Whole Genome Regulatory Variant Evaluation for Transcription Factor Binding

Haoyang Zeng[1], Tatsunori Hashimoto[1], Daniel D Kang[1], David K Gifford[1,2]*

[1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, United States of America

[2]Department of Stem Cell and Regenerative Biology, Harvard University and Harvard Medical School, Cambridge, MA, United States of America

*Corresponding author

E-mail: gifford@mit.edu (DG)

Abstract

Contemporary approaches to predict single nucleotide polymorphisms (SNPs) that alter transcription factor binding rely upon the sequence affinity of a transcription factor as represented by its canonical motif. WAVE (**W**hole-genome regul**A**tory **V**ariants **E**valuation) is a novel method for predicting more general regulatory variants that affect transcription factor binding, including those that fall outside of the canonical motif. WAVE learns a k-mer based generative model of transcription factor binding from ChIP-seq data and scores variants using its generative binding model. The k-mers learned by WAVE capture more sequence feature in transcription factor binding than a motif-based approach alone, including both a transcription factor's canonical motif as well as associated co-factor motifs. WAVE significantly outperforms motif-based methods in predicting SNPs associated with allele-specific binding.

Author Summary

Specific variations in our genome sequence can render us more susceptible to a genetic disease. Certain disease risks are caused by genetic variations that alter where transcription factors bind to the genome and regulate cellular function. Previous methods for identifying which genetic changes are significant have assumed that transcription factor binding is dependent on a short single sequence recognized by a transcription factor. Here we consider a more general model where the binding of a factor may be up or down regulated by any number of short DNA sequences that are proximal to a binding site. Our method substantially improves the detection of genomic changes that are important for factor binding.

Introduction

Genome-Wide Association Studies (GWAS) have proved to be a rich source of genetic polymorphisms that are strongly associated with complex traits and diseases [1–4]. Variants in protein coding sequences include missense and nonsense mutations are simple to characterize. However, many GWAS detected variations reside in non-coding regions with regulatory function [4,5]. The influence of such variation in partially annotated non-coding regions on gene expression and other cellular functions are not well understood. Previous work has observed that non-coding DNA changes in the recognition sequences of transcription factors (TF) can affect gene expression and cellular phenotypes [6]. Thus predicting the effect of genomic variants on TF binding is an essential part of interpreting the role of non-coding variants in pathogenesis. Existing computational approaches to predict the effect of SNPs on TF binding such as sTRAP, rSNP-MAPPER and HaplogReg2 are based on quantifying the difference of the reference and alternate alleles in matching to canonical TF motifs [7–13].

WAVE (Whole-genome regulAtory Variants Evaluation) learns the sequence features essential for TF binding *de novo* from whole-genome ChIP-seq data and scores variants by the predicted change in ChIP-seq read counts between the reference and alternate alleles. WAVE improves on the motif-based model in two ways. First, WAVE doesn't assume the existence of a canonical motif. Instead it provides more descriptive power in modeling transcription factor binding through *de novo* sequence effect learning. This allows us to capture more subtle sequence features underling transcription factor binding such as non-canonical motifs. Second, WAVE accounts for the spatial effect of the sequences and learns the effect of cis-regulatory regions outside of the motif. This enables us to model the role of important auxiliary sequences in transcription factor binding, such as cofactors.

We evaluate WAVE on the ChIP-seq data for transcription factors NF-κB and CTCF. In both datasets, the active k-mers identified by WAVE captures the canonical TF motifs as well as associated sequences such as known co-factors. WAVE also significantly outperforms existing approaches in prioritizing SNPs associated with NF-κB and CTCF allele-specific binding (ASB).

Results

WAVE learns a vocabulary of k-mers that regulate factor binding

WAVE is a fully generative model of ChIP-seq reads of the genome. We assume that the genome is a long regulatory sequence that contains k-mers as "code words" that induce invariant spatial effects on proximal transcription factor binding. Following this assumption, we modeled the read counts produced by transcription factor ChIP-seq at a given base as the log-linear combination of spatial effect of a set of learned k-mers whose effect range covers that base.

WAVE first learns the spatial effect of all k-mers (k=1 to 8) over a spatial window of ± 400 base pairs (bp) *de novo* from ChIP-seq data using a regularized Poisson regression (Fig. 1A). WAVE then computes the predicted ChIP-seq read counts for the reference and alternate allele of a variant from the log-linear combination of the spatial effect of the learned k-mers. WAVE predicts the effect of a genomic variant on transcription factor binding by the sum of squared per-base change of predicted reads between two alleles. (Materials and Methods)

Fig. 1. WAVE accurately predicts ChIP-seq signal across the genome
(**A**) The schematic of WAVE pipeline. The spatial effects of all the k-mers are learned from the ChIP-seq datasets. And then the k-mer-specific profiles (purple, cyan and green) corresponding to the k-mers underlining the reference and alternate alleles for a variant are aggregated by log-linear combination to yield a spatial prediction of local ChIP-seq reads for the reference (blue) and alternate allele (red). WAVE scores the variant by the change of predicted reads. (**B**) Example held-out genomic region on

chromosome 14 showing ChIP-seq reads (red), WAVE-predicted reads (black), and reads from a negative control model trained on rabbit IgG ChIP-seq data (green). (**C**) Comparison of WAVE-predicted (x-axis) and observed (y-axis) ChIP-seq reads in binned regions of held-out chromosome 14. Models were trained on combined ChIP-seq data from 10 ENCODE LCL individuals (black) or rabbit IgG ChIP-seq data (red).

