Bioinformatics, 2019, 1–9
doi: 10.1093/bioinformatics/btz612
Advance Access Publication Date: 2 August 2019
Original Paper

Genome analysis

Predicting the effects of SNPs on transcription factor binding affinity

Sierra S. Nishizaki [1], Natalie Ng[2], Shengcheng Dong[3], Robert S. Porter[1], Cody Morterud[3], Colten Williams[3], Courtney Asman[3], Jessica A. Switzenberg[3] and Alan P. Boyle[1,3,*]

[1]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA, [2]Department of Human Genetics, Stanford University, Stanford, CA 94305, USA and [3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on March 27, 2019; revised on July 15, 2019; editorial decision on July 30, 2019; accepted on August 1, 2019

Abstract

Motivation: Genome-wide association studies have revealed that 88% of disease-associated single-nucleotide polymorphisms (SNPs) reside in noncoding regions. However, noncoding SNPs remain understudied, partly because they are challenging to prioritize for experimental validation. To address this deficiency, we developed the SNP effect matrix pipeline (SEMpl).

Results: SEMpl estimates transcription factor-binding affinity by observing differences in chromatin immunoprecipitation followed by deep sequencing signal intensity for SNPs within functional transcription factor-binding sites (TFBSs) genome-wide. By cataloging the effects of every possible mutation within the TFBS motif, SEMpl can predict the consequences of SNPs to transcription factor binding. This knowledge can be used to identify potential disease-causing regulatory loci.

Availability and implementation: SEMpl is available from https://github.com/Boyle-Lab/SEM_CPP.

Contact: apboyle@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

To date, genome-wide association studies (GWAS) have identified over 100 000 loci associated with over 200 human diseases and phenotypic traits (Edwards *et al.*, 2013; Welter *et al.*, 2014). Though 95% of known single-nucleotide polymorphisms (SNPs) and 88% of GWAS SNPs fall into noncoding regions of the genome, most genetics studies focus on mutations within coding regions (Hindorff *et al.*, 2009; Sherry *et al.*, 2001). This large disparity in knowledge gained from big data initiatives is likely due to the more direct interpretability of genic variation even though noncoding variation is also strongly linked to human disease (VanderMeer and Ahituv, 2011; Zhang and Lupski, 2015). Identifying noncoding mutations leading to gene misregulation is critical to fully understand GWAS results and their impact on complex and polygenic disorders.

As noncoding GWAS variants are overwhelmingly abundant compared to coding variants, many methods have been developed to prioritize potentially disease-associated mutations in noncoding regions for further study (Nishizaki and Boyle, 2017). Generally, these tools focus on known regulatory regions of the genome, relying on variant overlap with experimental annotations, such as regions of open chromatin and transcription factor binding (Boyle *et al.*, 2012; Kircher *et al.*, 2014; Ward and Kellis, 2012). To date, these computational prioritization tools have assisted in identifying a handful of causal disease mutations from GWAS (He *et al.*, 2015; Higgins *et al.*, 2015). However, these tools have only shown up to a 50% concordance rate between predictions, highlighting the need for additional prioritization metrics (Nishizaki and Boyle, 2017). One way to improve these predictions is to investigate additional regulatory features to better understand a variant's mechanism of action.

Transcription factor binding sites (TFBSs) are a regulatory feature of particular interest as they make up 31% of GWAS SNPs, yet only comprise 8% of the genome (ENCODE Project Consortium *et al.*, 2012). Mutations in TFBSs influence transcription factor-binding affinity, alter gene expression, and have been associated with multiple human diseases including cancer and type 2 diabetes, as well as with increased total cholesterol (Fogarty *et al.*, 2014; Gaulton *et al.*, 2010; Musunuru *et al.*, 2010; Pomerantz *et al.*, 2009; Savic *et al.*, 2011; Stitzel *et al.*, 2010). However, altering different bases within a TFBS have been found to confer different effects on transcription factor binding (Kasowski *et al.*, 2010; McDaniell

et al., 2010). This finding has been reflected in cases of human disease, where certain bases in a sequence motif are more correlated with an associated disease than others (Umer *et al.*, 2016). Currently, the effect of mutations in a TFBS is estimated using a position weight matrix (PWM), which denotes a transcription factor's binding motif using *in silico* analyses to determine its predominant binding sequence using a competitive binding assay (Fig. 1A) (Stormo *et al.*, 1982). PWMs predict where a transcription factor may bind in the genome by acting as its most frequent binding sequence; however they may not recapitulate known binding activity and are not sufficient to predict which mutations within a motif may alter binding affinity (Weirauch *et al.*, 2013). Additionally, using PWMs to predict how a SNP may affect transcription factor binding can be challenging, as PWMs do not contain information on the potential direction of effect of a mutation.

While multiple tools have been developed to predict which mutations may lead to changes in binding affinity, many of these methods rely solely on information from PWMs and are thus subject to similar limitations (Andersen *et al.*, 2008; Barenboim and Manke, 2013; Khan *et al.*, 2018; Macintyre *et al.*, 2010; Manke *et al.*, 2010; Shrikumar *et al.*, 2017; Vorontsov *et al.*, 2015). More recent methods have incorporated additional measures of binding affinity, including protein-binding microarray data, systematic evolution of ligands by exponential enrichment (SELEX) data and/or chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) data (Alipanahi *et al.*, 2015; Foat *et al.*, 2006; Jolma *et al.*, 2010; Lee *et al.*, 2015; Riley *et al.*, 2015; Zhao and Stormo, 2011; Zhou and Troyanskaya, 2015). These methods represent a marked improvement over a strict input of PWMs, however they have their own pitfalls. As protein-binding microarray and SELEX data are generated outside of a native cell context they may not represent patterns of true intercellular binding. In addition, the majority of these models output *de novo* motifs similar in style to PWMs, which are not informative to direction of effect of a mutation. One tool of particular interest, the Intragenomic Replicates (IGR) method, was developed as a way to investigate FOXA1 involvement in breast cancer using GWAS data (Cowper-Sal Lari *et al.*, 2012). This method compares TFBSs containing putatively deleterious mutations to their wild-type counterparts using genome-wide ChIP-seq data to estimate predicted changes to transcription factor-binding affinity. The predictions generated by IGR were found to be highly correlated with ChIP-qPCR results and were successfully used to identify a risk allele associated with a 5-fold change in gene expression in breast cancer. IGR represents a marked improvement over other methods due to its specific calibration of variants to ChIP-seq data, an endogenous source of transcription factor-binding affinity information. Currently, IGR exists only as a method designed to probe individual mutations and must be reconstructed for each new mutation and transcription factor. However, the premise of using ChIP-seq data to predict transcription factor binding could be expanded to more quickly and accurately predict TBFS mutations.

