A general integrative genomic feature transcription factor binding site prediction method applied to analysis of USF1 binding in cardiovascular disease

*Tianyuan Wang,[1,5] Terrence S. Furey,[2] Jessica J. Connelly,[1] Shihao Ji,[3] Sarah Nelson,[1] Steffen Heber,[4] Simon G. Gregory[1] and Elizabeth R. Hauser[1]**

[1]Department of Medicine and Center for Human Genetics, Duke University Medical Center, Durham, NC 27710, USA
[2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27708, USA
[3]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA
[4]Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA
[5]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA
**Correspondence to:* Tel: +1 919 684 0603; Fax: +1 919 684 0913; E-mail: Elizabeth.Hauser@duke.edu

Date received (in revised form): 17th November 2008

Abstract

Transcription factors are key mediators of human complex disease processes. Identifying the target genes of transcription factors will increase our understanding of the biological network leading to disease risk. The prediction of transcription factor binding sites (TFBSs) is one method to identify these target genes; however, current prediction methods need improvement. We chose the transcription factor upstream stimulatory factor 1 (*USF1*) to evaluate the performance of our novel TFBS prediction method because of its known genetic association with coronary artery disease (CAD) and the recent availability of *USF1* chromatin immunoprecipitation microarray (ChIP-chip) results. The specific goals of our study were to develop a novel and accurate genome-scale method for predicting *USF1* binding sites and associated target genes to aid in the study of CAD. Previously published *USF1* ChIP-chip data for 1 per cent of the genome were used to develop and evaluate several kernel logistic regression prediction models. A combination of genomic features (phylogenetic conservation, regulatory potential, presence of a CpG island and DNaseI hypersensitivity), as well as position weight matrix (PWM) scores, were used as variables for these models. Our most accurate predictor achieved an area under the receiver operator characteristic curve of 0.827 during cross-validation experiments, significantly outperforming standard PWM-based prediction methods. When applied to the whole human genome, we predicted 24,010 *USF1* binding sites within 5 kilobases upstream of the transcription start site of 9,721 genes. These predictions included 16 of 20 genes with strong evidence of *USF1* regulation. Finally, in the spirit of genomic convergence, we integrated independent experimental CAD data with these USF1 binding site prediction results to develop a prioritised set of candidate genes for future CAD studies. We have shown that our novel prediction method, which employs genomic features related to the presence of regulatory elements, enables more accurate and efficient prediction of *USF1* binding sites. This method can be extended to other transcription factors identified in human disease studies to help further our understanding of the biology of complex disease.

Key words: *transcription factors, cardiovascular disease, human genetics, binding site prediction*

Background

Several transcription factors (TFs) have been characterised as mediators of complex disease processes.[1–3] Numerous publications have identified single nucleotide polymorphisms (SNPs) in TFs that are significantly associated with coronary artery disease (CAD).[2,4,5] This combined evidence suggests that the target genes of these TFs also may be associated with human complex disease. Identification of potential TF targets could further our understanding of gene–gene interactions underlying complex disease. Genome-wide experimental methods, such as chromatin immunoprecipitation microarray (ChIP-chip),[6,7] a technique combining chromatin immunoprecipitation and microarray analysis for identifying TF-interacting genomic regions, are time consuming and expensive. It would be more efficient to develop an *in silico* computational method for TF target prediction followed by less costly genotyping and more focused molecular biology experiments to identify the association between gene–gene interactions and complex disease.

TFs play important roles in the transcriptional regulation of genes by interacting with specific DNA sequences, called transcription factor binding sites (TFBSs), to control cell- and tissue-specific gene expression. Accurately identifying TFBSs is critical to our understanding of the biological regulation of the cell. Although many partially complete genome sequences are available, encoded functional elements such as TFBSs have not been fully characterised. This is due, in part, to the complexity of TF binding activity and the degeneracy of the DNA sequence in the core binding site.

Currently, the primary strategy for predicting TFBSs is by DNA motif scanning, which uses DNA sequence motifs to identify potential matching sequences across the genome.[8–10] The common approaches of motif scanning are based on either consensus sequences or binding site matrices. The consensus sequence approach works best on sites that have little degeneracy. The other approach is based on binding site matrices, which include the position weight matrix (PWM) and the position frequency matrix (PFM).[10] This approach takes degeneracy of the binding site motif into account when predicting TFBSs, and derives scoring matrices by using known binding sites to calculate a score for each possible nucleotide in each position within the TFBS. These matrices are then used to predict potential TFBSs by scoring DNA sequence in the target genome. The accuracy of the prediction is limited by the quality of the binding site matrix, which can vary based on the experimental input. It also lacks the flexibility to incorporate additional genomic information. In general, these methods lead to an inflated number of predicted TFBSs because of low specificity prediction, which leads to many false-positive results. Therefore, the reliability of prediction methods based on DNA sequence alone is low. An ideal prediction method needs to combine DNA sequence with additional genomic features to improve specificity.

Phylogenetic sequence conservation is an example of an additional genomic feature that can be used to study TFBSs. The phylogenetic approach presupposes that sequences are conserved between multiple species under selective pressure and may contain functional elements such as TFBSs.[11] This level of sequence conservation does not account for species specificity in either TF DNA-binding domains or TFBSs. Currently, many other genomic features related to regulatory elements are available at a genome scale. For example, the regulatory potential of a DNA sequence is measured by the frequency of known regulatory elements in short aligned regions across multiple species.[11] CpG islands are CG dinucleotide-rich regions of the genome commonly associated with transcription start sites and promoters.[12,13] These regions can also influence epigenetic control over gene expression via methylating cytosine within the CpG islands. Another genomic feature associated with gene regulation is a DNaseI hypersensitive (HS) site; these are hypersensitive to DNaseI cleavage. DNaseI HS sites (DNaseI HS) are nucleosome-free regions of open chromatin associated with regulatory elements, such as promoters, enhancers and silencers.[14] While some of these genomic features have been used individually to filter the predictions from sequence-based scoring methods,[8,9] TFBS prediction

methods would benefit from selecting and integrating these genomic features carefully. Although the number of genomic features available is fairly large, current prediction methods do not take full advantage of them.

Several linkage and association studies indicate that the transcription factor upstream stimulatory factor 1 (*USF1*) is genetically associated with CAD.[2] *USF1* is ubiquitously expressed in human tissues and is a key regulator of several biological processes, such as the stress and immune response, cell cycle and cell proliferation.[15] *USF1* belongs to the basic helix-loop-helix (*bHLH*) zipper transcription factor family. The binding sites of *USF1* share the same core DNA sequence, called the E-box (5′-CACGTG-3′), with some degeneracy.[16] The complete binding site of *USF1* is represented by 5′-RYCACGTGRY-3′.[16] The DNA-binding activity of *USF1* can be modulated through phosphorylation, homo- or heterodimerisation and variation in binding site sequence.[15]

We chose *USF1* to evaluate the performance of our novel TFBS prediction method because of its biological importance, particularly in regard to its known genetic association with CAD, and the recent availability of *USF1* ChIP-chip results for 1 per cent of the genome.[17] Our goals were to (1) develop a reliable and accurate method for *USF1* transcription factor binding site (*USF1*–BS) prediction; (2) make a genome-scale prediction of potential *USF1*–BSs and (3) identify *USF1* target genes. We have developed a novel prediction method incorporating additional genomic features related to the presence of regulatory elements, enabling a more accurate and efficient identification of *USF1*–BSs on a genome scale. The results of this study will help to prioritise CAD candidate genes, as well as provide biological information in evaluating gene–gene interactions with respect to this common complex disease.