Although WAVE fits a model with potentially large parameter space (± 400 bp window for 87380 k-mers when $k_{max} = 8$), it uses sparsifying regularization to avoid overfitting and to limit the number of active k-mers (Materials and Methods). For example, in the NF-κB WAVE model, most of the binding signal is predicted by 1% of the 87380 k-mers (S1 Fig.). WAVE is also robust to the choice of window sizes of k-mer's spatial effect, although we found that WAVE model with window size of ± 400 bp produced the best Pearson's correlation (S1 Table)

We first tested if WAVE could predict held-out ChIP-seq data. We trained a WAVE model on NF-κB ChIP-seq data from chromosomes 1-13 of 10 LCL ENCODE individuals and compared the predicted ChIP-seq signal from WAVE to ChIP-seq reads on a held-out chromosome (chromosome 14). The predicted ChIP-seq signals are very similar to actual ChIP-seq reads (Fig. 1B-C), with Pearson's correlations of 0.64 chromosome-wide and 0.44 restricted to regions within 2kb of a binding event identified by GEM [14]. Negative control WAVE models trained to capture biases such as chromatin state on rabbit IgG ChIP-seq datasets yield Pearson's correlations of only 0.51 chromosome-wide and 0.10 within 2kb of binding events.

We found the SNP scores generated by WAVE are consistent across similar training datasets. We trained four separate WAVE models on NF-κB ChIP-seq data from four different individuals GM12878 (CEU), GM12892 (CEU), GM18951 (JPT) and GM19193 (YRI). The SNP scores for the set of common (minor allele frequency ≥1%) SNPs from the 1000 Genomes Project (1KG) are consistent across the four different models (S2 Fig.). Moreover, we found the WAVE model trained on the combined ChIP-seq data from 10 LCL individuals had a clear improvement in Pearson's correlation between predicted and actual reads when compared with any model trained from a single individual (S1 Table).

WAVE captures sequence features of TF and its co-factors

We then examined if WAVE correctly learned the strongest expected sequence features from the binding data that correspond to the canonical motifs for NF-κB and CTCF. Both WAVE models were trained on combined ChIP-seq data from 10 LCL ENCODE individuals and position weight matrices were generated for visualization purposes by hierarchical clustering of the active k-mers in WAVE (Materials and Methods) and matched to known TF motifs in JASPAR and TRANSFAC with STAMP [14]. We found that the top two k-mer clusters for NF-κB were strongly matched to motifs from NF-κB family (Fig. 2A) and the top 6 k-mer clusters for CTCF were all strongly matched to the CTCF motif (Fig. 2B and S3 Fig.).

Fig. 2. Active k-mers detected by WAVE contains the canonical TF motif and associated cofactors

(A**)** The average k-mer effect and aggregated position weight matrix of the strongest k-mer cluster from the NF-κB WAVE model (columns 1 and 2) along with the motif for REL from NF-κB family (MA0107.1 from JASPAR). **(B)** The average k-mer effect and aggregated position weight matrix of the strongest k-mer cluster from the CTCF WAVE model (columns 1 and 2) along with the CTCF motif (MA0139.1 from JASPAR). **(C)** The average k-mer effect and aggregated position weight matrix of the 3rd, 4th, 6th, and 9th strongest k-mer clusters from the NF-κB WAVE model compared with motifs for NRF1 (M00652 from TRANSFAC), EST1 (M00032 from TRANSFAC), AP1 (MA0099.2 from JASPAR) and IRF1 (M00062 from TRANSFAC).

Many of the other k-mer clusters learned by WAVE correspond to co-factor binding motifs. The top k-mer clusters in the NF- κB WAVE model matched to ETS1, AP1, IRF1 and NRF1 (Fig. 2C), which have been associated with NF-κB regulation [15–18] . To validate the role of these transcription factors in NF-κB binding, we performed co-factor analysis on the same NF-κB data using GEM to search for transcription factors that have spatially binding constraint with NF-κB. This analysis identified AP-1 and IRF1 as the strongest co-factors of NF-κB binding. Thus, WAVE captures the sequence context of factor binding and provides additional descriptive power.

WAVE substantially outperforms motif-based approach in prioritizing ASB SNPs

We then compared WAVE's performance against motif-based approaches in discriminating SNPs that are known to alter transcription factor binding. Allele-specific binding (ASB) studies have identified SNPs associated with significantly imbalanced binding events on heterozygous sites, making these SNPs an ideal standard for benchmarking [19,20]. We collected ASB SNPs with known differential binding as positive sets, resulting in a total of 56 SNPs for NFKB and 60 SNPs for CTCF (Materials and Methods).