In order to improve current methods to be applicable to a wide range of transcription factors and to better predict which mutations within TFBSs may lead to changes in binding affinity, we have developed a new method: the SNP effect matrix pipeline (SEMpl). Our method uses endogenous ChIP-seq data and existing variants genome-wide similar to the IGR method, however SEMpl also includes a catalog of kmers separated by a single base change from a TFBS motif, allowing it to provide an estimate of the consequence of every possible mutation in a TFBS. We call these as SNP effect matrices (SEMs, Fig. 1). Here, we demonstrate that SEMs recapitulate known motifs, are robust to input data and cell type, and are better at predicting changes to transcription factor-binding affinity than the current standard, PWMs. By developing SEM scores, we aim to improve the prioritization of noncoding GWAS variants for further experimental validation, expand the understanding of noncoding genomic variation and further technology toward developing tools for personalized medicine.

2 Materials and methods

2.1 Usage/accessibility

SEMpl is open access and can be downloaded from github: https://github.com/Boyle-Lab/SEM_CPP. Over 200 precomputed SEMpl scores can be found in Supplementary Material.

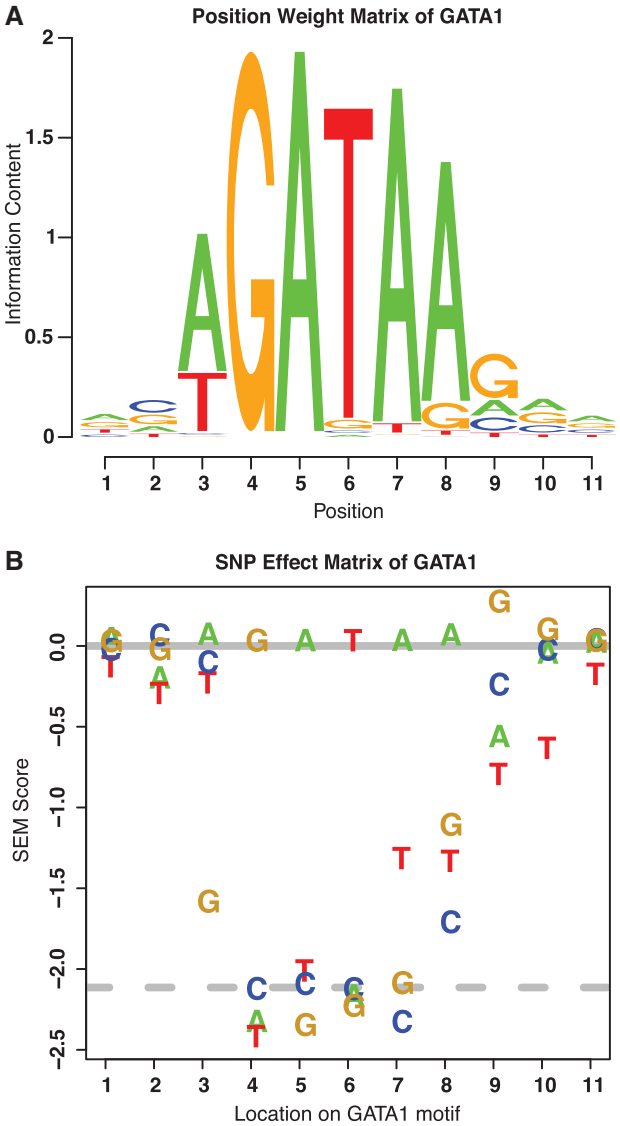

Fig. 1. PWM versus SEM of transcription factor GATA1. (**A**) The PWM can be read as likely nucleotides along a transcriptions factor's motif. (**B**) Similarly, the SEM can be read as nucleotides along a motif, but with additional information about the effect any given SNP may have on transcription factor-binding affinity. The solid gray line represents endogenous binding, the dashed gray line represents a scrambled background. We define anything above the solid gray line as predicted to increase binding on average, anything between the two lines as decreasing average binding and anything falling below the dashed gray line as ablating binding on average

2.2 SNP effect matrix pipeline

SEMpl utilizes three types of experimental evidence to make its predictions: ChIP-seq data, which provides a transcription factor's endogenous binding in the genome; DNase I hypersensitive site (DNase-seq) data, which represents regions of open chromatin where transcription factors are known to function and PWMs, which denote previous knowledge of the binding pattern of transcription factors (Fig. 1). We obtained ChIP-seq and DNase-seq data from the ENCODE project and PWMs from the JASPAR, Transfac, UniPROBE and Jolma databases (ENCODE Project Consortium *et al.*, 2012; Hume *et al.*, 2015; Jolma *et al.*, 2013; Khan *et al.*, 2018; Wang *et al.*, 2010).

SEMpl first enumerates a PWM of interest into a list of kmers using a permissive cutoff *P*-value threshold of 4^{-5} using the software transcription factor matrix *P*-value (TFM-PVALUE) (Fig. 2A) (Touzet and Varré, 2007). This first list of kmers, referred to as the endogenous kmer list, represents sequences where the transcription

factor of interest has an increased likelihood of binding. To observe additional sequences which may show distinct binding preferences, SEMpl next takes the endogenous kmer list and simulates all possible SNPs *in silico* to create lists of mutated kmers (Fig. 2B). For example, by changing all bases in position 6 to a G nucleotide in every kmer in the endogenous kmer list, SEMpl creates a mutated kmer list for G in position 6. These lists of mutated kmers are then aligned to the hg19 reference human genome in regions of open chromatin using bowtie, as determined by DNase-seq (Langmead *et al.*, 2009). The ChIP-seq score is then calculated as the highest signal value over the region 50 bp before and after the aligned site (Fig. 2C). Next the SEM score for each position is computed as the log2 of the average ChIP-seq signal to endogenous signal ratio for the mapped kmers for each mutated kmer list. Taken together, the SEM scores for each base form a matrix for each nucleotide at every position along the motif. Scores can be evaluated at individual nucleotides, or calculated across a full-length kmer by adding the nucleotide score for each position along the motif, similar to a PWM.

The above process is repeated, using a slightly more stringent TFM-PVALUE cutoff of $4^{-5.5}$ to generate kmers, until convergence using an estimation maximization (EM)-like method in order to correct for differences arising from unique starting kmers (Fig. 2D, Supplementary Fig. S1). This process continues until the number of kmers from the endogenous kmer list does not change or until 250 iterations, with the average run converging by iteration 117. To control for poor quality data and to identify background levels of binding, a final kmer list of randomly scrambled endogenous kmers is included to represent a random baseline where transcription factor binding would not be expected to occur (displayed as a dashed gray line on an SEM plot). Finally, we define scores above 0 as predicted to increase binding on average, scores between 0 and the scrambled background as decreasing average binding on average and scores falling below the scrambled background as ablating binding on average.