Methods

Genome sequence and features

All annotation and mapping locations of genomic features used to predict TFBSs were based on National Center for Biotechnology Information (NCBI) human genome build 35. ENCODE sequences[18] and the 5 kilobase (kb) regions upstream of the transcription start sites (TSSs) of 23,105 RefSeq mRNA sequences were obtained from the University of California, Santa Cruz (UCSC) Genome Browser.[19–21] These RefSeq mRNA sequences included the transcripts from alternative TSSs but did not include non-coding RNA. The ENCODE regions included promoter, intronic, exonic and intergenic regions from 44 genomic intervals on 20 chromosomes.

The values for genomic features (Table 1) for each potential 10 base pair (bp) *USF1*–BS are continuous

Table 1. Description of the five genomic features used for *USF1*–BS prediction method development

Name	Description	Score range
PhastCons8[a,b]	Conservation score across eight species (Human/chimp/mouse/rat/dog/chicken/fugu/zebrafish)	[−35, 0]
MostCons8[a]	Conserved region across eight species (Human/chimp/mouse/rat/dog/chicken/fugu/zebrafish)	[0, 1000]
RP5[a]	Regulatory potential across five species (Human/chimp/mouse/rat/dog)	[−0.1, 0.9]
CpG[a,c]	CpG island, CG dinucleotide-rich regions	[0.5, 1.6)
DNaseI HS[d]	Hypersensitive to DNaseI cleavage within human $CD4^+$ cell	[0, 17)

[a]Downloaded from the UCSC Genome Browser.[19–21]
[b]Base-10 logarithm of the product of base-by-base conservation score within 10 base pair (bp) *USF1*–BS.
[c]Used centre position of 10 bp *USF1*–BS to define overlapping CpG island score.
[d]Published results.[22]

variables. These were also obtained from the UCSC website.[19,20] The base-by-base conservation scores and predicted conserved elements (MostCons8) were generated by the program phastCons,[23] using genome-wide multiple alignments of eight species (human, chimpanzee, mouse, rat, dog, chicken, fugu and zebrafish). The total conservation score of each 10 bp *USF1*–BS (PhastCons8) was represented by the base-10 logarithm of the product of the conservation score of each bp within the *USF1*–BS. The regulatory potential (RP5) scores were computed from alignments of five species (human, chimpanzee, mouse, rat and dog). The RP5 score of each putative regulatory element indicates the frequency of known regulatory elements within short alignment regions using 100 bp windows.[11] CpG islands (CpG) were defined as CG dinucleotide-rich regions at least 200 bp long, with a ratio of observed to expected CG dinucleotides greater than 0.6.[13] The coordinates of DNaseI HS are the regions in the genome hypersensitive to DNaseI cleavage within human $CD4^+$ cells. The DNaseI HS score of each site was generated by kernel density estimation, and reflected the degree of chromatin accessibility at that site.[22]

USF1 ChIP-chip data

Known *USF1*-interacting genomic regions were used to evaluate our prediction method. A recently published *USF1* ChIP-chip study identified *USF1*-interacting genomic regions using chromatin immunoprecipitation from liver cells (HepG2) followed by microarray analysis.[17] The microarray contains approximately 18,000 loci, polymerase chain reaction (PCR) amplicons of 1.0–1.5 kb in length across the ENCODE regions. The authors classified the loci on the array according to the $\log_2$-ratio, the base-2 logarithm of the ratio of fluorescence intensities of immunoprecipitated chromatin to control chromatin for each spot on the array. The $\log_2$-ratios were in the range −1 to 4. Thirty-four loci with $\log_2$-ratio greater than 1.25 were considered to be bound, while 234 loci on the array with $\log_2$-ratio equal to −1 were considered not bound by *USF1*. For our experiment, these loci were used as positive and negative controls, respectively. The potential *USF1*–BSs from these control regions were used as the training dataset for the following method development.

Preliminary prediction based on PWM scoring method

The PWM scoring method was used to identify potential *USF1*–BSs from the target regions. The *USF1* binding matrix of 81 *USF1*–BSs generated *in vitro* by random sequence selection was obtained from the TRANSFAC database.[16,24] The Patser web application was used to convert the *USF1* binding matrix to PWM, and generated a numerically calculated cut-off score of 3.753 for predicting TFBSs based on the information content adjusted by sample size.[25] The average GC content of 47.1 per cent for the Patser analysis was calculated from 5 kb upstream sequences from 23,105 RefSeq mRNAs. The *USF1* PWM was used to score each 10 bp sliding window within target regions. A potential *USF1*–BS was defined as any 10 bp sequence with a score higher than the threshold of 3.753.

Prediction method based on genomic features

We initially applied the PWM scoring method to the ENCODE regions. This sequence-based prediction approach defined a set of potential *USF1*–BSs, each of which was mapped to a specific locus in the ENCODE genomic microarray used by *USF1* ChIP-chip experiments according to its genomic location, therefore allowing us to map the potential *USF1*–BSs to the *USF1* ChIP-chip results.

We carried out more specific *USF1*–BS predictions using five genomic features (PhastCons8, MostCons8, RP5, CpG and DNaseI HS). We implemented a kernel logistic regression algorithm[26] in MATLAB Version 7.0, using the radial basis function (RBF) as the kernel function. Various genomic features are used by the kernel function to map the input data to a high-dimensional space (see Additional file 1). This supervised statistical learning model was trained by the training dataset to select hyperparameters. These hyperparameters were then applied to the testing dataset. The model generated a score for

each potential *USF1*–BS within the testing dataset in the range from 0 to 1. The threshold of being a predicted *USF1*–BS was 0.5. Initially, we used each single feature and combinations of all features as variables in different binding site prediction models. We also performed backward stepwise linear regression in SAS Version 9.1, using the training dataset, to identify a subset of features significantly contributing to the model, using $p = 0.05$ as the threshold. Once the most significant features were identified, we implemented the prediction method using that model. The performance of each prediction model was evaluated based on sensitivity, specificity and area under the receiver operator characteristic curve (AUC) by performing a leave-one-locus-out (LOLO) cross-validation with the same training dataset. In many cases, multiple potential *USF1*–BSs were associated with a locus defined by the ChIP-chip study. Initially, these potential *USF1*–BSs were grouped by their loci. In each iteration, all potential *USF1*–BSs in one locus were held out for testing, while the remaining loci formed the training dataset for developing the prediction model that would be applied to potential *USF1*–BSs within the test locus. The test locus was classified as positive if it included at least one predicted *USF1*–BS; otherwise it was classified as a negative locus. Sensitivity was defined as the number of correctly predicted positive loci divided by the total positive loci, whereas specificity was defined as the number of correctly predicted negative loci divided by the total negative loci. AUC was calculated using the SPSS package.[27]

Genome-scale prediction and validation

Potential *USF1*–BSs within 5 kb upstream of the TSSs of 23,105 RefSeq mRNAs were first identified by the PWM scoring method. More specific *USF1*–BS predictions were then generated by the optimised model using the genomic features of each potential binding site. Lastly, the prediction results were evaluated by comparison with 20 robust *USF1* target genes — the overlap between the target genes obtained from the TRED database[28] and reported in the literature.[29]

Results

Prediction method development

We assessed the merits of predicting *USF1*–BSs using (1) DNA sequence alone; (2) sequence with single genomic features and (3) sequence with multiple genomic features to identify putative *USF1*–BSs within the ENCODE regions. Figure 1 summarises our general approach to method development and assessment.