We constructed three sets of negative SNPs that we assume do not exhibit differential factor binding. All of these negative sets are subsets of 1KG common (minor allele frequency ≥1%) SNPs. The first negative set is a random selection of 1000 1KG common SNPs from across the genome in order to sample overall background. To account for the non-uniform distribution of ASB SNPs on the genome, the second negative set was composed of 1KG common SNPs within 1kb from an ASB SNP. We found that 47 out of 56 NF-κB ASB SNPs reside in one of the 15522 NF-κB binding regions (BR) identified by previous work[21]. Thus the third negative set was constructed for NF-κB analysis only to control for the confounding effects arise from proximal binding strength. This final negative set is a subset of second negative set that are located in any BR that contains a positive ASB SNP.

We evaluated the performance of Haploreg2, sTRAP, rSNP-Mapper, and WAVE in discriminating our positive set from each of our three negative sets. The NF-κB and CTCF WAVE models were trained on combined data from 10 LCL ENCODE individuals. We found that WAVE outperformed all the other tested methods and had an AUC > 0.7 on the third and most stringent negative set (Fig. 3 and S4 Fig.). Since HaploReg2 does not provide a way to retrieve the actual score for each SNP, its binary classification performance was plotted as a point on the ROC curves.

Fig. 3. WAVE significantly outperforms motif-based approaches in prioritizing NF-κB ASB-SNPs

ROC curves for discriminating NF-κB ASB-SNPs from each of the three negative sets using WAVE, sTRAP, rSNP-MAPPER and HaploReg2. Dashed line indicates random chance. **(A)** Positive set chosen as all NF-κB ASB-SNPs (n=56)**.** Negative set randomly sampled from 1KG common SNPs (n=1000). **(B)** Negative set limited to common 1KG variants that are located within 1000 bp from any NF-κB ASB-SNP (n=828). **(C)** Positive set limited to NF-κB ASB-SNP that are in BR (n=47). Negative set limited to common 1KG variants that are located within same BR as any SNP in the positive set (n=372).

WAVE prioritizes SNPs that disrupt motifs

The power of motif-based approaches is constrained to evaluating variants that fall inside a factor's motif. To evaluate the power of WAVE on this narrower task we next tested WAVE's performance on SNPs in our positive set that could be detected by motif based methods.

As the first step, we classified our positive sets of NF-κB and CTCF ASB SNPs using sTRAP, rSNP-MAPPER and HaploReg2 with their default parameter and cutoff settings (Materials and Methods). As motif disruption has been considered the primary mechanism by which variants alter transcription factor binding, we would expect most of the ASB SNPs to be identified by these methods. Surprisingly we found that the best motif-based method (rSNP-MAPPER for NF-κB and HaploReg2 for CTCF) detected only 30% of our positive SNPs. The set of correctly classified positive SNPs by the approaches significantly overlapped, while the efficiency of each method varied across different transcription factors (S2 Table). A total of 21 out of 56 NF-κB ASB SNPs and 24 out of 60 CTCF ASB SNPs were detected by at least one of the three motif-based methods. We refer these SNPs as motif-disrupting (MD) SNPs. A large fraction (~60%) of positive SNPs were not detected by any of the traditional motif-based approaches. These results are consistent with our previous results (Fig. 3) which shows that the prediction power of motif-based approaches dramatically decreases to random after the top 30% positive targets.

We then compared the performance of WAVE, sTRAP, rSNP-MAPPER and HaploReg2 in discriminating the motif-disrupting SNPs from our three negative sets. For both

transcription factors, WAVE achieved performance equal to motif-based approaches with an AUC >0.85 (Fig. 4 and S5 Fig.) in all control scenarios.

Fig. 4. WAVE achieves performance equal to motif-based approaches in prioritizing NF-κB SNPs within binding motifs.
ROC curves for discriminating NF-κB SNPs identified by any motif-based method (MD SNPs) from the various negative sets using WAVE, sTRAP, rSNP-MAPPER and HaploReg2. Dashed line indicates random chance. **(A)** Positive set chosen as all NF-κB MD SNPs (n=21)**.** Negative set randomly sampled from common 1KG SNPs (n=1000). **(B)** Negative set limited to common 1KG variants that are located within 1000 bp from any NF-κB MD SNP (n=409). **(C)** Positive set limited to NF-κB MD SNPs that are in BR (n=18). Negative set limited to common 1KG variants that are located within same BR as any SNPs in the positive set (n=177).

Discussion

We have found the power of position weight matrices to be insufficient to properly score the effect of variants on factor binding. Motif-based approaches were only able to correctly annotate 30% of ASB SNPs in our test set. In addition, the performance of a motif-based model can vary dramatically across different transcription factors (Table S2). We expect that the poor performance of rSNP-MAPPER on CTCF might be the consequence of its use of a single CTCF position weight matrix (Materials and Methods). Thus motif-based methods are strongly constrained by their underling model.

WAVE significantly outperformed motif-based scoring methods in prioritizing ASB SNPs from negative controls. We find that WAVE's incorporation of a window of sequence context permits it to model the effect of other sequences involved in transcription factor binding including co-factors. These sequences are neglected by conventional motif-based motif modeling.

WAVE outputs a numeric score for each SNP that is easy to interpret as the predicted number of reads changed by the variant. We showed that WAVE scores are robust to the choice of window size, and are consistent across the individuals used for training. We further demonstrated that by combining ChIP-seq data from multiple individuals of the same cell lines type to increase the size of the training set, we could improve WAVE's goodness of fit to ChIP-seq reads.

