ABSTRACT

WANG, TIANYUAN. Identifying Transcription Factor Targets and Studying Human Complex Disease Genes. (Under the direction of Dr. Elizabeth R. Hauser and Dr. Steffen Heber.)

Transcription factors (TFs) have been characterized as mediators of human complex disease processes. The target genes of TFs also may be associated with disease. Identification of potential TF targets could further our understanding of gene-gene interactions underlying complex disease. We focused on two TFs, USF1 and ZNF217, because of their biological importance, especially their known genetic association with coronary artery disease (CAD), and the availability of chromatin immunoprecipitation microarray (ChIP-chip) results. First, we used USF1 ChIP-chip data as a training dataset to develop and evaluate several kernel logistic regression prediction models. Our most accurate predictor significantly outperformed standard PWM-based prediction methods. This novel prediction method enables a more accurate and efficient genome-scale identification of USF1 binding and associated target genes. Second, the results from independent linkage and gene expression studies suggest that ZNF217 also may be a candidate gene for CAD. We further investigated the role of ZNF217 for CAD in three independent CAD samples with different phenotypes. Our association studies of ZNF217 identified three SNPs having consistent association with CAD in three samples. Aorta expression profiling indicated that the proportion of the aorta with raised lesions was also positively correlated to ZNF217 expression. The combined evidence suggests that

ZNF217 is a novel susceptibility gene for CAD. Finally, we applied our previously developed TF binding site (TFBS) prediction method to ZNF217. The performance of the prediction models of ZNF217 and USF1 are very similar. We demonstrated that our TFBS prediction method can be extended to other TFs. In summary, the results of this dissertation research are (1) evaluation of two TFs, USF1 and ZNF217, as susceptibility factors for CAD; (2) development of a generalized method for TFBS prediction; (3) prediction of TFBSs and target genes of two TFs, and identification of SNPs within TFBSs. This research allows for the development of study design to access TF based interactions in genetic susceptibility to human complex disease.

Identifying Transcription Factor Targets and Studying Human Complex Disease Genes

by
Tianyuan Wang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2009

APPROVED BY:

_____	_____
Dr. Elizabeth R. Hauser Chair of Advisory Committee	Dr. Steffen Heber Chair of Advisory Committee
_____	_____
Dr. David McK. Bird	Dr. Jonathan M. Horowitz

_____
Dr. Jeffrey L. Thorne

DEDICATION

To my wife, daughter and parents

BIOGRAPHY

Tianyuan Wang was born in Beijing, People's Republic of China. He graduated from China Agricultural University in Beijing, China in 1991 with a B.S. in Plant Biochemistry and Physiology. He continued his education at the University of Georgia, where he received his M.S. in Plant Pathology in 1997 and M.S. in Pharmacy in 2001. He joined the Bioinformatics program at North Carolina State University in 2004. While working toward his doctoral degree, he worked as a full-time bioinformatics analyst at the Center for Human Genetics at Duke Medical Center under the direction of Dr. Elizabeth R. Hauser.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisors, Dr. Elizabeth R. Hauser and Dr. Steffen Heber, for their support, encouragement and guidance during my doctoral study. I feel very fortunate to work under Dr. Elizabeth R. Hauser's supervision for almost six years. I also would like to thank my other committee members, Dr. David McK. Bird, Dr. Jonathan M. Horowitz, and Dr. Jeffrey L. Thorne. They have guided my research projects and provided valuable suggestions to this dissertation. In addition to my advisory committee, I am very fortunate and grateful to collaborate with Dr. Terrence S. Furey at Duke University.

I would like to thank the staff at the Bioinformatics Research Center, in particular, Juliebeth Briseno, for her assistance with graduate school regulations. Further, I thank the staff at the Center for Human Genetics at Duke Medical Center. I really appreciate their supports and friendships.

I would like to give special thanks to my wife, Jun Yang, for her encouragement, support and love. Finally, I am very grateful for the supports of my family and many friends during the course of this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ADDITIONAL FILES

CHAPTER 1

Introduction

1. 1 Human complex diseases

Human complex or multifactorial diseases are defined as diseases which are caused by a number of genetic risk factors and environmental effects (Schork, 1997). In addition, genetic risk factors interact with environmental effects over a lifetime to change function at the molecular, cellular, tissue, and organ levels and to ultimately result in complex disease. The disease-gene mapping and the identification of genetic variations associated with complex disease will provide the opportunities for further understanding of diseases, clinical diagnosis, pharmacogenetics, and drug development in the future. Unlike "Mendelian" or "single gene" diseases caused by single mutations with strong phenotypic effects, complex diseases are caused by multiple variations, and each variation has a weak effect. In addition, these variations associated with complex disease might have been under little selective pressure over long evolution periods, so they are more common in the population. So our knowledge of mutation effects in Mendelian diseases cannot generally be applied to the variations associated to complex diseases.

Common human complex diseases include coronary heart disease, hypertension, diabetes, obesity, cancers, asthma, Alzheimer's disease, and Parkinson's disease. Among them, coronary heart disease, also called coronary artery disease (CAD) is the leading cause of death in the western world (Rosamond et al., 2007). CAD is the result of the accumulation of plaques within the walls of the arteries, which supply the heart with oxygen and nutrients. After long period of progression, some of these plaques may rupture and limit blood flow through the heart, and cause sudden death (Watkins and Farrall,

2006). There are two types of risk factors associated with CAD. Non-modifiable risk factors include age, gender (male), and family history; modifiable risk factors include hypercholesterolemia, smoking, hypertension, physical inactivity, obesity, high blood pressure, sleep deprivation, absence of key nutritional elements, stress, and depression. Prevention of CAD mainly focuses on these modifiable risk factors. In general, the combination of healthy diet and exercise will reduce risks of CAD.

1.2 Genome scan to identify candidate genes

Studying human complex diseases including CAD has been a challenging task. This is due to the complexity of the human body, population variation, genetic heterogeneity, late age of onset, gene-gene interaction, gene-environment interaction, and the difficulty of defining genetically-caused diseases phenotypes. Large scale family-based linkage studies and single marker association studies have been the main approaches applied to identify the locations of the disease genes and genetic variation associated with CAD. Recently, large-scale genome-wide approaches are becoming popular because of the completion of human genome sequence, the availability of a large number of single nucleotide polymorphisms (SNPs) and haplotype information, and improving technology (Chen et al., 2007; Borecki and Province, 2008). Linkage studies are used to identify specific genomic regions tagged by genetic markers, which do not recombine with the disease loci within families. By contrast, association studies are applied to seek correlation between specific genetic variations and traits among individuals in a sample, which may

suggest causal roles for the genetic variants. Linkage disequilibrium (LD) may make it such that we measure a marker that is not a casual variant but it is correlated with a casual variant. In linkage studies, the location of the trait locus is indicated by the genetic distance to the marker, which is measured by the recombination fraction. Linkage is often measured by the logarithm of the odds (LOD) score, which is the logarithm of odds of the recombination rate equal to θ estimated from the observed data with respect to the assumption that the recombination rate is 0.5 (Morton, 1998). By contrast, in association studies, the genetic variation itself or variations nearby `in LD have direct impact on trait variation. Association studies are generally more powerful than linkage studies to identify the causal risk factors (Borecki and Province, 2008).

Association studies are increasingly popular for identifying genetic factors associated with complex diseases (Rodriguez-Murillo and Greenberg, 2008). Especially, the availability of abundant SNP markers across the genome makes it possible to perform a genome-wide association study (GWAS). Association studies include two main categories, population-based (case-control) studies and family-based studies. The major advantage of population-based association studies is that the cases and controls are collected independently, which is relatively easier than collecting family samples (Rodriguez-Murillo and Greenberg, 2008). However, family-based association studies are robust to population stratification, which comes from the presence of multiple subgroups in a study sample with different disease prevalence and marker allele frequencies. The Association in the Presence of Linkage (APL) method has been applied in family-based association

analysis in order to take the advantage of the large number of affected sibling pairs in the sample, and to adjust for the correlation between transmission of parental SNP alleles to multiple affected offspring due to linkage (Martin et al., 2003; Chung et al., 2007).

Numerous genome scan studies have been applied to CAD studies, and the results suggest that CAD has a strong genetic component. However, the underlying genetic architecture of CAD has not been well understood. Ten independent linkage screens have been performed in CAD cohorts to identify the candidate genes. Cumulatively, these studies have identified regions of linkage on Chromosomes 1, 2, 3, 5, 7, 12, 13, 14, 16, 17, 19 and X. Only a single gene at chromosome 13q, ALOX5AP, has been replicated in linkage studies (Connelly et al., 2006; Hauser et al., 2004; Helgadottir et al., 2004). In addition, two independent GWASs had focused on identifying the regions of the genome showing evidence of association with CAD (The Wellcome Trust Case Control Consortium, 2007; Samani et al., 2007). Seven loci were identified by each study. The only common region was on 9p21, which included two well-annotated genes, CDKN2A and CDKN2B. Despite significant study, the role of this region in changing CAD susceptibility is unknown. Further this 9p21 locus may be only part of the regulatory network associated with CAD. Because of different populations, phenotypes, and limited SNP coverage of these two GWASs, it is difficult to make direct comparison of their results. Different results from two GWASs also illustrates a potential limitation of the GWAS, which may miss potentially important functional candidate genes.

1.3 Candidate gene analyses

In general large-scale genome-wide approaches are unbiased and efficient to find disease genes. However, these approaches very often produce a large number of candidates. It has been challenging to follow up these results and focus on a small number of meaningful genes. An alternative approach is focused candidate gene analysis to identifying susceptibility genes. "Genomic convergence" is an efficient strategy for prioritizing disease associated candidate genes. "Genomic convergence" integrates several independent separate lines of experimental evidence, including replication in phenotypically similar but independent populations, linkage, gene expression studies, proteomics, metabolomics, lipomics, etc, to assess the strength of the identified association (Hauser et al., 2003). In addition to simply overlapping multiple lists of genes, data integration could be performed at a higher level. For example, allele-specific expression is the assessment of allelic distortion in association study with differential gene expression in the same tissue or cell type. This approach has been successfully applied to a variety of genetic association studies of CAD, such as GATA2, PLA2G7, and ZNF217 (Connelly et al., 2006; Sutton et al., 2008; Chapter 3 from this dissertation).

In summary, genetic linkage, genome-wide association, and candidate gene studies all have their advantages and limitations. Genetic linkage studies have been successful in identifying genes associated with Mendelian diseases. However, finding CAD genes from linkage studies remain a challenge, which reflects the complexity of CAD including genetic heterogeneity, multiple genes with small effects, influence of environmental

effects. Candidate gene studies are more powerful in identifying the association of specific genes of interest. So this approach is biased and limited to genes with prior evidence for association. The significant associations from one study should be validated with additional independent studies to avoid false positives, which were discovered by random chance instead of real causal variants. In addition, detailed molecular biology experiments are necessary for understanding the biological roles of these genes and identifying gene-gene interaction underlying complex disease. By contrast, genome-wide association studies are unbiased approaches for finding novel disease candidate genes. As more SNPs are becoming available and technology is improving, this approach is becoming more popular in human complex disease studies (Chen et al., 2007). The extensive LD among variants across the genome may show the significantly associated genetic variations which are not necessarily the causative ones. Further identification and functional annotation of the causative variants can provide insights into the cause of the disease.

1.4 Transcription factors and human complex disease

The approximately 25,000 genes in the human genome demonstrate dramatic diversity in terms of expression levels, both temporally and spatially. Despite this diversity, the expression of all genes is controlled by a relatively small number (< 2,000) of transcription factors (TFs). These TFs usually work in specific combinations to regulate individual genes. Many TFs have been characterized as mediators of complex disease processes, and numerous publications have identified SNPs in TFs that are significantly

associated with human disease (Connelly et al., 2006; Mohlke et al., 2005; Pajukanta et al., 2004; Wang et al., 2003; Weedon et al., 2007). Complex genetic diseases, such as cardiovascular disease, are caused by genetic risk factors, environmental effects and the interaction of them. TFs (and their cognate binding sites) ultimately influence the expression of many downstream genes in temporal, cell type, or environmental specific ways. Slight changes to the level of a TF in the cell can have a significant effect on its downstream targets. Therefore TFs are likely to be important candidates in the dissection of complex human disease. 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 interaction underlying complex disease and its specific contribution to heart disease.

1.5 TF binding site and target gene prediction

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, therefore, critical to our understanding of the biological regulation of the cell. Although many genome sequences are available, encoded functional elements such as TFBSs have not been fully characterized. This is due, in part, to the complexity of TF binding activity and degeneration of the core binding site. Genome-wide gene expression arrays are used to identify TFBS motifs within transcription start sites of differentially expression genes.

However, it is difficult to differentiate direct or indirect target genes from clustered or co-expressed genes generated from gene expression microarrays. Alternatively, identification of TFBS by ChIP-chip, a technique which combines chromatin immunoprecipitation and microarray analysis (Hartman et al., 2005; Krig et al., 2007), allows for the identification of a large number of DNA fragments bound by TFs. However, ChIP-chip experiments are time-consuming, expensive, and can only identify subsets of all potential TFBS because of variation in tissue type, environmental conditions and biological cofactors necessary of TF binding. Thus there can be considerable experimental variability when experiments are performed under different conditions. It would be more efficient to develop an *in silico* computational method for TF target prediction followed by less costly genotyping and focused molecular biology experiments to identify association of gene-gene interaction and complex disease.

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 (Bulyk, 2003; MacIsaac et al., 2006; Stormo 2000). 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. However, because TFs can also bind to non-consensus sequences, individual base pairs (bp) within the binding site may be degenerate despite a relatively well-conserved consensus sequence (Stormo 2000). The other approach is based on binding site matrices, which include the position weight matrix (PWM) and the position frequency matrix (PFM)

(Stormo 2000). 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 sequences 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, and it also lacks the flexibility to incorporate additional genomic information. For example, a matrix from in vitro experiments may not reflect the true TF binding preference in vivo. Alternatively, matrices identified by in vivo experiments are reliable, but large numbers of such data are not readily available. Several tools have been developed to apply to all known TFs and include TFBSfinder (Tsai et al., 2006), BinomSite (http://wwwmgs3.bionet.nsc.ru/mgs/programs/binomsite2/), and TFSEARCH (http://molsun1.cbrc.aist.go.jp/research/db/TFSEARCH.html). However, universal prediction methods may not perform well for all TFs given variations in binding domain and binding sequence preference, and homology level across species and family members. In general, the low specificity of these methods leads to an inflated number of predicted TFBSs many of which are false positive results. Therefore, the reliability of prediction methods based on DNA sequence alone is low. An ideal prediction method combines 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 (King et al., 2005). 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-wide 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 (King et al., 2005). CpG islands are CG dinucleotide rich regions of the genome commonly associated with transcription start sites and promoters (Davuluri et al., 2001; Gardiner-Garden and Frommer, 1987). 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 DNaseI hypersensitive (HS) sites that are hypersensitive to DNaseI cleavage. DNaseI HS sites are nucleosome-free regions of open chromatin associated with regulatory elements, such as promoters, enhancers and silencers (Felsenfeld and Groudine, 2003). While some of these genomic features have been used individually to filter the predictions from sequence based scoring methods (Bulyk, 2003; MacIsaac et al., 2006), TFBS prediction methods would benefit from carefully selecting and integrating these genomic features. Although the number of genomic features available is quite large, current prediction methods do not take full advantage of these genomic features.

1.6 The scope of this dissertation

Many TFs have been characterized as mediators of complex disease processes. 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 interaction underlying complex disease and its specific contribution to heart disease. We focused on two TFs in this dissertation, specifically because of their biological importance, particularly in regard to their known genetic association with CAD, and the recent availability of ChIP-chip results.

First, Several linkage and association studies indicate that the transcription factor upstream stimulatory factor 1 (USF1) is genetically associated with CAD (Pajukanta et al., 2004). USF1 is ubiquitously expressed in human tissues and is a key regulator of several biological processes such as stress and immune response, cell cycle, and cell proliferation (Corre and Galibert, 2005). 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 (Bendall and Molloy, 1994). The complete binding site of USF1 is represented by 5'RYCACGTGRY 3' (Bendall and Molloy, 1994). DNA-binding activity of USF1 can be modulated through phosphorylation, homo- or heterodimerization, and variation in binding site sequence (Corre and Galibert, 2005). 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 chromatin ChIP-chip results for 1% of the genome (Rada-Iglesias et al., 2005). Our goals of Chapter 2 were to (1) develop a reliable and accurate method for USF1-BS prediction; (2) make a genome-wide prediction of potential USF1-BSs, and evaluate this prediction against a known set of experimentally defined robust genes; and (3) identify USF1 associated target genes to aid in the study of cardiovascular disease.

Second, Zinc finger protein 217 (ZNF217) is known to repress the transcription of many genes and is associated with cell proliferation, survival, and invasiveness of cancer cells (Quinlan et al., 2007). The ZNF217 region is selectively amplified during progression bof several cancers. These results suggest that ZNF217 may give tumor cells selective advantage by interfering with normal regulation of cell growth, cell death, differentiation, and DNA repair (Krig et al., 2007). ZNF217 locates to a modest linkage peak on chromosome 20q13 from a previous family based linkage study (GENECARD) of early-onset CAD (Hauser et al., 2004). Additionally, a study of gene expression signatures from human aortas identified ZNF217 among 229 genes to be differentially expressed in aortas with and without atherosclerosis (Seo et al., 2004). Given the convergence of this evidence, we hypothesized that as a TF with multiple target genes, ZNF217 may be associated with CAD. Therefore, in Chapter 3 we investigated genetic variation within the gene in three independent CAD samples: a case-control sample from the Duke CATHGEN cohort, GENECARD families, and human donor aorta samples, and evaluated the role of ZNF217 as a candidate gene for CAD/atherosclerosis. In addition, we performed non-parametric regression analysis of aorta disease burden (the rankings of the Sudan IV and

raised lesion) with ZNF217 expression, age, sex, and race. To further test the potential functional impact of the SNPs, we evaluated allele-specific expression in the aorta samples.

Third, we chose ZNF217 to apply our previously developed novel TFBS prediction method in Chapter 2 because of its biological importance (over-expression in many cancers and genetic association with CAD) and the availability of ZNF217 ChIP-chip results (Krig et al., 2007). The specific goals of our study in Chapter 4 were to (1) evaluate our novel TFBS prediction method with another independent TF, ZNF217; (2) make a genome-wide prediction of potential ZNF217-BSs; and (3) identify ZNF217 target genes. The results of this study will help prioritize CAD candidate genes as well as describe ZNF217's specific biological contribution to heart disease through the identification of its target genes. Finally, we proposed some ideas for future studies in Chapter 5.

1.7 References

Bendall AJ, Molloy PL. Base preferences for DNA binding by the bHLH-Zip protein USF: effects of MgCl2 on specificity and comparison with binding of Myc family members. Nucleic Acids Res. 1994; 22: 2801-2810.

Borecki IB, Province MA. Linkage and association: basic concepts. Adv Genet. 2008; 60: 51-74.

Bulyk ML. Computational prediction of transcription-factor binding site locations. Genome Biol. 2003; 5: 201.

Chen Y, Rollins J, Paigen B, Wang X. Genetic and genomic insights into the molecular basis of atherosclerosis. Cell Metab. 2007; 6(3): 164-179.

Chung RH, Hauser ER, Martin ER. Interpretation of simultaneous linkage and family-based association tests in genome screens. Genet Epidemiol. 2007; 31(2): 134-142.

Connelly JJ, Wang T, Cox JE, Haynes C, Wang L, et al. GATA2 Is Associated with Familial Early-Onset Coronary Artery Disease. PLoS Genet. 2006; 2: e139.

Corre S, Galibert MD. Upstream stimulating factors: highly versatile stress-responsive transcription factors. Pigment Cell Res. 2005; 18: 337-348.

Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. Nat Genet. 2001; 29: 412–417.

Felsenfeld G, Groudine M. Controlling the double helix. Nature. 2003; 421: 448-453.

Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. Mol Biol 1987, 196: 261-282.

Hartman SE, Bertone P, Nath AK, Royce TE, Gerstein M, et al. Global changes in STAT target selection and transcription regulation upon interferon treatments. Genes Dev. 2005; 19; 2953-2968.

Hauser ER, Crossman DC, Granger CB, Haines JL, Jones CJ, et al. A genomewide scan for early-onset coronary artery disease in 438 families: The GENECARD Study. Am J Hum Genet. 2004; 75: 436–447.

Hauser MA, Li YJ, Takeuchi S, Walters R, Noureddine M, et al. Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage,' Hum Mol Genet. 2003; 12: 671-677.

Helgadottir A, Manolescu A, Thorleifsson G, Gretarsdottir S, Jonsdottir H, et al. The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. Nature Genetics. 2004; 36: 233-239.

King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res. 2005; 15: 1051-1060.

Krig SR, Jin VX, Bieda MC, O'Geen H, Yaswen P, et al. Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays. J Biol Chem. 2007; 282(13): 9703-9712.

MacIsaac KD, Fraenkel E. Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol. 2006; 2: e36.

Martin ER, Bass MP, Hauser ER, Kaplan NL. Accounting for linkage in family-based tests of association with missing parental genotypes. Am J Hum Genet. 2003; 73: 1016–1026.

Mohlke KL, Boehnke M. The role of HNF4A variants in the risk of type 2 diabetes. Curr Diab Rep. 2005; 5: 149-156.

Morton NE. Significance Levels in Complex Inheritance. Am J Hum Genet. 1998; 62: 690-697.

Pajukanta P, Lilja HE, Sinsheimer JS, Cantor RM, Lusis AJ, Gentile M, et al. Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). Nat Genet. 2004; 36: 371-376.

Quinlan KG, Verger A, Yaswen P, Crossley M. Amplification of zinc finger gene 217 (ZNF217) and cancer: when good fingers go bad. Biochim Biophys Acta. 2007; 1775(2): 333-340**.**

Rada-Iglesias A, Wallerman O, Koch C, Ameur A, Enroth S, et al. Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. Hum Mol Genet. 2005; 14: 3435-3447.

Rodriguez-Murillo L, Greenberg DA. Genetic association analysis: a primer on how it works, its strengths and its weaknesses. Int J Androl. 2008; 31(6):546-556.

Rosamond W, Flegal K, Friday G, Furie K, Go A, Greenlund K, et al. Heart Disease and Stroke Statistics--2007 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation. 2007; 115: e69-171.

Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. WTCCC and the Cardiogenics Consortium. Genomewide association analysis of coronary artery disease. N Engl J Med. 2007; 357:443-453.

Schork NJ. Genetics of complex disease: approaches, problems, and solutions. Am J Respir Crit Care Med. 1997; 156(4): S103-109.

Seo D, Wang T, Dressman H, Herderick EE, Iversen ES, et al. Gene expression phenotypes of atherosclerosis,' Arterioscler Thromb Vasc Biol. 2004; 24: 1922-1927.

Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000; 16: 16-23.

Sutton BS, Crosslin DR, Shah SH, Nelson SC, Bassil A, et al. Comprehensive genetic analysis of the platelet activating factor acetylhydrolase (PLA2G7) gene and cardiovascular disease in case-control and family datasets. Hum Mol Genet. 2008; 17(9): 1318-1328.

Tsai HK, Huang GT, Chou MY, Lu HH, Li WH. Method for identifying transcription factor binding sites in yeast. Bioinformatics. 2006; 22(14):1675-1681.

Wang L, Fan C, Topol SE, Topol EJ, Wang Q. Mutation of MEF2A in an inherited disorder with features of coronary artery disease. Science. 2003; 302: 1578-1581.

Watkins H, Farrall M. Genetic susceptibility to coronary artery disease: from promise to progress. Nat Rev Genet. 2006; 7(3): 163-173.
Weedon MN. The importance of TCF7L2. Diabet Med. 2007; 24(10): 1062-1066.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678.

CHAPTER 2

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], Elizabeth R. Hauser[1]

Human Genomics. 2009; Volume 3(3).

[1]Department of Medicine and Center for Human Genetics, Duke University Medical Center, Durham, North Carolina 27710, USA

[2]Deptartment of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina 27708, USA

[3]Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina 27708, USA

[4]Department of Computer Science, North Carolina State University, Raleigh, North Carolina 27695, USA

[5]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695, USA

2.1 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 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% of the genome was 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 AUC (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 kb upstream of the transcription start site of 9,721 genes. These predictions included 16 out 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 prioritized set of candidate genes for future CAD studies. We have shown that our novel prediction method that 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.

2.2 Background

Several transcription factors (TFs) have been characterized 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 interaction underlying complex disease. Genome-wide experimental methods, such as ChIP-chip [6, 7], a technique combining chromatin immunoprecipitation and microarray analysis for identifying TFs 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 focused molecular biology experiments to identify association of gene-gene interaction 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 characterized. This is due, in part, to the complexity of TF binding activity and 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, 9, 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, and 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 of 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 DNaseI hypersensitive (HS) sites that are hypersensitive to DNaseI cleavage. DNaseI HS sites 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 quite large, current prediction methods do not take full advantage of these genomic features.

Several linkage and association studies indicate that the transcription factor upstream stimulatory factor 1 (USF1) is genetically associated with coronary artery disease (CAD) [2]. USF1 is ubiquitously expressed in human tissues and is a key regulator of several biological processes such as 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]. DNA-binding activity of USF1 can be modulated through phosphorylation, homo- or heterodimerization, 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 chromatin

immunoprecipitation microarray (ChIP-chip) results for 1% of the genome [17]. Our goals were to (1) develop a reliable and accurate method for 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 binding sites on a genome scale. The results of this study will help prioritize CAD candidate genes as well as provide biological information in evaluating gene-gene interactions with respect to this common complex disease.

2.3 Methods

2.3.1 Genome sequence and features

All annotation and mapping locations of genomic features used to predict TFBSs were based on NCBI human genome build 35. ENCODE sequences [18] and the 5 kilobases (kb) regions upstream of the transcription start sites (TSSs) of 23,105 RefSeq mRNA sequences were obtained from the UCSC Genome Browser [19, 20, 21]. These RefSeq mRNA 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 2.1) for each potential 10 bp USF1-BS are continuous variables, and they 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 [22], 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 each base-pair's conservation score within the USF1-BS. The regulatory potential (RP5) scores were computed from alignments of 5 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 sites (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 [23].

2.3.2 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, 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 log2-ratio, base-2 logarithm of the ratio of fluorescence intensities of immunoprecipitated chromatin versus control chromatin for each spot on the array. The log2-ratios were in the range of -1 to 4. Thirty four loci with the log2-ratio greater than 1.25 were considered to be bound, while 234 loci on the array with the log2-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.

2.3.3 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 web application Patser was used to convert the USF1 binding matrix to PWM, and generated a numerically calculated cutoff score of 3.753 for predicting TFBSs based on the information content adjusted by sample size [25]. The average GC content of 47.1% 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.

2.3.4 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 2.1: Kernel logistic regression). 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*-value 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 AUC (area under the receiver operator characteristic curve) 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 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].

2.3.5 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 optimized model using the genomic features of each potential binding site. Lastly, the prediction results were evaluated by comparing to 20 robust USF1 target genes - the overlap between the target genes obtained from the TRED database [28] and reported from the literature [29].

2.4 Results

2.4.1 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 2.1 summarizes our general approach for 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 Mb). 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. Out of 234, 57 negative loci did not include any potential USF1-BSs and were excluded from the training dataset.