SEMpl output files include error messages during the run (.err), the cache, a tally of kmer similarity between iterations (kmer_similarity.out) and an output file containing information on run time and where the program is in the run (.out). Additionally, within each iteration, output files include the alignments for the SNP kmer lists (alignment folder) and endogenous and scrambled kmer lists (baseline folder) which include the aligned loci and ChIP-seq signal. A quality control file is also provided within each iteration file that provides the number of kmers mapped within the iteration, as well as a –log10(*P*-value) representing the average of 100 *t*-tests from 1000 randomly chosen kmers from the SNP signal files versus 1000 randomly chosen kmers from the scrambled signal file. We used a threshold of 2.5 to report confidence in a SEM run. Resulting aligned loci and ChIP-seq values are stored in a cache, which allows for a quick lookup of nonunique kmers without realignment. SEMpl options include –readcache, which can be used to speed up a run for which a cache has already been created. SEMpl is written in C++ and R. PWMs were created using the R package seqLogo (Bembom, 2019).

2.2.1 Scoring a variant or sequence with a SEM

Scoring variants or sequences using SEMpl are as straightforward as scoring using a PWM. A score can be computed in two ways. First, a single base change can be scored by subtracting the wild-type nucleotide score from the variant score using the SEM matrix to determine the total predicted difference between the two nucleotides. Second, a k-mer sequence can be scored in a manner very similar to a PWM. Because the matrix is log transformed, the score of each nucleotide can be added to reflect the predicted binding of the full sequence. In this way the effect of multiple variants can be calculated for a single sequence. In either case, the final value represents the expected change compared to endogenous binding levels.

2.3 Correlation with ChIP-seq data

All possible kmers from the original transcription factor PWMs were generated. For each unique kmer, average ChIP-seq signal and standard error were calculated. PWM, SEM, DeepBind and LS-GKM scores were calculated for each kmer. DeepBind scores were calculated from precomputed models, and LS-GKM scores were computed using the options $l=10$ and $k=6$ for motifs with length $\geq$10—as recommended by the author. For LS-GKM motifs length 9, $l=9$ and $k=6$, and motifs length 8 were run using $l=8$ and $k=5$. Correlations cutoffs were calculated for PWMs above the standard TFM-PVALUE cutoff (P-value $=4^{-8}$) typically used for PWM visualization. Correlation cutoffs for SEM, DeepBind and LS-GKM scores were defined as the average scrambled baseline across all iterations for a single transcription factor run.

2.4 SEM correlation across runs

SEM outputs from different starting ChIP-seq or PWM data were compared using least square regression in R. More details about the

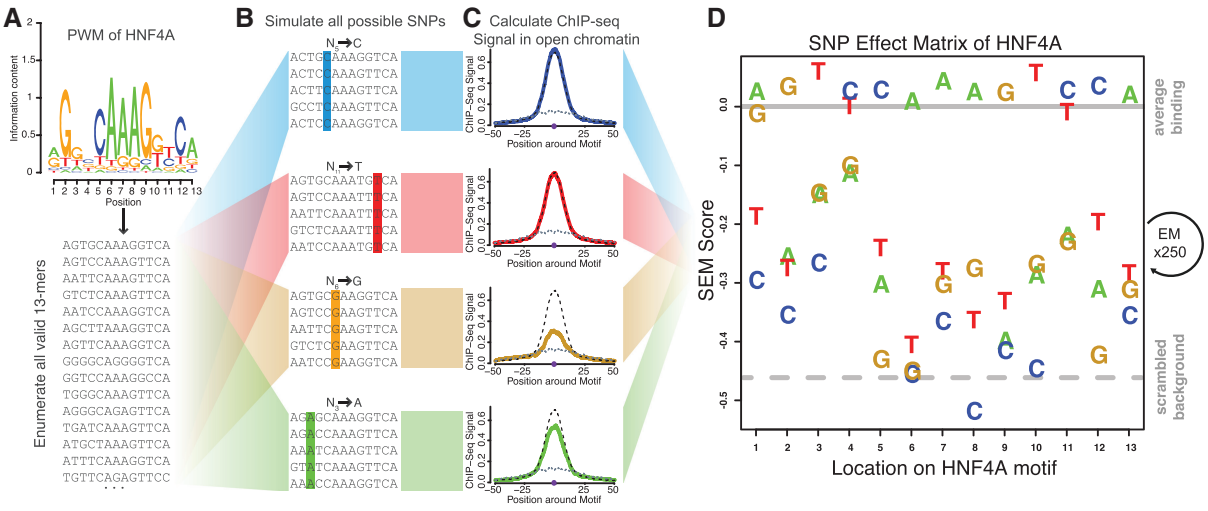

Fig. 2. SEM methods pipeline. (**A**) All kmers with a PWM score below the TFM-PVALUE are generated for a single transcription factor. (**B**) All possible SNPs are introduced *in silico* for each kmer. (**C**) All enumerated kmers are then aligned to the genome, and filtered for regions of open chromatin by DNase-seq. The average ChIP-seq scores are then calculated for each alignment (dashed line represents endogenous binding, dotted line represents scrambled background). (**D**) Final SEM scores are log2 transformed and normalized to the average binding score of the original kmers (solid gray line). A scrambled baseline, representing the binding score of randomly scrambled kmers of the same length is also added (dashed gray line). Once a SEM score is calculated, the output can be used to generate a new PWM. This iterative process can correct for disparities introduced by the use of different starting PWMs. The HepG2 cell line data were used for the ChIP-seq and DNase data for HNF4a

datasets used for analysis can be found in Supplementary Table S1. Overlapping DNase-seq peaks were downloaded from ENCODE and calculated using bedtools (Quinlan and Hall, 2010). SEMpl runs from the same cell type, and therefore using the same DNase dataset, share 100% DNase peak overlap.

2.5 Allele-specific CTCF-binding pattern analysis

Allele-specific binding sites were defined as loci containing one or more heterozygous SNPs while showing significant differences in ChIP-seq signal from two alleles. We applied the AlleleDB pipeline to count the number of ChIP-seq reads from two alleles respectively for each heterozygous site and identified 468 allele-specific binding sites at an FDR of 5% (Chen *et al.*, 2016). CTCF ChIP-seq data from GM12878 cell line was used in this analysis (accession number: ENCSR000DZN). For all heterozygous sites within CTCF ChIP-seq peaks in GM12878 cell line, 240 of them also have matching CTCF PWMs, which we further used for the comparison of SEM and PWM scores. For those 240 heterozygous sites, we calculated the allelic ratio defined by the ratio between the number of ChIP-seq reads from the maternal allele and the total number of reads from two alleles. We then evaluated the correlation between the change of SEM or PWM scores and allelic ratios.