We are training WAVE on many transcription factor ChIP-Seq data to form a library of models that can be used to score the importance of candidate SNPs on a library of transcription factors. With WAVE's superior performance in modeling transcription factor binding and predicting regulatory non-coding variants, we expect WAVE to play an important role in annotating and prioritizing putative causal variants for further downstream analysis.

Materials and Methods

WAVE Model Overview

The WAVE procedure of variant scoring consists of the following three steps:
Step1: Learn the spatial effect of k-mers on TF binding from ChIP-seq datasets.
Step2: Predict the TF binding signal around the reference and alternate allele of each SNP of interest
Step3: Score a SNP by the sum of squared per-base change of binding signal between alleles

Learning the Spatial Effect of K-mers

The effect profile of a k-mer is defined as a real-valued vector of length $2M$ that corresponds to a spatial window of $[-M, M-1]$ relative to the start position of the k-mer. Specifically, the j-th entry of the profile for a k-mer is the expected log-change in read counts at the j-th base relative to the start of the k-mer. Here we consider k-mers with k from 1 to 8 ($k_{max} = 8$) as this is the maximum length learnable with a typical ChIP-seq dataset. As ChIP-seq signals are relatively sparse and spikey, we chose an effect range of ± 400 bp for each k-mer ($M = 400$).

For notational convenience we will use i for genomic coordinate, k for k-mer length, and j for coordinate offset from the start of a k-mer. We assume that the genome consists of one large chromosome with coordinate 0 to N. In practice we will construct this by concatenating chromosomes with the telomeres acting as a spacer. We represent the effect vector of all k-mer of length k as a parameter matrix θ^k of size $4^k \times 2M$. For any particular k-mer of length k starting at base i on the reference genome, we define g_i^k as its row index in θ^k. So $\theta^k_{(g_i^k,\ j)}$ would denote the effect of this kmer at offset $j \in [-M, M-1]$. Additionally, a special parameter θ_0 is used to set the average read rate of the genome globally.

Given these definitions, we define a generative model for ChIP-seq reads on the genome. Observed counts C_i are generated from a Poisson distribution with rate parameter λ_i which is defined as:

$$\lambda_i = \exp\left(\left(\sum_k \sum_{j \in [-M, M-1]} \theta^k_{(g^k_{i+j},\ j)} \right) - \theta_0 \right)$$

The problem we solve is a regularized Poisson regression. Particularly, we would like to maximize the following:

$$\min_\theta \left(\sum_i C_i \log(\lambda_i) - \lambda_i \right) - \eta \Sigma \left| \theta^k \right|_1$$

To efficiently optimize this objective function, we performed minibatch-gradient descent, a variant of stochastic gradient descend method where in each iteration the gradient and error was calculated against a "mini-batch" of all the samples [22]. The detail of implementation can be found in the supplementary text (S1 Text).

ChIP-seq Signal Prediction for Reference and Alternate Allele

In step 2, given the effect profile θ^k of all the k-mers trained from step1, we first predict the ChIP-seq count at each position i across the reference genome by combining the effect of proximal k-mers into the log-linear model:

$$\lambda_i = \exp\left(\left(\sum_k \sum_{j \in [-M, M-1]} \theta^k_{\left(g^k_{i+j},\ j\right)} \right) - \theta_0 \right)$$

Then in similar manner, we predict the read counts of the alternate allele $\lambda_i^{'}$ after replacing the k-mers that are affected by the variant. If we assume a Single Nucleotide Polymorphism (SNP), at most $\frac{4}{3} \times \left(4^{k_{max}} - 1\right)$ k-mers will change.

Variant Scoring

In step 3, we score a SNP at locus i on the genome by the sum of squared per-base change of binding signal at all bases within the effect range of any k-mers affected by the variant:

$$s_i = \sum_{j \in [-M - k_{max} + 1, M-1]} \left(\lambda^{'}_{i+j} - \lambda_{i+j}\right)^2$$

Collapsing WAVE Profiles into PWM

We interpret the active k-mers captured by WAVE with a post-processing framework that aggregates similar k-mers into position weight matrixes after filtering on effect size:

1. We filter k-mers based on the sum of effect to eliminate inactive k-mers.
2. We calculate the pair-wise Levenstein distance of the remaining k-mers.
3. We perform UPGMA hierarchical clustering over the candidate k-mers until the minimal distance among clusters is larger than 2.
4. For each cluster, we define its key k-mer as the one with the largest aggregate effect. We obtain the position weight matrix for this cluster by aligning all k-mers in the cluster against the key k-mer.

Comparing with sTRAP, rSNP-MAPPER, HaploReg2

sTRAP. We used the R version of sTRAP downloaded from the website for scalability. We used motif data from the JASPAR (included in sTRAP R package) and TRANSFAC (2013.1) databases, including MA0105.1, MA0107.1, MA0061.1, M00054, M00194, M00052, M00051, M03557, M00208, M03563 for NFKB and MA0139.1, M01200,

M01259 for CTCF. For binary classification, we chose an absolute log ratio cutoff of 0.21 and min-pvalue cutoff of 0.1 as was suggested by the sTRAP paper. To plot the sTRAP ROC curve we ranked the SNPs by their absolute log ratio.

rSNP-MAPPER. We scored SNPs with rSNP-MAPPER using the models associated with target TF in rSNP-MAPPER model library, including MA0105, MA0107, MA0061, M00774, M00054, M00052, T00595, T00594, T00606, T00593, T00592, T00591, T00588, T00587, T00590, M00051 for NF-κB and T00284 for CTCF. For binary classification, we used a score cutoff of 0 and score change cutoff of 2, as suggested in the rSNP-MAPPER paper. To plot the rSNP-MAPPER ROC curve we ranked the SNPs by their score change.