To more accurately refine predictions, we evaluated several models trained to identify USF1-BSs using kernel logistic regression with genomic features as variables: PhastCons8, MostCons8, RP5, and CpG and DNaseI HS (Table 2.1), 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 has the highest AUC (0.672), however its sensitivity (0.088) is much lower than the DNaseI HS model (0.235) (Table 2.2). 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*-value = 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, PhastCons8, RP5, and DNaseI HS) achieved 55.9% sensitivity and 87.6% 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 features or other combinations of features (Table 2.2, Figure 2.2, Additional file 2.2: Comparison of USF1 binding sites prediction methods). This model was considered as the optimal predictor among the models tested, and this model 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. 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 predicted USF1-BSs. The average $\log_2$-ratio (base-2 logarithm of ChIP enrichment ratio) of these predicted positive loci (0.1034) in the ChIP-chip experiment is significantly higher (p-value < 10^{-23}) than that of predicted negative loci (0.0085). This result indicates that the predicted positive loci are enriched among the loci with $\log_2$-ratio higher than 0 that are more likely to include USF1-BSs (Figure 2.3).

2.4.2 Genome-scale prediction and validation

We obtained DNA sequence 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% of the initial potential USF1-BSs (See Additional file 2.3: Predicted USF1-BSs and associated target genes in the human genome). We created a set of 20 robust USF1 target genes obtained from the TRED database and from literature to validate the genome-scale prediction results. Our prediction method was able to identify 16 out of these 20 genes (80%) as USF1 targets (Table 2.3).

2.4.3 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 2.4). 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 further reduce the number of predicted UFS1 target genes 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 2.5). 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 1kb sequence upstream of the transcription start site (data not shown). This could explain the concentration of predictions in this region, however, it did not alone account for the concentration of predicted USF1-BSs.

To better understand what 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 indicates 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 to the location of the USF1-BS in the region upstream of TSS ($r^2 = 0.41$).

2.5 Discussion

2.5.1 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 a large number of false positive results. One strategy for improving specificity of TFBS prediction is by using phylogenetic footprinting, which is based on the assumption that regions of multi-species sequence conservation are more likely to include regulatory elements. We hypothesized 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. Then we 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 the 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%) and AUC (0.827) among the models tested while still achieving high specificity (87.6%) (Figure 2.2). 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 is not sufficient to accurately distinguish between sample classes (See Additional file 2.4: Distribution of PWM and DNaseI HS scores in the training dataset and Additional files 2.5-2.8). Kernel logistic regression modeling maps the training data to high-dimensional space by considering all features jointly, and 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 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 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.

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 at each step during the prediction process. We find that: 1) the common PWM scoring method was sufficient to identify potential USF1-BSs for most of these genes; 2) these 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 (Additional file 2.9: PWM scores distribution of USF1-BSs). 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. The 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. However, we believe our general model building framework has the potential to be extended to other TFs for which there is available data detailing locations of a sufficient number of binding sites to be used 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 (SVM). 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]. However, USF1-BSs could occur beyond 5 kb upstream of TSSs, implying that a wider range of genomic regions could be considered in the future.

2.5.2 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 it represents 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-BS within

each locus was 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 that include promoter, intronic, exonic and intergenic regions. USF1-BSs in all these regions may have different properties than 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.

2.5.3 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 find 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 c-Myc 25,000 binding sites. A genome-wide STAT1 binding site study using ChIP-sequencing (ChIP-seq) 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 given 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 even higher using our method as compared to 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 very 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.

2.5.4 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. Our supplemental data lists predicted USF1-BSs and their prediction score (See Additional file 2.3: Predicted USF1-BSs and associated target genes in the human genome). 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 emphasize a particular subset of interest. "Genomic convergence", a strategy that integrates several independent separate lines of experimental evidence to prioritize disease associated candidate genes [33], is being used by our CAD study to combine the USF1 binding site 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 as differentially expressed in aortas with and without atherosclerosis and found these genes to be highly predictive of 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 aorta (See Additional file 2.10: CAD candidate genes identified by the "genomic convergence" approach, and Additional file 2.11: PWM score distribution of USF1-BSs within CAD candidate genes). 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.

Single nucleotide polymorphisms (SNPs) are the most abundant molecular marker 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 very efficiently. However, it is still remains a challenge to select SNPs with potential functional impact, especially from the large number of identified non-coding SNPs. One type of variant of particular interest are 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 base pair resolution of our USF1-BS predictions enable 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 NCBI in dbSNP build 126 (See Additional file 2.3: Predicted USF1-BSs and associated target

genes in the human genome). The experimental approaches to 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 specific USF1 regulatory network and ultimately study of USF1's association with complex disease in humans.

2.6 Conclusions

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 that 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.

2.7 Acknowledgements

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).

2.8 References

1. Mohlke, K.L., 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., Lusis, A.J., Gentile, 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., Wang, Q. (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., Wang, L., Shah, S.H. *et al*. (2006), 'GATA2 Is Associated with Familial Early-Onset Coronary Artery Disease,' *PLoS Genet*, Vol 2, e139.
5. Komulainen, K., Alanne, M., Auro, K., Kilpikari, R., Pajukanta, P., Saarela, J. *et al*., (2006), 'Risk Alleles of USF1 Gene Predict Cardiovascular Disease of Women in Two Prospective Studies,' *PLoS Genet* Vol 2, e69.
6. Krig S.R., Jin V.X., Bieda M.C., O'Geen H., Yaswen P., Green R., Farnham P.J. (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., Gerstein M., Weissman S., Snyder M. (2005), 'Global changes in STAT target selection and transcription regulation upon interferon treatments,' *Genes Dev*, Vol 19, pp 2953-2968.
8. MacIsaac, K.D., Fraenkel, E. (2006), 'Practical strategies for discovering regulatory DNA sequence motifs,' *PLoS Comput Biol*, Vol 2, 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., Miller, W., Hardison, R.C. (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., 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., Frommer, M. (1987), 'CpG islands in vertebrate genomes,' *Mol Biol*, Vol 196, pp 261-282.
14. Felsenfeld, G., Groudine, M. (2003), 'Controlling the double helix,' *Nature*, Vol 421, pp 448-453.

15. Corre, S., 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., Molloy, P.L. (1994), 'Base preferences for DNA binding by the bHLH-Zip protein USF: effects of MgCl2 on specificity and comparison with binding of Myc family members,' *Nucleic Acids Res*, Vol 22, pp 2801-2810.

17. Rada-Iglesias, A., Wallerman, O., Koch, C., Ameur, A., Enroth, S., Clelland, G. *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., Hinrichs, A., Lu, Y.T. *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., Pringle, T.H., Zahler, A.M., Haussler, D. (2002), 'The Human Genome Browser at UCSC', *Genome Res*, Vol 12, pp 996-1006.

22. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K. *et al*. (2005), 'Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes,' *Genome Res*, Vol 15, pp 1034-1050.

23. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z. *et al*., (2008), 'High-resolution mapping and characterization of open chromatin across the genome,' *Cell,* Vol, 132, pp 311-322.

24. TRANSFAC [http://www.gene-regulation.com/pub/databases.html].

25. Hertz, G.Z., Stormo, G.D., (1999), 'Identifying DNA and protein patterns with statistically significant alignments of multiple sequences,' *Bioinformatic,* Vol 15, pp 563-577.

26. Minka, T. (2003), 'A comparison of numerical optimizers for logistic regression,' Department of Statistics, Carnegie Mellon University.

27. Norusis, M. (2004), 'SPSS 13.0 Statistical Procedures Companion', Upper Saddle River, NJ, Prentice Hall, In.

28. Jiang, C., Xuan. Z., Zhao, F., 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., Koistinen, H.A., Pajukanta, P. *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., Sekinger, E.A., Kampa, D. *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., Zhao, Y., Zeng, T. *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., Noureddine, M., Maready, M. *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., Iversen, E.S., Dong, C, *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., Eriksson, P., Lenhard B, (2008), 'In silico detection of sequence variations modifying transcriptional regulation,' *PLoS Comput Biol*, Vol 4, e5.

2.9 Figures

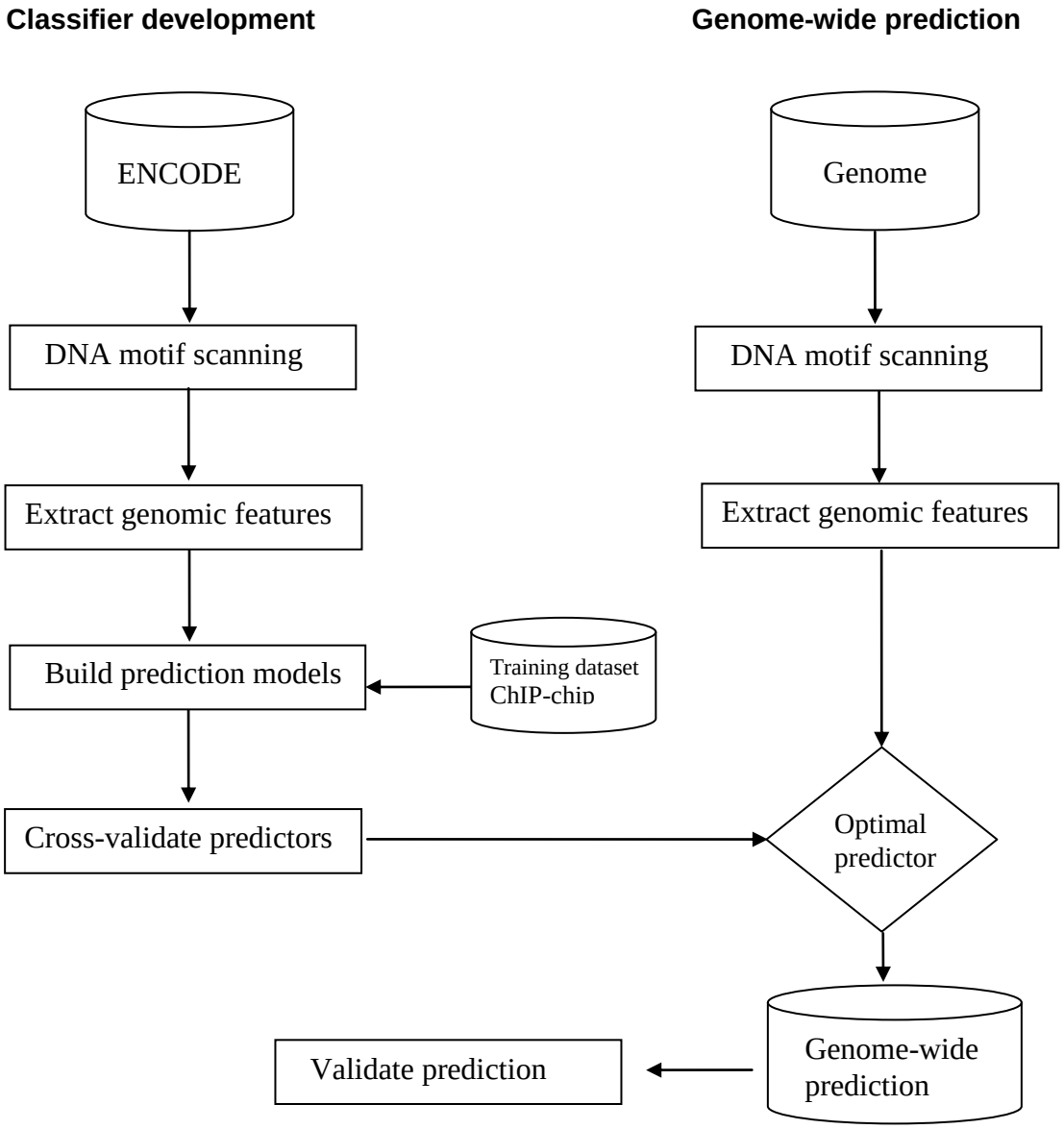


Figure 2.1 - Schematic representation of USF1 binding site prediction

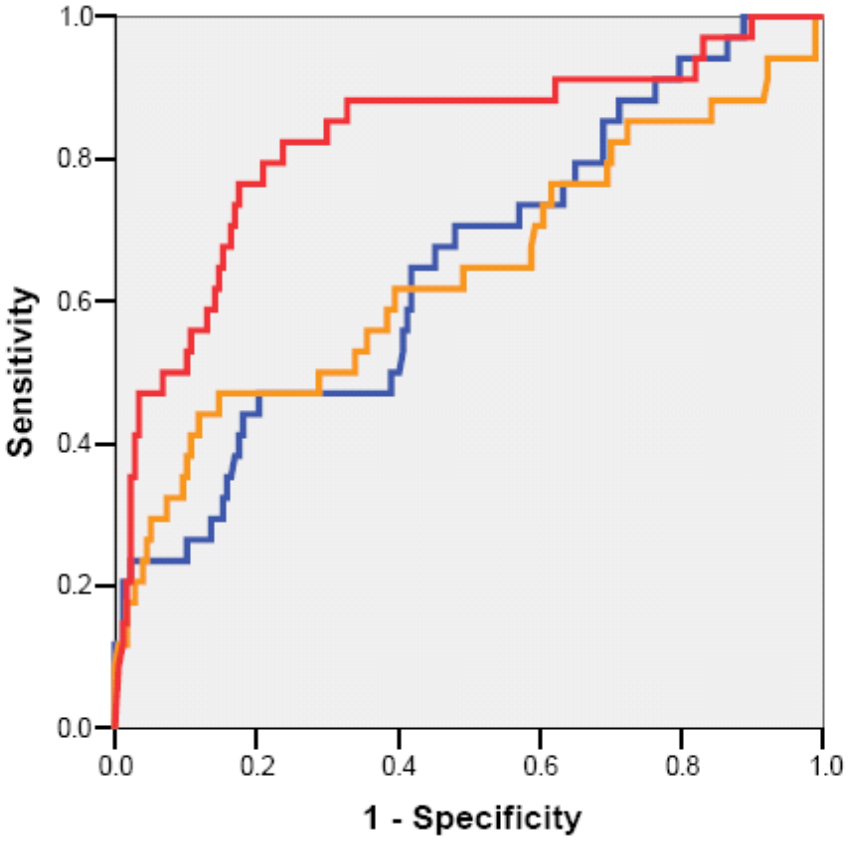


Figure 2.2 - Receiver operator characteristic curve (ROC) of USF1 prediction models

The curves generated by the SPSS package [27] with different colors indicate the sensitivity and specificity of three different prediction models. The sensitivity and specificity were calculated from LOLO cross-validation (see Methods). All features model used all five genomic features (PhastCons8, MostCons8, RP5, CpG, DNaseI HS) and PWM score, and selected features model only included three genomic features (PhastCons8, RP5, DNaseI HS) and PWM score.

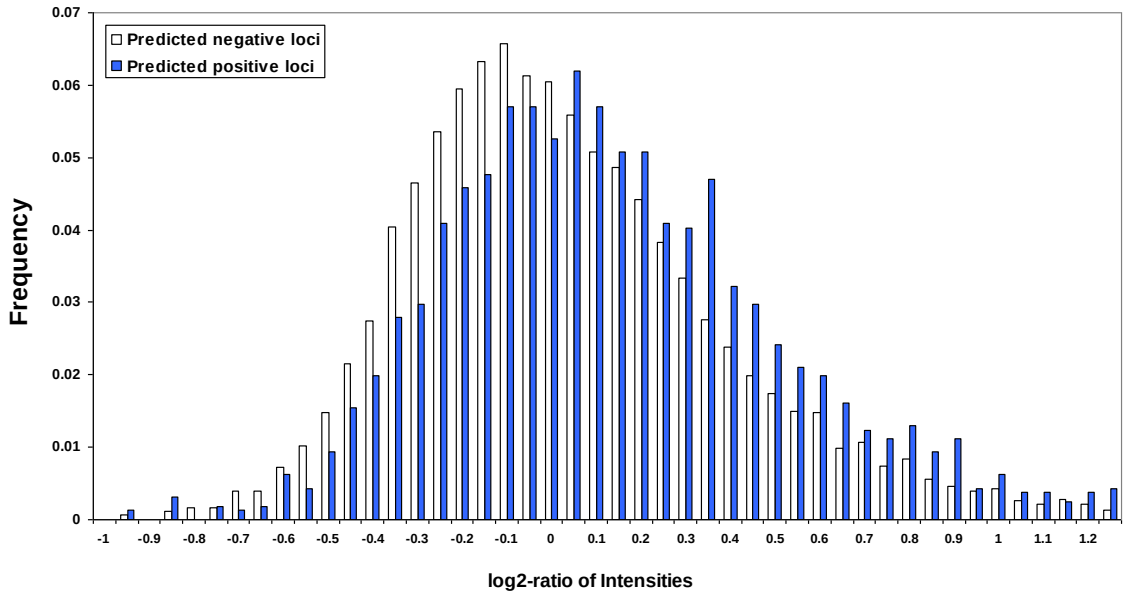


Figure 2.3 - Prediction evaluation with USF1 ChIP-chip results from the ENCODE regions

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 each 0.1 intensities $\log_2$-ratio interval.

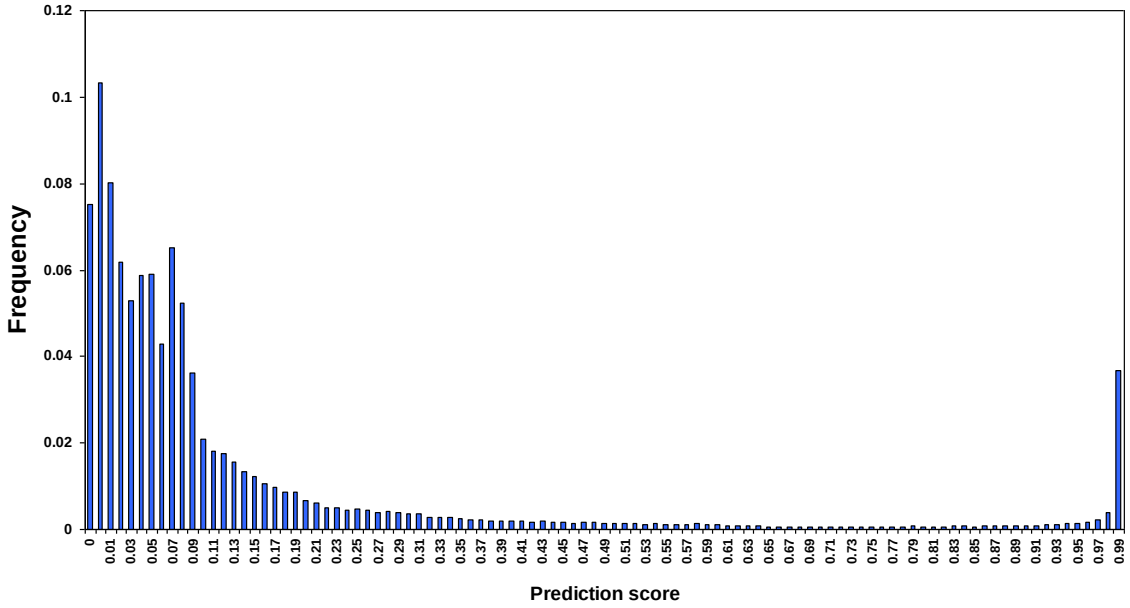


Figure 2.4 - Prediction scores distribution of potential USF1-BSs

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.

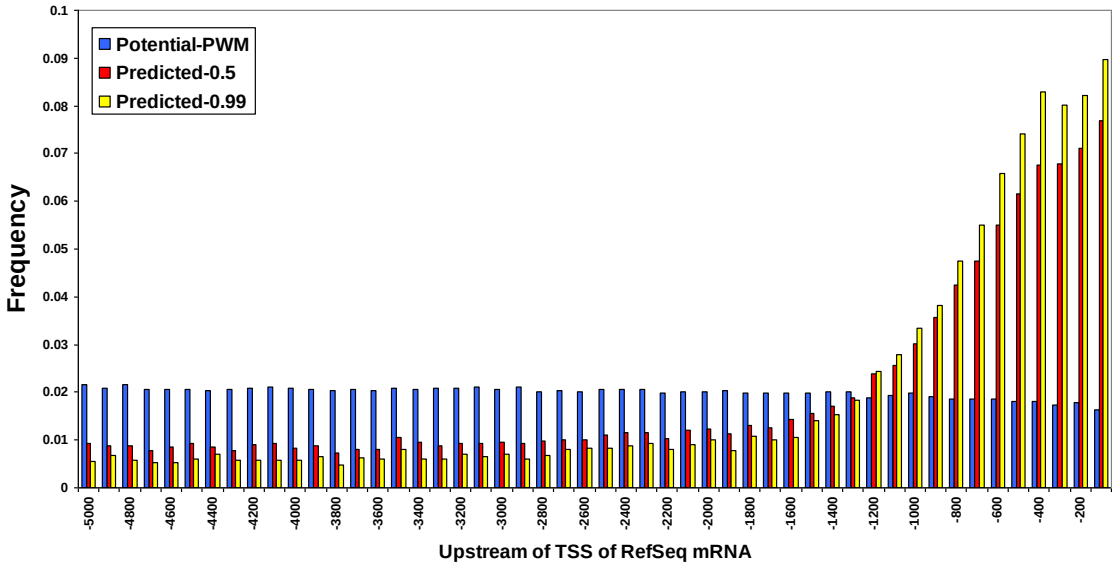


Figure 2.5 - Location distribution of USF1-BSs

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 stringent prediction threshold (0.99). The data are represented as histograms of frequency at each 100 bp interval.

2.10 Tables

Table 2.1 - Description of the five genomic features used for USF1-BS prediction method development

Name	Description	Score range
PhastCons8[1,2]	Conservation score across 8 species (Human/chimp/mouse/rat/dog/chicken/fugu/zebrafish)	[-35, 0]
MostCons8[1]	Conserved region across 8 species (Human/chimp/mouse/rat/dog/chicken/fugu/zebrafish)	[0, 1000]
RP5[1]	Regulatory potential across 5 species (Human/chimp/mouse/rat/dog)	[-0.1, 0.9]
CpG[1,3]	CpG island, CG dinucleotide rich regions	[0.5, 1.6)
DNaseI HS[4]	Regions hypersensitive to DNaseI cleavage within human CD4+ cell	[0, 17)

[1] Downloaded from the UCSC Genome Browser [19, 20, 21].

[2] Base-10 logarithm of the product of base-by-base conservation score within 10-bp USF1-BS.

[3] Used center position of 10 bp USF1-BS to define overlapping CpG island score.

[4] Published results [23].

Table 2.2 - Comparison of USF1 binding sites prediction models

Variables in the model	AUC	Std. 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

The threshold of the score to define a predicted USF1-BS was 0.5. Sensitivity was the proportion of the correctly predicted true positive loci, whereas specificity was the proportion of the true negative loci predicted as negative loci. The area under the curve (AUC) was calculated from LOLO cross-validation (see Methods).

Table 2.3 - Validation of 20 robust USF1 target genes

We used 20 robust USF1 target genes obtained from TRED database [28] and reported from 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	APOA2	Yes
2	ABCA1	Yes
3	ACACA	No
4	APOE	Yes
5	BRCA2	Yes
6	CCNB1	Yes
7	CYP19	Yes
8	CYP1A1	Yes
9	EFP	Yes
10	FMR1	Yes
11	FSHR	Yes
12	GCK	Yes
13	GHRL	No
14	HOXB4	Yes
15	HOXB7	Yes
16	hTERT	Yes
17	PF4	No
18	PIGR	No
19	PTPN6	Yes
20	SERPINE1	Yes

2.11 Additional files

Additional file 2.1 – Kernel logistic regression

Suppose we have a training set $\{x_i, y_i\}_{i=1}^N$ of input samples x_i where $x_i \in R^d$ and corresponding targets y_i. x_i denotes each input data point, which is mapped to space based on the vector of various features.

Considering a binary classification problem with targets $y_i \in \{-1,1\}$. The kernel logistic regression model is

$$p(y \mid x, w) = \sigma(yw^T\phi(x)) = \frac{1}{1+\exp(-yw^T\phi(x))} \qquad (1)$$

where $\sigma(x) = 1/(1+\exp(-x))$ is the logistic sigmoid function, $\phi(x) = [1, k(x_1, x), \cdots, k(x_N, x)]^T$ is a kernel mapping function that maps the original input space to the high-dimensional feature space.

One of the most common choices for kernel function is the radial basis function (RBF):

$$k(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \qquad (2)$$

Given a training data set $\{x_i, y_i\}_{i=1}^N$, the goal of logistic regression training is to estimate the parameter vector w to maximize the following penalized log likelihood function:

$$l(w) = -\sum_{i=1}^{N} \log(1+\exp(-y_i w^T \phi(x_i))) - \frac{\lambda}{2} w^T w \qquad (3)$$

where the term $\lambda / 2w^T w$ is used to avoid over-fitting the training data. Since (3) is a convex function that has a unique optima, the general local optimization methods guarantee to find the global solution. In the experiments, we use the conjugate gradient method [Tom Minka] to optimize (3).

Additional file 2.2 – Comparison of USF1 binding site prediction methods

Variables in the model	AUC	Std. 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
MostCons8	0.115	0.051	0.017	0.213	0.029	0.966
CpG	0.289	0.077	0.138	0.439	0.206	0.972
DNaseI HS	0.483	0.083	0.390	0.716	0.235	0.960
PWM, PhastCons8	0.661	0.051	0.561	0.761	0.382	0.819
PWM, RP5	0.746	0.046	0.657	0.836	0.235	0.977
PWM, MostCons8	0.461	0.060	0.343	0.579	0.088	0.949
PWM, CpG	0.694	0.054	0.589	0.799	0.353	0.972
PWM, DNaseI HS	0.690	0.057	0.578	0.802	0.382	0.960
PWM, PhastCons8, RP5	0.710	0.051	0.611	0.81	0.353	0.864
PWM, PhastCons8, DNaseI HS	0.810	0.045	0.723	0.899	0.500	0.881
PWM, RP5, DNaseI HS	0.791	0.044	0.705	0.877	0.382	0.960
PWM, PhastCons8, RP5, DNaseI HS	0.827	0.044	0.740	0.913	0.559	0.876
All features	0.639	0.060	0.522	0.756	0.382	0.898

The threshold of being a predicted USF1-BS was 0.5.
Sensitivity, specificity and AUC were defined in the Method.

Additional file 2.3 – 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. We only show a subset here as an example.