2.6 Electrophoretic mobility shift assay (EMSA) analysis

The DNA-binding domains of CTCF (F1–F9) were amplified from Addgene plasmid 102859 and cloned into a bacterial expression vector with a GST tag (pGEX4T) (Zuo *et al.*, 2017). This construct was transformed into BL21(DE3) cells. 1L LB liquid bacteria cultures were induced by 0.25 mM IPTG at OD600 = 0.6 and incubated at 12°C for 24 h. Cells were lysed by sonication, and GST-CTCF was pulled down by a glutathione column. Following five washes with wash buffer (20 mM HEPES-KOH, pH 7.2, 150 mM KCl, 0.05% NP-40, 10% glycerol), the sample was cleaved by thrombin and run through a column, resulting in purified, cleaved CTCF (F1–F9) protein (Supplementary Fig. S2A).

For our EMSA analysis, we tested a 20-bp genomic binding region to CTCF flanked by 200 bp upstream and downstream of endogenous sequence (hg19, chr9: 135045357–135045377). We introduced mutations to create 10 variable regions containing a single mutation and one scrambled region. We completed EMSAs as previously reported (Levitsky *et al.*, 2014), incubating 50 nM DNA fragments with 0, 50, 100, 250 and 500 nM purified CTCF protein fragments for 30 min. EMSA reactions were then run on 4–12% TBE gels (EC62352BOX) for 3 h at 80 V and 4°C. EMSA analysis was completed as previously reported using densiometric scanning by ImageJ and an Excel Solver Package (Aghera *et al.*, 2011; Schneider *et al.*, 2012). EMSA scores were normalized to the genomic background (+) and scaled between 0 and 1.

3 Results

3.1 SEM scores better recapitulate endogenous binding than PWMs

SEM scores are expected to be more representative of endogenous binding patterns than PWMs as these predictions are generated using an endogenous measure of genome-wide binding affinity. We demonstrate this by correlating SEM and PWM scores across full-length kmers for transcription factor FOXA1 to their average ChIP-seq signals at corresponding sequences genome wide (Fig. 3). When comparing predictions with experimentally generated binding affinity data above standard cutoffs (see Section 2), SEMs had a stronger correlation than PWMs (SEM: $R^2 = 0.66$, PWM: $R^2 = 0.24$), demonstrating our predictions represent a more robust measure of endogenous binding affinity. This pattern holds true when allowing a very lenient PWM cutoff of 11 ($R^2 = 0.28$) as well as for the entire datasets (SEM: $R^2 = 0.19$; PWM: $R^2 = 0.03$) (Supplementary Fig. S3).

These findings indicate that SEM plots better recapitulate known patterns of transcription factor binding beyond the information detailed in a PWM. Of note, there are cases where the PWM shows approximately equal information content for distinct bases sharing a position, yet the SEM plot reveals a wide margin of binding differences between the two bases fueled by differences in predicted direction of effect on binding affinity (i.e. position 3 or 10 of HNF4a in Fig. 2).

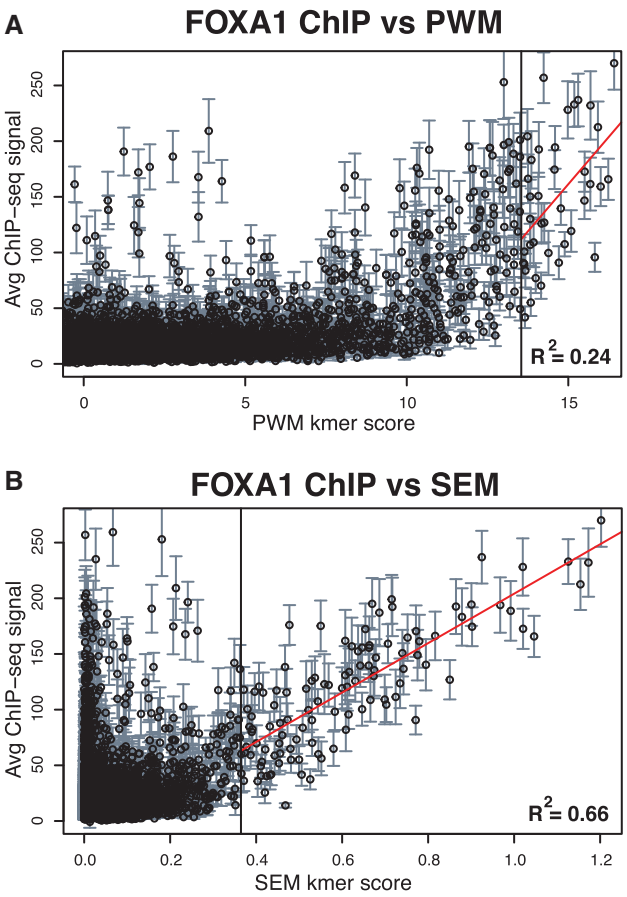

Fig. 3. SEMs show a better correlation with whole kmer ChIP-seq signal (**B**, $R^2 = 0.66$) than PWMs (**A**, $R^2 = 0.24$). The line dividing the plot represents a standard cutoff for PWM visualization (*P*-value = 4^{-8}). Coefficient of determinations (R^2) were calculated to the right of the vertical lines, representing the TFM-PVALUE cutoff for PWMs and the average scrambled background cutoff for SEMs (0.36 for FOXA1). SEM values are displayed as 2^n for visualization purposes. PWM values only shown >0, a full plot can be found in Supplementary Figure S3

3.2 Ubiquitous transcription factors show cell type and dataset independence

To determine if SEM results show a dataset-specific dependence, we evaluated the transcription factor FOXA1 using ChIP-seq data from two different ENCODE datasets gathered in the same HepG2 cell line (ENCFF658RGX; ENCFF898FCL) (Fig. 3). We found nearly identical SEMpl outputs (*P*-value = 4.14e-56, RMSD = 0.0178) using least-squares regression analysis.