HaploReg2. We used HaploReg2's default parameters. As HaploReg2 is not able to give a numeric score for each SNPs, we performed binary classification of each SNP by looking for "NF-kappaB" or "CTCF" in the Motif column of the result for SNP sets of NF-κB and CTCF respectively.

ChIP-seq Data

The NF-κB and CTCF ChIP-seq data used in this paper are both from ENCODE (GEO accession GSE31477 and GSE33213). NF-κB ChIP-seq data are from GM10847, GM12878, GM12891, GM12892, GM15510, GM18505, GM18526, GM18951, GM19099 and GM19193. CTCF ChIP-seq data are from GM10248, GM10266, GM12878, GM12891, GM12892, GM13976, GM13977, GM19238, GM19239 and GM19240

Allele-Specific Binding (ASB) SNPs

As a gold standard for SNPs that affect TF binding, we used the list of SNPs that are reported to induce allele-specific binding (ASB) of NF-κB and CTCF in GM12878. Our NF-κB positive SNP set consists of 70 ASB SNPs combined from [19] and [20]. Our CTCF positive SNP set consists of 1336 ASB SNPs from [19]. After filtering on minor allele frequency (≥0.01), we are left with 54 SNPs for NF-κB and 1247 SNPs for CTCF, from the latter of which we further down-sampled 60 SNPs as our final CTCF positive SNP set to accommodate the limited scoring throughput of motif-based approaches evaluated in this study.

WAVE Software

The implementation of WAVE and related data are available at http://wave.casil.mit.edu/

Acknowledgements

We thank Yuchun Guo for technical support in co-factor analysis using GEM. We also thank Matt Edwards for many helpful comments and discussions.

References

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP a, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9: 356–69. doi:10.1038/nrg2344

2. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010; doi:10.1056/NEJMra0905980

3. Stranger BE, Stahl E a, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187: 367–83. doi:10.1534/genetics.110.120907

4. Hindorff L a, Sethupathy P, Junkins H a, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106: 9362–7. doi:10.1073/pnas.0903103106

5. Frazer K a, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009;10: 241–51. doi:10.1038/nrg2554

6. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol. Nature Publishing Group; 2012;30: 1095–106. doi:10.1038/nbt.2422

7. Andersen MC, Engström PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, et al. In silico detection of sequence variations modifying transcriptional regulation. PLoS Comput Biol. 2008;4: e5. doi:10.1371/journal.pcbi.0040005

8. Macintyre G, Bailey J, Haviv I, Kowalczyk A. is-rSNP: a novel technique for in silico regulatory SNP detection. Bioinformatics. 2010;26: i524–30. doi:10.1093/bioinformatics/btq378

9. Manke T, Heinig M, Vingron M. Quantifying the effect of sequence variation on regulatory interactions. Hum Mutat. 2010;31: 477–83. doi:10.1002/humu.21209

10. Riva A. Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. BMC Genomics. BioMed Central Ltd; 2012;13 Suppl 4: S7. doi:10.1186/1471-2164-13-S4-S7

11. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40: D930–4. doi:10.1093/nar/gkr917

12. Teng M, Ichikawa S, Padgett LR, Wang Y, Mort M, Cooper DN, et al. Regsnps: A strategy for prioritizing regulatory single nucleotide substitutions. Bioinformatics. 2012;28: 1879–1886. doi:10.1093/bioinformatics/bts275

13. Molineris I, Schiavone D, Rosa F, Matullo G, Poli V, Provero P. Identification of functional cis-regulatory polymorphisms in the human genome. Hum Mutat. 2013;34: 735–42. doi:10.1002/humu.22299

14. Guo Y, Mahony S, Gifford DK. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput Biol. 2012;8: e1002638. doi:10.1371/journal.pcbi.1002638

15. Sgarbanti M, Remoli AL, Marsili G, Ridolfi B, Borsetti A, Perrotti E, et al. IRF-1 is required for full NF-kappaB transcriptional activity at the human immunodeficiency virus type 1 long terminal repeat enhancer. J Virol. 2008;82: 3632–3641. doi:10.1128/JVI.00599-07

16. Fujioka S, Niu J, Schmidt C, Guido M, Peng B, Uwagawa T, et al. NF- κ B and AP-1 Connection : Mechanism of NF- κ B-Dependent Regulation of AP-1 Activity NF- B and AP-1 Connection : Mechanism of NF- B-Dependent Regulation of AP-1 Activity. Society. 2004;24: 7806–7819. doi:10.1128/MCB.24.17.7806

17. Bartels M, Schweda AT, Dreikhausen U, Frank R, Resch K, Beil W, et al. Peptide-mediated disruption of NFkappaB/NRF interaction inhibits IL-8 gene activation by IL-1 or Helicobacter pylori. J Immunol. 2007;179: 7605–7613. doi:10.4049/jimmunol.179.11.7605