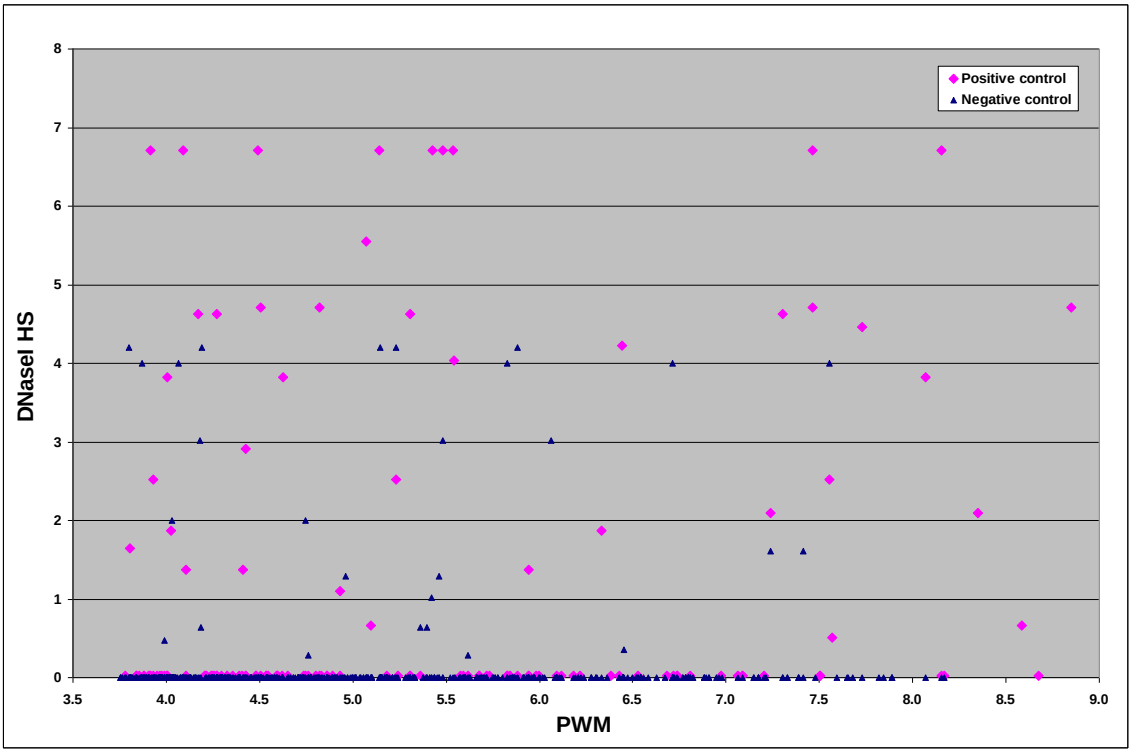
RefSeq mRNA	Gene	Location[1]	USF1-BS	Prediction score	chr[2]	start	end	SNP
NM_198943	MGC52000	-3508	ATCCCCTGTC	0.999	1	18,254	18,263	
NM_198943	MGC52000	-4014	AGCTCCTGGT	1	1	18,760	18,769	
NM_024796	FLJ22639	-284	CTCACGCGCC	0.9993	1	803,025	803,034	
NM_024796	FLJ22639	-551	AACAGCTGCC	1	1	803,292	803,301	
NM_024796	FLJ22639	-604	CTCAGGTGCG	1	1	803,345	803,354	
NM_198317	KLHL17	-1030	GGCACGCGCC	0.9948	1	935,079	935,088	
NM_021170	HES4	-796	GCCTGGTGAC	0.9988	1	976,325	976,334	
NM_148902	TNFRSF18	-1554	GTCAGCTGCG	1	1	1,183,558	1,183,567	
NM_004195	TNFRSF18	-1973	GACACGGGGC	1	1	1,183,977	1,183,986	RS3753349
NM_148902	TNFRSF18	-2023	GCCCCGTGGG	0.9993	1	1,184,027	1,184,036	
NM_004195	TNFRSF18	-2222	GCCCCGTGGG	0.9963	1	1,184,226	1,184,235	
NM_148902	TNFRSF18	-2543	GCCACGCGCC	0.9995	1	1,184,547	1,184,556	
NM_153339	PUSL1	-2820	GCCAGGCGGT	1	1	1,281,096	1,281,105	
NM_017871	FLJ20542	-457	CTCACCTGGG	1	1	1,300,401	1,300,410	
NM_181870	DVL1	-76	CCCCCGTGAC	1	1	1,324,751	1,324,760	
NM_004421	DVL1	-619	ACCCCGTGGG	1	1	1,325,026	1,325,035	

[1] The upstream of the TSS of RefSeq mRNA.

[2] Chromosome.

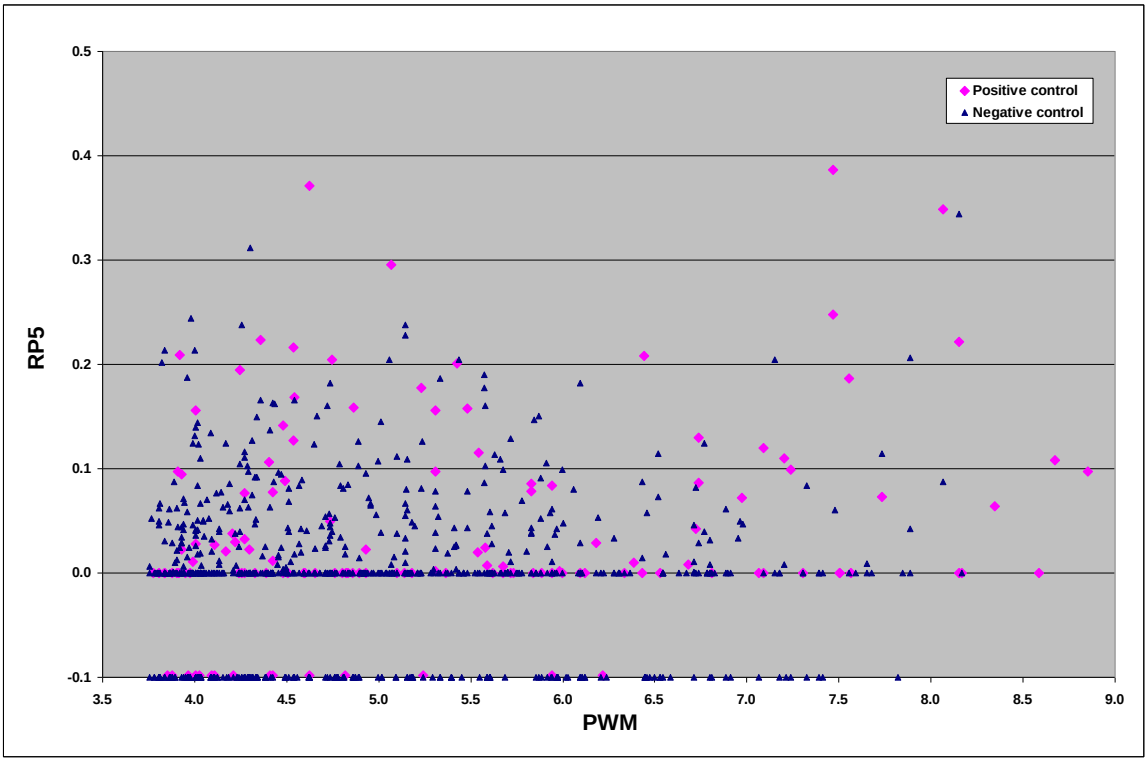
Additional file 2.4 – Distribution of PWM and DNaseI HS scores in the training dataset (correlation coefficient = 0.121)

Our training dataset includes 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.



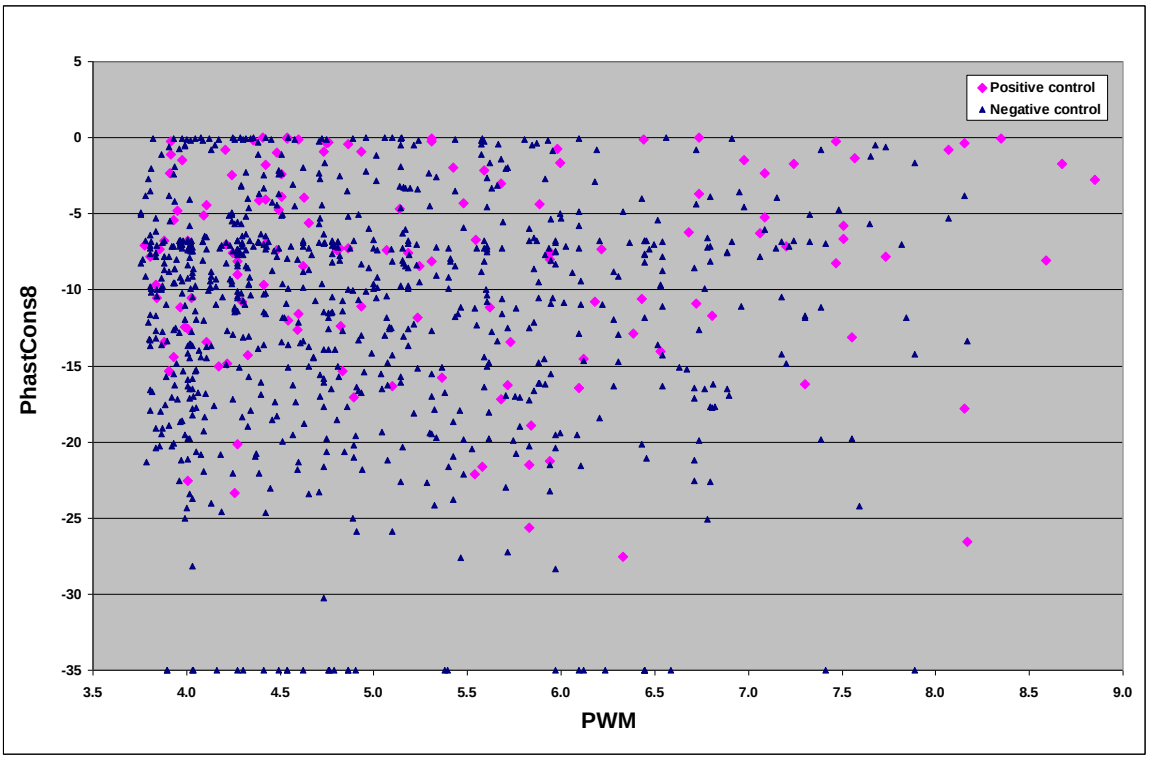
Additional file 2.5 - Distribution of PWM and RP5 scores in the training dataset (correlation coefficient = 0.117)

Our training dataset includes 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.



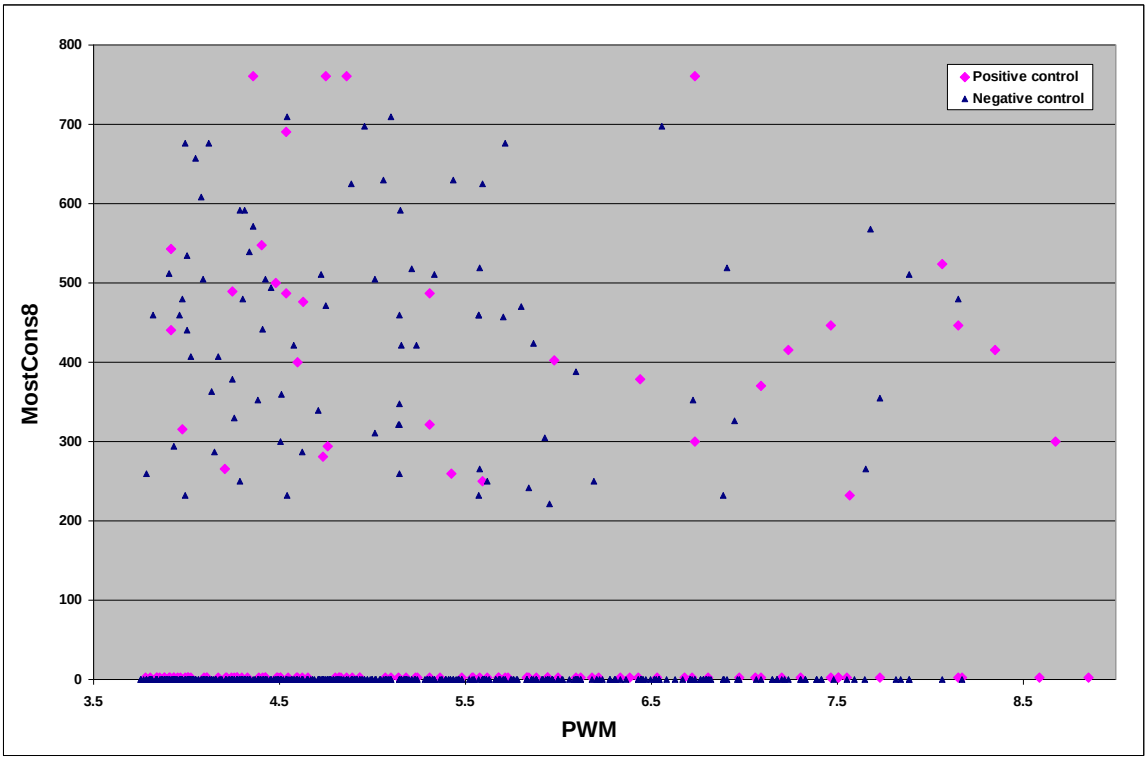
Additional file 2.6 – Distribution of PWM and PhastCons8 scores in the training dataset (correlation coefficient = -0.007)

Our training dataset includes 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.



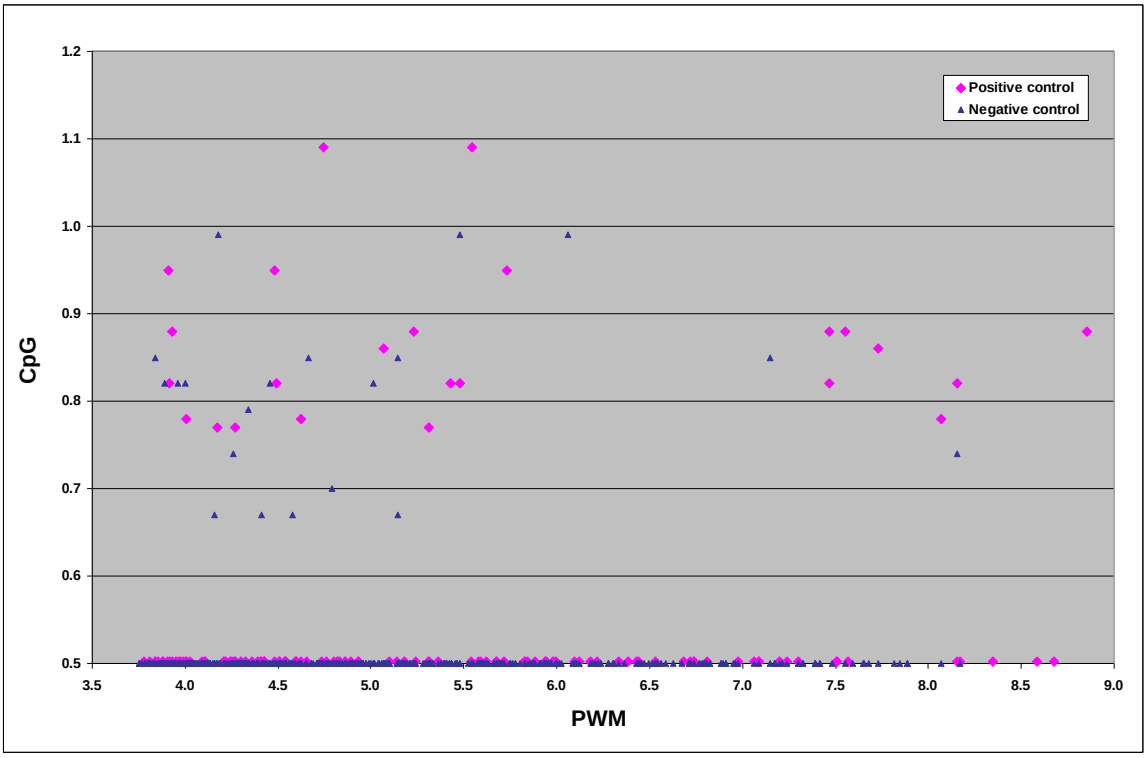
Additional file 2.7 – Distribution of PWM and MostCons8 scores in the training dataset scores (correlation coefficient = 0.060)

Our training dataset includes 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.



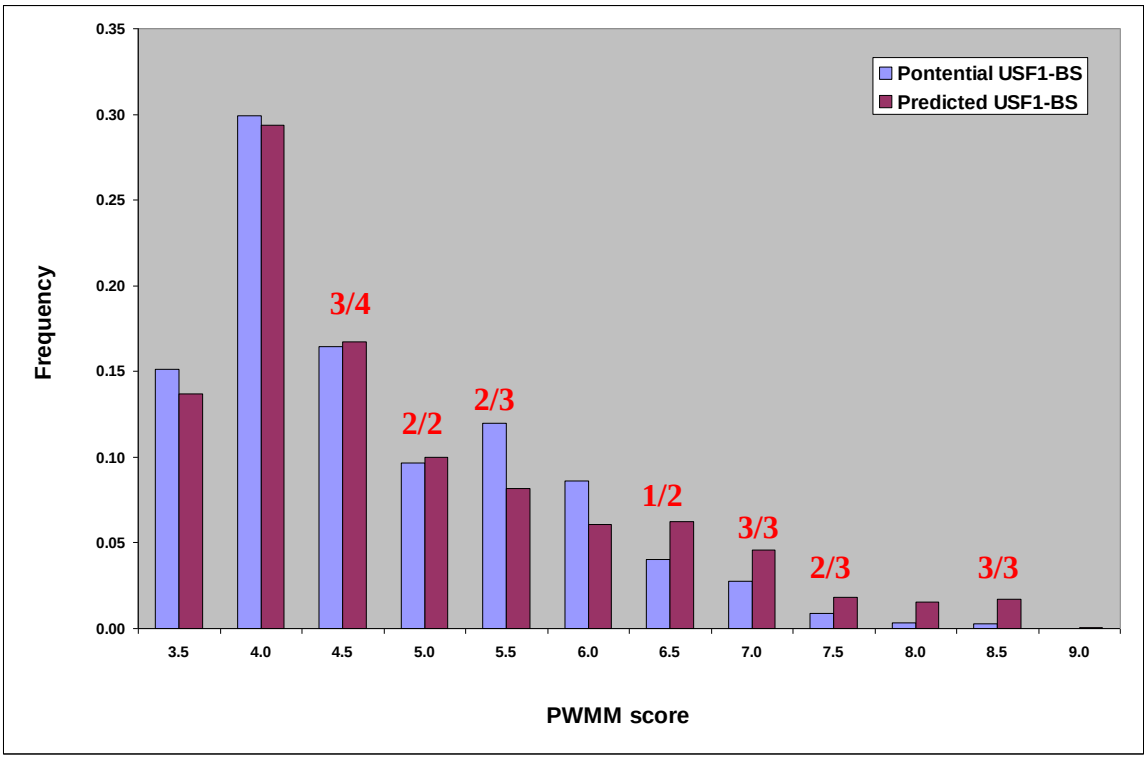
Additional file 2.8 – Distribution of PWM and CpG scores in the training dataset scores (correlation coefficient = 0.064)

Our training dataset includes 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 2.9 – 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, twenty experimentally identified USF1 target genes had been used to validate our prediction. The numbers on the top of each 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 of 4.5 to 5, and three of these genes are correctly identified by the optimal kernel-based prediction method.



Additional file 2.10 – CAD candidate genes identified by the "genomic convergence" approach

A previously published study of gene expression signatures from human aortas identified 229 genes to be 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 UFS1 target genes to 5,801 to be used as candidate genes for further analysis. 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 aorta. The prediction score of each USF1-BS within these genes can be found in Additional file 2.3.

No.	Genbank	Gene	RefSeq
1	J04430	ACP5	NM_001611
2	M12529	APOE	NM_000041
3	AB016811	ARL7	NM_005737
4	AF055024	ASB1	NM_016114
5	L35249	ATP6V1B2	NM_001693
6	D49400	ATP6V1F	NM_004231
7	AL050008	BRMS1	NM_015399
8	U35451	CBX1	NM_006807
9	AL031846	CBX7	NM_175709
10	AF014958	CCRL2	NM_003965
11	M16336	CD2	NM_001767
12	U10906	CDKN1B	NM_004064
13	U46692	CSTB	NM_000100
14	M63138	CTSD	NM_001909
15	L06797	CXCR4	NM_003467
16	M21186	CYBA	NM_000101
17	X81109	BCAP31	NM_005745
18	M63193	ECGF1	NM_001953

Additional file 2.10 Continued

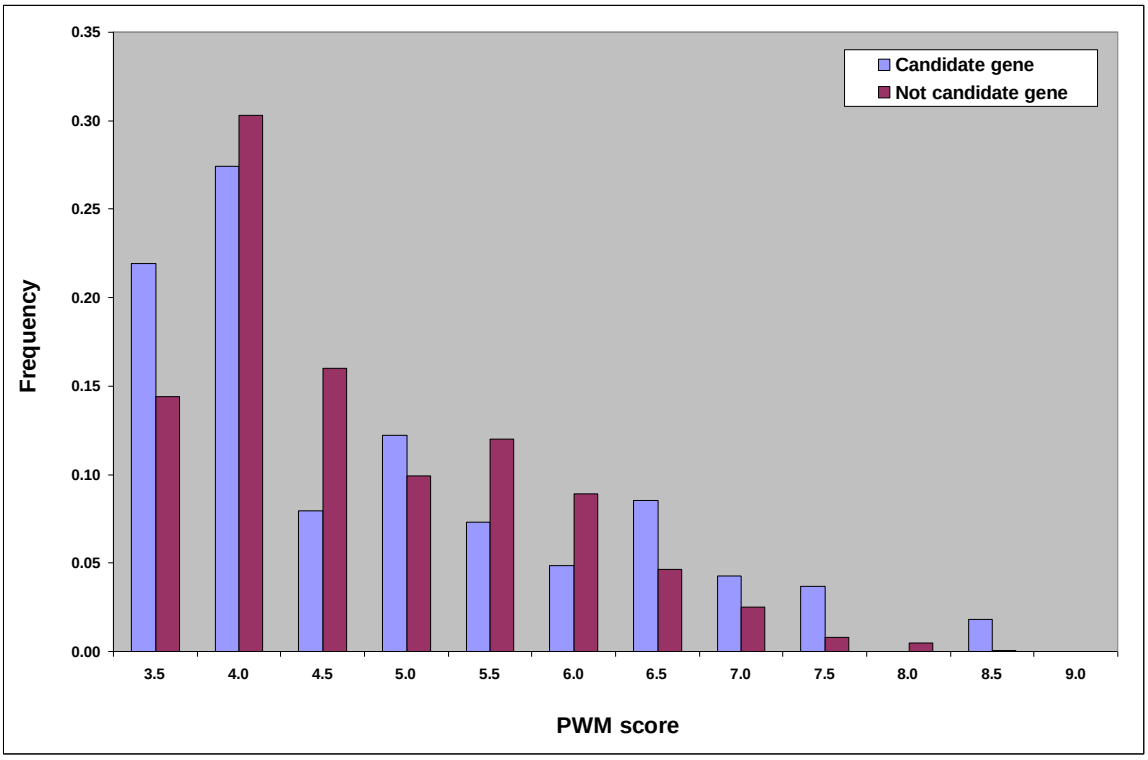
19	M80634	FGFR2	NM_181356
20	AL050139	FLJ13910	NM_022780
21	AC005546	FLJ20244	NM_017722
22	L49169	FOSB	NM_006732
23	U59831	FOXD1	NM_004472
24	X63657	FVT1	NM_002035
25	M68891	GATA2	NM_032638
26	L15388	GRK5	NM_018135
27	AF055008	GRN	NM_002087
28	D64142	H1FX	NM_006026
29	X96752	HADHSC	NM_005327
30	X16663	HCLS1	NM_005335
31	X62744	HLA-DMA	NM_006120
32	U15085	HLA-DMB	NM_002118
33	J00194	HLA-DRA	NM_019111
34	M32578	HLA-DRB1	NM_002124
35	M74297	HOXA4	NM_002141
36	X16665	HOXB2	NM_002145
37	M16937	HOXB7	NM_004502
38	X51757	HSPA6	NM_002155
39	J03909	IFI30	NM_006332
40	U00672	IL10RA	NM_001558
41	D26350	ITPR2	NM_002223
42	AF044253	KCNAB2	NM_003636
43	AF022797	KCNN4	NM_002250
44	D87434	KIAA0247	NM_014734
45	U51240	LAPTM5	NM_006762
46	AB019527	LDOC1	NM_012317
47	D86961	LHFPL2	NM_005779
48	AF055581	LNK	NM_005475
49	AF062075	LPXN	NM_004811
50	M33552	LSP1	NM_002339
51	Y14768	LTA	NM_000595
52	AC005546	LYL1	NM_005583
53	M64571	MAP4	NM_002375

Additional file 2.10 Continued

54	AF004709	MAPK13	NM_002754
55	U09578	MAPKAPK3	NM_004635
56	J05070	MMP9	NM_004994
57	M64925	MPP1	NM_002436
58	X75918	NR4A2	NM_006186
59	U00952	PBXIP1	NM_020524
60	U89606	PDXK	NM_003681
61	L10678	PFN2	NM_002628
62	AB007972	PPP1R12B	NM_002481
63	X71874	PSMB10	NM_002801
64	S59049	RGS1	NM_002922
65	X91809	RGS19	NM_005873
66	D87074	RIMS3	NM_014747
67	X90976	RUNX1	NM_0010018
68	S59184	RYK	NM_002958
69	X79204	ATXN1	NM_000332
70	AF051323	SCAP2	NM_003930
71	Z22555	SCARB1	NM_005505
72	AJ000534	SGCE	NM_003919
73	D89077	SLA	NM_006748
74	U81800	SLC16A3	NM_004207
75	AF030409	SLC9A6	NM_006359
76	U78095	SPINT2	NM_021102
77	AF051851	SVIL	NM_003174
78	AF095791	TACC2	NM_012120
79	U13991	TAF10	NM_006284
80	AF029750	TAPBP	NM_003190
81	U45285	TCIRG1	NM_006053
82	AL050262	TLR1	NM_003263
83	M32315	TNFRSF1B	NM_001066
84	AJ006973	TOM1	NM_005488
85	U58334	TP53BP2	NM_005426
86	AF022789	USP12	NM_182488
87	AF084481	WFS1	NM_006005

Additional file 2.11 – PWM score distribution of USF1-BSs within CAD candidate genes

Previously published study of gene expression signatures from human aortas identified 229 genes to be 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 aorta. 142 genes were excluded because they did not have predicted USF1-BSs.



CHAPTER 3

ZNF217 is associated with coronary artery disease in multiple samples

3.1 Abstract

Zinc finger protein 217 (ZNF217) locates to the linkage peak on chromosome 20q13 from our previous family based linkage study (GENECARD) of early-onset coronary artery disease (CAD). Additionally, a study of gene expression signatures from human aortas identified ZNF217 to be differentially expressed in aortas with and without atherosclerosis. ZNF217 is known to repress the transcription of many genes and is associated with cell proliferation, survival, and invasion. We investigated the role of ZNF217 as a candidate gene for CAD in three independent CAD samples: GENECARD families (n = 880 families, 713 probands from US families), CATHGEN case-control samples (cases = 905 and controls = 405), and human donor aorta samples (n = 205). We selected 15 single nucleotide polymorphisms (SNPs) for genotyping based on their linkage disequilibrium (LD), function and tagging potential. We identified three SNPs (rs2741372, rs2766671 and rs1056948) having association with CAD in all these three datasets. We performed non-parametric regression analysis of aorta disease burden (the rankings of the Sudan IV and raised lesion) with ZNF217 expression, age, sex, and race. Raised lesion was positively correlated with ZNF217 expression ($p = 0.0270$). In addition, age and sex (male) increased ZNF217 expression by 0.2277 and 0.010 normalized units respectively. To further test the potential functional impact of the SNPs, we evaluated allele-specific expression in the aorta samples adjusted for sex, race and age. The results indicated that three SNPs (rs16998248/A342A, rs35720349/I548T, and rs6097488) were significantly associated with the level of ZNF217 expression ($p = 0.0045 – 0.042$). These combined

results from independent genome-wide linkage, association, and gene expression studies suggest that ZNF217 is a novel susceptibility gene for CAD. We are currently studying the gene's specific contribution to heart disease through the identification of its target genes.

3.2 Introduction

Coronary artery disease (CAD) is the leading cause of death in the western world (Rosamond et al., 2007). As for other complex human diseases, CAD has both genetic and environmental risk factors. Recent technological advances have enabled the undertaking of large-scale genome-wide association studies (GWAS) in order to find genomic regions associated with disease. Two independent GWASs had focused on identifying the regions of the genome showing evidence of association with CAD (The Wellcome Trust Case Control Consortium, 2007; Samani et al., 2007). Seven loci were identified by each study. The only common region was on 9p21, which included two well-annotated genes, CDKN2A and CDKN2B. Despite significant study, the role of this region in changing CAD susceptibility is unknown. And this 9p21 locus may be only part of the regulatory network associated with CAD.