We next expanded this to investigate if SEM results were dependent on the cell line used and thus included three additional ChIP-seq datasets (ENCFF699KBP; ENCFF845PAS; ENCFF723DLM) from distinct cell types (Fig. 4). It is important to note that while some of the regions tested in the cell lines are at the same locations, there are large differences in the open chromatin regions (and thus site accessibility) across these cell types, often with >50% unique sites between cell types (bottom half of Fig. 4). We saw high levels of correlation using these additional cell types, with R^2 values over 0.97 for HepG2, A549 and T47D (*P*-values < 1e-32, RMSD < 0.0717). We also saw this trend between SEMs run on different cell lines for additional transcription factors including in MYC, NKFB1 and FOS, suggesting that for ubiquitous transcription factors, we expect there to be no

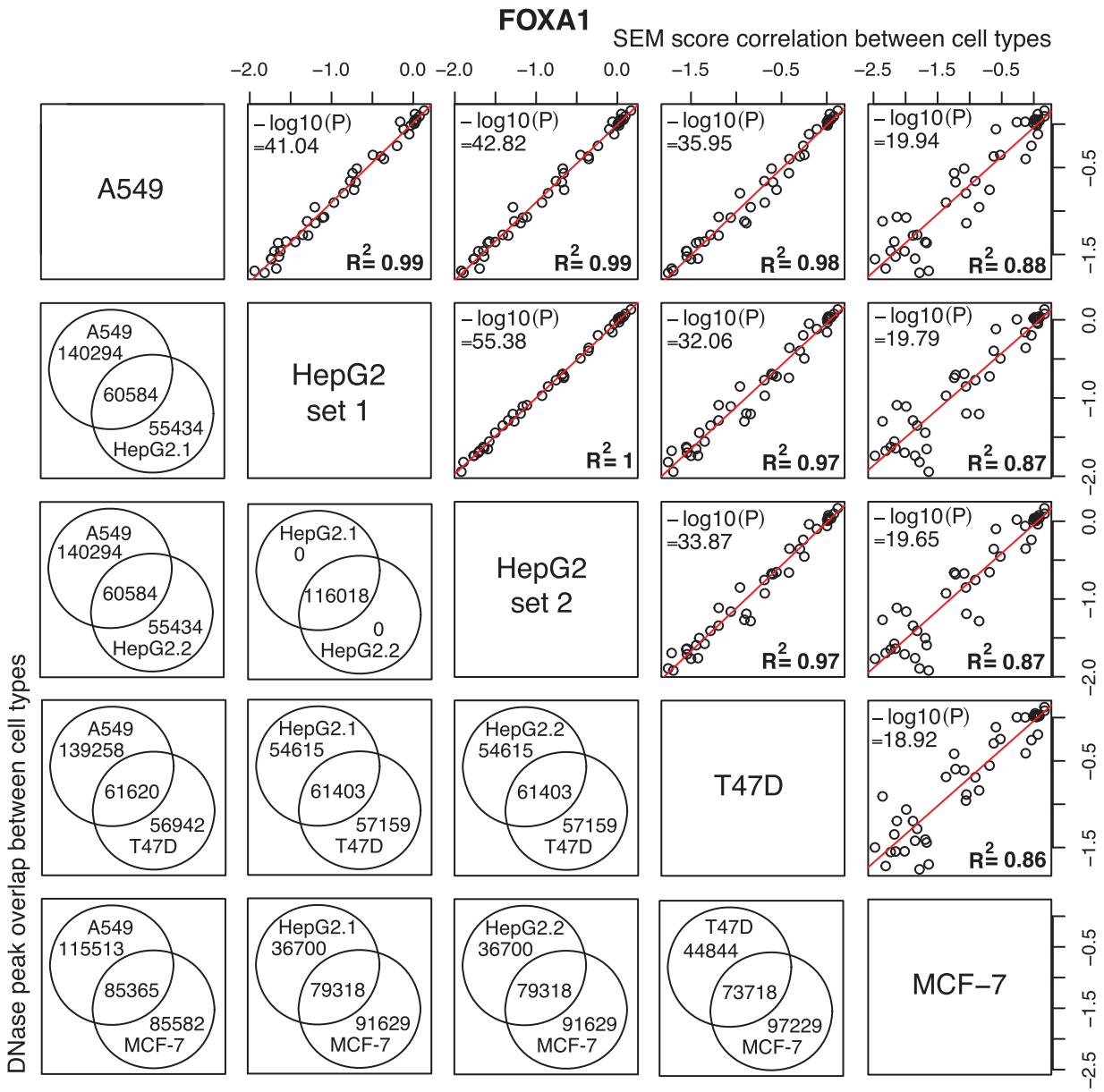

Fig. 4. Different ChIP-seq input produce similar SEMs. The top right half of the table shows a least square regression analysis which reveals that FOXA1 SEMs are highly correlated across four cell types and one pair of biological replicates with correlations between samples ranging from R^2 =0.86 and R^2 = 1. The bottom left half of the table shows overlapping DNase peaks between cell types. A549, lung carcinoma cell line; HepG2, hepatocellular carcinoma cell line; T47D, breast tumor cell line; MCF-7, breast adenocarcinoma cell line

appreciable difference between SEMpl outputs (Supplementary Figs S4–S6).

It has been proposed that there may be binding affinity differences between cell types when a transcription factor has known cell type-specific functions or cofactors. To address this, we investigated the proto-oncogene *MYC*, which encodes for the transcription factor c-myc known to have distinct functions and cofactors between differing cell types (Cappellen *et al.*, 2007). Interestingly, we found that c-myc yielded a highly similar pattern between almost all cell types observed, but a distinct SEM plot in HeLa cells that cannot be explained by low data quality (Supplementary Fig. S4). This suggests that SEMpl can also be used to identify transcription factors that have distinct cell type-specific functions. However, this seems to be the exception rather than the rule as the majority of SEMs we observed were cell-type agnostic.

Finally, we asked if the starting PWM for a TF would influence the final SEM output. We found no appreciable difference in SEMpl outputs when using different starting PWMs, given that the starting PWMs represent the general binding of the transcription factor of interest (Supplementary Fig. S1). However, certain PWMs and/or datasets do not contain enough information about the binding of a TF and so do not produce any significant enrichments in the final SEM output and are thus discarded (Supplementary Table S1).

3.3 SEMpl recapitulates known allele-specific binding patterns

Allele-specific binding differences in noncoding regions of the genome have long been associated with regulatory sequence (Kasowski *et al.*, 2010; McDaniell *et al.*, 2010). To compare SEM scores against known allele-specific binding data, we annotated heterozygous sites in the GM12878 cell line with ChIP-seq read counts from two alleles using ENCODE CTCF ChIP-seq datasets. Least-squares regression analysis of SEM or PWM score changes against ChIP-seq