18. Thomas RS, Tymms MJ, McKinlay LH, Shannon MF, Seth a, Kola I. ETS1, NFkappaB and AP1 synergistically transactivate the human GM-CSF promoter. Oncogene. 1997;14: 2845–2855. doi:10.1038/sj.onc.1201125

19. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. Nature Publishing Group; 2011;7: 522. doi:10.1038/msb.2011.54

20. Karczewski KJ, Dudley JT, Kukurba KR, Chen R, Butte AJ, Montgomery SB, et al. Systematic functional regulatory assessment of disease-associated variants. Proc Natl Acad Sci U S A. 2013;110: 9607–12. doi:10.1073/pnas.1219099110

21. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, et al. Variation in transcription factor binding among humans. Science. 2010;328: 232–5. doi:10.1126/science.1183621

22. Duchi J, Hazan E, Singer Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. J Mach Learn Res. 2011;12: 2121–2159. Available: http://jmlr.org/papers/v12/duchi11a.html

Supporting Information

S1 Fig. Distribution of k-mer effect size predicted by WAVE
Distribution of the summed effect-size of all 87380 k-kmers with length less than or equal to 8 from NF-κB WAVE model

S2 Fig. WAVE is consistent cross LCL cell lines
Scatter plots of SNP scores assigned by WAVE models trained on ChIP-seq data from different LCL individuals. **(A)** GM12878 vs. GM12892. **(B)** GM12878 vs. GM18951. **(C)** GM12878 vs. GM19193

S3 Fig. Active set of k-mers detected by WAVE contained different parts of longer CTCF motif
The average k-mer effect and aggregated position weight matrix of the 2rd, 3rd, 5th, 6th and 7th strongest k-mer cluster from CTCF WAVE model compared with CTCF motif (MA0139.1).

S4 Fig. WAVE significantly outperformed motif-based approaches in prioritizing CTCF ASB-SNPs
ROC curves for discriminating CTCF ASB-SNP from the various negative sets using WAVE, sTRAP, rSNP-MAPPER and HaploReg2. Dashed line indicates random chance.
(A) Positive set chosen as all CTCF ASB-SNPs (n=60)**.** Negative set randomly sampled from common SNPs from 1KG (n=1000). **(B)** Negative set limited to common 1KG variants that are located within 1000 bp from any CTCF ASB-SNP (n=1095).

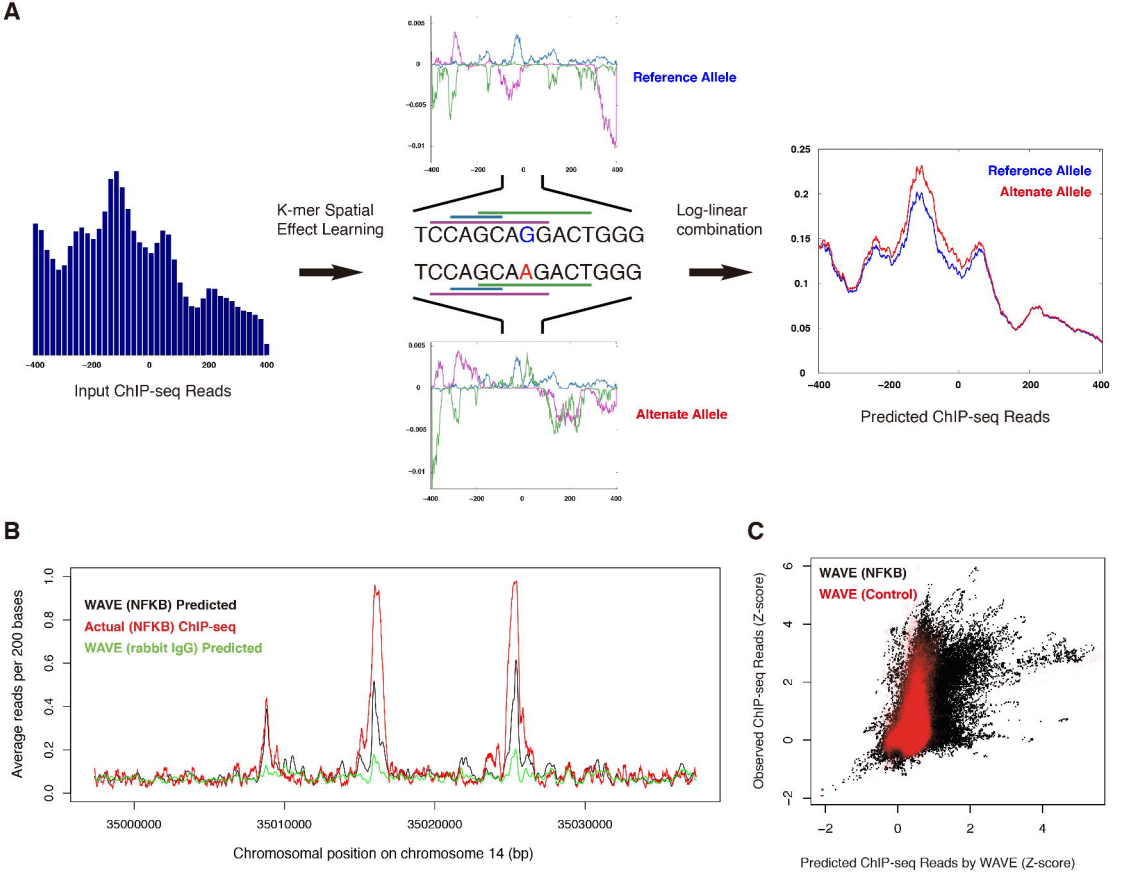
S5 Fig. WAVE achieved similar performance as motif-based approaches in prioritizing CTCF MD SNPs
ROC curves for discriminating CTCF MD SNPs from the various Negative sets using WAVE, sTRAP, rSNP-MAPPER and HaploReg2. Dashed line indicates random chance.
(A) Positive set chosen as all CTCF MD SNPs (n=24)**.** Negative set randomly sampled from common SNPs from 1KG(n=1000). **(B)** Negative set limited to common variants from 1KG variants that are located within 1000 bp from any CTCF MD SNP (n=421).