We started by using the PWM scoring method to identify potential *USF1*–BSs (see Methods). A total of 99,013 potential *USF1*-BSs were identified within the ENCODE regions (30 megabases). Among these potential *USF1*–BSs, 135 were associated with the 34 positive loci and 800 with the 177 negative loci identified by the *USF1* ChIP-chip study. These 935 *USF1*–BSs were then used to construct the training dataset for further prediction method development. Of 234 negative loci, 54 did not include any potential *USF1*–BSs and were excluded from the training dataset.

To refine predictions more accurately, we evaluated several models trained to identify *USF1*–BSs using kernel logistic regression with genomic features (PhastCons8, MostCons8, RP5, CpG and DNaseI HS [Table 1]) as variables, in addition to PWM scores. Using LOLO cross-validation with the same training dataset, we examined the sensitivity, specificity and AUC of the models using different sets of variables, including the PWM score alone, a single feature, all features and selected features. Among the prediction models based on a single feature, the RP5 model had the highest AUC (0.672); however, its sensitivity (0.088) is much lower than the DNaseI HS model (0.235) (Table 2). We performed backward stepwise feature selection for model building, starting with all five genomic features. This procedure calculated the contribution of each feature to classification. We removed MostCons8 and CpG features from the model based on a $p = 0.05$ threshold. In the final model, DNaseI HS had the lowest *p*-value (<0.0001), followed by PhastCons8 (0.0022), RP5 (0.0067) and PWM score (0.0081). The prediction model based on these selected features (PWM,

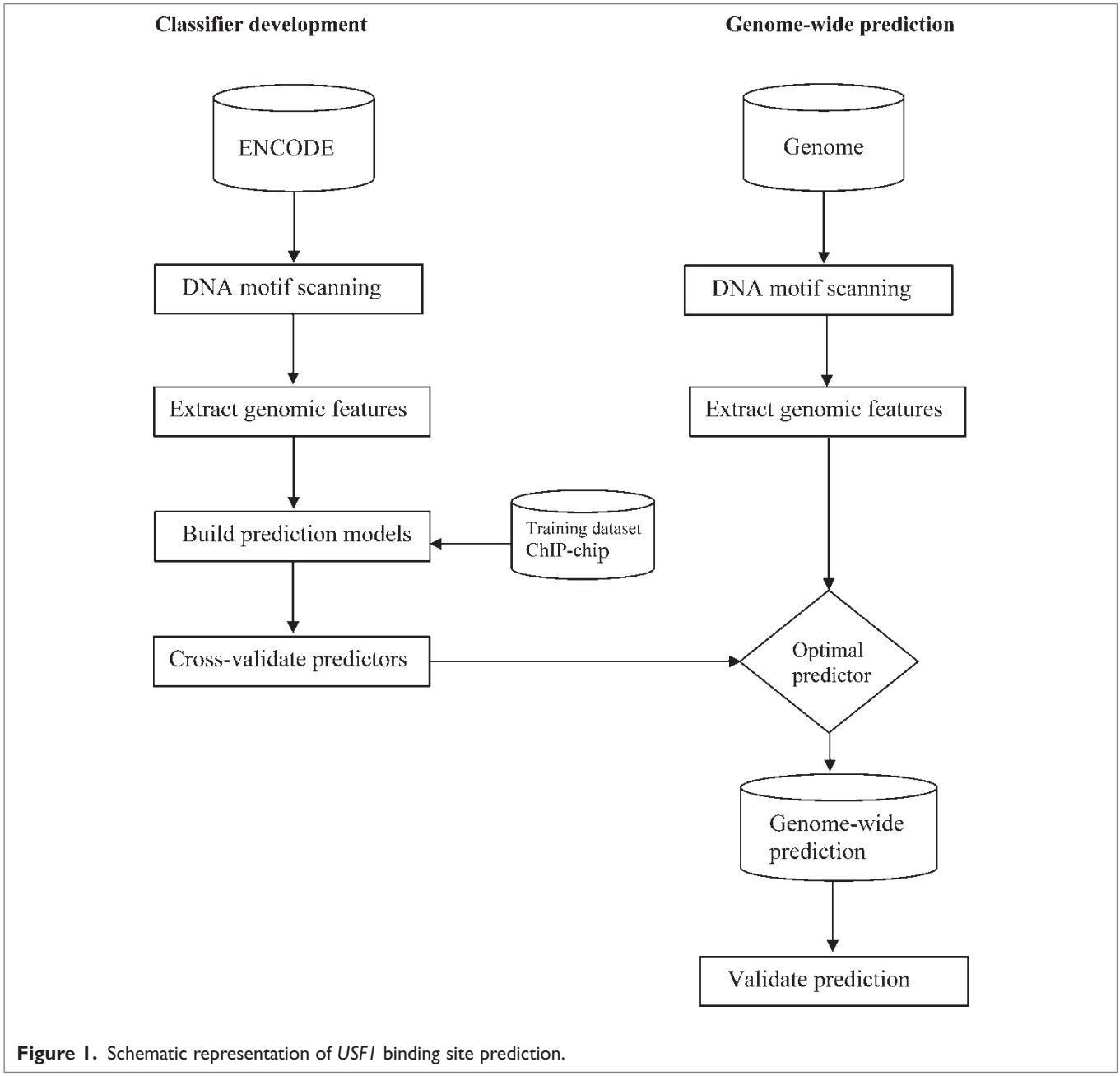

Figure 1. Schematic representation of *USF1* binding site prediction.

PhastCons8, RP5 and DNaseI HS) achieved 55.9 per cent sensitivity and 87.6 per cent specificity when using a 0.5 scoring threshold for predicting each *USF1*–BS. With the highest AUC (0.827), this model outperforms all other models based on any single feature or combination of features (Table 2; Figure 2; Additional file 2). This model was considered as the optimal predictor among the models tested, and was used to predict the genome-wide binding sites.

We applied the optimal prediction model, based on selected features, to the same regions used by *USF1* ChIP-chip experiments. There were 16,405 loci from the ENCODE ChIP-chip annotation with potential USF1–BSs identified by the PWM scoring method. Among them, 34 positive and 177 negative loci were used to construct the training dataset for developing the prediction method (see Methods). The remaining 16,194 loci were used as an independent testing dataset.

