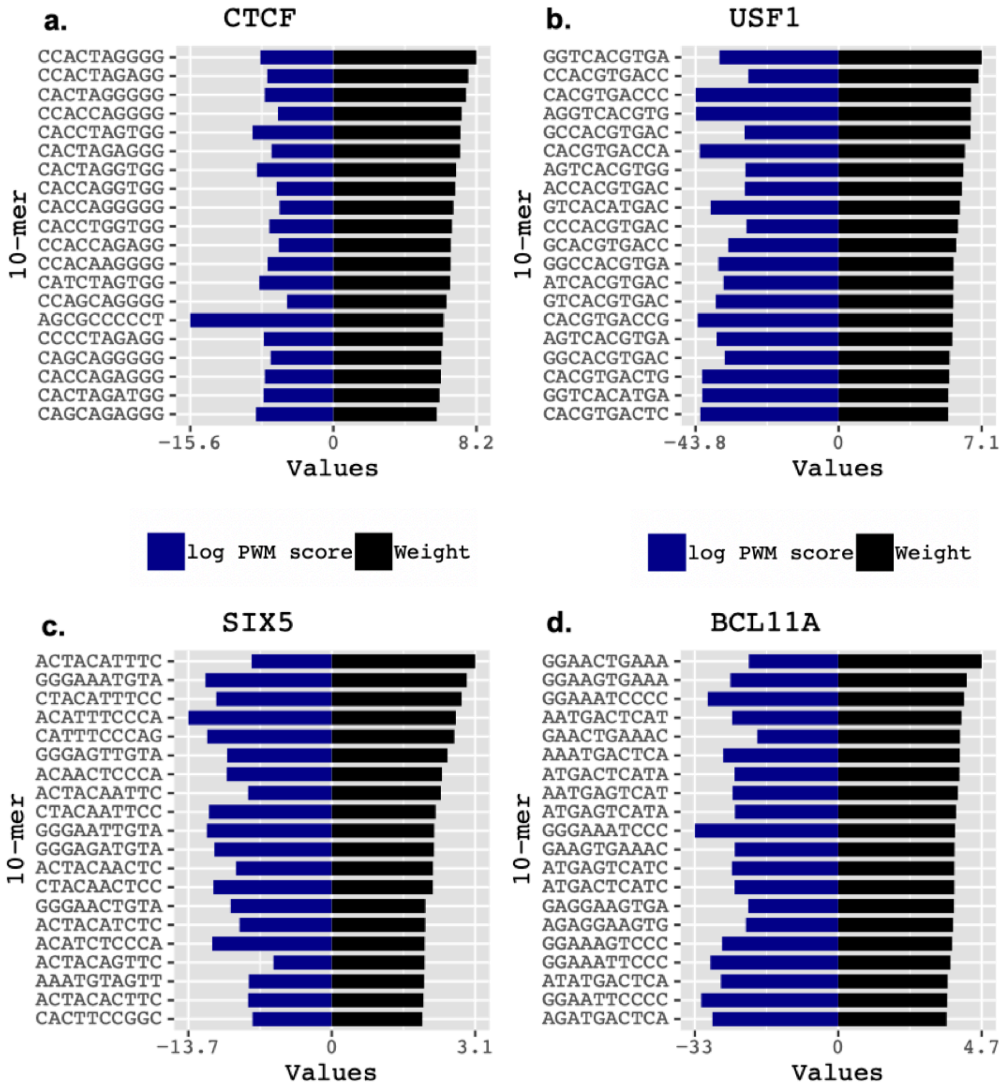

Fig2. Correlation between PWM scores and gkm-SVM weights for top 20 10-mers

The top 20 10-mers comparison between PWM scores (blue bar) and gkm-SVM weights (black bar) for (a) CTCF, (b) USF1, (c) SIX5, and (d) BCL11A.