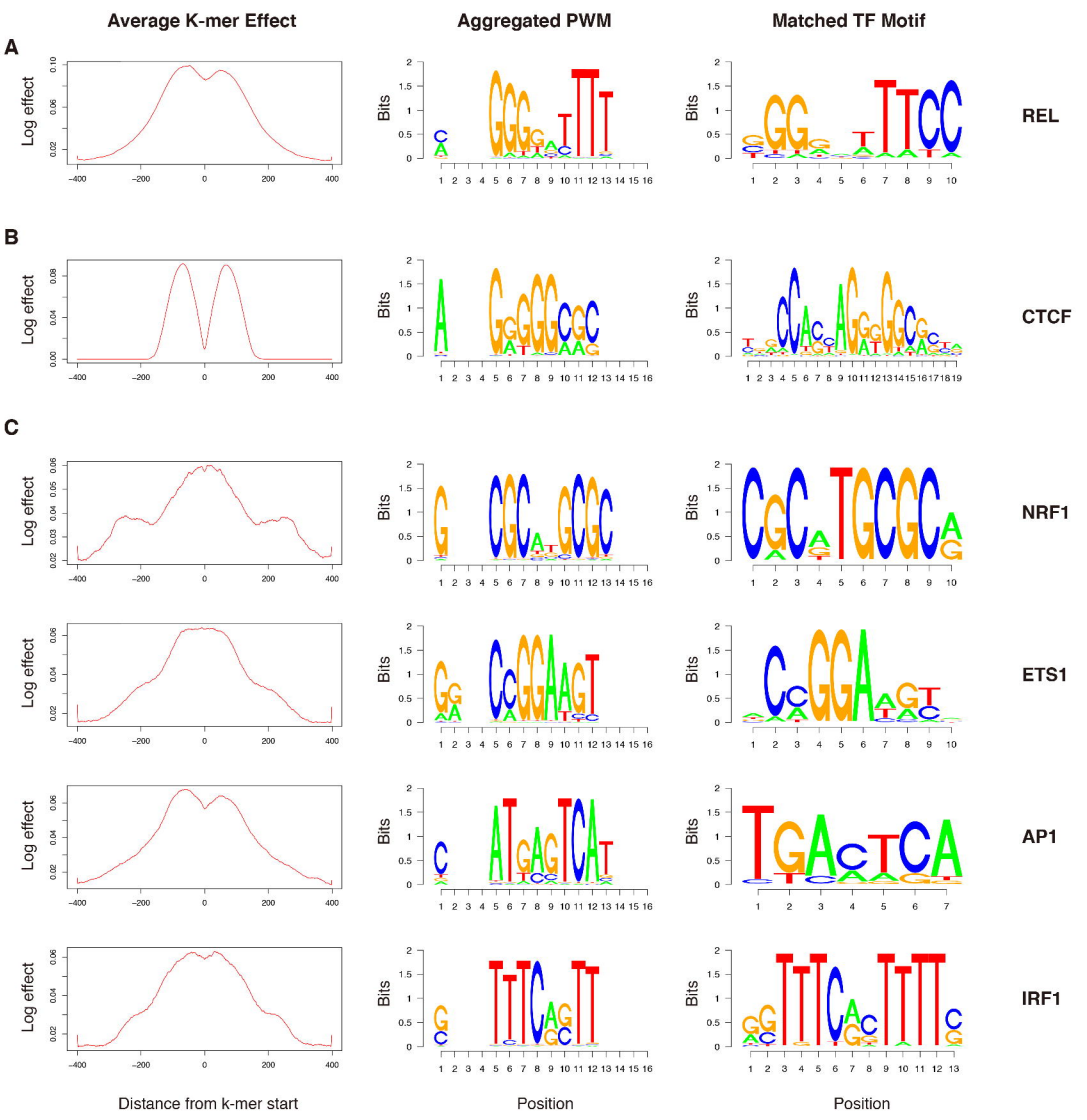
S1 Table. Pearson's correlation between binding signal predicted by WAVE and actual ChIP-seq read counts on held-out chromosome 14