Table 2. Comparison of *USF1*–BS prediction models. The threshold score that defined a predicted *USF1*–BS was 0.5. Sensitivity was the proportion of correctly predicted true positive loci, whereas specificity was the proportion of true negative loci predicted as negative loci. The AUC was calculated from LOLO cross-validation (see Methods)

Variables in the model	AUC	Standard error	Asymptotic 95% confidence interval		Sensitivity	Specificity
			Lower bound	Upper bound		
PWM	0.648	0.053	0.544	0.752	0.176	0.989
PhastCons8	0.599	0.053	0.496	0.702	0.000	0.955
RP5	0.672	0.058	0.559	0.786	0.088	0.994
DNaseI	0.553	0.083	0.390	0.716	0.235	0.960
All features	0.639	0.060	0.522	0.756	0.382	0.898
Selected features	0.827	0.044	0.740	0.913	0.559	0.876

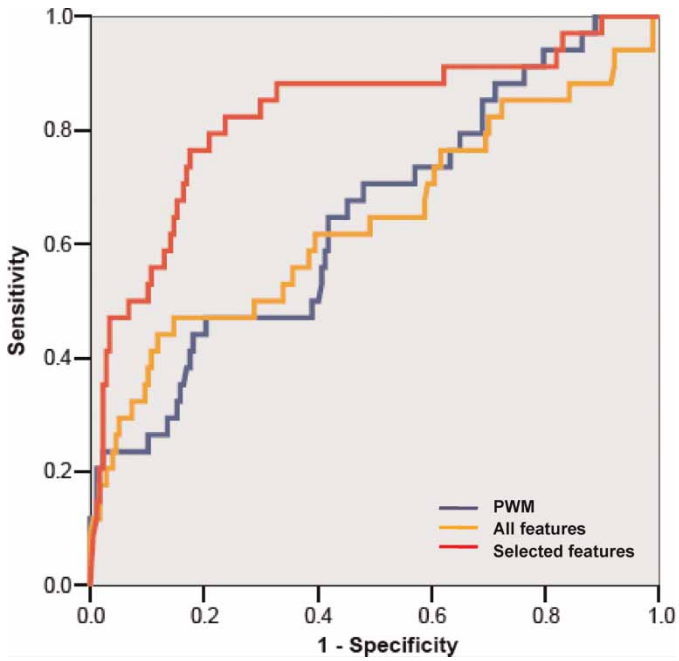

Figure 2. Receiver operator characteristic curve (ROC) of *USF1* prediction models. The curves generated by the SPSS package[27] with different colours indicate the sensitivity and specificity of three different prediction models. The sensitivity and specificity were calculated from LOLO cross-validation (see Methods). Sensitivity was the proportion of correctly predicted positive loci whereas specificity was the proportion of correctly predicted negative loci. The all features model used all five genomic features (PhastCons8, MostCons8, RP5, CpG and DNaseI HS) and PWM score, and the selected features model only included three genomic features (PhastCons8, RP5 and DNaseI HS) and PWM score.

Our prediction method was able to divide these unclassified loci into two groups: positive loci (1,615) with a predicted *USF1*–BS and negative loci (14,579) without a predicted *USF1*–BS. The average $\log_2$-ratio of these predicted positive loci (0.1034) in the ChIP-chip experiment is significantly higher ($p < 10^{-23}$) than that of predicted negative loci (0.0085). This result indicates

that the predicted positive loci are enriched among the loci with a $\log_2$-ratio higher than 0, which are more likely to include *USF1*–BSs (Figure 3).

Genome-scale prediction and validation

We obtained DNA sequences from 5 kb upstream regions of the TSSs of 23,105 RefSeq mRNAs. The PWM scoring method identified 290,614 potential *USF1*–BSs in these sequences. We applied our most robust model of *USF1*–BS prediction, kernel logistic regression using three genomic features (PhastCons8, RP5 and DNaseI HS) and the PWM score, to improve the specificity of these predictions. 24,010 *USF1*–BSs from 9,721 genes were predicted as *USF1* targets, representing 8.3 per cent of the initial potential *USF1*–BSs (see Additional file 3). We created a set of 20 robust *USF1* target genes obtained from the TRED database and from the literature to validate the genome-scale prediction results. Our prediction method was able to identify 16 of these 20 genes (80 per cent) as *USF1* targets (Table 3).

Distributions of predicted *USF1*–BSs

Our prediction method generates a score for each potential *USF1*–BS identified by the PWM scoring method. Prediction scores range from 0 to 1 and correspond to the confidence of the model's prediction. The score distribution of our genome-scale prediction showed that a large portion of predicted sites had scores higher than 0.99 (Figure 4). Selecting *USF1*–BSs with the highest scores dramatically reduces the number of predicted target genes. Based on the score distribution, we chose a stringent threshold (0.99) to reduce the number of predicted *USF1* target genes further, from 9,721 to 5,801, to be used as candidate genes for further analysis.

Potential *USF1*–BSs identified solely by the PWM scoring method are evenly distributed across 5 kb upstream regions of the TSSs of 23,105 RefSeq mRNAs (Figure 5). Our predicted *USF1*–BSs using a 0.5 scoring threshold are concentrated within 1 kb upstream of TSS, the region most likely to contain TFBSs.[30] Predicted *USF1*–BSs using the higher threshold (0.99) are even more enriched within 1 kb upstream of TSS. The most significant feature in the prediction model, DNaseI HS, is over-represented in the first 1 kb sequence upstream of

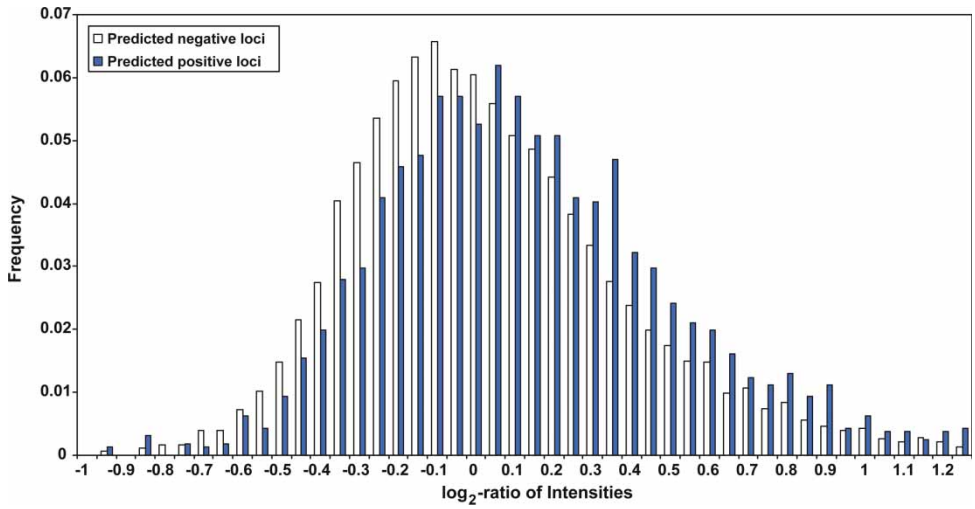

Figure 3. Prediction evaluation with *USF1* ChIP-chip results from the ENCODE regions. There are 16,405 loci on the ENCODE microarray with potential *USF1*–BSs identified by the PWM scoring method. Among them, 34 positive and 177 negative loci were used to construct the training dataset for developing the prediction method (see Methods). The remaining 16,194 loci were used as an independent testing dataset. Our prediction method divided these unclassified loci into two groups: predicted positive loci (1,615) with predicted *USF1*–BSs, and predicted negative loci (14,579) without predicted *USF1*–BSs. The data are represented as histograms of frequency at intervals of 0.1 $\log_2$-ratio of intensities.