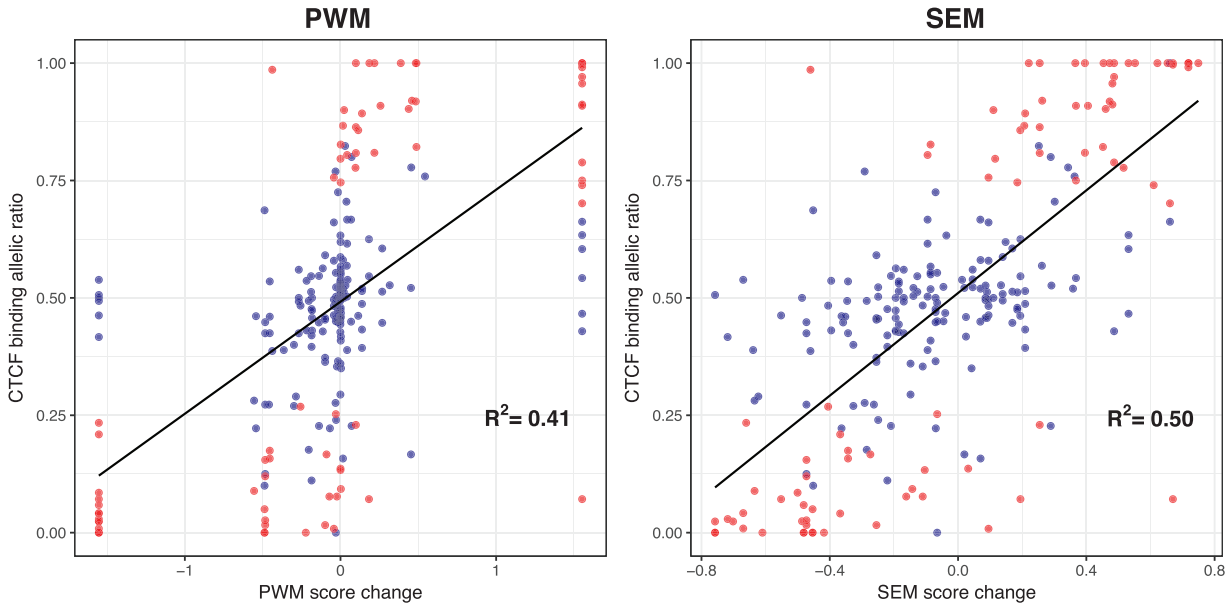

Fig. 5. SEMs reflect allele-specific CTCF-binding patterns. Linear regression reveals a higher correlation between SEM score change and binding affinity change in two alleles of heterozygous sites ($R^2 = 0.50$) than PWM scores ($R^2 = 0.41$). Allele-binding affinity change was measured by allelic ratio, which is the ratio between CTCF ChIP-seq read counts from maternal allele and total read counts from two alleles. Allele-specific binding sites (red/light gray points) generally have larger changes on SEM scores. (Color version of this figure is available at *Bioinformatics* online.)

signal changes of these 240 heterozygous sites in CTCF-binding sites revealed a higher correlation for SEM score changes with an R^2 of 0.50 compared to a PWM R^2 of 0.41 (Fig. 5). We also observed a more dispersed distribution of SEM score changes, where the allele-specific binding sites have overall larger changes between two alleles (red points in Fig. 5). These indicate that the SEM score is more able to capture the change of TF-binding affinity compared to PWM.

To validate that SEMpl scores accurately predict transcription factor-binding affinity changes *in vitro*, we compared SEMpl scores to previously generated ChIP-qPCR data, which measures endogenous transcription factor-binding affinity (Cowper-Sal Lari *et al.*, 2012). ChIP-qPCR was generated from 10 allele-specific FOXA1-binding sites in the genome. Regression analysis comparing SEMpl scores to changes in transcription factor binding by ChIP qPCR analysis reveal that SEM scores are a better predictor of SNP changes ($R^2 = 0.64$) than PWMs ($R^2 = 0.44$) (Supplementary Fig. S7).

We examined SEMpl predictions further by comparing them to *in vitro* binding data generated by EMSA of purified protein of the DNA-binding domains of CTCF to engineered DNA consensus sequences. EMSAs of 10 CTCF-binding sites containing a mutation, which we define here as variable regions, compared to a known CTCF-binding site along with the endogenous sequence and scrambled background reveals a better correlation with SEM predictions ($R^2 = 0.76$) than PWM predictions ($R^2 = 0.65$) (Fig. 6A, Supplementary Fig. S2). This is further supported by comparing SEM and PWM scores to previously published EMSA data for the mouse transcription factor FoxA1 (Levitsky *et al.*, 2014). This analysis showed a marked improvement of SEM scores ($R^2 = 0.75$) compared to PWM scores ($R^2 = 0.6$), and machine learning models DeepBind ($R^2 = 0.66$) and LS-GKM ($R^2 = 0.67$), and suggests that the SEMs of highly conserved transcription factors may be comparable between species (Fig. 6B) (Alipanahi *et al.*, 2015; Lee *et al.*, 2015). Together, these results suggest that SEMpl has the ability to return biologically meaningful results and can be used to predict the direction and magnitude of allele-specific changes.

3.4 SEMpl predictions agree with experimentally validated SNPs from the literature

To verify that SEMpl would allow researchers to identify variants potentially leading to transcription factor-binding changes associated with gene expression changes, we validated our method against four published TFBS SNPs found to disrupt transcription factor binding (Supplementary Fig. S8). In most cases, we found that SEMpl predictions agreed with the direction of the validated changes, as well as the magnitude, when available. For example, a T to G change in position 12 of a TCF7L2-binding site was found to increase binding affinity by 1.3-fold by mass spec (Pomerantz *et al.*, 2009), where SEMpl predicted a 1.27-fold increase. Only one of the four SEMpl predictions that we identified did not match the experimentally determined variant. This C/T allele in position 11 of a FOXA2-binding site was predicted to decrease binding affinity by FAIRE-seq, however SEMpl predicted no difference in binding between the two alleles (data not shown). Interestingly, PWMs also predicted no difference in binding between the two alleles, suggesting additional factors may be at play.

We also compared SEMpl predictions to predicted variant effects measured through a massively parallel reporter assay (MPRA) (Kheradpour *et al.*, 2013). We found a correlation between these previously published expression changes and SEM score changes (Supplementary Fig. S9). However, this relationship was not as strong ($R^2 = 0.23$), though still outperforming PWMs ($R^2 = 0.16$), possibly due to the nonlinear relationship between transcription factor binding, regulatory element use and gene expression.

3.5 SEMpl outperforms other methods in predicting changes to transcription factor binding

In order to compare SEMpl to current state-of-the-art methods, we compared SEMpl and PWMs to methods utilizing machine learning able to predict the consequence of variants to transcription factor binding, DeepBind and LS-GKM (Alipanahi *et al.*, 2015; Lee *et al.*, 2015). Both tools use models trained on ChIP-seq datasets to generate predictions of function variation. Here, we compared scores for all methods (PWM, SEMpl, Deepbind and LS-GKM) against ChIP-seq scores for all kmers from 13 transcription factors (Supplementary Fig. S10).