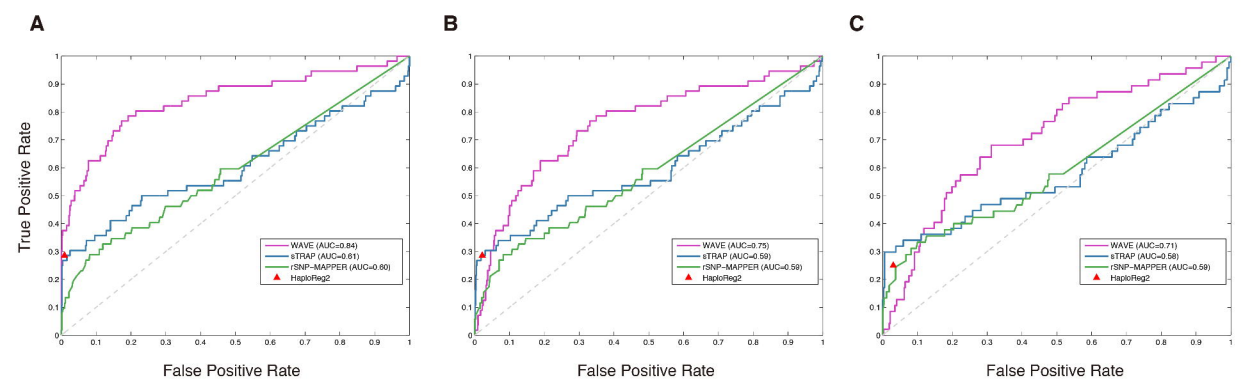
S2 Table. Number of correctly annotated ASB SNP by motif-based approaches

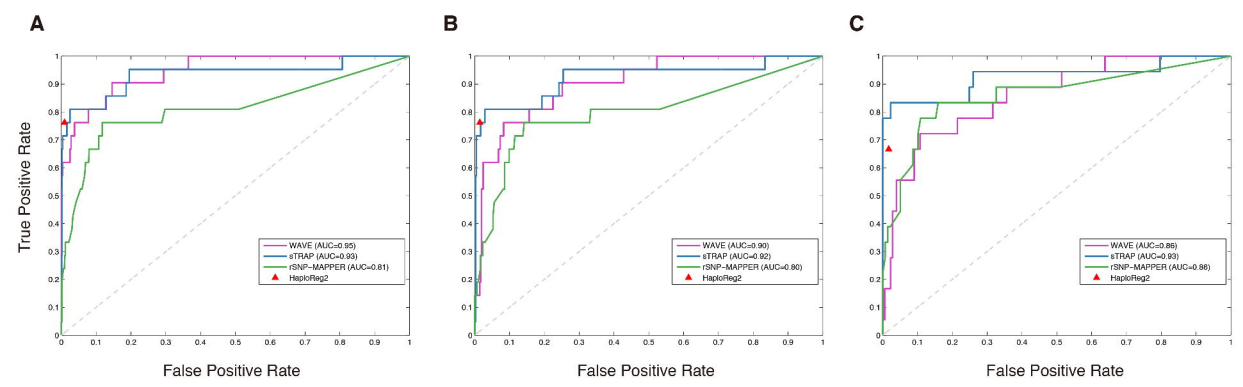
S3 Table. List of ASB SNP included in the analysis

S1 Text. The detailed implementation of parameter optimization in WAVE


A
Input ChIP-seq Reads
K-mer Spatial
Effect Learning
TCCAGCAGGACTGGG
TCCAGCAAGACTGGG
Reference Allele
Altenate Allele
Log-linear
combination
Reference Allele
Altenate Allele
Predicted ChIP-seq Reads
B
WAVE (NFKB) Predicted
Actual (NFKB) ChIP-seq
WAVE (rabbit IgG) Predicted
Average reads per 200 bases
Chromosomal position on chromosome 14 (bp)
C
WAVE (NFKB)
WAVE (Control)
Observed ChIP-seq Reads (Z-score)
Predicted ChIP-seq Reads by WAVE (Z-score)


Average K-mer Effect
Aggregated PWM
Matched TF Motif
A
B
C
Log effect
Bits
REL
CTCF
NRF1
ETS1
AP1
IRF1
Distance from k-mer start
Position
Position


A
B
C
True Positive Rate
False Positive Rate
False Positive Rate
False Positive Rate
WAVE (AUC=0.84)
sTRAP (AUC=0.61)
rSNP–MAPPER (AUC=0.60)
HaploReg2
WAVE (AUC=0.75)
sTRAP (AUC=0.59)
rSNP–MAPPER (AUC=0.59)
HaploReg2
WAVE (AUC=0.71)
sTRAP (AUC=0.58)
rSNP–MAPPER (AUC=0.59)
HaploReg2


A
B
C
True Positive Rate
False Positive Rate
False Positive Rate
False Positive Rate
WAVE (AUC=0.95)
sTRAP (AUC=0.93)
rSNP−MAPPER (AUC=0.81)
HaploReg2
WAVE (AUC=0.90)
sTRAP (AUC=0.92)
rSNP−MAPPER (AUC=0.80)
HaploReg2
WAVE (AUC=0.86)
sTRAP (AUC=0.93)
rSNP−MAPPER (AUC=0.86)
HaploReg2