Table 3. Validation of 20 robust *USF1* target genes. We used 20 robust *USF1* target genes obtained from the TRED database[28] and reported the literature[29] to evaluate the prediction method. Our optimal prediction model was able to identify 16 out of these 20 genes as *USF1* targets

No.	Gene	Correctly predicted
1	Apolipoprotein A-II (*APOA2*)	Yes
2	ATP-binding cassette, sub-family A (*ABC1*)	Yes
3	Acetye-coenzyme A carboxylase alpha (*ACACA*)	No
4	Apolipoprotein (*APOE*)	Yes
5	Breast cancer 2 (*BRCA2*) early onset	Yes
6	*CCNB1*	Yes
7	Cytochrome p450, family19 (*CYP19*)	Yes
8	Cytochrome p450, family1, subfamily A, polypeptide1 (*CYP1A1*)	Yes
9	*EFP*	Yes
10	Fragile x mental retardation 1 (*FMR1*)	Yes
11	Follicle stimulating hormone receptor (*FSHR*)	Yes
12	Glucokinase (hexokinase 4) (*GCK*)	Yes
13	Ghrelin/obestatin prepropeptide (*GHRL*)	No
14	Homeobox B4 (*HOXB4*)	Yes
15	Homeobox B7 (*HOXB7*)	Yes
16	*h*-telomerase reverse transcriptase (*hTERT*)	Yes
17	Platelet factor 4 (*PF4*)	No
18	Polymeric immunoglobulin receptor (*PIGR*)	No
19	Protein tyrosine phosphatase, non-receptor type 6 (*PTPN6*)	Yes
20	Serpin peptidase inhibitor, clade 5, member1 (*SERPINE1*)	Yes

the transcription start site (data not shown). This could explain the concentration of predictions in this region; however, it alone did not account for the concentration of predicted *USF1*–BSs.

To understand better which factor contributed most to *USF1*–BSs predictions at the highest thresholds, we divided the predicted *USF1*–BSs into two groups: one with prediction scores ranging between 0.99 and 1, and the second with scores ranging between 0.5 and 0.99. We then compared the value of each genomic feature between these two groups. This analysis indicated that DNaseI HS is the most distinguishing feature. On average, *USF1*–BSs with higher scores have higher DNaseI HS values than *USF1*–BSs with lower scores. The DNaseI HS value was also closely correlated with the location of the *USF1*–BS in the region upstream of TSS ($r^2 = 0.41$).

Discussion

USF1 binding site prediction method

We focused on *USF1* to develop a novel TFBS prediction method because of its genetic association with CAD and the availability of *USF1* ChIP-chip results from the ENCODE regions. Common TFBS prediction methods based on DNA sequence alone generate large numbers of false-positive results. One strategy for improving specificity of TFBS prediction is to use phylogenetic footprinting, which is based on the assumption that regions of multi-species sequence conservation are more likely to include regulatory elements. We hypothesised that combining multiple genomic features with regions of sequence conservation could increase the accuracy of TFBS predictions. To test our hypothesis, we began by using PWM, the most common binding motif search method, to identify potential *USF1*–BSs. We then incorporated several genomic features related to TFBSs, including sequence conservation, regulatory potential, and the presence of CpG islands and DNaseI HS sites. Using a training dataset constructed from published *USF1* ChIP-chip results,[17] we were able to compare the sensitivity, specificity and AUC of prediction models trained with different sets of features, such as PWM score alone, single features, all features and the features generated by feature selection. Prediction models based on single genomic features performed poorly, with low sensitivity and AUC. A prediction model using four selected features (PhastCons8, RP5, DNaseI HS and PWM score) produced the highest sensitivity (55.9 per cent) and AUC (0.827) among the models tested, while still achieving high specificity

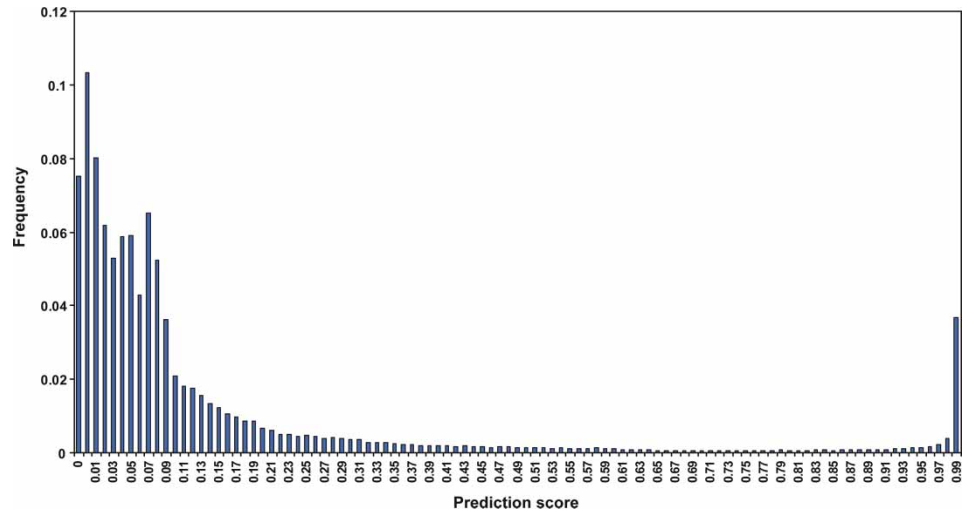

Figure 4. Prediction scores distribution of potential *USF1*–BSs. By scanning 5 kb upstream of TSSs of 23,105 RefSeq mRNAs in the human genome, 290,614 potential *USF1*–BSs were identified by the PWM scoring method. The prediction method generates a score for each potential *USF1*–BS identified by the PWM scoring method. These prediction scores range from 0 to 1 and correspond to the confidence of the model's prediction. A total of 24,010 predicted *USF1*–BSs were generated using the optimal prediction model with default prediction threshold (0.5). The data are represented as histograms of frequency at each 0.01 score interval.

(87.6 per cent) (Figure 2). These results show that the prediction model using selected features outperforms models based on a single feature and all features. That the performance of the model using all features is not better than others might be due to the noise introduced by the irrelevant and redundant features.