Using a performance comparison, we found that SEMpl better correlates with ChIP-seq data than both DeepBind and LS-GKM for 6/13 of the transcription factors tested, and comparably to 3/13 (Fig. 7). Of the final four transcription factors, two were better predicted by PWMs (EGR1 and MEF2A), HNF4a was poorly predicted by all methods and FOXA2 was best predicted by DeepBind. However, we note that, with some exception, all methods do have

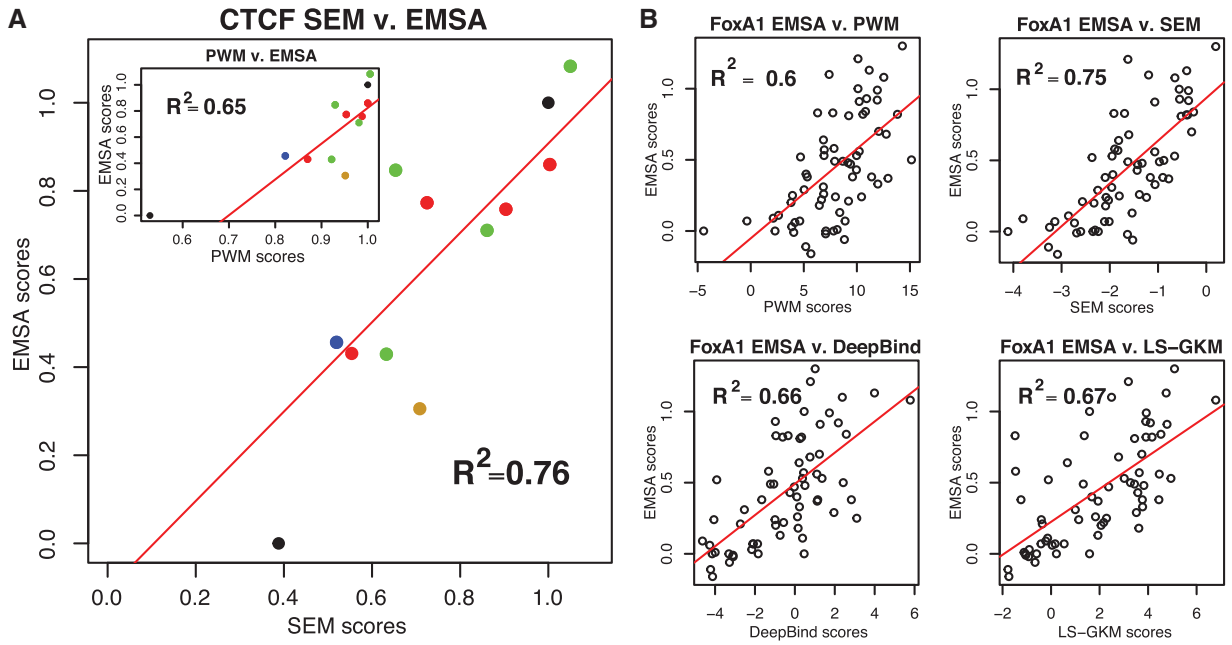

Fig. 6. SEMpl scores agree with *in vitro* transcription factor-binding results. (**A**) Electrophoretic mobility shift assay (EMSA) for CTCF correlated to SEMpl and PWM predictions. Correlations are calculated without the inclusion of the genomic and scrambled controls (black points). Additional colors correspond to the SNP change made to the variable region. (**B**) FoxA1 EMSA data from Levitsky *et al.* correlated to PWM, SEM, DeepBind and LS-GKM predictions (Levitsky *et al.*, 2014)

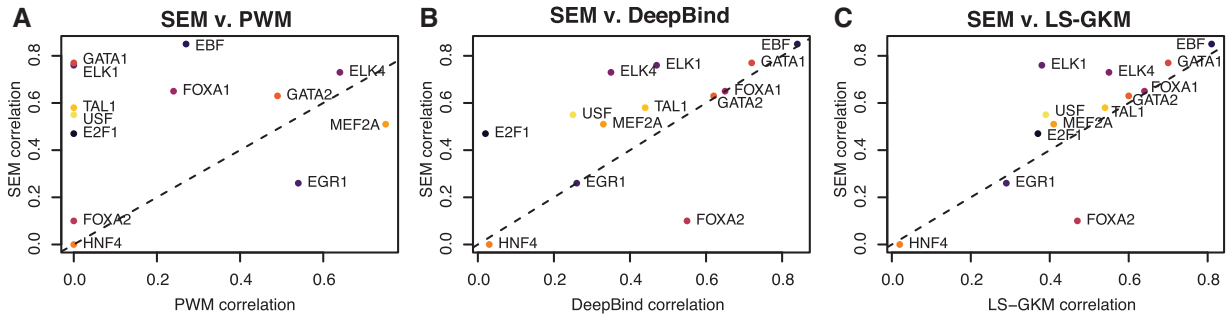

Fig. 7. Performance comparison of SEMpl to other noncoding SNP prediction methods. Predictions for 13 TFs were generated using PWM (**A**), SEM, DeepBind (**B**), and LS-GKM (**C**) and compared to the average ChIP-seq score for the analogous kmer sequence. Correlations for each transcription factor were then compared across methods. SEMpl produced better or comparable correlations for 9/13 transcription factors tested. PWMs performed better for EGR1 and MEFF2A, and DeepBind performed best for FOXA2. All methods performed poorly for HNF4. The colors/shades of gray of points are unique to each transcription factor. (Color version of this figure is available at *Bioinformatics* online.)

good apparent correlation with ChIP-seq data and provide some indication of the effect of variation on TF binding. We would expect transcription factors with binding strongly dependent on sequence outside of the central motif to be better predicted by machine leaning models such as DeepBind and LS-GKM, however for the majority of transcription factors examined here SEMpl predictions based on the central motif were sufficient. This is interesting as it suggests reasonable predictions for transcription factor-binding affinity changes can be made using a much simpler scoring system, analogous to scoring using a PWM, while avoiding the pitfalls and computational effort required to train a machine learning model. Indeed, by providing pregenerated predictions for many transcription factors, we hope to make using SEMpl as fast and straightforward as possible.

4 Discussion

A deeper understanding of the role noncoding variants play in altering gene expression is critical to fully illustrate the regulatory complexity of our genome and is an important first step toward developing tools for personalized medicine. Approaches such as the IGR method have expanded our ability to use currently available data to predict SNPs that play a regulatory role and have successfully been implemented in multiple studies to link human disease to specific transcription factors and their binding sites. Since its release, the IGR method has been used to successfully identify functional SNPs in TFBSs from GWAS data for breast cancer, atrial fibrillation and lupus (Bailey *et al.*, 2015; Ye *et al.*, 2016; Zhang *et al.*, 2018). Functional predictions for these SNPs were experimentally validated, suggesting that the IGR process can be a robust method for functional noncoding GWAS SNP prediction. Unfortunately, this method is not accessible for widespread use. By developing a tool which generalizes the IGR methodology to predict the magnitude and direction of effect of all SNPs within a TFBS, we can identify novel variants associated with human disease in TFBSs genome-wide.