Commonly transcription factors (TFs) play a role in susceptibility to complex diseases (Weedon et al., 2007). Numerous publications have identified single nucleotide polymorphisms (SNPs) in TFs that are significantly associated with CAD, including GATA2, USF1, and MEF2A (Connelly et al., 2006; Komulainen et al., 2006; Mohlke et al., 2005; Pajukanta et al., 2004; Wang et al., 2003). Zinc finger protein 217 (ZNF217) is

known to repress the transcription of many genes and is associated with cell proliferation, survival, and invasiveness of cancer cells (Quinlan et al., 2007). The ZNF217 region is selectively amplified during progression of several cancers. These results suggest that ZNF217 may give tumor cells selective advantage by interfering with normal regulation of cell growth, cell death, differentiation, and DNA repair (Krig et al., 2007).

ZNF217 locates to a modest linkage peak on chromosome 20q13 from a previous family based linkage study (GENECARD) of early-onset CAD (Hauser et al., 2004). Additionally, a study of gene expression signatures from human aortas identified ZNF217 among 229 genes to be differentially expressed in aortas with and without atherosclerosis (Seo et al., 2004). Given the convergence of the results, we hypothesized that as a TF with multiple target genes ZNF217 may be associated with CAD. Therefore, we investigated genetic variation within the gene in three independent CAD samples: a case-control sample from the Duke CATHGEN cohort, GENECARD families, and human donor aorta samples, and evaluated the role of ZNF217 as a candidate gene for CAD/atherosclerosis. The results suggest that ZNF217 is a novel susceptibility gene for CAD. We are currently studying the gene's specific contribution to heart disease through the identification of its target genes and regulatory network important for CAD.

3.3 Results

3.3.1 ZNF217 SNP selection and genotyping

SNPs for genotyping were selected using SNPselector (Xu et al., 2005). We covered the region from 5 kb upstream to 5 kb downstream of ZNF217 to include potential regulatory elements. Tagging SNPs were selected for unique linkage disequilibrium (LD) bins based on HapMap (http://www.hapmap.org/) and Perlegen (http://genome.perlegen.com/) databases using the criteria $r^2 > 0.7$ and minor allele frequency (MAF) > 5% in the Caucasian population. Additional SNPs were preferentially included based on being a coding SNP, regulatory potential and also to cover gaps greater than 2 kb.

We genotyped a total of 15 SNPs in our association studies (Table 3.2). The pairwise LD plot was calculated based on Caucasian control samples from CATHGEN (Figure 3.1). We also measured the pairwise LD in Caucasian affected samples. The LD patterns in these two groups were very similar (unpublished data). By design, the majority of the SNPs tested were not highly correlated with each other ($r^2 <= 0.7$). It was expected that these SNPs captured most of the common genetic variation of ZNF217.

3.3.2 Single –marker association of ZNF217 SNP in CATHGEN case-control sample

Genotyping was performed in the CATHGEN case-control sample. The clinical characteristics in CATHGEN participants (n = 905) and unaffected controls (n = 405) are presented in Table 3.1. As expected there are differences in the known CAD risk factors

for cases and controls. Thus we adjusted for sex, age, and cardiovascular risk factors stratified by Caucasian vs. African American in order to control for these differences in our regression model. We identified significant associations ($p < 0.05$) in the young affected (YA, n = 642) and old affected (OA, n = 263) samples separately in addition to all affected (AA) samples, which was the combined OA and YA groups. Within the Caucasian group, one SNP (rs16998248) was associated in the YA subgroup, four SNPs (rs2741372, rs34323943, rs2766671, and rs1056948) in the OA subgroup, and two SNPs (rs16998248 and rs2741372) in the AA samples (Table 3.3). None of these associated SNPs ($p = 0.0256 – 0.0426$) withstood the significance threshold value of the stringent Bonferroni correction for multiple comparisons ($\alpha = 0.05$, n = 15, $p \leq 0.003$). These five associated SNPs are located in exon 2 (rs16998248 coding A342A), intron 2 (rs2741372), exon 4 (rs34323943 coding G889D), and 3' UTR of the gene (rs2766671 and rs1056948). Two of these SNPs are in the same LD bin, rs2766671 and rs1056948 ($r^2 = 0.98$). The significant SNPs identified by using the combined Caucasian and African American were similar to the ones identified from Caucasian group only (data not shown).

3.3.3 Single –marker family-based association in GENECARD

We genotyped ten SNPs in the GENECARD sample to explore SNPs with significant association in CATHGEN and aorta samples (Table 3.4). We prepared the case dataset using US GENECARD probands (age of CAD onset (AOO) men < 51 and women < 56, n = 713) to compare with CATHGEN controls. We identified one SNP (rs2741372)

with significant association (Table 3.4). In addition, we performed family-based association analysis using the association in the presence of linkage test (APL) in order to take the advantage of the large number of affected sibling pairs in the GENECARD sample, and to adjust for the correlation between transmission of parental SNP alleles to multiple affected offspring due to linkage (Martin et al., 2003; Chung et al., 2007). APL analysis of these ten SNPs identified two additional SNPs (rs2766671 and rs1056948) besides rs2741372 with significant associations ($p = 0.0102 – 0.0149$) with early-onset CAD (Table 3.4). We noted that none of these three SNPs would meet the Bonferroni correction threshold value of significance ($\alpha = 0.05$, $n = 15$, $p \leq 0.003$). However, all three SNPs showed the same direction of effect as the association identified in the CATHGEN sample. Thus the case-control association studies and the family-based association study provide consistent results with respect to associated SNPs and direction of effect.

3.3.4 Aorta association analysis

Human aorta samples were collected from donor hearts, and categorized by atherosclerotic disease burden (see Materials and Methods). Using logistic regression analysis, we identified three SNPs (rs2741372, rs2766671 and rs1056948) significantly associated with atherosclerotic disease burden ($p = 0.0082$ - 0.0235) (Table 3.5). After stratifying the aorta samples based on the amount of aorta atherosclerosis, the significant association appears to be driven by the early atherosclerotic lesions measured by Sudan IV staining according to non-parametric regression analysis ($p = 0.0072$ - 0.0238) (Table 3.5).

All three SNPs had ORs less than 1 and negative effects of Sudan IV in the model (Table 3.5). This result indicated that the major alleles of these SNPS are potential susceptibility alleles.

Aorta expression profiling was performed using Affymetrix GeneChip U95Av2. ZNF217 expression represented by tag 32034_at was used as phenotype in the regression model. We performed non-parametric regression analysis of disease burden (the rankings of the Sudan IV and raised lesion) with ZNF217 expression, age, sex, and race. Age positively correlated with CAD in both Sudan IV ($p = 0.0040$) and raised lesion ($p < 0.0001$).

To further test the potential functional impact (cis-effects) of the SNPs, we evaluated allele-specific ZNF217 expression in the aorta samples adjusted for sex, race and age. A mixed model was used to account for repeated expression measures (i.e., multiple expression values per sample). The analysis result indicated that three SNPs (rs16998248/A342A, rs35720349/I548T, and rs6097488) were significantly associated with the level of ZNF217 expression ($p = 0.0045 – 0.042$) (Table 3.5). Each copy of the minor alleles of two coding SNPs (rs16998248 and rs35720349) reduced ZNF217 expression by 0.25 and 1.07 normalized units respectively. Rs16998248 was also significantly associated with CAD in the CATHGEN sample, and the minor allele was protective (odds ratios (OR) = 0.559 and $p = 0.0426$). The minor allele of rs6097488 in intron 4 increased ZNF217 expression by 0.169 normalized units. Raised lesion rank was also positively correlated to ZNF217 expression with p value 0.0270. In addition, age and

sex (male) increased ZNF217 expression by 0.2277 and 0.010 normalized units respectively.

3.3.5 Integration of three datasets

We identified three SNPs (rs2741372, rs2766671 and rs1056948) having association with CAD in all three independent datasets: CATHGEN, GENECARD, and aorta samples. Although none of these three associated SNPs met the Bonferroni threshold value of significance ($\alpha = 0.05$, $n = 15$, $p \leq 0.003$), the associations were consistent in all three samples and the directions of association identified by the odds ratios were the same too. The minor allele of all three SNPS are protective alleles (OR < 1) in three datasets (Table 3.3, 3.4 and 3.5). These three SNPs are located in intron 2 (rs2741372) and 3' UTR (rs2766671 and rs1056948) of ZNF217. Rs2766671 and rs1056948 are in the same LD bin ($r2 = 0.98$) (Figure 3.1 and Table 3.2). While there were a number of SNPs that were significantly associated with ZNF217 expression, the three consistent SNPs were not significantly associated with expression level ($p = 0.2807 – 0.4584$).

3.4 Discussion

In this genetic study of ZNF217, we genotyped fifteen SNPs in three independent CAD samples, and analyzed ZNF217 expression in human donor aorta samples. We identified ZNF217 as a susceptibility gene candidate for CAD, and captured the genetic variations of ZNF217 associated with cardiovascular disease phenotypes.

We took advantage of three independent CAD samples available for us. CATHGEN was based on clinical presentations, and included a large size case-control cohort with well documented cardiovascular disease. GENECARD used a family-based cohort with documented family history of early age of CAD onset. Human donor aorta samples allowed us to visualize the disease phenotype on the tissue/cellular level, and to compare ZNF217 expression in the aorta samples with and without atherosclerosis. These independent datasets allow us to evaluate our findings in different disease presentations including clinical outcome, tissue specificity, and gene expression. We identified three SNPs having consistent association with CAD in all three independent datasets with multiple different measures, and the directions of association identified by the odds ratios are the same too. These results strongly support ZNF217 as a novel susceptibility gene for CAD.

Our data suggest that the minor allele of these three SNPS (rs2741372, rs2766671 and rs1056948) associated with CAD are protective alleles (OR < 1), and may decrease susceptibility to developing CAD (Table 3.3, 3.4 and 3.5). Further we investigated the potential functional relevance of these three SNPs. Interestingly, we found that rs1056948 is highly conserved between multiple species (human, chimp, rhesus, rat, dog, house, and chicken), and has high regulatory potential. Additionally, rs1056948 is located near one known and three predicted polyadenylation sites according to the UCSC Genome Browser on Human (Kent et al., 2002). We hypothesize that rs1056948 may play a role in controlling tissue-specific polyadenylation of the transcript of ZNF217. Another SNP

(rs2766671) is in the same LD bin with rs1056948 (r2 = 0.98). And rs2741372 is located in intron 2 (Figure 3.1 and Table 3.2). The functional relevance of the significantly-associated intron 2 SNPs remains unknown.

In addition, we evaluated allele-specific ZNF217 expression in the aorta samples. We identified three SNPs (rs16998248/A342A, rs35720349/I548T, and rs6097488) that were significantly associated with the level of ZNF217 expression (Table 3.5). Two of them are coding SNPs, rs16998248/A342A and rs35720349/I548T. A342A is in the CoREST binding domain, and I548T is 12 amino acids away from DNA binding domain. The minor alleles of these two coding SNPs reduced ZNF217 expression. And the minor allele of rs16998248 was a protective allele for CAD in the CATHGEN sample. These SNPs may have potential functional impact (cis-effects) on CAD through changing the level of ZNF217 expression.

Chromosome 20q13 is highly amplified in human cancer. ZNF217 is a candidate oncogene on 20q13.2. The copy number and expression of ZNF217 have been intensively studied in many human cancer and cell lines. The results indicated that the selected expression of ZNF217 may cause 20q13 amplification during critical early stages of cancer progression (Quinlan et al., 2007). However, the association between ZNF217 and CAD is limited, except that ZNF217 locates to the linkage peak on 20q13 from our previous family based linkage study (GENECARD) of early-onset CAD and is differentially expressed in human aortas with and without atherosclerosis. In this study, we focused on ZNF217 and used a candidate gene approach to identify association with CAD. Our results strongly

support that ZNF217 is a novel susceptibility gene for CAD. We identified three SNPs (rs2741372, rs2766671 and rs1056948) having consistent association with CAD in all three independent datasets with multiple different disease measures. Aorta expression profiling indicated that raised lesion was also positively correlated to ZNF217 expression ($p = 0.0270$). Further allele-specific ZNF217 expression analysis identified three SNPs (rs16998248/A342A, rs35720349/I548T, and rs6097488) significantly associated with the level of ZNF217 expression ($p = 0.0045 – 0.042$) (Table 3.5). In addition, sex (male) and age were found to be positively correlated to ZNF217 expression. Based on these results, we hypothesize that ZNF217 plays a role in susceptibility to CAD. Genetic variations such as SNPs within the regulatory elements and the coding regions of the functional domain of ZNF217 may have impact on the expression and protein function of the gene. Additionally, age and sex (male) are also important confounders of CAD through influencing ZNF217 expression (Figure 3.3).

Several other TFs have been identified to be associated with cardiovascular phenotypes, such as MEF2A (myocardial infarction) and USF1 (lipid traits). Complex genetic diseases, such as cardiovascular disease, are caused by genetic risk factors, environmental effects and the interaction of them. TFs (and their cognate binding sites) ultimately influence the expression of many downstream genes in temporal, cell type, or environmental specific ways. Slight changes to the level of a TF in the cell can have a significant effect on its downstream targets. Therefore TFs are likely to be important candidates in the dissection of complex human disease. We are currently studying

ZNF217's specific contribution to cardiovascular disease through the identification of its target genes and regulatory network.

Two recent GWASs have been performed to identify association with CAD. One of them is from the Wellcome Trust Case Control Consortium, which examined ~2,000 cases of each of seven major human complex diseases including CAD and ~3,000 shared controls using the Affymetrix GeneChip 500K Mapping Array Set (The Wellcome Trust Case Control Consortium, 2007). This study included six SNPs in ZNF217, but none of them showed significant association with CAD ($p < 0.05$). The three ZNF217 SNPs (rs3748501, rs6063966 and rs2766672) in common between the GWAS and our own study were not strongly associated with CAD in any of our three sample sets. However, one of these SNPs, rs6063966, trends toward association ($p = 0.0805$). More interestingly, rs6063966 is in the same LD bin with rs34323943 ($r^2 = 0.87$), which was associated with CAD in the OA subgroup of our CATHGEN sample ($p = 0.0302$). Both of them are nonsynonymous coding SNPs in the functional region of ZNF217. Rs6063966/I739V is in the CoREST binding domain, and rs34323943/G889D is in the proline-rich region of transcription activation domain. The other two SNPs, Rs3748501 and rs2766672, are not in LD ($r^2 > 0.7$) with any of the highly associated SNPs identified in our study. Another recent GWAS from the Framingham Heart Study examined 1,345 adult participants from 310 pedigrees to identify the associations of SNPs with 987 phenotypes, including cardiovascular disease and cardiovascular risk factor. The 100K Affymetrix GeneChip Human Mapping Array Set used for genotyping in this study only includes one SNP

(rs10485444) from ZNF217. No significant association was identified between this SNP and any phenotype related to cardiovascular disease. Because of the limited coverage of these two GWAS with respect to ZNF217, we cannot make direct comparison of their results with our most significant results. This also illustrates a potential limitation of the GWAS, which may miss potentially important functional candidate genes.

In summary, our association studies of ZNF217 identified three SNPs having consistent association with cardiovascular disease in three independent samples with different phenotypes. Age and sex are also important variables influencing ZNF217 expression. The combined results from independent genome-wide linkage, association, and gene expression studies suggest that ZNF217 is a novel susceptibility gene for CAD. We are currently studying the gene's specific contribution to heart disease through the identification of its target genes and gene-gene interaction between ZNF217 and its target genes.

3.5 Materials and Methods

3.5.1 Early-onset CAD case-control sample (CATHGEN)

CATHGEN participants were recruited through the cardiac catheterization laboratories at Duke University Hospital with approval from the Duke Institutional Review Board. All participants undergoing catheterization were offered participation in the study and signed informed consent. Medical history and clinical data of the participants were collected and

stored in the Duke Information System for Cardiovascular Care database maintained at the Duke Clinical Research Institute (Fortin et al., 1995).

The total number of CATHGEN participants was 4,855. Controls and cases were chosen from this CATHGEN cohort based on amount of CAD measured by the CAD index (CADi). CADi is a numerical summary of coronary angiographic data that incorporates the extent and anatomical distribution of coronary disease (Smith et al., 1991). CADi has been shown to be a better predictor of clinical outcome than extent of CAD (Kong et al., 2002). Affected status was determined by the presence of significant CAD defined as a CADi $\geq$ 32 (Felker et al., 2002). For patients older than 55 years of age, a higher CADi threshold (CADi $\geq$ 74) was used to adjust for the higher baseline extent of CAD in this group. Medical records were reviewed to determine the AOO of CAD, i.e., the age at first documented surgical or percutaneous coronary revascularization procedure, myocardial infarction, or cardiac catheterization meeting the above-defined CADi thresholds. The CATHGEN cases were stratified into a young affected group (AOO $\leq$ 55 years), which provides a consistent comparison for the GENECARD family study. Controls were defined as $\geq$ 60 years of age, with no CAD as demonstrated by coronary angiography and no documented history of cerebrovascular or peripheral vascular disease, myocardial infarction, or interventional coronary revascularization procedures. A comparison of clinical characteristics between CATHGEN cases and unaffected CATHGEN controls is presented in Table 3.1.

3.5.2 Early-onset CAD family-based sample (GENECARD)

The Genetics of Early Onset Cardiovascular Disease (GENECARD) is a collaborative study involving investigators at the Duke Center for Human Genetics, the Duke University Center for Living, the Duke Clinical Research Institute, the Duke University Consortium for Cardiovascular Studies, and additional investigative sites of the GENECARD Study Network. The study is coordinated at Duke and includes five other international sites, and the study design has been previously described (Hauser et al., 2003). In brief, collection of families began in March 1998 and was completed in March, 2002. All study participants signed a consent form approved by the responsible institutional review board or local ethics committee.

Qualified participants were required to have medical record documentation of at least one of the following: myocardial infarction or unstable angina (acute coronary syndrome), coronary catheterization demonstrating significant disease (at least a 50% stenosis in one major epicardial coronary vessel), interventional coronary revascularization procedure (percutaneous transluminal coronary intervention or coronary artery bypass grafting), or a functional test documenting reversible myocardial ischemia with cardiac imaging. The qualifying event or procedure must have occurred at or before age 51 years in men and 56 years in women. For families to be eligible for the GENECARD study, they were required to include at least two siblings who met these diagnostic criteria and were available for sampling and data collection. For family based association studies, unaffected siblings were also collected, defined as siblings and relatives who have not been diagnosed with

CAD and are older than 55 years of age (males) or older than 60 years of age (females). This additional collection increased the sample size to 3,036. Participants came from 880 families with two or more affected sibs. Among them, there were 713 probands from the southeastern United States. The characteristics of this study group are summarized in Table 3.1 and elsewhere (Hauser et al., 2003; Hauser et al., 2004).

3.5.3 SNP selection and genotyping

The information from HapMap (http://www.hapmap.org/) and Perlegen (http://genome.perlegen.com/) was used for Tagging SNPs selection. We covered the region from 5 kb upstream to 5 kb downstream of the gene, chr20:51,612,012 – 51,648,043 (NCBI build 36), to include potential regulatory elements. A minimal set of tagging SNPs with an MAF > 5% and r^2 > 0.7 in Caucasians was chosen by SNPselector to cover the predicted LD structure in ZNF217 (Xu et al., 2005). Coding SNPs were preferentially selected in each LD bin. Additional SNPs were chosen based on coding status and regulatory potential or to cover gaps greater than 2 kb.

Genomic DNA for the GENECARD and CATHGEN samples was extracted from whole blood using the PureGene system (Gentra Systems, Minneapolis, Minnesota, United States). Genotyping in GENECARD was performed using the ABI 7900HT Taqman SNP genotyping system (Applied Biosystems, Foster City, California, United States), which is the standard PCR-based, dual fluor, allelic discrimination assay in a 384-well plate format with a dual laser scanner. Allelic discrimination assays were ordered from Applied

Biosystems. A total of 15 quality control samples (six reference genotype controls in duplicate, two Centre d'Etude du Polymorphisme Humain (CEPH) pedigree individuals, and one no-template sample) were included in each quadrant of the 384-well plate. Genotyping in CATHGEN was performed using the Illumina BeadStation 500G SNP genotyping system (Illumina, San Diego, California, United States). Each Sentrix Array generates 1,536 genotypes of 96 individuals. Four quality control samples (two CEPH pedigree individuals and two identical in-plate controls) were included within each individual array experiment. Results of these quality-control samples were used to identify possible sample plating errors and genotype calling inconsistencies. The genotype call rates of all the SNPs genotyped in this study were higher than 95%. The estimated error rate of SNPs meeting the quality control benchmarks was less than 0.2%.

3.5.4 Human donor aorta samples collection and expression

A collection of aorta tissue samples (n = 205) were harvested and prepared as described previously (Seo et al., 2004). Because these aorta samples were obtained from deceased heart donors, the clinical data associated with these aortas are limited to age, sex and race. DNA and mRNA were extracted from aorta tissue. Disease burden in the aorta was measured using the procedures from the Pathobiological Determinants of Atherosclerosis in the Young study (PDAY) (Cornhill et al., 1995). The procedures included measuring a proportion of the aorta with Sudan IV staining (early atherosclerotic lesions) and a proportion of the aorta with raised lesions (severe disease).

In addition, mRNAs were extracted from aorta samples for expression analysis. There were ninety samples in total representing seventy individuals, because some individuals had more than one sample from different aorta sections (Seo et al., 2004). The expression level of ZNF217 was measured by the abundance of tag 32034_at from the Affymetrix U95Av2 microarray. The raw signal intensity value of expression was $\log_2$ transformed and normalized using quantile normalization. This normalized value was modeled using linear regression model with additive genotype, age, sex, and race. To take into account repeated expression measures for multiple sections per aorta sample, a mixed model as implemented in the SAS MIXED procedure was used. We also performed linear/non-parametric regression analysis of disease burden with additive genotype, ZNF217 expression, age, sex, and race. The values of disease burden represented by the proportion of the aorta with Sudan IV staining and raised lesions were not normally distributed. And thus instead of using the Sudan IV and raised lesion measures, we used the rankings of the Sudan IV and raised lesion to implement the non-parametric regression model. The sample with the lowest value was ranked as no. 1.

3.5.5 Statistical analysis

LD between pairs of SNPs was assessed using the Graphical Overview of Linkage Disequilibrium package (Abecasis et al., 2000) and was displayed using Haploview (Barrett et al., 2005). Genotypic and allelic association in CATHGEN, aorta samples and the GENECARD probands from the United States was examined using multivariable

logistic regression modeling adjusted for race and sex, and also for race, sex, and known CAD risk factors (history of hypertension, history of diabetes mellitus, body mass index, history of dyslipidemia, and smoking history) overall. In addition, the CATHGEN sample was stratified by ethnicity (Caucasian vs. African American). SAS 9.1 (SAS Institute, Cary, North Carolina, United States) was used for statistical analysis. The software MERLIN (Multipoint Engine for Rapid Likelihood Inference) was used for two-point and multipoint nonparametric linkage analysis (Abecasis et al., 2000). Family-based association was tested using the APL test (Martin et al., 2003; Chung et al., 2007). The APL test incorporates data from affected sibling pairs with available parental data and unaffected siblings in the analyses, effectively using all available information in the GENECARD families. The APL software appropriately accounts for the non-independence of affected siblings and calculates a robust estimate of the variance.

3.6 Funding

This work was supported by the National Institutes of Health [HL073389 to Hauser, MH059528 to Hauser, and HL73042 to Goldschmidt, Kraus].

3.7 Acknowledgements

We are deeply appreciative of the subject volunteers' participation in the GENECARD and CATHGEN studies. We would also like to acknowledge the essential

contributions of the following individuals to making this publication possible: Elaine Dowdy, the GENECARD Investigators Network, the CATHGEN Steering Committee Members, Neil Freedman, and the staff at the Center for Human Genetics at Duke Medical Center for their innumerable contributions to this manuscript.

3.8 References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—Rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002; 30: 97–101.

Abecasis GR, Cookson WO. GOLD—Graphical overview of linkage disequilibrium. Bioinformatics. 2000; 16: 182–183.

Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21: 263–265.

Chung RH, Hauser ER, Martin ER. Interpretation of simultaneous linkage and family-based association tests in genome screens. Genet Epidemiol. 2007; 31(2): 134-142.

Connelly JJ, Wang T, Cox JE, Haynes C, Wang L, Shah SH, et al. GATA2 Is Associated with Familial Early-Onset Coronary Artery Disease. PLoS Genet. 2006; 2: e139.

Cornhill JF, Herderick EE, Vince DG. The clinical morphology of human atherosclerotic lesions. Lessons from the PDAY Study. Pathobiological Determinants of Atherosclerosis in Youth. Wien. Klin. Wochenschr. 1995; 107: 540-543.

Felker GM, Shaw LK, O'Connor CM. A standardized definition of ischemic cardiomyopathy for use in clinical research. J Am Coll Cardiol. 2002; 39: 210–218.

Fortin DF, Califf RM, Pryor DB, Mark DB. The way of the future redux. Am J Cardiol. 1995; 76: 1177–1182.

Hauser ER, Mooser V, Crossman DC, Haines JL, Jones CH, et al. Design of the genetics of early-onset cardiovascular disease (GENECARD) study. Am Heart J. 2003; 145:602–613.

Hauser ER, Crossman DC, Granger CB, Haines JL, Jones CJ, et al. A genomewide scan for early-onset coronary artery disease in 438 families: The GENECARD Study. Am J Hum Genet. 2004; 75: 436–447.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. The human genome browser at UCSC. Genome Res. 2002; 12: 996-1006.

Komulainen K, Alanne M, Auro K, Kilpikari R, Pajukanta P, Saarela J, et al. Risk Alleles of USF1 Gene Predict Cardiovascular Disease of Women in Two Prospective Studies. PLoS Genet. 2006; 2: e69.

Kong DF, Shaw LK, Harrell FE, Muhlbaier LH, Lee KL, et al. Predicting survival from the coronary arteriogram: An experience-based statistical index of coronary artery disease severity. J Am Coll Cardiol. 2002; 39 (Suppl A): 327A.

Krig SR, Jin VX, Bieda MC, O'Geen H, Yaswen P, Green R, Farnham PJ. Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays. J Biol Chem. 2007; 282(13): 9703-9712.

Martin ER, Bass MP, Hauser ER, Kaplan NL. Accounting for linkage in family-based tests of association with missing parental genotypes. Am J Hum Genet. 2003; 73: 1016–1026.

Mohlke KL, Boehnke M. The role of HNF4A variants in the risk of type 2 diabetes. Curr Diab Rep. 2005; 5: 149-156.

Pajukanta P, Lilja HE, Sinsheimer JS, Cantor RM, Lusis AJ, Gentile M, et al. Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). Nat Genet. 2004; 36: 371-376.

Quinlan KG, Verger A, Yaswen P, Crossley M. Amplification of zinc finger gene 217 (ZNF217) and cancer: when good fingers go bad. Biochim Biophys Acta. 2007; 1775(2): 333-340**.**

Rosamond W, Flegal K, Friday G, Furie K, Go A, Greenlund K, et al. Heart Disease and Stroke Statistics--2007 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation. 2007; 115: e69-171.

Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. WTCCC and the Cardiogenics Consortium. Genomewide association analysis of coronary artery disease. N Engl J Med. 2007; 357:443-453.

Seo D, Wang T, Dressman H, Herderick EE, Iversen ES, Dong C, et al. Gene Expression Phenotypes of Atherosclerosis. Arterioscler. Thromb. Vasc. Biol. 2004; 24: 1922-1927.
Smith LR, Harrell FE, Rankin JS, Califf RM, Pryor DB, et al. Determinants of early versus late cardiac death in patients undergoing coronary-artery bypass graft-surgery. Circulation. 1991; 84: 245–253.

Wang L, Fan C, Topol SE, Topol EJ, Wang Q. Mutation of MEF2A in an inherited disorder with features of coronary artery disease. Science. 2003; 302: 1578-1581.

Weedon MN. The importance of TCF7L2. Diabet Med. 2007; 24(10): 1062-1066.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678.

Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, Vance JM, Zuchner S, Hauser MA. SNPselector: a web tool for selecting SNPs for genetic association studies. Bioinformatics. 2005; 21: 4181-4186.

3.9 Figures

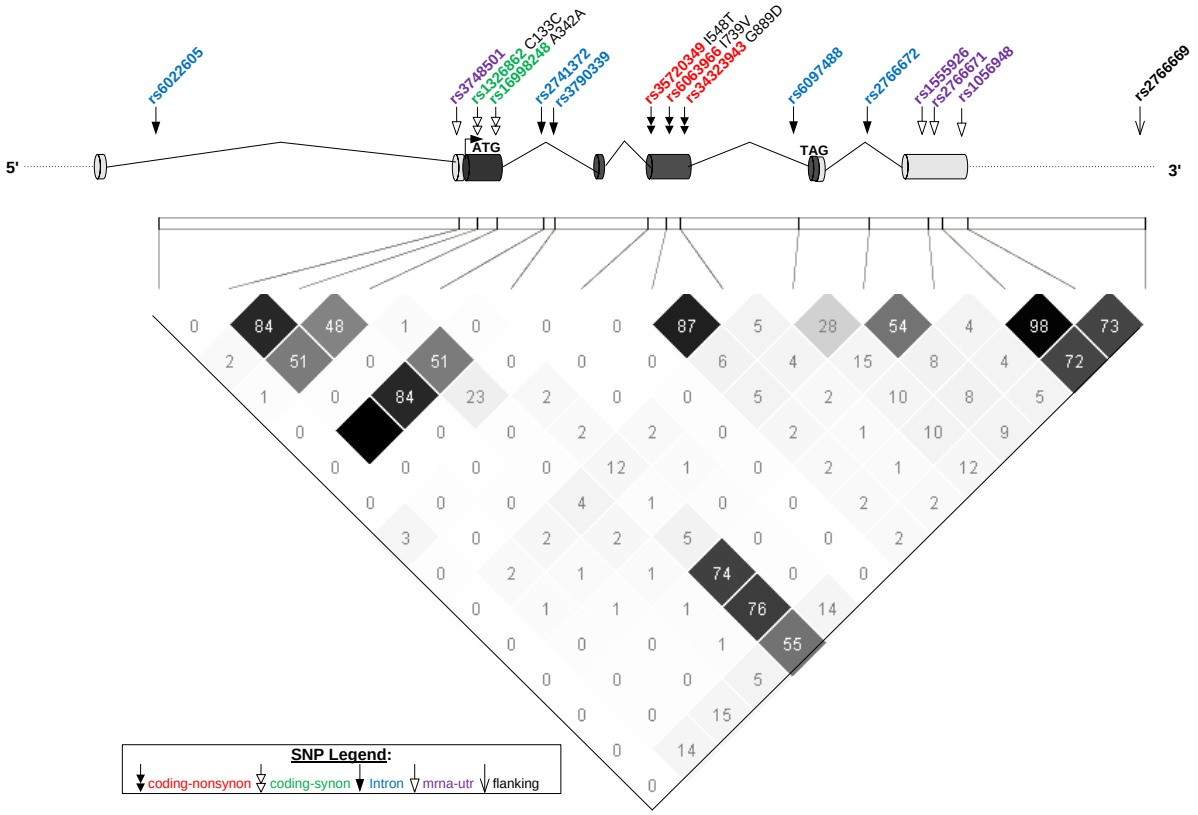


Figure 3.1 ZNF217 gene schematic and LD view of Caucasian control samples from CATHGEN

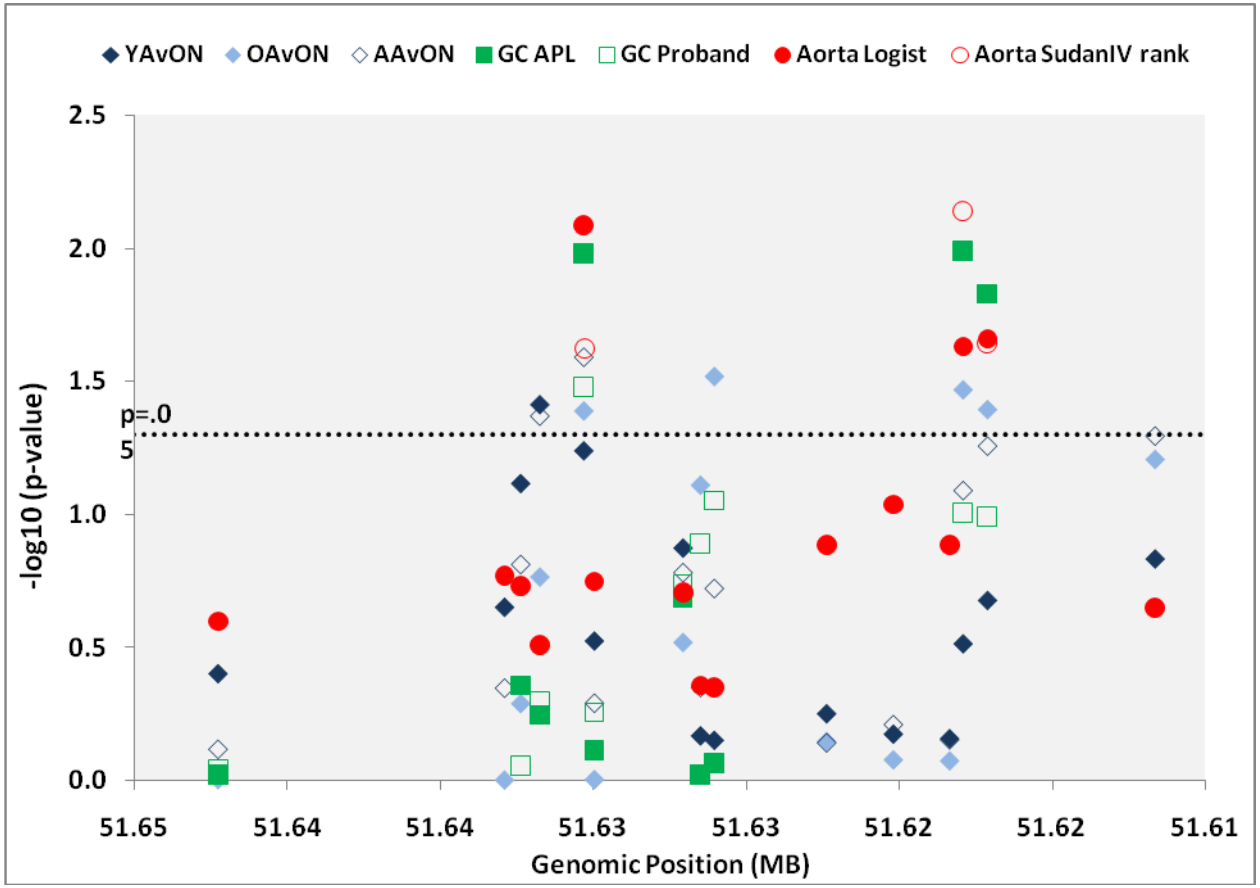


Figure 3.2 Combined results of three independent association studies

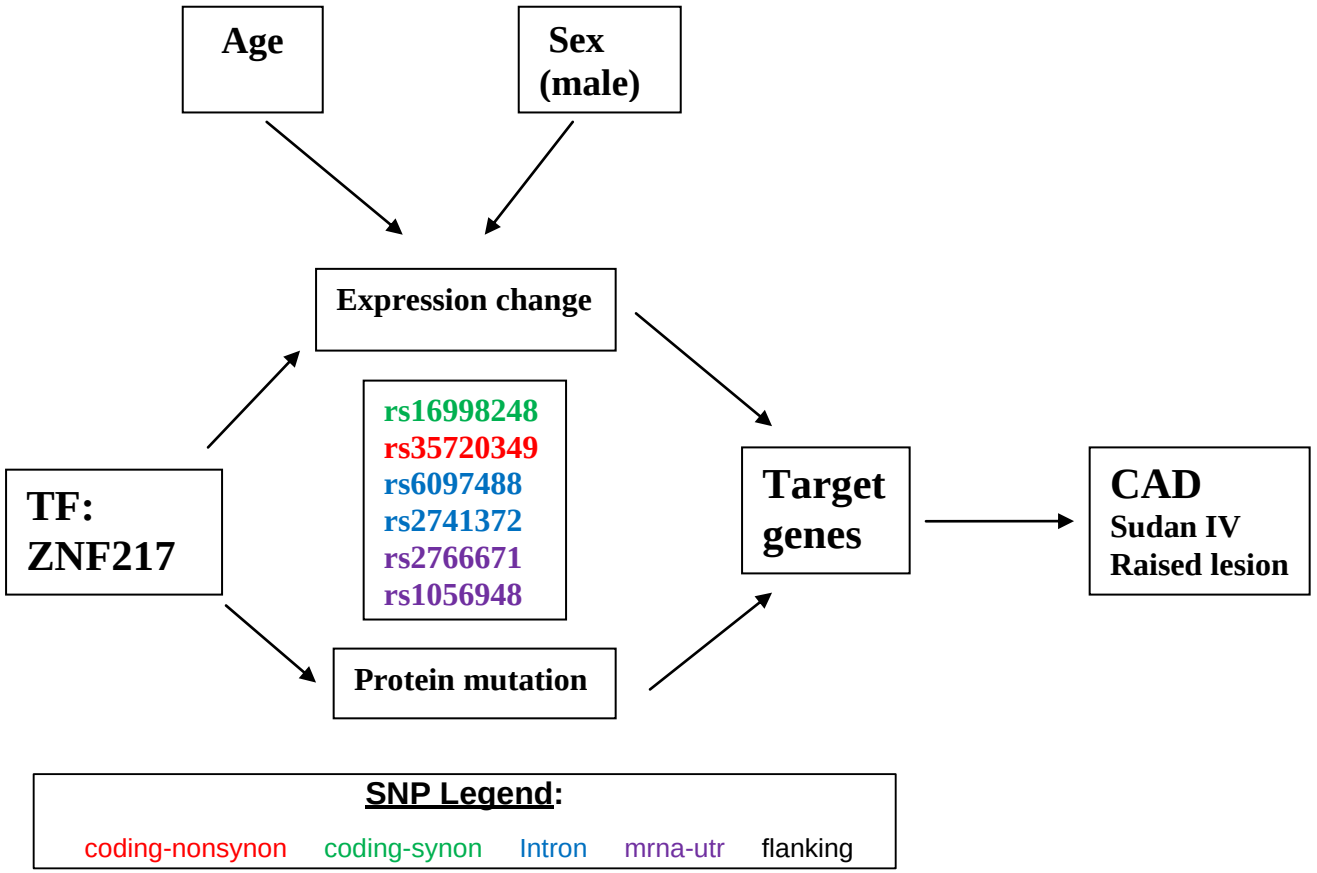


Figure 3.3 Hypothesis about CAD (Sudan IV and Raised lesion) association with ZNF217 (SNP and expression), Age, and Sex

3.10 Tables

Table 3.1 Clinical characteristics of CATHGEN subjects and southeastern US GENECARD probands

Category	CATHGEN Young affected (n = 642)	Old affected (n = 263)	Unaffected controls (n = 405)	GENECARD Probands (n = 713)
Mean age at onset (SD)	46.4 (6.3)	65.9 (7.7)	69.7 (6.6)	44.0 (5.7)
Race: Caucasian	71.3%	87.1%	71.4%	89.0%
Sex: Male	79.1%	72.3%	42.9%	70.7%
Family History of CAD	54.7%	43.2%	26.3%	100.0%
Body Mass Index (BMI), kg/m^2	30.7 (6.4)	28.0 (6)	29.1 (7.0)	30.7 (6.9)
Ever-smoked	69.2%	51.9%	40.0%	78.2%
History of Diabetes	33.2%	31.4%	21.0%	22.3%
History of Hypertension	69.7%	75.4%	67.1%	58.6%
History of Myocardial Infarction	52.1%	37.9%	0.0%	63.1%
History of Coronary Artery Bypass Graft	39.8%	51.9%	0.0%	47.8%
Mean systolic blood pressure, mm Hg (SD)	140 (24)	151 (26)	150 (24)	138 (23)
Mean diastolic blood pressure, mm Hg (SD)	78 (14)	78 (13)	77 (14)	81 (12)
Total Cholesterol, mg/dL (SD)	193 (62)	173 (41)	192 (49)	230 (61)
LDL, mg/dL (SD)	110 (43)	98 (34)	107 (36)	139 (50)
HDL, mg/dL (SD)	41 (12)	45 (14)	52 (18)	39 (14)
Triglycerides mg/dL (SD)	222 (260)	163 (96)	157 (126)	219 (147)

Table 3.2 Information for SNPs genotyped in ZNF217

				Control	CATHGEN		GC proband	Aorta
				ON (n = 405)	YA (n = 642)	OA (n =263)	n = 713	n = 205
SNP	Gene Location	Location (bp[1])	Alleles[2]	MAF[3]	MAF[3]	MAF[3]	MAF[3]	MAF[3]
rs6022605	Intron 1	51,642,268	G/A	0.013	0.018	0.002	0.008	0.015
rs3748501	Exon 2, 5' UTR	51,632,902	T/C	0.035	0.025	0.035	NA	0.059
rs1326862	Exon 2, C133C	51,632,374	A/G	0.059	0.046	0.045	0.046	0.081
rs16998248	Exon 2, A342A	51,631,747	T/A	0.060	0.052	0.044	0.056	0.087
rs2741372	Intron 2	51,630,310	G/A	0.156	0.120	0.127	0.139	0.110
rs3790339	Intron 2	51,629,964	G/A	0.035	0.026	0.035	0.036	0.062
rs35720349	Exon 4, I548T	51,627,067	G/A	0.013	0.006	0.004	0.009	0.010
rs6063966	Exon 4, I739V	51,626,495	G/A	0.115	0.117	0.148	0.129	0.117
rs34323943	Exon 4, G889D	51,626,044	T/C	0.077	0.086	0.129	0.102	0.090
rs6097488	Intron 4	51,622,372	A/G	0.333	0.335	0.392	NA	0.372
rs2766672	Intron 5	51,620,191	G/A	0.291	0.293	0.284	NA	0.301
rs1555926	Exon 6, 3' UTR	51,618,351	C/T	0.170	0.167	0.169	NA	0.198
rs2766671	Exon 6, 3' UTR	51,617,909	C/T	0.139	0.106	0.104	0.123	0.096
rs1056948	Exon 6, 3' UTR	51,617,112	C/T	0.127	0.100	0.105	0.121	0.099
rs2766669	Downstream	51,611,633	A/G	0.182	0.148	0.148	NA	0.1708

[1] NCBI build 35

[2] Major/minor allele

[3] Minor allele frequency

Table 3.3 ZNF217 SNPs associated with CAD in CATHGEN Caucasian group. OR is for the comparison of the additive genetic model. 1 copy of the minor allele vs. 0 copy.

	YA vs ON		OA vs ON		AA vs ON	
SNP	OR (95% CI)	*P*-value	OR (95% CI)	*P*-value	OR (95% CI)	*P*-value
rs6022605	2.582 (0.3, 23.0)	0.3957	0.001 (0, 1000)	0.9862	1.382 (0.2, 11.0)	0.7597
rs3748501	0.585 (0.2, 1.4)	0.2225	1.005 (0.4, 2.4)	0.9904	0.754 (0.4, 1.6)	0.4476
rs1326862	0.486 (0.2, 1.1)	0.0762	0.755 (0.3, 1.7)	0.5111	0.605 (0.3, 1.2)	0.1537
rs16998248	0.500 (0.3, 1.0)	**0.0385**	0.608 (0.3, 1.2)	0.1710	0.559 (0.3, 1.0)	**0.0426**
rs2741372	0.714 (0.5, 1.0)	0.0574	0.676 (0.5, 1.0)	**0.0407**	0.707 (0.5, 1.0)	**0.0256**
rs3790339	0.639 (0.3, 1.5)	0.2980	1.008 (0.4, 2.4)	0.9856	0.784 (0.4, 1.6)	0.5098
rs35720349	0.336 (0.1, 1.4)	0.1333	0.421 (0.1, 2.2)	0.3011	0.414 (0.1, 1.4)	0.1649
rs6063966	0.912 (0.6, 1.4)	0.6776	1.468 (1.0, 2.2)	0.0773	1.158 (0.8, 1.7)	0.4451
rs34323943	1.087 (0.7, 1.7)	0.7047	1.617 (1.0, 2.5)	**0.0302**	1.294 (0.9, 1.9)	0.1892
rs6097488	1.100 (0.8, 1.5)	0.5598	1.068 (0.8, 1.5)	0.7092	1.054 (0.8, 1.4)	0.7183
rs2766672	1.064 (0.8, 1.4)	0.6676	1.033 (0.8, 1.4)	0.8314	1.066 (0.8, 1.4)	0.6133
rs1555926	1.065 (0.8, 1.5)	0.6942	0.966 (0.7, 1.4)	0.8389	1.056 (0.8, 1.4)	0.7000
rs2766671	0.823 (0.6, 1.2)	0.3054	0.635 (0.4, 1.0)	**0.0339**	0.746 (0.5, 1.0)	0.0811
rs1056948	0.788 (0.5, 1.1)	0.2098	0.648 (0.4, 1.0)	**0.0402**	0.725 (0.5, 1.0)	0.0552
rs2766669	0.772 (0.5, 1.1)	0.1465	0.691 (0.5, 1.0)	0.0619	0.734 (0.5, 1.0)	0.0506

Table 3.4 ZNF217 SNPs associated with CAD in the GENECARD sample for family-based association (APL) and case-control association of US-born GENECARD probands versus CATHGEN controls

	GENECARD (APL)	GENECARD proband vs. ON	
SNP	*P*-value	OR (95% CI)	*P*-value
rs6022605	0.9552	1.108 (0.2, 6.4)	0.9092
rs1326862	0.4399	1.053 (0.5, 2.0)	0.8786
rs16998248	0.5697	0.841 (0.5, 1.4)	0.5284
rs2741372	**0.0105**	0.722 (0.5, 1.0)	**0.0331**
rs3790339	0.7736	1.239 (0.6, 2.5)	0.5524
rs35720349	0.2065	0.483 (0.2, 1.4)	0.1828
rs6063966	0.9558	1.249 (0.8, 1.8)	0.2596
rs34323943	0.8651	1.307 (0.9, 1.9)	0.184
rs2766671	**0.0102**	0.762 (0.6, 1.1)	0.0983
rs1056948	**0.0149**	0.766 (0.6, 1.1)	0.1017

Table 3.5 ZNF217 SNPs associated with CAD in aorta samples with the mixed models

Model	Diseased vs. non-diseased[1]		SudanIV_rank[2]	RaisedLesion_rank[2]	ZNF217 expression
Variable	OR (95% CI)	*P*-value	β (*P*-value)	β (*P*-value)	β (*P*-value)
rs6022605	3.697 (0.4, 34.8)	0.2531	25.496 (0.4644)	-6.409 (0.7568)	-0.477 (0.1631)
rs3748501	1.977 (0.7, 5.2)	0.1699	11.452 (0.5298)	8.214 (0.4452)	-0.321 (0.0738)
rs1326862	1.693 (0.8, 3.7)	0.1867	3.013 (0.8385)	8.706 (0.3190)	-0.207 (0.1275)
rs16998248	1.480 (0.7, 3.1)	0.3092	-1.568 (0.9165)	2.657 (0.7639)	**-0.250 (0.0320)**
rs2741372	0.079 (0, 0.5)	**0.0082**	**-30.279 (0.0238)**	-9.890 (0.2120)	0.092 (0.4584)
rs3790339	1.845 (0.8, 4.5)	0.1791	16.024 (0.3298)	3.260 (0.7375)	-0.169 (0.2567)
rs35720349	6.339 (0.4, 104.7)	0.1968	-34.007 (0.4144)	-7.443 (0.7636)	**-1.070 (0.0045)**
rs6063966	0.673 (0.2, 1.8)	0.4403	8.608 (0.4785)	-12.666 (0.0773)	-0.094 (0.2597)
rs34323943	0.648 (0.2, 2.0)	0.4485	-0.714 (0.6575)	-13.086 (0.0997)	-0.025 (0.7856)
rs6097488	2.394 (0.8, 7.4)	0.1308	15.103 (0.2822)	1.832 (0.8277)	**0.169 (0.0420)**
rs2766672	0.684 (0.4, 1.1)	0.0918	0.259 (0.9772)	5.951 (0.2645)	-0.020 (0.7688)
rs1555926	1.85 (0.8, 4.1)	0.1306	16.920 (0.1089)	9.375 (0.1299)	0.020 (0.8188)
rs2766671	0.108 (0, 0.7)	**0.0235**	**-36.657 (0.0072)**	0.072 (0.9929)	0.119 (0.2807)
rs1056948	0.106 (0, 0.7)	**0.0220**	**-31.178 (0.0227)**	-0.259 (0.9745)	0.110 (0.3480)
rs2766669	0.484 (0.2, 1.6)	0.2248	-21.042 (0.0605)	3.196 (0.6285)	0.029 (0.7613)

[1] Logistic regression model

[2] Non-parametric regression model

CHAPTER 4

Application of a novel transcription factor binding site prediction method to ZNF217

4.1 Abstract

Zinc finger protein 217 (ZNF217) is known to repress the transcription of many genes and is associated with cell proliferation, survival, and invasiveness of cancer cells. Our recent association study of ZNF217, the combined results from independent genome-wide linkage, and gene expression studies suggest that ZNF217 is a novel susceptibility gene for CAD. We chose ZNF217 to apply our previously developed novel TFBS prediction method because of its biological importance and the availability of ZNF217 ChIP-chip results. The specific goals of our study were to (1) evaluate our novel TFBS prediction method with another independent TF, ZNF217; (2) make a genome-wide prediction of potential ZNF217-BSs; and (3) identify ZNF217 target genes. Published ZNF217 ChIP-chip results were used to construct the training dataset and optimize the kernel logistic regression prediction model. Based on the cross-validation, this prediction model achieved 60.8% sensitivity, 97.5% specificity, and an AUC (area under the receiver operator characteristic curve) of 0.918. When applied to the human genome, we predicted 49,703 ZNF217 binding sites within 5 kb upstream of the transcription start site of 10,786 genes. These predictions included 32 out of 51 experimentally identified ZNF217 target genes. We demonstrated that our TFBS prediction method can be extended to other transcription factors identified in human disease studies. In addition, identification of potential ZNF217 targets could further our understanding of gene-gene interaction underlying complex disease and ZNF217's specific contribution to heart disease.

4.2 Introduction

Transcription factors (TFs) play a key role in transcriptional regulation by controlling gene expression in temporal, cell type, or environmental specific ways. Several TFs have been characterized as mediators of complex disease processes (Connelly et al., 2006; Mohlke et al., 2005; Pajukanta et al., 2004; Wang et al., 2003; Weedon et al., 2007). Zinc finger protein 217 (ZNF217) is known to repress the transcription of many genes and is associated with cell proliferation, survival, and invasiveness of cancer cells (Quinlan et al., 2007). The region containing ZNF217 is selectively amplified during progression of several cancers suggesting that ZNF217 may give tumor cells selective advantage by interfering with normal regulation of cell growth, cell death, differentiation, and DNA repair (Krig et al., 2007). Our recent association study of ZNF217, that combined results from independent genome-wide linkage, association, and gene expression studies suggest that ZNF217 is a novel susceptibility gene for CAD (Hauser et al., 2004; Seo et al., 2004; Chapter 3 from this dissertation). We proposed that the target genes of these TFs also may be associated with human complex disease. Identification of potential ZNF217 targets could further our understanding of gene-gene interaction underlying complex disease and ZNF217's specific contribution to heart disease.

TFs interact with specific DNA elements called transcription factor binding sites (TFBSs) to control cell and tissue-specific gene expression. Accurately identifying TFBSs is, therefore, critical to our understanding of the biological regulation of the cell. Although many genome sequences are available, encoded functional elements such as TFBSs have

not been fully characterized. This is due, in part, to the complexity of TF binding activity and degeneration of the core binding site. ChIP-chip, a technique combining chromatin immunoprecipitation and microarray analysis (Hartman et al., 2005; Krig et al., 2007), allows for the identification of a large number of genomic regions bound by TFs. However, ChIP-chip experiments are time-consuming, expensive, only able to identify a subset of potential TFBSs due to variation in cell or tissue type, environmental conditions, and the biological cofactors necessary for TF binding. It would be more efficient to develop a computational method for TF target prediction followed by less costly genotyping and focused molecular biology experiments to identify association of gene-gene interaction and complex disease. 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 (Bulyk, 2003; MacIsaac et al., 2006; Stormo 2000). The common approaches of motif scanning are based on either consensus sequences or binding site matrices. In general, the low specificity of these methods leads to an inflated number of predicted TFBSs many of which are false positive results. Therefore, the reliability of prediction methods based on DNA sequence alone is low. An ideal prediction method combines DNA sequence with additional genomic features to improve specificity. Although the number of genomic features available is quite large, current prediction methods do not take full advantage of these genomic features.

Previously we have developed a novel TFBS prediction method based on a kernel logistic regression model (Wang et al., 2009). 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 building a prediction model. Our most accurate predictor achieved an AUC (area under the receiver operator characteristic curve) of 0.827 during cross-validation experiments, significantly outperforming standard PWM-based prediction methods.

We chose ZNF217 to apply our previously developed novel TFBS prediction method because of its biological importance (over-expression in many cancers and genetic association with CAD) and the availability of ZNF217 ChIP-chip results (Krig et al., 2007). The specific goals of our study were to (1) evaluate our novel TFBS prediction method with another independent TF, ZNF217; (2) make a genome-wide prediction of potential ZNF217-BSs; and (3) identify ZNF217 target genes. We demonstrated that our TFBS prediction method could be generalized to other TFs resulting in similar or better performance metrics. The results of this specific study will help prioritize CAD candidate genes as well as describe ZNF217's specific biological contribution to heart disease through the identification of its target genes.

4.3 Methods

We applied our previously developed novel TFBS prediction method to ZNF217 as described previously (Wang et al., 2009).

4.3.1 Genome sequence and features

All annotation and mapping locations were based on NCBI human genome build 35. The 5 kilobase (kb) region upstream of the transcription start sites (TSSs) of 23,105 RefSeq mRNA sequences were downloaded from the UCSC Genome Browser (Kent et al., 2002). Numerical values for each of these features were assigned to each 8 bp ZNF217 binding site, and were also obtained from the UCSC website (Kent et al., 2002). The conservation scores and predicted conserved elements (MostCons8) were generated by the program phastCons (Siepel et al., 2005), based on genome-wide multiple alignments of eight species (human, chimpanzee, mouse, rat, dog, chicken, fugu, and zebrafish). The total conservation score of each 8 bp ZNF217-BS (PhastCons8) was represented by the base-10 logarithm of the product of each base-pair's conservation score within the ZNF217-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 (King et al., 2005). 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 (Gardiner-Garden et al., 1987). The coordinates of DNaseI HS sites (DNaseI HS) are the regions in the genome hypersensitive to DNaseI cleavage within human CD4+ cells. The DNaseI HS score of each site reflected the degree of chromatin accessibility at that site (Boyle et al., 2008).

4.3.2 ZNF217 ChIP-chip data

Known ZNF217 interacting genomic regions were used to develop and evaluate our prediction method. A recently published ZNF217 ChIP-chip study identified ZNF217 interacting genomic regions using chromatin immunoprecipitation from three tumor cell lines (MCF7, SW480, and Ntera2) followed by microarray analysis (Krig et al., 2007). The ENCODE oligonucleotide microarray regions included promoter, intronic, exonic and intergenic regions from 44 genomic intervals on 20 chromosomes. We used 53 ZNF217 binding regions as positive controls. Each positive region was identified as a binding site in at least two of three biological independent experiments in all three of the different cell types. The average size of these positive regions is 466 bp. A set of 506 genomic intervals from the ENCODE regions, which were not bound by ZNF217 in any of these three tumor cell lines, were selected as negative controls. The size of these negative regions is 500 bp. The potential ZNF217-BSs from these control regions were used as the training dataset to develop the TFBS prediction model for ZNF217.