Kernel-based classifiers allow for the development of non-linear classifiers in cases where simple linear combinations of features are not sufficient accurately to distinguish between sample classes (see Additional file 4). Kernel logistic regression modelling maps the training data to high-dimensional space by considering all features jointly, and

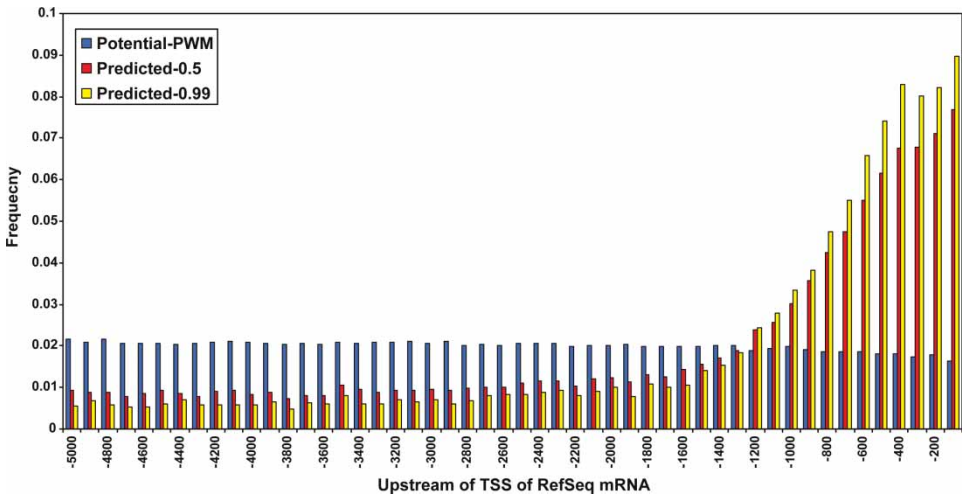

Figure 5. Location distribution of *USF1*–BSs. By scanning the 5 kb upstream of the TSSs of 23,105 RefSeq mRNAs in the human genome, 290,614 potential *USF1*-BSs were identified by the PWM scoring method. A total of 24,010 predicted *USF1*–BSs were generated using the optimal prediction model with default prediction threshold (0.5) and 10,296 predicted *USF1*-BSs were generated using a stringent prediction threshold (0.99). The data are represented as histograms of frequency at each 100 bp interval.

generates a non-linear decision boundary to separate two classes. The data within each class depend on a specific combination of the features learned from the training dataset. For example, a site with a relatively low PWM score may still be predicted as a *USF1*–BS if it has high DNaseI HS, conservation or regulatory potential scores. Conversely, if each feature contributes to the prediction method as an independent filter, the predicted *USF1*–BS will be based on a limited range of values of each feature. For example, if the prediction method only relies on a stringent PWM threshold to improve the specificity, it will be biased toward the binding site with high affinity, and only the target genes of *USF1* with strong binding sites will be identified by this method. Comparisons of PWM scores with the scores of each genomic feature are included in Additional files 5—8.

To test whether our prediction method may be biased toward sites with higher binding affinity, as indicated by higher PWM scores, we examined the PWM scores from the 20 robust *USF1* target genes at each step during the prediction process. As shown in Additional file 9, we find that: 1) the common PWM scoring method was sufficient to identify potential *USF1*–BSs for most of these genes; 2) the potential sites identified spanned a wide range of PWM scores. Further, the distributions of PWM scores among the initial 290,614 potential sites and the final 24,010 predicted *USF1*–BSs were not significantly different (see Additional file 9). These results suggest that our prediction method is not biased toward binding sites within any specific range of PWM scores, and if these scores do correlate with binding affinities, predictions are also not biased towards sites with high affinities.

For this study, we focused on five genomic features related to regulatory elements currently available on a genome scale. Backward stepwise feature selection during model building indicated that DNaseI HS was the most important predictor of *USF1*–BS among the features considered. We will consider other relevant genomic annotations, such as histone modifications, in future prediction method development. One important caveat is that the reliability and accuracy of these individual features will influence the performance of the prediction method. Feature selection during model building will become even more important when we integrate more genomic features in the future.

Each TF is unique in its binding site preference. Universal prediction methods may not perform well for all TFs, given inherent variation in binding domains, binding sequence preferences, homology level across species and family members. We believe, however, that our general model-building framework has the potential to be extended to other TFs for which there are available data detailing locations of a sufficient number of binding sites for use as a training dataset. As more results from genome-wide ChIP-chip studies become publicly available, it will become feasible to apply this prediction method to many other TFs.

Several aspects of this method can be improved in the future, such as using additional TFBS-related genomic features, evaluating other motif scanning methods, incorporating protein–DNA interactions, including binding site cluster information, using different gene annotations and exploring additional computational prediction models, such as support vector machines. We focused on a region 5 kb upstream of the TSSs of RefSeq mRNAs because the published *USF1* ChIP-chip study indicated that most of the *USF1* binding regions were found in proximal promoters.[17] *USF1*-BSs could occur beyond 5 kb upstream of TSSs, however, implying that a wider range of genomic regions could be considered in the future.

Training dataset from published *USF1* ChIP-chip results

A reliable training dataset is crucial for the development of an accurate and reliable prediction method. We chose published *USF1* ChIP-chip results[17] as our training dataset because they represented the largest publicly available *USF1* binding dataset; however, the exact location of *USF1*–BSs from these data is confounded by the common noise of ChIP-chip experiments and by a large average locus size on the ENCODE microarray — approximately 1 kb. To circumvent these

problems, we used the potential *USF1*–BSs identified by the PWM scoring method from the positive and negative loci to construct the training dataset. Each positive locus might include multiple potential *USF1*–BSs; however, it is unlikely that every potential *USF1*–BS from each positive locus interacts with *USF1*. Accordingly, we expect that our training dataset includes some false positives. To address this problem, we grouped all the potential *USF1*–BSs in the training dataset by their locus on the microarray and performed a LOLO cross-validation to evaluate the prediction models. The prediction scores of these potential *USF1*–BSs within each locus were used to predict that locus. If the locus has at least one predicted *USF1*–BS, it would be scored as a positive locus, otherwise it would be scored as a negative locus. This allows us to compare our prediction with *USF1* ChIP-chip results directly. We believe that LOLO cross-validation retains the underlying biological correlations while avoiding over-fitting the prediction models.

The training dataset was derived from the ENCODE regions, which included promoter, intronic, exonic and intergenic regions. *USF1*–BSs in all these regions may have different properties to the genomic features within the 5 kb upstream region of TSSs from the RefSeq mRNAs. These differences may cause the model based on this training dataset to behave differently on the full ENCODE regions.

Predicted *USF1* binding sites and target genes

Scanning 5 kb upstream of the TSSs of 23,105 RefSeq mRNAs, we identified a list of potential *USF1*–BSs and target genes that can be used as candidates for studying susceptibility to CAD and other complex human diseases. Our genome-scale prediction includes 24,010 *USF1*–BSs and 5,801 candidate genes. The numbers of *USF1* binding sites and target genes in the genome were expected to be large, since *USF1* is widely expressed in many tissues and developmental stages.[15] Other large-scale *in vivo* experiments have found that many other TFs are associated with an unexpectedly large numbers of target genes. For example, an investigation of *c-Myc*, which belongs to the same *bHLH* family as *USF1* and shares a similar core binding site, identified 756 *c-Myc* binding sites on chromosomes 21 and 22.[31] Extrapolating this to the whole genome provides an estimate of 25,000 *c-Myc* binding sites. A genome-wide study of the transcription factor signal transducer and activator of transcription 1 (*STAT1*) binding sites using ChIP-sequencing technology also identified 41,582 and 11,004 putative binding regions in stimulated and unstimulated cells, respectively.[32] These data are similar in magnitude to our genome-scale estimate of *USF1*–BSs. ChIP-based studies can only identify genes that are targets under specific cellular or environmental conditions. It is important to note, however, that our *in silico* prediction method will identify potential *USF1*–BSs independent of cell type, stage or environment. Thus, the numbers of predicted *USF1* binding sites and target genes might be higher using our method than using *in vivo* experiments. As more data become available, especially DNaseI HS identified from multiple cell lines, we will be able to evaluate the tissue specificity of the genomic features and our predictions. We acknowledge that DNA binding domains of the two members of the *USF* family (*USF1* and *USF2*) are highly conserved across multiple species, and often *USF1* and *USF2* form heterodimers to bind DNA, suggesting that the two proteins may share target genes. Although we used experimentally defined *USF1*–BSs to construct the PWM, other *bHLH* family members also have the same core binding sequence, 5′-CACGTG-3′; therefore, our prediction results might include the binding sites of other *bHLH* family members.