In this article, we introduced SEMpl, a new tool designed to identify putative deleterious mutations in TFBSs. SEMpl predictions reflect known patterns of transcription factor binding while providing additional information about magnitude and direction of predicted change. We demonstrate that SEMpl provides more robust

and consistent predictions both on a single variant and a TFBS kmer level than the current standard, PWMs. The method leverages simulation and real data to better model strength of binding rather than a consensus sequence. Additionally, SEMpl scores correlate with known allele-specific binding sites and agree with *in vitro* binding analysis via ChIP qPCR and EMSA as well as previously published variants known to alter transcription factor-binding affinity. Importantly, we found that SEMpl predictions outperform popular machine learning methods for the majority of transcription factors tested.

SEMpl was designed to be easy to use and accessible. In addition to being available as an open source application, precompiled SEM plots for 90 transcription factors from over 200 PWMs are available online. While SEMpl is currently limited to transcription factors with available ChIP-seq and PWM data, we may be able to eliminate the use of PWMs to guide TFBS loci in future versions of our pipeline, potentially by using overrepresented kmers from the ChIP-seq data, which would reduce bias and expand our list of compatible transcription factors. In addition, we are working to include additional genomic features, such as DNA methylation which would allow the inclusion of additional bases to SEM plots and a more nuanced understanding of transcription factor binding.

SEMpl's ability to better predict the impact of genomic variation on transcription factor binding has broad implications to the cross-disciplinary study of the regulatory genome. SEMpl has great usability for prioritizing GWAS SNPs for experimental follow-up, in individual studies or through the evaluation of noncoding GWAS catalog SNPs. With the increased need for experimental validations following large-scale genomics studies, we anticipate that annotation tools, such as SEMpl, will be critical in revealing developmental and disease-associated regulatory SNPs.

Funding

This work was supported by the National Institutes of Health [U41 HG009293 to A.P.B., T32 HG00040 to S.S.N.]. C.M. and C.W. were supported by the University of Michigan Undergraduate Research Opportunity Program.

Conflict of Interest: none declared.

References

Aghera,N. *et al.* (2011) Equilibrium unfolding studies of monellin: the double-chain variant appears to be more stable than the single-chain variant. *Biochemistry*, **50**, 2434–2444.

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.

Andersen,M.C. *et al.* (2008) In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput. Biol.*, **4**, e5.

Bailey,S.D. *et al.* (2015) ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun*, **6**, 6186–6194.

Barenboim,M. and Manke,T. (2013) ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation. *Bioinformatics*, **29**, 2197–2198.

Bembom,O. (2019) seqLogo: Sequence logos for DNA sequence alignments.

Boyle,A.P. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, **22**, 1790–1797.

Cappellen,D. *et al.* (2007) Novel c-MYC target genes mediate differential effects on cell proliferation and migration. *EMBO Rep.*, **8**, 70–76.

Chen,J. *et al.* (2016) A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat. Commun.*, **7**, 1–6.

Cowper-Sal Lari,R. *et al.* (2012) Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.*, **44**, 1191–1198.

Edwards,S.L. *et al.* (2013) Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.*, **93**, 779–797.

ENCODE Project Consortium. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Foat,B.C. *et al.* (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141–e149.

Fogarty,M.P. *et al.* (2014) Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. *PLoS Genet.*, **10**, e1004633-10.

Gaulton,K.J. *et al.* (2010) A map of open chromatin in human pancreatic islets. *Nat. Genet.*, **42**, 255–259.

He,H. *et al.* (2015) Multiple functional variants in long-range enhancer elements contribute to the risk of SNP rs965513 in thyroid cancer. *Proc. Natl. Acad. Sci. USA*, **112**, 6128–6133.

Higgins,G.A. *et al.* (2015) Epigenomic mapping and effect sizes of noncoding variants associated with psychotropic drug response. *Pharmacogenomics*, **16**, 1565–1583.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA*, **106**, 9362–9367.

Hume,M.A. *et al.* (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.

Jolma,A. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.

Jolma,A. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.

Kasowski,M. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.

Khan,A. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.

Kheradpour,P. *et al.* (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.*, **23**, 800–811.

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lee,D. *et al.* (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, **47**, 955–961.

Levitsky,V.G. *et al.* (2014) Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genomics*, **15**, 80.

Macintyre,G. *et al.* (2010) is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, **26**, i524–i530.

Manke,T. *et al.* (2010) Quantifying the effect of sequence variation on regulatory interactions. *Hum. Mutat.*, **31**, 477–483.

McDaniell,R. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235–239.

Musunuru,K. *et al.* (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, **466**, 714–719.

Nishizaki,S.S. and Boyle,A.P. (2017) Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends Genet.*, **33**, 34–45.

Pomerantz,M.M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Riley,T.R. *et al.* (2015) Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. *Elife*, **4**, 307.

Savic,D. *et al.* (2011) Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Res.*, **21**, 1417–1425.

Schneider,C.A. *et al.* (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Shrikumar,A. *et al.* (2017) Learning important features through propagating activation differences. *arXiv*, cs.CV. https://arxiv.org/abs/1704.02685v1.

Stitzel,M.L. *et al.* (2010) Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab.*, **12**, 443–455.

Stormo,G.D. *et al.* (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.*, **10**, 2997–3011.

Touzet,H. and Varré,J.-S. (2007) Efficient and accurate P-value computation for position weight matrices. *Algorithms Mol. Biol.*, **2**, 316–312.

Umer,H.M. *et al.* (2016) A significant regulatory mutation burden at a high-affinity position of the CTCF motif in gastrointestinal cancers. *Hum. Mutat.*, **37**, 904–913.

VanderMeer,J.E. and Ahituv,N. (2011) cis-regulatory mutations are a genetic cause of human limb malformations. *Dev. Dyn.*, **240**, 920–930.

Vorontsov,I.E. *et al.* (2015) PERFECTOS-APE predicting regulatory functional effect of SNPs by approximate P-value estimation. *Bioinformatics*. **1**, 102–108.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Ward,L.D. and Kellis,M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930–D934.

Weirauch,M.T. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.

Ye,J. *et al.* (2016) A functional variant associated with atrial fibrillation regulates PITX2c expression through TFAP2a. *Am. J. Hum. Genet.*, **99**, 1281–1291.

Zhang,F. and Lupski,J.R. (2015) Non-coding genetic variants in human disease. *Hum. Mol. Genet.*,

Zhang,H. *et al.* (2018) Meta-analysis of GWAS on both Chinese and European populations identifies GPR173 as a novel X chromosome susceptibility gene for SLE. *Arthritis Res. Ther.*, 20, 92.

Zhao,Y. and Stormo,G.D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.*, **29**, 480–483.

Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.

Zuo,Z. *et al.* (2017) Measuring quantitative effects of methylation on transcription factor-DNA binding affinity. *Sci. Adv.*, **3**, eaao1799.