4.3.3 Preliminary prediction based on PWM scoring method

The PWM scoring method was used to identify potential ZNF217-BSs from the target regions. The ZNF217 binding matrix was obtained from a published ZNF217 ChIP-chip study (Krig et al., 2007). The web application Patser was used to convert the ZNF217 binding matrix to the PWM, generating a calculated cutoff score of 4.256 for predicting TFBSs based on the information content adjusted by sample size (Hertz et al., 1999). The

average GC content of 47.1% for the Patser analysis was calculated from 5 kb upstream sequences from 23,105 RefSeq mRNAs. The ZNF217 PWM was used to score each 8 bp sliding window within both strands of the target regions. A potential ZNF217-BS was defined as any 8 bp sequence with a score higher than the threshold of 4.256. The result of this step is the set of all 8-bp sequences defined as a PWM-based binding site. We called these potential binding sites.

4.3.4 Development of the prediction model incorporating features

To the set of sequence-based potential ZNF217-BSs, we then applied a more stringent prediction method using five genomic features (PhastCons8, MostCons8, RP5, CpG and DNaseI HS). Our goal is to identify the model that best predicts TFBS (positive/negative) using any combination of PWM and five features. We implemented a kernel logistic regression algorithm (Minka 2003) with the radial basis function (RBF) in MATLAB Version 7.0. Multiple genomic features were used by the kernel function to map the input data to a high-dimensional space. This supervised statistical learning model was trained by the positive and negative controls to find optimal hyperparameters with the best separation of these two groups. This optimized model was then applied to each potential ZNF217-BS and generated a score in the range from 0 to 1. The threshold for being a predicted ZNF217-BS was 0.5. Initially we 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 a p-value less than 0.05 as the threshold. Then we

implemented the prediction method using the most significant features identified from this feature selection approach. The performance of each prediction model was evaluated using a leave-one-locus-out (LOLO) cross-validation procedure based on sensitivity, specificity, and AUC (area under the receiver operator characteristic curve). And the model with the highest AUC was chosen as best. One difficulty in the model building exercise is that in most cases, several potential ZNF217-BSs were included within a positive locus in the training dataset. Initially, these potential ZNF217-BSs were grouped by their loci. In each iteration, all potential ZNF217-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 ZNF217-BSs within the test locus. The test locus was classified as positive if it included at least one predicted ZNF217-BS, otherwise it was classified as negative. 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 (Norusis 2004).

4.3.5 Genome-scale prediction and validation

Potential ZNF217-BSs within 5 kb upstream of the TSSs of 23,105 RefSeq mRNAs were first identified by the PWM scoring method. Using the optimized model developed from the training dataset, specific ZNF217-BS prediction scores were generated for each potential binding site using the genomic features of each site. The result for each potential

ZNF217-BS is the probability of being a binding site calculated from the optimized model. We evaluated the predicted ZNF217-BSs by their distributions of the prediction scores and the locations across 5 kb upstream regions of the TSSs. The prediction results were also evaluated by comparing to 54 published robust ZNF217 target genes, which were identified by both ZNF217 ChIP-chip assay with a promoter array and ZNF217 knockdown RNA analysis with siRNA in three tumor cell lines (Krig et al., 2007).

4.3.6 Functional annotation of ZNF217 target genes

To determine if the predicted ZNF217 target genes are over-represented in any functional categories, functional annotations were performed using the web application DAVID (The Database for Annotation, Visualization and Integrated Discovery). We focused on a few functional categories specifically, including gene ontology (GO), protein domain, pathway, disease, and tissue specificity. We used *p*-value 0.05 as the threshold for significant enrichment.

4.4 Results

4.4.1 Prediction method optimization

We applied the same TFBS prediction method developed previously to ZNF217 (Wang et al., 2009). We started by using the PWM scoring method to identify potential ZNF217-BSs from the positive and negative controls (see Methods). Among these potential ZNF217-BSs, 199 were associated with the 51 positive loci and 1,170 with the 439

negative loci identified by the ZNF217 ChIP-chip study from three tumor cell lines. These 1,369 ZNF217-BSs were then used to construct the training dataset. Two positive loci and 67 negative loci did not include any potential ZNF217-BSs and were excluded from the training dataset.

To further refine predictions, we started with all five genomic features in addition to the PWM score for model building and performed backward stepwise feature selection based on the contribution of each feature to classification of positive and negative site. We removed PhastCons8, MostCons8 and CpG features from the model based on a *p*-value = 0.05 threshold. The remaining variables (PWM, RP5, and DNaseI HS) had *p*-values less than 0.0001. Using these variables identified from the feature selection procedure, we developed a supervised statistical learning model using kernel logistic regression and applied it to the training dataset for LOLO cross-validation. Based on the LOLO cross-validation, this prediction model achieved 60.8% sensitivity, 97.5% specificity, and an AUC of 0.918 when using a 0.5 scoring threshold for predicting each ZNF217-BS. This optimized prediction model based on the selected features (PWM, RP5, and DNaseI HS) was then used to predict the ZNF217-BSs in the human genome (Table 4.1).

4.4.2 Genome-scale prediction and validation

In order to make genome-wide prediction of ZNF217 binding sites and associated target genes in the human genome, we focused on 5 kb upstream regions of the TSSs of

23,105 RefSeq mRNAs. At first, the PWM scoring method was used to identify 360,601 potential ZNF217-BSs in these sequences. We then applied our optimized ZNF217-BS prediction model, kernel logistic regression using two genomic features (RP5 and DNaseI HS) in addition to the PWM score, to improve the specificity of the prediction. Based on the default threshold (0.5) of the prediction scores, 49,703 ZNF217-BSs from 10,786 genes were predicted as ZNF217 targets, representing 13.8% of the initial potential ZNF217-BSs (Table 4.2). There were 1,346 SNPs mapped to these predicted ZNF217-BSs according to dbSNP build 129.

In addition, the target gene prediction results were evaluated by comparing to 54 robust ZNF217 target genes identified from both the ZNF217 ChIP-chip study and ZNF217 knockdown analysis (Krig et al., 2007). Only 51 out of 54 robust genes had related RefSeq mRNAs included in our genome-wide prediction. Our prediction method was able to identify 32 out of these 51 ZNF217 target genes (62.7%) (Table 4.3).

4.4.3 Distributions of predicted ZNF217-BSs

The final optimized prediction model generated a score for each potential ZNF217-BS identified using the PWM scoring alone. Prediction range from 0 to 1 and correspond to the predicted probability that the potential TFBS is a binding site. The score distribution between 0 and 1 with 0.01 intervals of our genome-wide prediction of ZNF217-BS indicated that a large portion of predicted sites had scores higher than 0.99 (Figure 4.1). Selecting ZNF217-BSs with the highest prediction scores significantly

reduced the number of predicted target genes. Based on the prediction score distribution, we chose a stringent threshold (0.99) to further reduce the number of predicted ZNF217 target genes in the human genome from 10,786 to 5,213 to be defined as candidate genes for further analysis.

Potential ZNF217-BSs identified only by the PWM scoring method were evenly distributed across 5 kb upstream regions of the TSSs of 23,105 RefSeq mRNAs (Figure 4.2). Our predicted ZNF217-BSs using a 0.5 scoring threshold were concentrated within 1 kb upstream of TSS, the region most likely to contain TFBSs (Zhang et al., 1998). Predicted ZNF217-BSs using the stringent threshold (0.99) are even more enriched within 1 kb upstream of TSS.

To explore whether the prediction of ZNF217-BS was biased toward sites with higher binding affinity indicated by higher PWM scores, we compared the PWM score distributions of all the initial 360,601 potential ZNF217-BSs identified by the PWM scoring method and predicted ZNF217-BSs with scores higher than 0.5 and 0.99, which were generated from the optimized prediction model. The results indicated that there was no significant difference among these three groups (Figure 4.3). In addition, the PWM score and the prediction score were not highly correlated with each other within the predicted ZNF217-BSs group with score higher than 0.5 (correlation coefficient = -0.095**).**

4.4.4 Functional annotation

To further understand the biological roles of these 5,213 ZNF217 candidate genes (with prediction score higher than 0.99), we used the web application, DAVID, to perform

functional annotations of these genes. DAVID provides a measure of significance for the identified categories, which are more highly enriched in the target set than would be expected by random chance. The DAVID analysis showed enrichment for genes with twelve diseases in the Genetic Association Database (p = 0.00098 – 0.05) (Becker et al., 2004). And nine of these diseases were different type of cancers (Table 4.4). In addition, these 5,213 ZNF217 candidate genes were most significantly enriched in cell cycle pathway with p-value 1.2×10^{-13}.

4.5 Discussion

4.5.1 Applying novel prediction method to ZNF217

Previously we used USF1 as a test case to develop a novel TFBS prediction method, which was based on a kernel logistic regression model including multiple genomic features (phylogenetic conservation, regulatory potential, presence of a CpG island and DNaseI hypersensitivity) as well as PWM scores. According to the cross-validation results using the training data from USF1 ChIP-chip study, our optimized predictor significantly outperformed standard PWM-based prediction methods. However, the prediction method developed from one TF may not perform well for all TFs given inherent variations, such as binding domains, binding sequence preferences, homology levels across species, and family members. One of the specific goals of our study was to apply our TFBS prediction method to another independent TF (ZNF217), and to evaluate if our general model building framework has the potential to be extended to other TFs.

Table 4.1 provided detailed comparison between ZNF217 and USF1 predictions (Table 4.1). The cross-validation results indicated that the performance of the prediction models for each TF was similar, and the ZNF217 prediction model was slightly better in terms of sensitivity, specificity, and AUC. The genome-wide prediction of ZNF217 produced more potential and predicted binding sites than USF1. This might due to the shorter ZNF217 core binding site (8 bp) than USF1 (10 bp). In general, the PWM of the shorter binding site is less informative and predicts more potential sites. Another difference was that the ZNF217 prediction model was based on three variables (PWM, RP5, and DNaseI HS) instead of four variables (PhastCons8, PWM, RP5, and DNaseI HS). These variables were generated by backward stepwise feature selection in order to identify a subset of features significantly contributing to the model. In the training dataset of ZNF217, PhastCons8 was correlated with DNase HS and RP5. So PhastCons8 was removed from the prediction model. The remaining three variables (PWM, RP5, and DNaseI HS) were efficient to build the model. By applying our prediction method to ZNF217; we demonstrated that this novel prediction method could be successfully applied to other TFs.

In order to evaluate the performances of the prediction method, besides performing cross-validation with the training dataset we compared our prediction results with 54 robust ZNF217 target genes. Both of the datasets used for validation were generated from the same ZNF217 ChIP-chip study (Krig et al., 2007). And the binding sites from these robust genes were not validated by individual experiments. It would be helpful to compare

our prediction with independent external robust genes. However, the target genes of ZNF217 have not been well identified. According to the TRED database (Jiang et al., 2007) and PubMed, currently there is no individually experimentally identified ZNF217 target gene. On the contrary, we were able to find 20 robust USF1 target genes - the overlap between the target genes obtained from the TRED database (Jiang et al., 2007) and reported from the literature (Naukkarinen et al., 2005). As more experimentally identified ZNF217 target genes become publicly available, we will be able to further evaluate our prediction and improve the prediction method. To evaluate 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 51 robust ZNF217 target genes at each step during the prediction process. As shown in Figure 4.3, we find that: 1) the common PWM scoring method was sufficient to identify potential ZNF217–BSs for most of these genes; 2) these potential sites identified spanned a wide range of PWM scores. Further, the distributions of PWM scores among the initial 360,601 potential sites, the 49,703 predicted ZNF217–BSs with default prediction threshold (0.5), and 20,790 predicted ZNF217-BSs generated using the stringent prediction threshold (0.99) were not much different (Figure 4.3). These results are consistent with our previous study using USF1 (Wang et al., 2009). This evidence suggests 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.

4.5.2 Training dataset from published ZNF217 ChIP-chip results

This prediction method was based on a supervised statistical learning model. A robust training dataset is crucial for the development of an accurate and reliable prediction method. Recently published ZNF217 ChIP-chip results allowed us to construct the training dataset and apply the prediction method. The ZNF217 interacting genomic regions identified by the ZNF217 ChIP-chip study were generated from the breast cancer line MCF7, the colon cancer line SW480, and a teratocarcinoma line Ntera2 followed by an unbiased location analysis using the ENCODE microarray (Krig et al., 2007). Each of 53 positive regions was identified in at least two of three independent biological experiments in all three of the different cell types. The 506 negative intervals were not bound by ZNF217 in any of these three tumor cell lines. One of the limitations of ChIP-chip technique is that it is hard to differentiate between loci directly bound by the TF or indirectly bound by the TF's partners. This problem may be overcome by performing multiple independent experiments in several different cell lines. In addition, in contrast to USF1, the size of the training dataset was larger and the average size of the control regions was smaller (Table 4.1), which made the ZNF217 the training dataset more robust and accurate than USF1. This may be one of the reasons why the cross-validation result of the ZNF217 prediction method was slightly better than USF1 in terms of sensitivity, specificity, and AUC.

4.5.3 Predicted ZNF217 binding sites and target genes

As the result of this study, we made a genome-wide prediction of ZNF217 targets, which includes 49,703 ZNF217-BSs and 5,213 candidate genes. The published ZNF217 ChIP-chip study also indentified thousands of promoters bound by ZNF217 (Krig et al., 2007). These large numbers are consistent with the ChIP-chip studies of other human TFs, such as USF1, c-Myc, STAT1, Men1, and E2F1 (Bieda et al., 2006; Cawley et al., 2004; Rada-Iglesias et al., 2008; Robertson et al., 2007; Scacheri et al., 2006). ChIP based studies can only identify genes that are targets given specific cellular or environmental conditions. These ZNF217 TFBSs and target genes identified by ChIP-chip experiments are likely to be a subset of all potential ZNF217 targets. In contrast, our *in silico* prediction method will identify potential ZNF217-BSs independent of cell type, stage, or environment. Thus the numbers of predicted ZNF217 binding sites and target genes are expected to be higher than *in vivo* experiments. In this study, we focused on the region 5 kb upstream of the TSSs of RefSeq mRNAs, because of the enrichment of ZNF217 binding regions in proximal promoters indicated by the published ZNF217 ChIP-chip study (Krig et al., 2007). However, ZNF217-BSs could occur beyond 5 kb upstream of TSSs. Thus a wider range of genomic regions could be considered in the future. As result, more ZNF217-BSs are expected to be identified.

4.5.4 Application to human disease study

Experimental evidence suggests that ZNF217 is associated with many cancers and CAD (Quinlan et al., 2007; Chapter 3 from this dissertation). The target genes of ZNF217 also may be associated with these diseases. One of the major goals of this study is to make genome-wide prediction of ZNF217 targets and to obtain insight into how ZNF217 can contribute to human diseases through its target genes. These predicted ZNF217 target genes that can be used as candidates for studying susceptibility to complex human diseases. Our supplemental file includes a list of predicted ZNF217-BSs and their prediction score (Table 4.2). This large number of predicted ZNF217-BSs along with the scores allow for adjusting the stringency of the prediction score threshold to refine candidate genes and also for choosing specific filters to emphasize a particular subset of interest.

High-throughput technologies, such as gene expression microarray, ChIP-chip and bioinformatics scanning approach, usually result in a large interesting gene list. It is important to extract meaningful biological signals from these genes. Gene set enrichment analysis provides a promising way for systematic functional annotation of the given gene list. For example, the web application DAVID provides a measure of significance for the identified category with a *p*-value, which indicates the probability that the identified category is more enriched in the target set than would be expected by random chance. The DAVID analysis 5,213 ZNF217 candidate genes showed enrichment in nine cancers and cell cycle pathway. These results are consistent with its role in cancers reported in

many publications (Quinlan et al., 2007). Further studying these predicted ZNF217 target genes may provide important information about cancer formation and progression.

Our recent association study of ZNF217, that combined results from independent genome-wide linkage, association, and gene expression studies, suggests that ZNF217 is a novel susceptibility gene for CAD (Chapter 3 from this dissertation). Identification of potential ZNF217 targets could further our understanding of ZNF217's specific contribution to heart disease. "Genomic convergence", a strategy that integrates several independent separate lines of experimental evidence to prioritize disease associated candidate genes (Hauser et al., 2003), is being used by our CAD study to combine the ZNF217 binding site 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 atherosclerosis (Seo et al., 2004). By combining our *in silico* ZNF217-BS prediction method with this expression result we identified 43 ZNF217 target genes that were differentially expressed between cases and controls in aorta (Table 4.5). Using the same approach, we identified 282 ZNF217 target genes that located to a modest linkage peaks from a previous family based linkage study (GENECARD) of early-onset CAD (Hauser et al., 2004) (data not shown). In addition, we have been gathering a list of candidate genes for CAD study based on prior knowledge of associations, such as the results from genome-wide linkage, association, gene expression studies, and literature reports. The current list includes 450

candidate genes (Chapter 3 from this dissertation). We found 94 predicted ZNF217 target genes from this list (data not shown). Note that it does not appear that CAD genes are enriched or rather that ZNF217-BSs are enriched among known candidates. These genes generated from "genomic convergence" analysis define a list of important candidate genes from a large list of initial targets. The intersection of these lists also provide a start analysis of statistical interactions between single nucleotide polymorphisms (SNPs) in ZNF217 and its target genes.

SNPs are the most abundant molecular marker in the human genome. SNPs are commonly used for large-scale genetic association studies to identify genetic factors responsible for complex genetic diseases. However, it is still remains a challenge to select SNPs with potential functional impact, especially from the large number of identified non-coding SNPs. One type of variant of particular interest are 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. The base pair resolution of our ZNF217-BS predictions enable 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 1,346 SNPs within our predicted USF1-BSs in the human genome based on the genomic locations of the SNPs released by NCBI in dbSNP build 129 (Table 4.2). The experimental approaches to 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 ZNF217-BSs have the potential to influence the regulation of ZNF217 target genes; they enable study of ZNF217's association with complex disease in humans.

4.6 Conclusions

We applied our previously developed novel TFBS prediction method to ZNF217. Published ZNF217 ChIP-chip results were used to construct the training dataset and optimize the kernel logistic regression prediction model. We demonstrated that our TFBS prediction method can be extended to other TFs identified in human disease studies. Because of the biological importance of ZNF217 (over-expression in many cancers and genetic association with CAD), the predicted ZNF217 targets from this study will help prioritize candidate genes as well as describe ZNF217's specific biological contribution to human diseases through its target genes.

4.7 Acknowledgements

We thank the staff at the Center for Human Genetics at Duke Medical Center. This study was supported by NIH grants HL073389 (Hauser), MH059528 (Hauser) and HL73042 (Goldschmidt, Kraus).

4.8 References

Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004; 36:431-432.

Bieda M, Xu X, Singer M, Green R, Farnham P J. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res. 2006; 16: 595–605.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008; 132: 311-322.

Bulyk ML. Computational prediction of transcription-factor binding site locations. Genome Biol. 2003; 5: 201.

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell. 2004; 116: 499-509.

Connelly JJ, Wang T, Cox JE, Haynes C, Wang L, et al. GATA2 Is Associated with Familial Early-Onset Coronary Artery Disease. PLoS Genet. 2006; 2: e139.

Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. Mol Biol. 1987; 196: 261-282.

Hartman SE, Bertone P, Nath AK, Royce TE, Gerstein M, et al. Global changes in STAT target selection and transcription regulation upon interferon treatments. Genes Dev. 2005; 19; 2953-2968.

Hauser MA, Li YJ, Takeuchi S, Walters R, Noureddine M, et al. Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage,' Hum Mol Genet. 2003; 12: 671-677.

Hauser ER, Crossman DC, Granger CB, Haines JL, Jones CJ, et al. A genomewide scan for early-onset coronary artery disease in 438 families: The GENECARD Study. Am J Hum Genet. 2004; 75: 436–447.

Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatic. 1999; 15: 563-577.

Huang D, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4(1): 44-57.

Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res. 2007; 35: 137–140.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. The human genome browser at UCSC. Genome Res. 2002; 12: 996-1006.

King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, et al. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res. 2005; 15: 1051-1060.

Krig SR, Jin VX, Bieda MC, O'Geen H, Yaswen P, et al. Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays. J Biol Chem. 2007; 282(13): 9703-9712.

MacIsaac KD, Fraenkel E. Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol. 2006; 2: e36.

Minka T. A comparison of numerical optimizers for logistic regression. Department of Statistics, Carnegie Mellon University. 2003.

Mohlke KL, Boehnke M. The role of HNF4A variants in the risk of type 2 diabetes. Curr Diab Rep. 2005; 5: 149-156.

Naukkarinen J, Gentile M, Soro-Paavonen A, Saarela J, Koistinen HA, et al. USF1 and dyslipidemias: converging evidence for a functional intronic variant. Hum Mol Genet. 2005; 14: 2595-2605.

Norusis M. SPSS 13.0 Statistical Procedures Companion. Upper Saddle River, NJ, Prentice Hall, Inc. 2004.

Pajukanta P, Lilja HE, Sinsheimer JS, Cantor RM, Lusis AJ, et al. Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). Nat Genet. 2004; 36: 371-376.

Quinlan KG, Verger A, Yaswen P, Crossley M. Amplification of zinc finger gene 217 (ZNF217) and cancer: when good fingers go bad. Biochim Biophys Acta. 2007; 1775(2): 333-340.

Rada-Iglesias A, Ameur A, Kapranov P, Enroth S, Komorowski J, et al. Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. Genome Res. 2008; 18(3): 380-92.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007; 8: 651-657.

Scacheri PC, Davis S, Odom DT, Crawford GE, Perkins S, et al. Genome-wide analysis of menin binding provides insights into MEN1 tumorigenesis. PLoS Genet. 2006; 2: 406–419

Seo D, Wang T, Dressman H, Herderick EE, Iversen ES, et al. Gene expression phenotypes of atherosclerosis,' Arterioscler Thromb Vasc Biol. 2004; 24: 1922-1927.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15: 1034-1050.

Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000; 16: 16-23.

Wang L, Fan C, Topol SE, Topol EJ, Wang Q. Mutation of MEF2A in an inherited disorder with features of coronary artery disease. Science. 2003; 302: 1578-1581.

Wang T, Furey TS, Connelly JJ, Ji S, Nelson S, et al. A general integrative genomic feature transcription factor binding site prediction method applied to analysis of USF1 binding in cardiovascular disease. Human Genomics. 2009; 3(3) (In press).

Weedon MN. The importance of TCF7L2. Diabet Med. 2007; 24(10): 1062-1066.

Zhang MQ. Identification of human gene core promoters in silico. Genome Res. 1998; 8: 319-326.

4.9 Figures

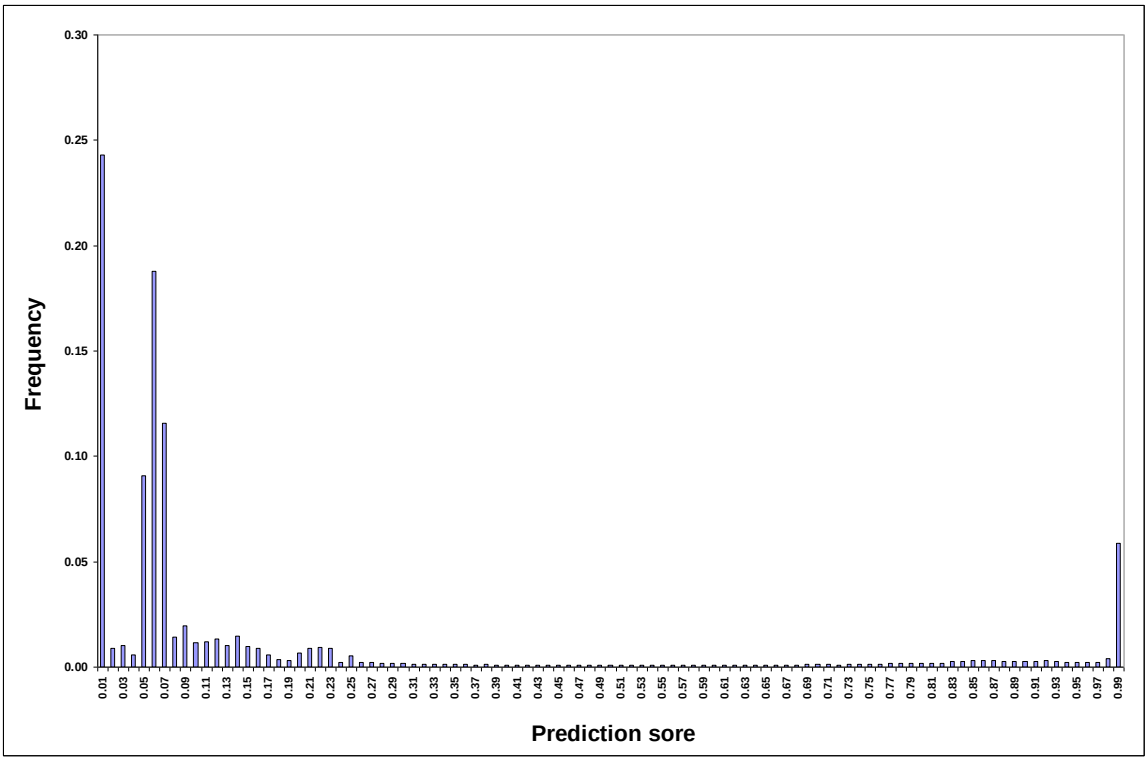


Figure 4.1 - Prediction scores distribution of potential ZNF217-BSs

By scanning 5 kb upstream of TSSs of 23,105 RefSeq mRNAs in the human genome, 360,601 potential ZNF217-BSs were identified by the PWM scoring method. The refined prediction method generates a score for each potential ZNF217-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 49,703 predicted ZNF217-BSs were generated using the optimized prediction model with default prediction threshold (0.5). The data are represented as histograms of frequency at each 0.01 score interval.

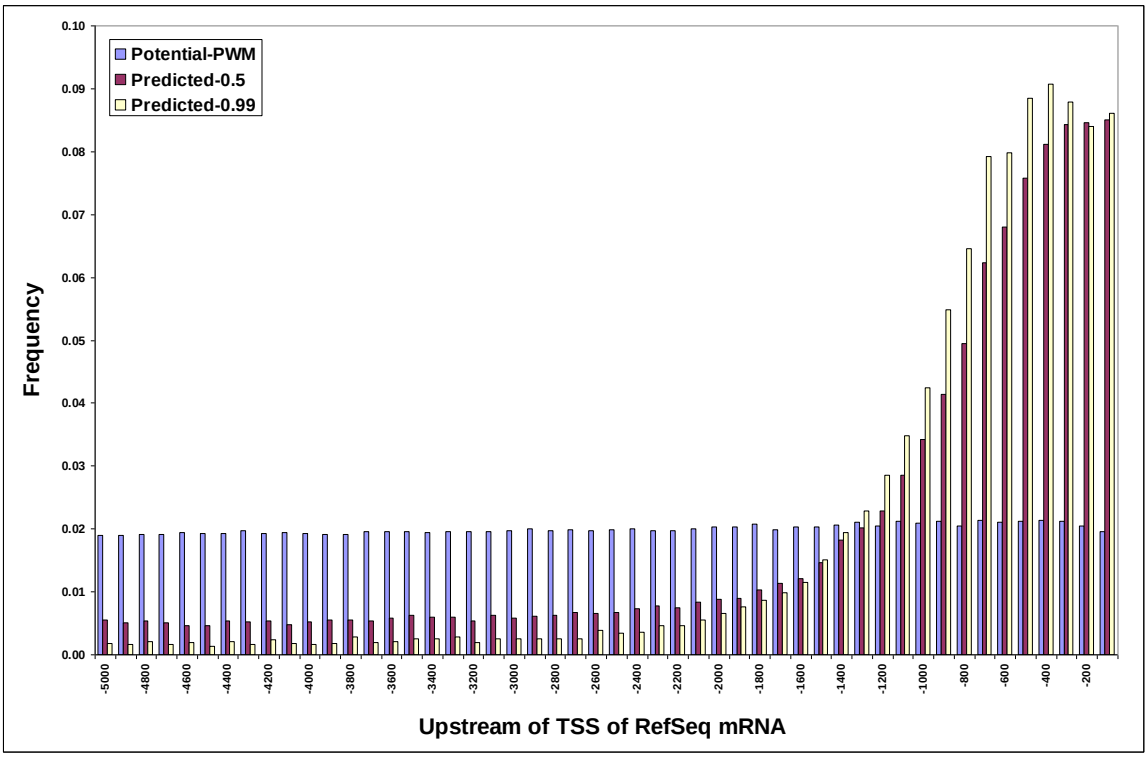


Figure 4.2 - Location distribution of ZNF217-BSs

By scanning the 5 kb upstream of the TSSs of 23,105 RefSeq mRNAs in the human genome, 360,601 potential ZNF217-BSs were identified by the PWM scoring method. A total of 49,703 predicted ZNF217-BSs were generated using the optimized prediction model with default prediction threshold (0.5) and 20,790 predicted ZNF217-BSs were generated using stringent prediction threshold (0.99). The data are represented as histograms of frequency at each 100 bp interval.