Application to human disease study

The main goals for this study were to predict genome-scale binding sites of *USF1* and to identify a novel group of CAD candidate genes regulated by *USF1* that give us the opportunity to evaluate gene–gene interactions. In Additional file 3, we have provided the predicted *USF1*–BSs and their prediction scores. By identifying a large number of predicted *USF1*–BSs, our results allow for adjusting the stringency of the prediction score threshold

to refine gene targets and also for choosing specific filters to emphasise a particular subset of interest. 'Genomic convergence', a strategy that integrates several independent and separate lines of experimental evidence to prioritise disease-associated candidate genes,[33] is being used by our CAD study to combine the *USF1*–BS prediction results with other information related to CAD to identify candidate genes. For example, a previously published study of gene expression signatures from human aortas identified 229 genes to be differentially expressed in aortas with and without atherosclerosis and found these genes to be highly predictive of the condition.[34] By combining our *in silico USF1*–BS prediction method with this expression result, we identified 87 *USF1* target genes that were differentially expressed between cases and controls in aorta (see Additional files 10 and 11). This approach highlights the potential for combining information from two distinct and methodologically diverse genome-scale investigations to define a list of important candidate genes from an unmanageably large list of initial targets.

SNPs are the most abundant molecular markers in the human genome. SNPs are commonly used for large-scale genetic association studies to identify genetic factors responsible for complex genetic diseases. Current high-throughput genotyping technologies enable researchers to genotype large numbers of SNPs efficiently. It remains a challenge to select SNPs with potential functional impact, however, especially from the large number of identified non-coding SNPs. One variant of particular interest are the SNPs within *cis*-regulatory elements, such as TFBS, because changing the TFBS sequence could alter the TF binding affinity within this region and further may influence the transcriptional regulation of the corresponding gene. These *cis*-regulatory variations are not necessarily deleterious. They might have subtle effects on gene expression and may contribute to the disease through interacting with other alleles and/or environmental factors, thereby playing important roles in the pathogenesis of many complex diseases in humans.[35] The bp resolution of our *USF1*–BS predictions enables us to isolate potential functional variations that may be used to select candidate variants for further testing for a functional impact and relation to disease. We have identified 751 SNPs within our predicted *USF1*–BSs in the human genome based on the genomic locations of the SNPs released by the NCBI in dbSNP build 126 (see Additional file 3). The experimental approaches that distinguish functional from neutral variations among these SNPs include, but are not limited to, well-designed case-control or family-based genetic association studies, allele-specific gene expression analysis and focused molecular biology studies. In summary, these SNPs within predicted *USF1*–BSs have the potential to influence the regulation of *USF1* target genes; they enable identification of a specific *USF1* regulatory network and, ultimately, study of the association of *USF1* with complex disease in humans.

Conclusion

This novel prediction method makes use of additional genomic features besides the PWM score and enables a more accurate and efficient genome-scale identification of specific *USF1*–BSs and associated target genes. The results of this study will help to identify *USF1*-regulated genes which might, in turn, be associated with CAD. We suggest that this method be generally applied to other transcription factors identified in human disease studies to further the understanding of encoded functional elements in the genome and their role in complex disease pathways.

Acknowledgments

We thank the staff at the Center for Human Genetics at Duke Medical Center. We would also like to give special thanks to the following individuals: Deqiong Ma, David Crosslin and Andrew Dellinger for their contribution to this publication. This study was supported by NIH grants HL073389 (Hauser), MH059528 (Hauser) and HL73042 (Goldschmidt, Kraus).

References

1. Mohlke, K.L. and Boehnke, M. (2005), 'The role of HNF4A variants in the risk of type 2 diabetes', *Curr. Diab. Rep.* Vol. 5, pp. 149–156.
2. Pajukanta, P., Lilja, H.E., Sinsheimer, J.S., Cantor, R.M. *et al.* (2004), 'Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1)', *Nat. Genet.* Vol. 36, pp. 371–376.