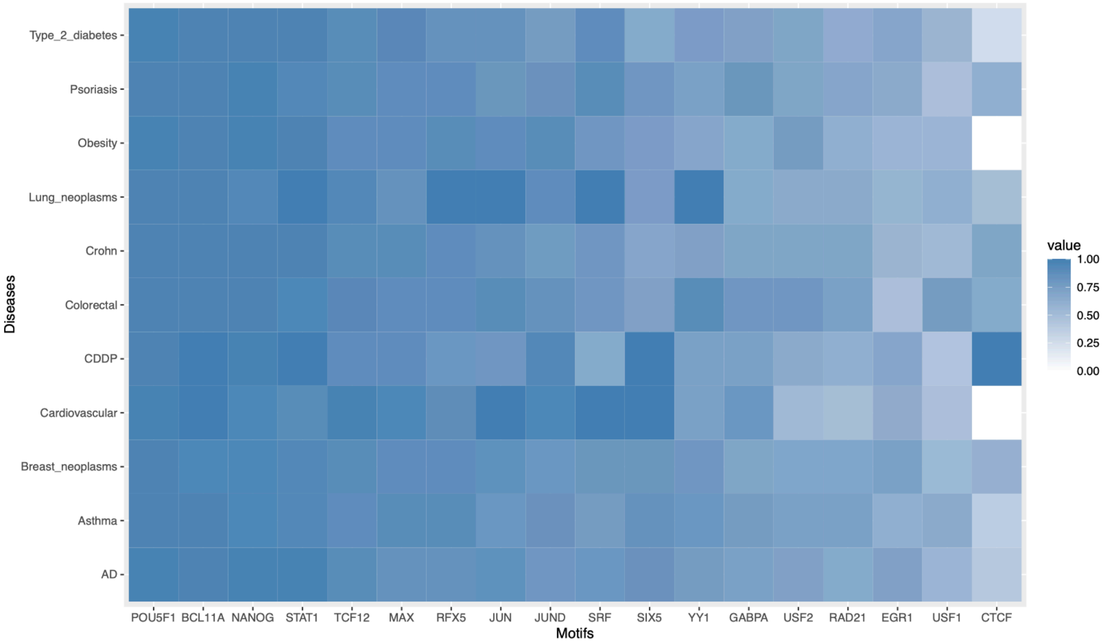

Fig3. Heatmap of the false positive rates of all 11 diseases for each motif

7 Supplementary Materials

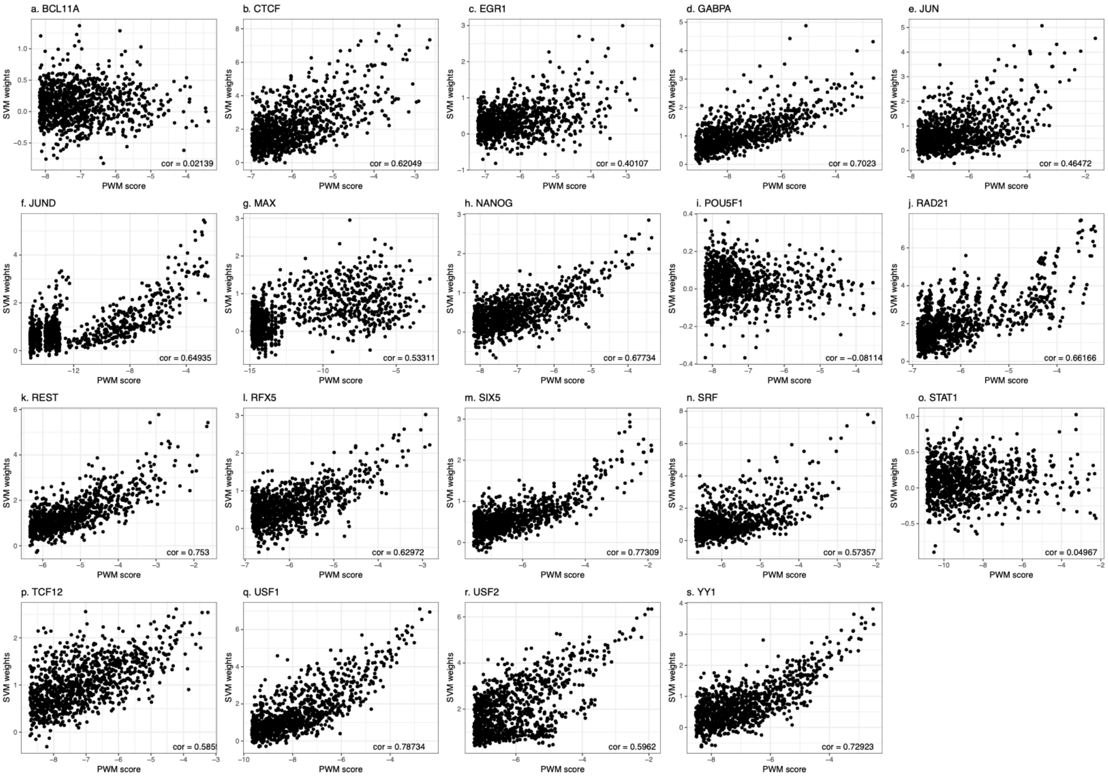

Fig4. Correlation between PWM scores and gkm-SVM weights for top 1000 10-mers for 18 TFs