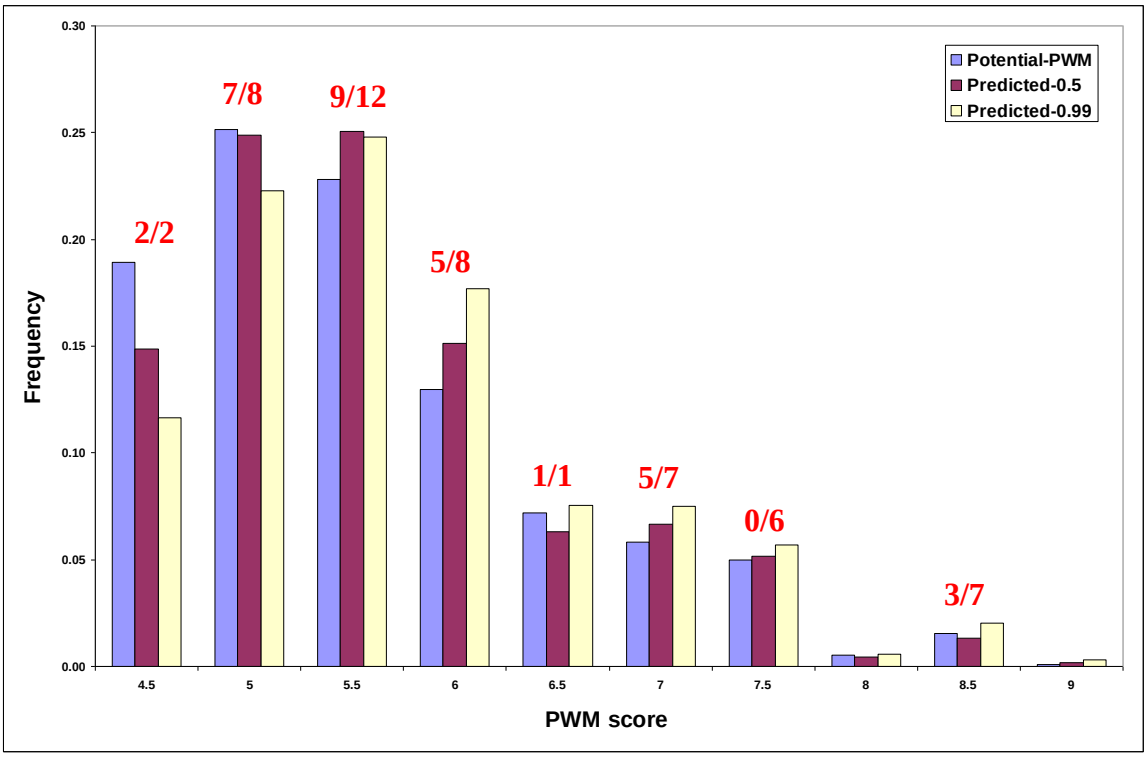


Figure 4.3 - PWM scores distribution of ZNF217-BSs

By scanning the 5 kb upstream of the TSSs of 23,105 RefSeq mRNAs in the human genome, 360,601 potential ZNF217-BSs were identified by the PWM scoring method. A total of 49,703 predicted ZNF217-BSs were generated using the optimized prediction model with default prediction threshold (0.5). In addition, fifty one experimentally identified ZNF217 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, "7/8" means that there are four ZNF217 target genes with PWM scores in the range of 4.5 to 5, and seven of these genes are correctly identified by the optimized kernel-based prediction model.

4.10 Tables

Table 4.1 – Comparison of ZNF217 and USF1 predictions

	ZNF217	USF1
TFBS[1] size	8 bp	10 bp
PWM Threshold	4.256	3.753
Training dataset		
Positive loci	51	34
Positive TFBS	199	135
Negative loci	439	177
Negative TFBS	1,170	800
Locus length	0.5 Kb	1 - 1.5 Kb
Feature selection		
Feature (p value)	DNaseI HS (< 0.0001)	DNaseI HS (< 0.0001)
		PhastCons8 (0.0022)
	RP5 (< 0.0001)	RP5 (0.0067)
	PWM score (< 0.0001)	PWM score (0.0081)
Cross-validation		
AUC	0.918	0.827
Sensitivity	60.8 %	55.9 %
Specificity	97.5 %	87.6 %
Genome-wide prediction		
TFBS (PWM)	360,601	290,614
TFBS (optimized)	49,703	24,010
SNP in TFBS	1,346	751
Gene	10,786	9,721
Stringent threshold	0.99	0.99
Candidate genes	5,213	5,801
Validation		
Robust genes	51	20
Predicted %	62.7 %	80.0 %

[1] Transcription factor binding site

Table 4.2 - Predicted ZNF217-BSs and associated target genes in the human genome

The optimized prediction model was applied to the 5 kb regions upstream of the TSSs of 23,105 RefSeq mRNAs. 10,786 genes with 49,703 ZNF217-BSs were predicted to be the targets of ZNF217. We only show a subset here as an example.

RefSeq mRNA	Gene	Location[1]	ZNF217-BS	Prediction score	chr[2]	start	end	SNP
NM_024796	FLJ22639	-254	CTTCCACC	0.8275	1	802,997	803,004	
NM_024796	FLJ22639	-826	AATCCACC	0.7665	1	803,569	803,576	
NM_015658	NOC2L	-22	CTTCCGGC	0.7165	1	934,790	934,797	
NM_015658	NOC2L	-622	CTTCCCGC	0.7929	1	935,390	935,397	
NM_198317	KLHL17	-592	AATCCGCC	0.9102	1	935,517	935,524	
NM_198317	KLHL17	-450	CTTCCTCC	0.8256	1	935,659	935,666	
NM_198317	KLHL17	-93	ATTCTTCC	0.909	1	936,016	936,023	
NM_015658	NOC2L	-3497	CTTCCTGC	0.8592	1	938,265	938,272	
NM_032129	PLEKHN1	-1356	TTTCCCGC	0.768	1	940,587	940,594	
NM_005101	ISG15	-262	ATTCCAGA	0.8696	1	988,683	988,690	
NM_207356	C1orf174	-47	ATTCCCCC	1	1	3,840,047	3840054	rs4072725
NM_207356	C1orf174	-698	CTTCCCGC	1	1	3,840,698	3840705	rs35759189
NM_207356	C1orf174	-1006	TTTCCCGC	1	1	3,841,006	3841013	

[1] The upstream of the TSS of RefSeq mRNA.

[2] Chromosome.

Table 4.3 - Validation of 51 robust ZNF217 target genes

We evaluated the target gene prediction results by comparing to 54 robust ZNF217 target genes identified from both the ZNF217 ChIP-chip study and ZNF217 knockdown analysis (Krig et al., 2007). Only 51 out of 54 robust genes had related RefSeq mRNAs included in our genome-wide prediction. Our prediction method was able to identify 32 out of these 51 ZNF217 target genes (62.7%).

No	Gene	Regulated by ZNF217	Correctly predicted
1	ARMCX5	Activated	Yes
2	CCNE2	Activated	Yes
3	GAD1	Activated	Yes
4	GATA4	Activated	No
5	KRT18	Activated	Yes
6	PUNC	Activated	No
7	SOCS2	Activated	Yes
8	ST3GAL6	Activated	No
9	ST6GAL1	Activated	Yes
10	STRA6	Activated	No
11	WNT5B	Activated	No
12	ZNF616	Activated	Yes
13	ABHD7	Repressed	No
14	ADM	Repressed	Yes
15	ANK3	Repressed	Yes
16	ATP10D	Repressed	Yes
17	CCL2	Repressed	No
18	COL8A1	Repressed	No
19	CREB5	Repressed	Yes
20	CXXC4	Repressed	Yes

Table 4.3 Continued

21	DSCR1	Repressed	Yes
22	EVA1	Repressed	Yes
23	GPRC5A	Repressed	No
24	HAPLN1	Repressed	Yes
25	HOXC6	Repressed	Yes
26	IFI16	Repressed	Yes
27	IGFBP3	Repressed	Yes
28	ITM2A	Repressed	Yes
29	KLHL4	Repressed	No
30	LMO3	Repressed	No
31	LYPD1	Repressed	Yes
32	MAP2K5	Repressed	Yes
33	MYCBP2	Repressed	No
34	NLGN1	Repressed	Yes
35	NMNAT2	Repressed	No
36	NMU	Repressed	No
37	NRK	Repressed	No
38	NRXN3	Repressed	Yes
39	PAK3	Repressed	Yes
40	PIPOX	Repressed	No
41	PKP2	Repressed	No
42	PLAT	Repressed	Yes
43	RGS20	Repressed	Yes
44	SEC14L2	Repressed	Yes
45	SEMA3A	Repressed	Yes
46	SH3RF2	Repressed	No
47	SLC6A15	Repressed	No
48	SPAG9	Repressed	Yes
49	SPG3A	Repressed	Yes
50	TP53AP1	Repressed	Yes
51	VSNL1	Repressed	Yes

Table 4.4 – Functional annotation of ZNF217 candidate genes

To further understand the biological roles of 5,213 ZNF217 candidate genes (with prediction score higher than 0.99), we used the web application, DAVID, to perform functional annotations of these genes. The DAVID analysis showed enrichment for genes with twelve diseases in Genetic Association Database (p = 0.00098 – 0.05) (Becker et al., 2004). And nine of these diseases were different type of cancers.

No	Disease	Fold Enrichment	P-Value
1	bladder cancer leukemia lung cancer	4.5	0.00098
2	bladder cancer	1.6	0.0029
3	head and neck cancer lung cancer	3.6	0.0052
4	skin cancer, non-melanoma	2.2	0.0057
5	lung cancer	1.4	0.0065
6	esophageal cancer	1.8	0.012
7	leukemia; bladder cancer; radiotherapy	4.3	0.015
8	cytogenetic studies	2	0.02
9	breast cancer	1.2	0.022
10	diurnal preference	5.2	0.024
11	ALS/amyotrophic lateral sclerosis	2.1	0.037
12	bladder cancer; cytogenetic studies	3.2	0.05

Table 4.5 - Predicted ZNF217 candidate genes differentially expressed in aorta

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 atherosclerosis (Seo et al., 2004). By combining our *in silico* ZNF217-BS prediction method with this expression result we identified 43 ZNF217 target genes that were differentially expressed between cases and controls in aorta.

No	Gene	Description
1	AKR1B1	ALDO-KETO REDUCTASE FAMILY 1, MEMBER B1 (ALDOSE REDUCTASE)
2	ARHGAP4	RHO GTPASE ACTIVATING PROTEIN 4
3	ASB1	ANKYRIN REPEAT AND SOCS BOX-CONTAINING 1
4	ATP1B1	ATPASE, NA+/K+ TRANSPORTING, BETA 1 POLYPEPTIDE
5	ATP6V1B2	ATPASE, H+ TRANSPORTING, LYSOSOMAL 56/58KDA, V1 SUBUNIT B2
6	ATP6V1F	ATPASE, H+ TRANSPORTING, LYSOSOMAL 14KDA, V1 SUBUNIT F
7	ATXN1	ATAXIN 1
8	BCAP31	B-CELL RECEPTOR-ASSOCIATED PROTEIN 31
9	CBX7	CHROMOBOX HOMOLOG 7
10	CDKN1B	CYCLIN-DEPENDENT KINASE INHIBITOR 1B (P27, KIP1)
11	ELL2	ELONGATION FACTOR, RNA POLYMERASE II, 2
12	FBLN5	FIBULIN 5
13	FOSB	FBJ MURINE OSTEOSARCOMA VIRAL ONCOGENE HOMOLOG B
14	GM2A	GM2 GANGLIOSIDE ACTIVATOR
15	HADH	L-3-HYDROXYACYL-COENZYME A DEHYDROGENASE, SHORT CHAIN
16	ITPR2	INOSITOL 1,4,5-TRIPHOSPHATE RECEPTOR, TYPE 2
17	KIAA0247	KIAA0247
18	LSP1	LYMPHOCYTE-SPECIFIC PROTEIN 1

Table 4.5 Continued

19	MAGED2	MELANOMA ANTIGEN FAMILY D, 2
20	MAPKAPK3	MITOGEN-ACTIVATED PROTEIN KINASE-ACTIVATED PROTEIN KINASE 3
21	NR4A2	NUCLEAR RECEPTOR SUBFAMILY 4, GROUP A, MEMBER 2
22	PDXK	PYRIDOXAL (PYRIDOXINE, VITAMIN B6) KINASE
23	PKD2	POLYCYSTIC KIDNEY DISEASE 2 (AUTOSOMAL DOMINANT)
24	PLXND1	PLEXIN D1
25	PPIF	PEPTIDYLPROLYL ISOMERASE F (CYCLOPHILIN F)
26	PSMB10	PROTEASOME (PROSOME, MACROPAIN) SUBUNIT, BETA TYPE, 10
27	RUNX1	RUNT-RELATED TRANSCRIPTION FACTOR 1
28	RYK	RYK RECEPTOR-LIKE TYROSINE KINASE
29	SCARB1	SCAVENGER RECEPTOR CLASS B, MEMBER 1
30	SEC14L1	SEC14-LIKE 1 (S. CEREVISIAE)
31	SH2B3	LYMPHOCYTE ADAPTOR PROTEIN
32	SH3BGR	SH3 DOMAIN BINDING GLUTAMIC ACID-RICH PROTEIN
33	SOD1	SUPEROXIDE DISMUTASE 1, SOLUBLE (AMYOTROPHIC LATERAL SCLEROSIS 1
34	SPINT2	SERINE PEPTIDASE INHIBITOR, KUNITZ TYPE, 2
35	STK10	SERINE/THREONINE KINASE 10
36	SVIL	SUPERVILLIN
37	TAF10	TAF10 RNA POLYMERASE II, TATA BOX BINDING PROTEIN (TBP)-ASSOCIATED FACTOR
38	TOM1	TARGET OF MYB1 (CHICKEN)
39	TP53BP2	TUMOR PROTEIN P53 BINDING PROTEIN, 2
40	UBE2N	UBIQUITIN-CONJUGATING ENZYME E2N (UBC13 HOMOLOG, YEAST)
41	UPP1	URIDINE PHOSPHORYLASE 1
42	WFS1	WOLFRAM SYNDROME 1 (WOLFRAMIN)
43	ZIC1	ZIC FAMILY MEMBER 1 (ODD-PAIRED HOMOLOG, DROSOPHILA)

CHAPTER 5

Conclusions

5.1 Summary

We focused on two TFs, USF1 and ZNF217, because of their biological importance, particularly in regard to their known genetic association with CAD, and the recent availability of ChIP-chip results. First, we used USF1 ChIP-chip data as a training dataset to develop and evaluate several kernel logistic regression prediction models. Our most accurate predictor achieved an AUC of 0.827 during cross-validation experiments, significantly outperforming standard PWM-based prediction methods. This final prediction model makes use of additional genomic features (phylogenetic conservation, regulatory potential, and DNaseI hypersensitivity) besides the PWM score. These variables were generated by backward stepwise feature selection in order to identify a subset of features significantly contributing to the model. This novel prediction method enables a more accurate and efficient genome-scale identification of specific USF1-BSs and associated target genes.

Other TFs show consistent evidence of association with complex disease. ZNF217 is known to repress the transcription of many genes and is associated with cell proliferation, survival, invasion, and over-expressed in multiple cancers. The results from independent genome-wide linkage and gene expression studies suggest that ZNF217 also may be a candidate gene for CAD. We further investigated the role of ZNF217 for CAD in three independent CAD samples. Our association studies of ZNF217 identified three SNPs (rs2741372, rs2766671 and rs1056948) having consistent association with cardiovascular disease in three independent samples with different phenotypes. Aorta expression profiling

indicated that the proportion of the aorta with raised lesions was also positively correlated to ZNF217 expression ($p = 0.0270$). Further allele-specific ZNF217 expression analysis identified three SNPs (rs16998248/A342A, rs35720349/I548T, and rs6097488) significantly associated with the level of ZNF217 expression ($p = 0.0045 – 0.042$) (Table 3.5). In addition, sex (male) and age were found to be positively correlated to ZNF217 expression. The combined evidence suggests that ZNF217 is a novel susceptibility gene for CAD with potential action through variation in the protein itself or allele-specific regulation of gene expression (Figure 5.1).

Finally, we applied our previously developed novel TFBS prediction method to ZNF217. Published ZNF217 ChIP-chip results were used to construct the training dataset and optimize the kernel logistic regression prediction model. Based on the cross-validation, this prediction model achieved 60.8% sensitivity, 97.5% specificity, and an AUC of 0.918. This optimized prediction model makes use of two additional genomic features (regulatory potential, and DNaseI hypersensitivity) besides the PWM score. The performance of the prediction models of ZNF217 and USF1 are very similar. We demonstrated that our TFBS prediction method can be extended to other TFs identified in human disease studies to further the understanding of encoded functional elements in the genome and their role in complex disease pathways. Because of the biological importance of ZNF217 (over-expression in many cancers and genetic association with CAD), the predicted ZNF217 targets from this study will help prioritize candidate genes as well as describe ZNF217's specific biological contribution to human diseases through its target genes.

In summary, the results of this dissertation research are (1) evaluation of two TFs, USF1 and ZNF217, as susceptibility factors for CAD with the evidence for ZNF217's role developed from prior data collection and analysis; (2) development of a generalized method for TFBS prediction; (3) prediction of TFBSs and target genes of two TFs, and identification of SNPs within TFBSs. This research allows for the development of study design to access TF-based gene-gene interactions in genetic susceptibility to human complex disease.

5.2 Future studies

The results of our TFBS prediction method applied on ZNF217 and USF1 indicate that this novel method significantly outperforms the standard PWM-based prediction method. We believe that our prediction method can be generalized and extended to other TFs identified in human complex disease studies. However, 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 methods, such as support vector machines. We focused on a region 5 kb upstream of the TSSs of RefSeq mRNAs because the published ChIP-chip studies indicated that most of the TF binding regions were found in proximal promoters. TFBSs could occur beyond 5 kb upstream of TSSs, implying that a wider range of genomic regions could be considered in the future. It

would be very helpful if we can develop a user-friendly software or web application based on our TFBS prediction method to allow easy application to other TFs of interest.

For this study, we focused on five genomic features related to regulatory elements currently available on a genome-wide scale. Backward stepwise feature selection during model building indicated that DNaseI HS was the most important predictor of USF1–BS among the features considered. As more data become available, especially DNaseI HS identified from multiple cell lines, we will be able to evaluate the cell/tissue specificity of the genomic features and our predictions. In addition, 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. The feature selection during model building will become even more important when we integrate more genomic features in the future.

In order to evaluate the performances of the prediction of ZNF217 targets, besides performing cross-validation with the training dataset we compared our prediction results with 51 robust ZNF217 target genes. Both of the datasets used for validation were generated from the same ZNF217 ChIP-chip study. Unlike the USF1 study, the binding sites from these robust genes were not validated by individual experiments. It would be helpful to compare our prediction with independent external robust genes. However, the target genes ZNF217 have not been well identified. According to the TRED database and PubMed, currently there is no individually experimental identified ZNF217 target gene.

We could compare our prediction with the ZNF217 target genes identified by ZNF217 ChIP-chip assay with a promoter array and ZNF217 knockdown RNA analysis with siRNA separately. We plan to perform these analyses in the near future. As more experimentally identified ZNF217 target genes become publicly available, we will be able to further evaluate our prediction and improve the prediction method.

As the result of this study, we made a genome-wide prediction of USF1 and ZNF217 targets. These predicted ZNF217 target genes can be used as candidates for studying susceptibility to complex human diseases. Our supplemental file includes a list of predicted ZNF217-BSs and their prediction score. This large number of predicted ZNF217-BSs along with the scores allow for adjusting the stringency of the prediction score threshold to refine candidate genes and also for choosing specific filters to emphasize a particular subset of interest. We also demonstrated the "genomic convergence" approach, which integrates several independent separate lines of experimental evidence to prioritize disease associated candidate genes. Further detailed functional annotation will help us obtain insight into how TFs contribute to human diseases through their target genes. Additional focused molecular biology experiments are necessary for understanding the biological roles of these candidate genes and identifying gene-gene interactions underlying complex disease. This work helps narrow the list from a huge one to a prioritized list, even for the focused experiments.

5.3 Figure

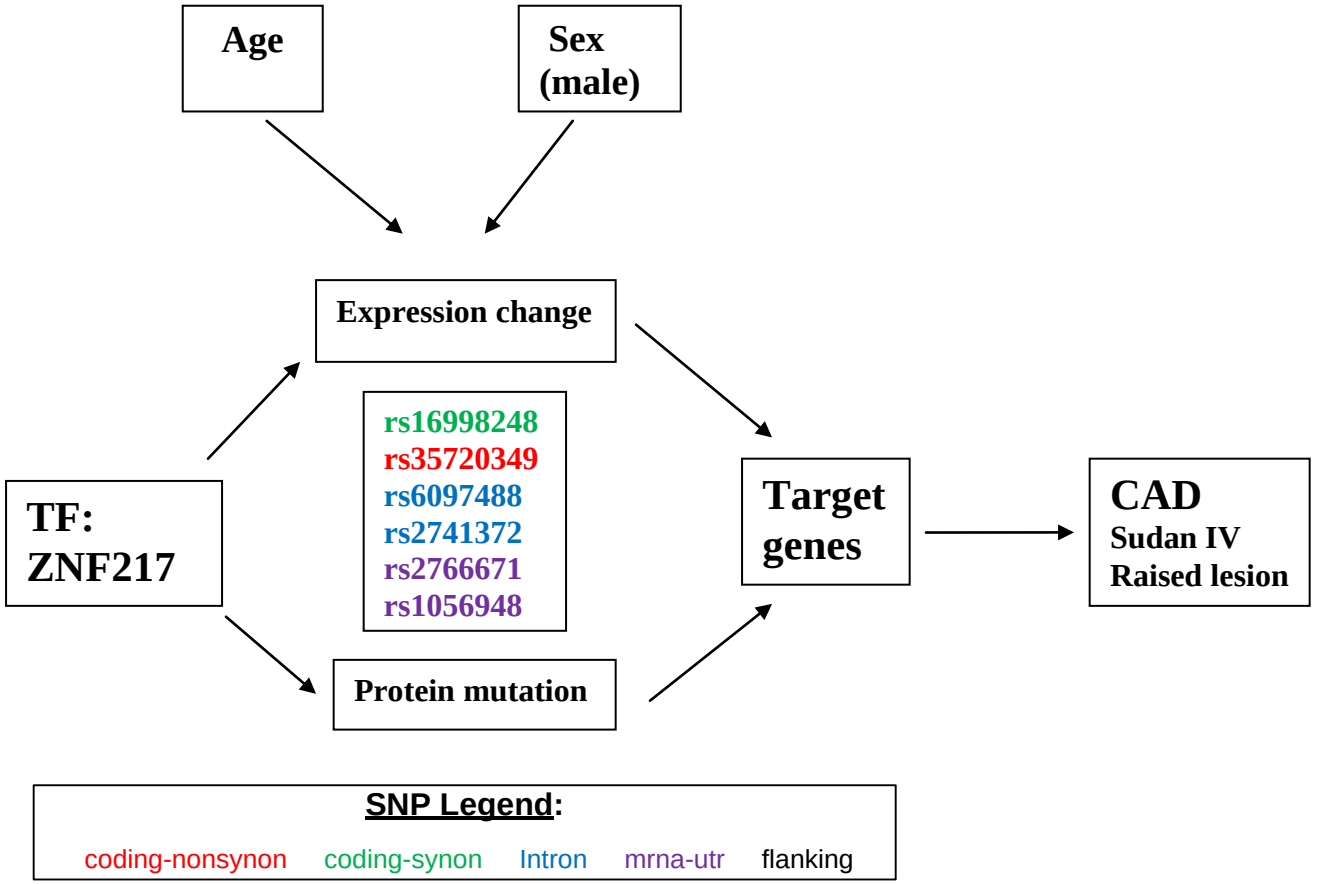


Figure 5.1 Hypothesis about CAD (Sudan IV and Raised lesion) association with ZNF217 (SNP and expression), Age, and Sex

APPENDIX

Analysis of Complex Disease Association and Linkage Studies Using the UCSC Genome Browser

Tianyuan Wang[1], Terrence S. Furey[2]

Circulation: Cardiovascular Genetics. 2009.

[1] Center for Human Genetics, Duke University Medical Center

[2] Institute for Genome Sciences and Policy, Duke University;

Abstract: Complex disease association and linkage studies, including genome-wide association studies (GWAS), generate multiple potentially large regions of interest that require further investigation to identify potential causative genes and elements in regions of association. The UCSC Genome Browser provides a tool, Genome Graphs, for this specific purpose providing efficient access to a wealth of publicly available genomic and disease-related information. This review employs data from a recently published GWAS concerning coronary artery disease to illustrate with detailed step-by-step procedures how to use the Genome Graphs tool to guide analyses based on association and linkage results in order to identify potential candidate genes.

Key words:

Computers, mapping, statistics, cardiovascular diseases, coronary disease

Introduction

The sequencing of the human genome, the identification of common SNPs and haplotype blocks, and advances in microarray technology have enabled the study of complex diseases at a level of detail not previously imaginable. These have aided in the design and analyses of association and linkage studies of many complex diseases including cardiovascular disease. Recent technological advances have enabled the undertaking of large-scale genome-wide association studies (GWAS) that can assay hundreds of thousands of polymorphic sites on hundreds to thousands of individuals in order to find genomic regions associated with disease. While results from these experiments enable the identification of smaller regions of association than previous studies, as with all linkage and association studies, there is the need for the further investigation of regions of interest for the casual gene or variant.

The purpose of this review is to present a detailed demonstration as to how publicly available resources can be utilized to easily guide more detailed research into genomic regions of interest identified in linkage and association study data. Large-scale projects, such as the Human Genome Sequencing project[1, 2], have generated large volumes and varieties of annotated genomic data necessitating the development of Internet-based tools to organize and make practically available these public data. One important tool in human disease research is web-based graphical genome browsers that use the human genome sequence as the framework on which to organize genomic annotations providing various ways for researchers to view and extract important information. Currently, there are three

human genome browsers that have been developed for public use: 1) the National Center for Biotechnology Information (NCBI) Map Viewer[3]; 2) the University of California Santa Cruz (UCSC) Genome Browser[4]; and 3) the European Bioinformatics Institute's Ensembl system[5]. Although these genome browsers share common features and genomic information, each being built on top of the same reference genome sequence, they each have a different look and feel and provide unique capabilities[6].