3. Wang, L., Fan, C., Topol, S.E., Topol, E.J. *et al.* (2003), 'Mutation of MEF2A in an inherited disorder with features of coronary artery disease', *Science* Vol. 302, pp. 1578–1581.
4. Connelly, J.J., Wang, T., Cox, J.E., Haynes, C. *et al.* (2006), 'GATA2 is associated with familial early-onset coronary artery disease', *PLoS Genet*. Vol. 2, p. e139.
5. Komulainen, K., Alanne, M., Auro, K., Kilpikari, R. *et al.* (2006), 'Risk alleles of USF1 gene predict cardiovascular disease of women in two prospective studies', *PLoS Genet*. Vol. 2, p. e69.
6. Krig, S.R., Jin, V.X., Bieda, M.C., O'Geen, H. *et al.* (2007), 'Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays', *J. Biol. Chem.* Vol. 282, pp. 9703–9712.
7. Hartman, S.E., Bertone, P., Nath, A.K., Royce, T.E. *et al.* (2005), 'Global changes in STAT target selection and transcription regulation upon interferon treatments', *Genes Dev.* Vol. 19, pp. 2953–2968.
8. MacIsaac, K.D. and Fraenkel, E. (2006), 'Practical strategies for discovering regulatory DNA sequence motifs', *PLoS Comput. Biol.* Vol. 2, p. e36.
9. Bulyk, M.L. (2003), 'Computational prediction of transcription-factor binding site locations', *Genome Biol.* Vol. 5, p. 201.
10. Stormo, G.D. (2000), 'DNA binding sites: Representation and discovery', *Bioinformatics* Vol. 16, pp. 16–23.
11. King, D.C., Taylor, J., Elnitski, L., Chiaromonte, F. *et al.* (2005), 'Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences', *Genome Res.* Vol. 15, pp. 1051–1060.
12. Davuluri, R.V., Grosse, I. and Zhang, M.Q. (2001), 'Computational identification of promoters and first exons in the human genome', *Nat. Genet*. Vol. 2, pp. 412–417.
13. Gardiner-Garden, M. and Frommer, M. (1987), 'CpG islands in vertebrate genomes', *Mol. Biol*. Vol. 196, pp. 261–282.
14. Felsenfeld, G. and Groudine, M. (2003), 'Controlling the double helix', *Nature* Vol. 421, pp. 448–453.
15. Corre, S. and Galibert, M.D. (2005), 'Upstream stimulating factors: Highly versatile stress-responsive transcription factors', *Pigment Cell Res.* Vol. 18, pp. 337–348.
16. Bendall, A.J. and Molloy, P.L. (1994), 'Base preferences for DNA binding by the bHLH-Zip protein USF: Effects of $MgCl_2$ on specificity and comparison with binding of *Myc* family members', *Nucl. Acids Res*. Vol. 22, pp. 2801–2810.
17. Rada-Iglesias, A., Wallerman, O., Koch, C., Ameur, A. *et al.* (2005), 'Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays', *Hum. Mol. Genet*. Vol. 14, pp. 3435–3447.
18. The ENCODE (ENCyclopedia Of DNA Elements) Project (2004), *Science* Vol. 306, pp. 636–640.
19. UCSC Genome Bioinformatics. http://genome.ucsc.edu/.
20. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S. *et al.* (2003), 'The UCSC Genome Browser Database', *Nucl. Acids Res.* Vol. 31, pp. 51–54.
21. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M. *et al.* (2002), 'The Human Genome Browser at UCSC', *Genome Res*. Vol. 12, pp. 996–1006.
22. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P. *et al.* (2008), 'High-resolution mapping and characterization of open chromatin across the genome', *Cell* Vol. 132, pp. 311–322.
23. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S. *et al.* (2005), 'Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes', *Genome Res.* Vol. 15, pp. 1034–1050.
24. TRANSFAC. http://www.gene-regulation.com/pub/databases.html.
25. Hertz, G.Z. and Stormo, G.D. (1999), 'Identifying DNA and protein patterns with statistically significant alignments of multiple sequences', *Bioinformatics* Vol. 15, pp. 563–577.
26. Minka, T. (2003), 'A comparison of numerical optimizers for logistic regression', Department of Statistics, Carnegie Mellon University. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.85.7017.
27. Norusis, M. (2004), 'SPSS 13.0 Statistical Procedures Companion', Prentice Hall, Inc., Upper Saddle River, NJ, USA.
28. Jiang, C., Xuan, Z., Zhao, F. and Zhang, M.Q. (2007), 'TRED: A transcriptional regulatory element database, new entries and other development', *Nucleic Acids Res.* Vol. 35, pp. 137–140.
29. Naukkarinen, J., Gentile, M., Soro-Paavonen, A., Saarela, J. *et al.* (2005), 'USF1 and dyslipidemias: Converging evidence for a functional intronic variant', *Hum. Mol. Genet*. Vol. 14, pp. 2595–2605.
30. Zhang, M.Q. (1998), 'Identification of human gene core promoters *in silico*', *Genome Res.* Vol. 8, pp. 319–326.
31. Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P. *et al.* (2004), 'Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs', *Cell* Vol. 116, pp. 499–509.
32. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M. *et al.* (2007), 'Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing', *Nat. Methods* Vol. 8, pp. 651–657.
33. Hauser, M.A., Li, Y.J., Takeuchi, S., Walters, R. *et al.* (2003), 'Genomic convergence: Identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage', *Hum. Mol. Genet*. Vol. 12, pp. 671–677.
34. Seo, D., Wang, T., Dressman, H., Herderick, E.E. *et al.* (2004), 'Gene expression phenotypes of atherosclerosis', *Arterioscler. Thromb. Vasc. Biol.* Vol. 24, pp. 1922–1927.
35. Andersen, M.C., Engström, P.G., Lithwick, S., Arenillas, D. *et al.* (2008), 'In silico detection of sequence variations modifying transcriptional regulation', *PLoS Comput. Biol*. Vol. 4, p. e5.

Additional files

Additional file 1: Kernel logistic regression

We provide a brief introduction to the kernel logistic regression. For more details, please refer to the study by Minka.[26]

Additional file 2: Comparison of *USF1* binding sites prediction methods

The threshold of being a predicted *USF1*–BS was 0.5. Sensitivity was the proportion of correctly predicted true positive loci, whereas specificity was the proportion of true negative loci predicted as negative loci. The AUCs were calculated from LOLO cross-validation (see Methods).

Additional file 3: Predicted *USF1*–BSs and associated target genes in the human genome

The optimal prediction model was applied to the 5 kb regions upstream of the TSSs of 23,105 RefSeq mRNAs. 9.721 genes with 24,010 *USF1*–BSs are predicted to be the targets of *USF1*.

Additional file 4: Distribution of PWM and DNaseI HS scores in the training dataset (correlation coefficient = 0.121)
Our training dataset included 935 potential *USF1*–BSs; 135 were associated with the 34 positive loci and 800 with the 177 negative loci identified by the *USF1* ChIP-chip study.

Additional file 5: Distribution of PWM and RP5 scores in the training dataset (correlation coefficient = 0.117)
Our training dataset included 935 potential *USF1*–BSs; 135 were associated with the 34 positive loci and 800 with the 177 negative loci identified by the *USF1* ChIP-chip study.

Additional file 6: Distribution of PWM and PhastCons8 scores in the training dataset (correlation coefficient = −0.007)
Our training dataset included 935 potential *USF1*–BSs; 135 were associated with the 34 positive loci and 800 with the 177 negative loci identified by the *USF1* ChIP-chip study.

Additional file 7: Distribution of PWM and MostCons8 scores in the training dataset scores (correlation coefficient = 0.060)
Our training dataset included 935 potential *USF1*–BSs; 135 were associated with the 34 positive loci and 800 with the 177 negative loci identified by the *USF1* ChIP-chip study.

Additional file 8: Distribution of PWM and CpG scores in the training dataset scores (correlation coefficient = 0.064)
Our training dataset included 935 potential *USF1*–BSs; 135 were associated with the 34 positive loci and 800 with the 177 negative loci identified by the *USF1* ChIP-chip study.

Additional file 9: PWM scores distribution of *USF1*–BSs
By scanning the 5 kb upstream of the TSSs of 23,105 RefSeq mRNAs in the human genome, 290,614 potential *USF1*–BSs were identified by the PWM scoring method. A total of 24,010 predicted *USF1*–BSs were generated using the optimal prediction model with the default prediction threshold (0.5). In addition, 20 experimentally identified *USF1* target genes had been used to validate our prediction. The numbers on the top of the bar indicate the number of these genes having PWM scores within that range. For example, '3/4' means that there are four *USF1* target genes with PWM scores in the range 4.5 to 5, and three of these genes are correctly identified by the optimal kernel-based prediction method.

Additional file 10: CAD candidate genes identified by the 'genomic convergence' approach
A previously published study of gene expression signatures from human aortas identified 229 genes that are differentially expressed in aortas with and without atherosclerosis.[34] Based on the score distribution of our prediction, we chose a stringent threshold (0.99) to reduce the number of predicted *USF1* target genes to be used as candidate genes for further analysis to 5,801. By combining our predicted *USF1* candidate genes with the published expression result, we identified 87 *USF1* target genes that were differentially expressed between cases and controls in the aorta. The prediction score of each *USF1*–BS within these genes can be found in Additional file 3.

Additional file 11: PWM score distribution of *USF1*–BSs within CAD candidate genes
A previously published study of gene expression signatures from human aortas identified 229 genes that are differentially expressed in aortas with and without atherosclerosis.[34] By combining our *in silico* *USF1*–BS prediction method with this expression result, we identified 87 *USF1* target genes that were differentially expressed between cases and controls in the aorta. 142 genes were excluded because they did not have predicted *USF1*–BSs.