In particular, the UCSC Genome Browser has a tool called Genome Graphs that is especially suited for linkage and association study analyses. The following sections will demonstrate the capabilities of this tool focusing on a recently published GWAS result from the Wellcome Trust Case Control Consortium (WTCCC). As with all studies of this type, regions of disease association were identified encompassing large numbers of genes that are candidates for further studies. In order to prioritize future research, genes in each region need to be investigated for a possible role in a particular disease. The following step-by-step instructions demonstrate a straightforward and efficient method using the Genome Graphs tool within the UCSC Genome Browser that can help to prioritize a small number of meaningful candidates from a large-scale association study. Figure 1 provides an illustrated outline of this method with a more detailed description of each step below (Note: Each of the individual figures in Figure 1 is also available as larger figures in the Supplementary material).

Data

Numerous previous family-based linkage studies and case-control single marker association studies have indicated a strong genetic component to cardiovascular disease. Currently, about 40 quantitative trait loci (QTL) for human atherosclerotic disease have been identified by genetic linkage studies[7]. Large-scale association studies also identified several genes, such as LTA[8], VAMP 8 and HNRPUL1[9], and CDKN2A and CDKN2B[10, 11, 12].

Coronary artery disease (CAD) is one of the complex diseases studied by the WTCCC[12]. In their study, Affymetrix GeneChip 500K Mapping Arrays were used to identify seven regions of the genome showing evidence of association with CAD. The full dataset is only available with permission of the WTCCC, so we created a synthetic dataset (Supplemental file 1) for the NCBI build 36 human genome sequence based on the most significant SNP within each of the seven regions displayed in Tables 3 and 4 in their manuscript. Each of these SNPs is assigned its reported –log10 of *P* value to represent the statistical significance of the SNP and is calculated based on the allele distribution in cases and controls. To reflect the full extent of the identified region of association, we added SNPs at the edges of their identified regions to our dataset, each with a value of 3.51. Lastly, we added background SNPs (~2 Mb away from each side) with value 0. Readers are encouraged to download this synthetic dataset (Supplemental file 1) and to use it to actively perform each of the following steps in order to better understand the functionality of the Genome Graphs tool.

Step 1: Upload linkage/association results

First, proceed to the **UCSC Genome Browser** homepage (http://genome.ucsc.edu - Step 1, Figure 1, top image). Links to several tools available at this site can be seen in blue horizontal and vertical tool bars. For more information on the functionality of these other tools, we encourage exploring the **FAQ** and **Help** pages and reading a recent review describing features of this browser[13].

Click on the **Genome Graphs** link on the left vertical pane. In the page that appears (Step 1, Figure 1, middle image), data from association or linkage studies can be input. Up to two datasets can be uploaded and displayed simultaneously. The bottom section of this web page describes briefly the page controls, and there is a link to the **Genome Graphs User's Guide** that provides a more detailed set of instructions for this tool.

Click the **upload** button in the upper box to display the **Upload Data to Genome Graphs** page (Step 1, Figure 1, right image). On this page, information about the association data may be input such as a name and description. This tool will accept files of association or linkage data that are tab-delimited, comma delimited, or space delimited (see **file format** menu). Our test file is tab-delimited and simply consists of lines consisting of the name of a SNP and a corresponding value. The intent is that in general these values reflect some significance measure for that SNP, but there are no restrictions. This tool is aware of locations of several types of markers including SNPs denoted by *rs* values, and probes on several genome-wide genotyping microarrays from Affymetrix, Illumina, and Agilent (see **markers are** menu).

The association or linkage information to be displayed can be copied and pasted into the text box shown on this web page or can be uploaded as a file. The latter is recommend for very large datasets. Upload the test dataset and press the **submit** button. This will input the association results to be displayed in a graphical output page (Step 2, top figure). By default, the range of the dataset to be plotted will be obtained from the minimum and maximum values in the data. Alternatively, this display range can be specified by setting **display min value/max value** on the **Upload Data** page, or can be adjusted later (see next step).

Step 2: Display significant regions

Once the data is uploaded, first a summary text page appears indicating how many (%) markers within the data file were successfully mapped to the genome. Click the **OK** button to proceed. Next, the main **Genome Graphs** page appears again where the uploaded data can be selected for display in a genome-wide manner. Using the **graph** drop-down menu, select the track name corresponding to the newly input data. This will cause these data to be displayed on this same web page as a line graph on top of ideograms of each chromosome (Figure 1, Step 2, top panel). Seven peaks corresponding to the regions of significance are displayed directly above the appropriate chromosomes for this specific dataset. The height of the peak indicates its statistical significance, in this case the $-\log_{10}P$ value described above. Different features of this ideogram graph can be customized by clicking the **configure** button including the range of values to be shown.

From this display, clicking on any point of interest on any chromosome will open the main **Genome Browser** tool (described in the next section) displaying a 1 Mb region around that chromosomal base pair position. Alternatively, regions of association can be displayed in the **Genome Browser** by first specifying a **significance threshold** (3.5 for this dataset), and then pressing the **browse regions** button. The **Genome Browser** tool is displayed with a pane on the left containing links to significant regions that are above the given significance threshold (Figure 1, Step 2, bottom panel). In this dataset, it includes a total of 1.7 Mb from seven regions sorted by their genomic positions. Each region can be displayed on the **Genome Browser** on the right pane by clicking on the corresponding link. The first region on the list is shown by default. Click on the link for the region **chr9 21.9M to 22.2M**, the region with the most significant association, to show this area of the genome in the browser.

Step 3: Investigate significant regions

The **Genome Browser** tool presents a graphical view of a wide variety of annotations, particularly those related to genes, for a specific span of genomic sequence in the form of horizontal annotation "tracks" (Figure 1, Step 3). The genome is as presented runs from left to right with the shorter p-arm on the left. Genes are represented by solid boxes (exons) connected by lines (introns) with arrows indicating the direction of transcription. Scrolling down this page shows numerous drop-down menus that control the multitude of tracks that can be displayed. Currently, there are more than 200 annotations, some developed using

public data and/or research performed at UCSC, and others contributed from outside resources by third party researchers. These annotations are organized into categories, such as **Genes and Gene Prediction**, **Regulation, Comparative Genomics,** and **Variation and Repeats**[13, 14]. Annotation tracks most relevant to linkage and association studies include **UCSC Genes, 7X Reg Potential** (regulatory potential based on cross-species alignments)**, Conservation** (can select what other species to view)**, Most Conserved, SNPs (129)** (from dbSNP), and **HapMap LD Phased** (the association of alleles on chromosomes)**.** Any specific track can be displayed by selecting any visibility option (**dense, squish, pack, full)** other than **hide** from the drop-down list under that track name and then pressing any of the **refresh** buttons to update the display. These options primarily control whether each element in the track is distinctly displayed or is summarized in a single line (Please see http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#TRACK_CONT for more detailed descriptions of the different display options). In addition, navigation buttons are present along the top of this page that allow the zooming in and out of regions displaying more or fewer bases, and scrolling left and right along the chromosome. More detailed instructions concerning the functionality of this tool can be found in the **Help** section and also in recent reviews[13, 14].

Individual genes in regions of interest can be easily investigated within the **Genome Browser**. The currently displayed region on chromosome 9 (Figure 1, Step 3) indicates that two well-annotated genes, a pair of cyclin dependent kinase inhibitors, CDKN2A and

CDKN2B, are fully contained in the corresponding sequence representing potential candidate genes related to CAD. In fact, the original WTCCC analysis of this region focuses on these two genes. The multiple instances of each gene in the browser display correspond to alternatively spliced isoforms. Darker shades of blue indicate strong supporting evidence for the correct annotation of that isoform, while the lighter shades indicate less confidence. A general feature of the browser is that clicking on any element in any annotation track will display more a detailed description of that element. Genes in particular have a rich collection of information available including links to other databases and online resources as described in the following section.

Step 4: Research individual genes

To investigate these genes of interest further, click on one of the genes (CDKN2A) within the graphic on the browser page to open the **Human Gene Description and Page Index** page (Figure 1, Step 4, top and middle). At the top of this page are a brief description of the gene and a summary of what is currently known about its biological function taken from the Reference Sequence (RefSeq) project[15]. In addition, to facilitate investigation into potential associations with the disease in question, several diverse types of detailed information are provided such as links to other genomic tools and databases, results from genetic association studies, microarray gene expression data, mRNA and protein structure models and information, Gene Ontology (GO) annotations, and biochemical and signaling pathways in which the gene participates. Each of these sections

on this web page can be viewed either by scrolling down or using the **Page Index** table of links to directly jump to information of interest.

For CDKN2A and CDKN2B, the **Genetic Association Studies of Complex Diseases and Disorders** section (**Genetic Associations** link in the **Page Index** table) indicates that these have been previously linked to many types of disease including several cancers (Figure 1, Step 4, middle). Of particular interest, though, are the genetic association studies have linked both to myocardial infarction (click on "click on here to view complete list" link, see item 12[10]). To further understand the potential CAD association of CNKN2A and CNKN2B, clicking on the **myocardial infarct** link in the **Positive Disease Associations** list will open a page in the **Genetic Association Database** (GAD)[16] (Figure 1, Step 4, bottom). The GAD contains several published independent studies supporting an association between CAD and these two genes.

In addition to the GAD, several other publicly available resources contain valuable information related to disease association such as PubMed, OMIM, Entrez Gene, and GeneCards. Information in these can be easily accessed through links provided within the **UCSC Browser** gene description web page (Figure 4, Step 4, top panel) under the **Sequence and Links to Tools and Databases** section. Other sections in this page that may also be of interest are **Comparative Toxicogenomics Database (CTD)** that reports what chemicals have been shown to interact with this gene, **Microarray Gene Expression** that displays in what tissues and cell types the gene is expressed, and **Biochemical and Signaling Pathways** that lists in what general cellular processes the gene is involved.

Note, not all genes are necessarily as well-annotated as these and may not contain information in all of these sections.

In summary, by following the above described analysis pipeline within genome browser, we quickly and easily find two meaningful candidate genes, CNKN2A and CNKN2B, in one of the seven regions with evidence of association generated from a GWAS. Obviously, the other regions could and should be further investigated in a similar manner, and further experimentation is necessary to confirm and better understand the potential role of any particular gene in the disease.

Discussion

The UCSC Genome browser is a powerful online tool that integrates a large and diverse set of genomic data efficiently and intuitively providing much needed support for biomedical research, especially in the age of large-scale data intensive experiments such as genome-wide association studies. Using a specific example, we have illustrated how to use this Genome Browser to obtain well-supported candidate genes from a GWAS concerned with coronary artery disease. Admittedly, not all investigations using this method will quickly yield such informative results and is largely dependent on the accuracy of the association or linkage data and the previous research and annotations of genes. Nonetheless, we feel that this provides a concrete way in which to integrate the results of GWAS with the wealth of publicly available genomic data for further discovery.

This tutorial has only briefly introduced some of the functionality of the UCSC Genome Browser. A recent review provides a more in-depth description of this browser[13, 14, 17]. We also do not describe other human genome browsers hosted at the NCBI and EnsEMBL[3, 5, 18]. These also provide rich sets of genomic annotations and functionality that greatly but not completely overlap those available at UCSC. A general review comparing these three browsers is available[6, 19]. We encourage the further exploration of these websites to better understand these alternatives and their strengths. Researchers should decide for themselves the one that addresses their needs the best. We also caution that the quality of publicly available data displayed in the genome browsers and available in other public resources is highly variable. All these data should be viewed carefully and critically.

In the specific example we discussed here, we only included data in our synthetic results file from seven regions of the genome showing evidence of association with CAD. We surmise that researchers may want to upload more complete sets of raw or processed data generated from linkage and association studies to analyze within the Genome Graphs tool. Therefore, we need to point out that it takes a similar amount of time, about two minutes, to upload results that consist of 1 SNP or 500,000 SNPs. Therefore, analyses of large datasets are well with the capabilities of this tool.

Acknowledgments

We would like to give special thanks to Elizabeth Hauser for her thoughtful comments and discussions of this manuscript.

Funding sources

This study was supported by the Duke Institute for Genome Sciences & Policy (IGSP), and NIH grants HL073389 (Hauser), MH059528 (Hauser) and HL73042 (Goldschmidt, Kraus).

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001; 409:745-964.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan

M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E,

Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science. 2001; 291:1145-1434.

3. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2008; 36:D13-21 (Database issue).

4. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002; 12:996-1006.

5. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S. Ensembl 2008. Nucleic Acids Res. 2008; 36:D707-714 (Database issue).

6. Furey TS. Comparison of human (and other) genome browsers. Human Genomics. 2006; 2:266-270.

7. Chen Y, Rollins J, Paigen B, Wang X. Genetic and genomic insights into the molecular basis of atherosclerosis. Cell Metab. 2007; 6:164-79.

8. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet. 2002; 32:650–654.

9. Shiffman D, Rowland CM, Louie JZ, Luke MM, Bare LA, Bolonick JI, Young BA, Catanese JJ, Stiggins CF, Pullinger CR, Topol EJ, Malloy MJ, Kane JP, Ellis SG, Devlin JJ. Gene variants of VAMP8 and HNRPUL1 are associated with early-onset myocardial infarction. Arterioscler Thromb Vasc Biol. 2006; 26:1613–1618.

10. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A, Masson G, Gudbjartsson DF, Magnusson KP, Andersen K, Levey AI, Backman VM, Matthiasdottir S, Jonsdottir T, Palsson S, Einarsdottir H, Gunnarsdottir S, Gylfason A, Vaccarino V, Hooper WC, Reilly MP, Granger CB, Austin H, Rader DJ, Shah SH, Quyyumi AA, Gulcher JR, Thorgeirsson G, Thorsteinsdottir U, Kong A, Stefansson K. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science. 2007; 316:1491–1493.

11. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, Boerwinkle E, Hobbs HH, Cohen JC. A

common allele on chromosome 9 associated with coronary heart disease. Science. 2007; 316:1488–1491.

12. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678.

13. Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. UCSC genome browser tutorial. Genomics. 2008; 92:75-84.

14. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC genome browser Database: update 2009. Nucleic Acids Res. 2008 (Epub).

15. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005; 33:D501-504 (Database issue).

16. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004; 36:431-432.

17. Mangan ME, Williams JM, Lathe SM, Karolchik D, Lathe WC 3rd. UCSC genome browser: deep support for molecular biomedical research. Biotechnol Annu Rev. 2008; 14:63-108.

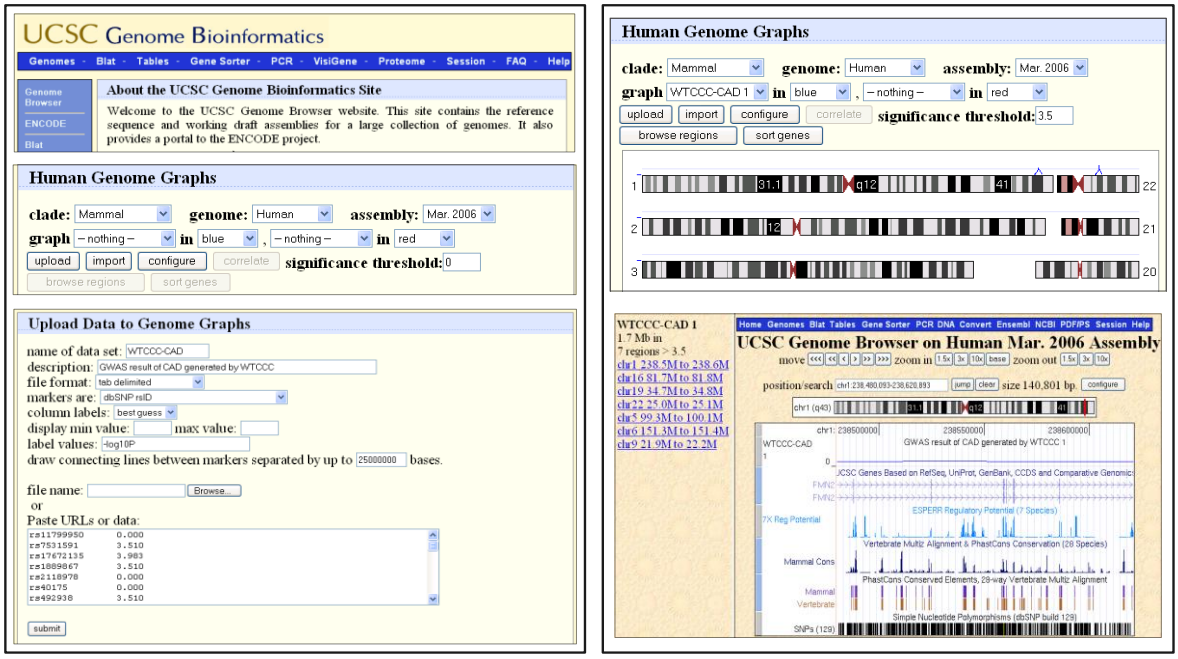
18. Spudich G, Fernández-Suárez XM, Birney E. Genome browsing with Ensembl: a practical overview. Brief Funct Genomic Proteomic. 2007; 6:202-219.

19. Baxevanis AD. Using genomic databases for sequence-based biological discovery. Mol Med. 2003; 9:185–192.

Figure legend

Step 1: Upload association result

Step 2: Display significant regions



Step 3: Investigate significant region

Step 4: Research individual gene

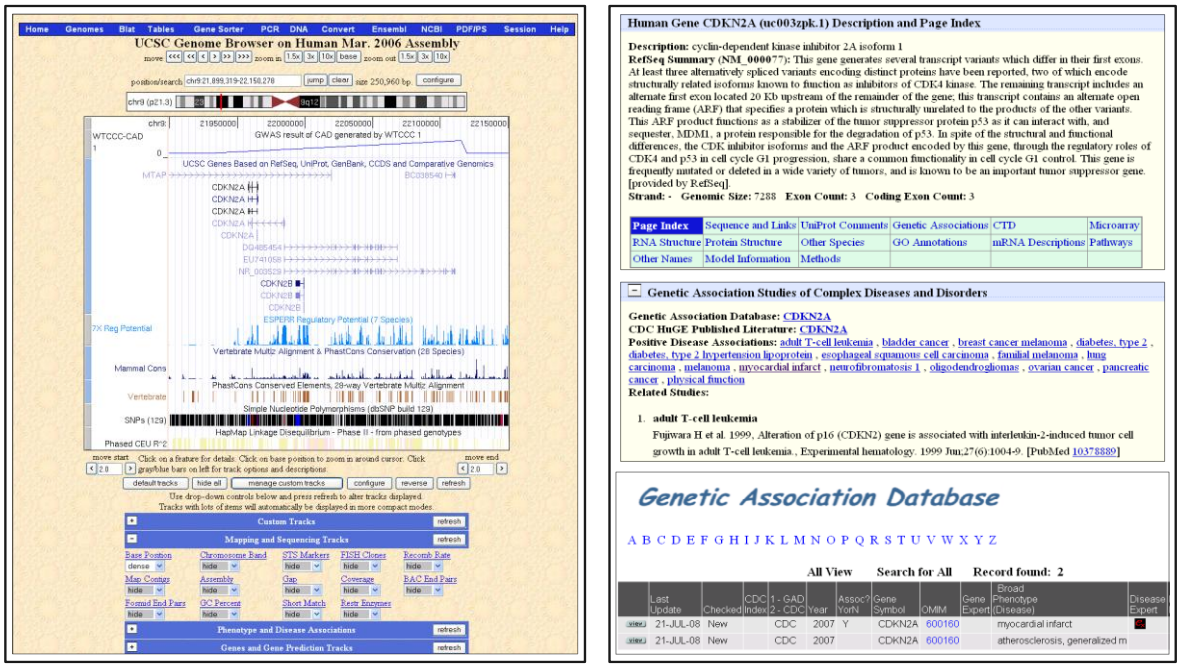
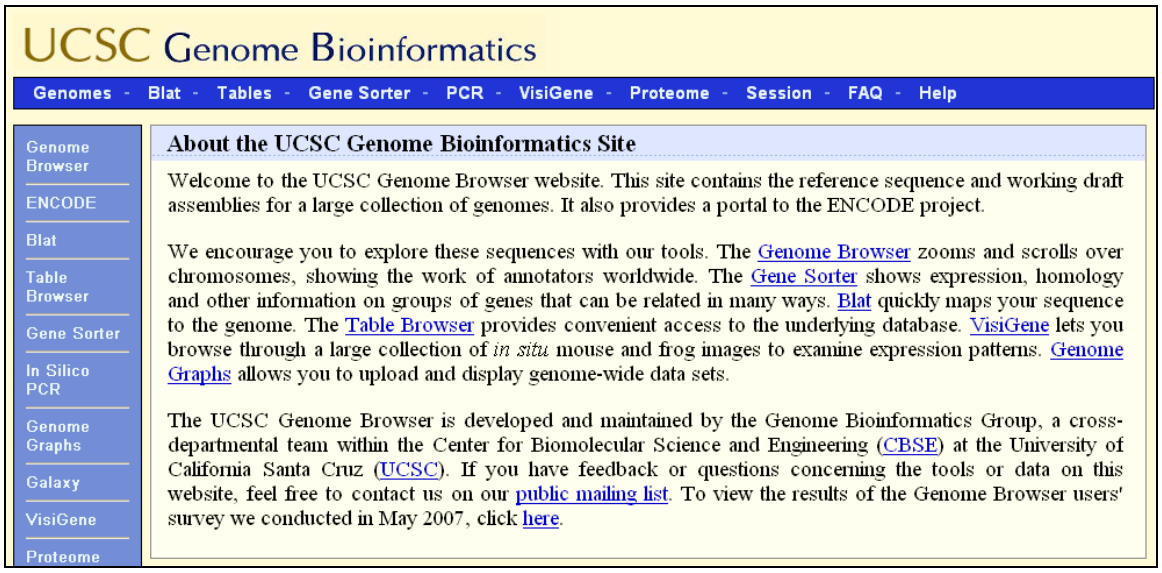


Figure 1: Flow chart of using UCSC genome browser to analyze GWAS results.

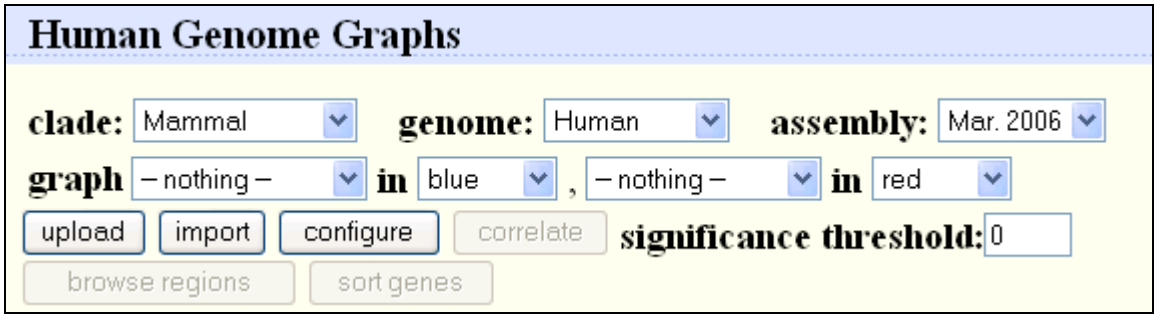
Supplementary material

Supplemental file 1: Example dataset for practicing the **Genome Graphs** tool

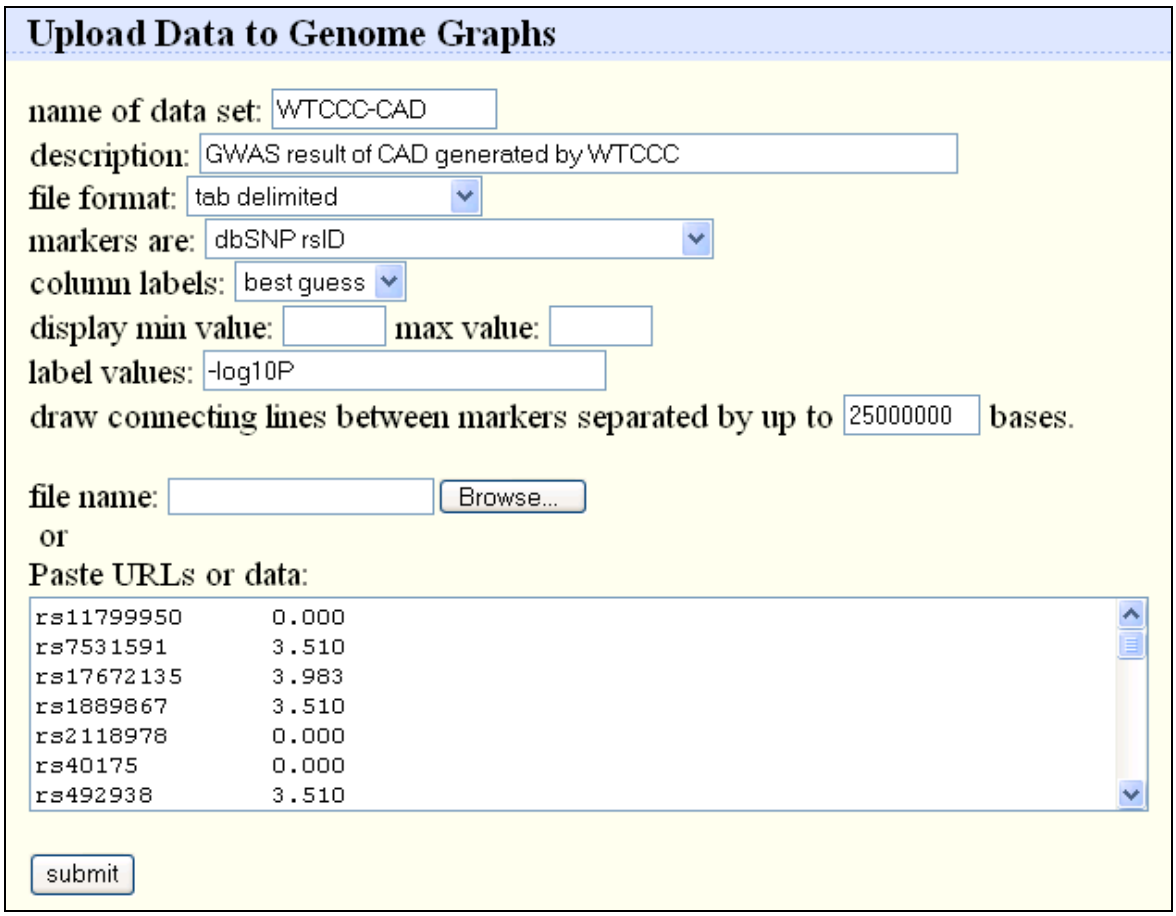
SNP	$-\log_{10}P$
rs11799950	0.000
rs7531591	3.510
rs17672135	3.983
rs1889867	3.510
rs2118978	0.000
rs40175	0.000
rs492938	3.510
rs383830	5.243
rs34162536	3.510
rs455144	0.000
rs11961921	0.000
rs36082661	3.510
rs6922269	5.199
rs505358	3.510
rs34091791	0.000
rs10124918	0.000
rs2811716	3.510
rs1333049	13.747
rs7020996	3.510
rs13300968	0.000
rs11860434	0.000
rs11648346	3.510
rs8055236	5.012
rs16959735	3.510
rs4782691	0.000
rs7339484	0.000
rs8100086	3.510
rs7250581	5.040
rs4805440	3.510
rs10403126	0.000
rs10610555	0.000
rs28643474	3.510
rs688034	5.161
rs5761483	3.510
rs11913617	0.000



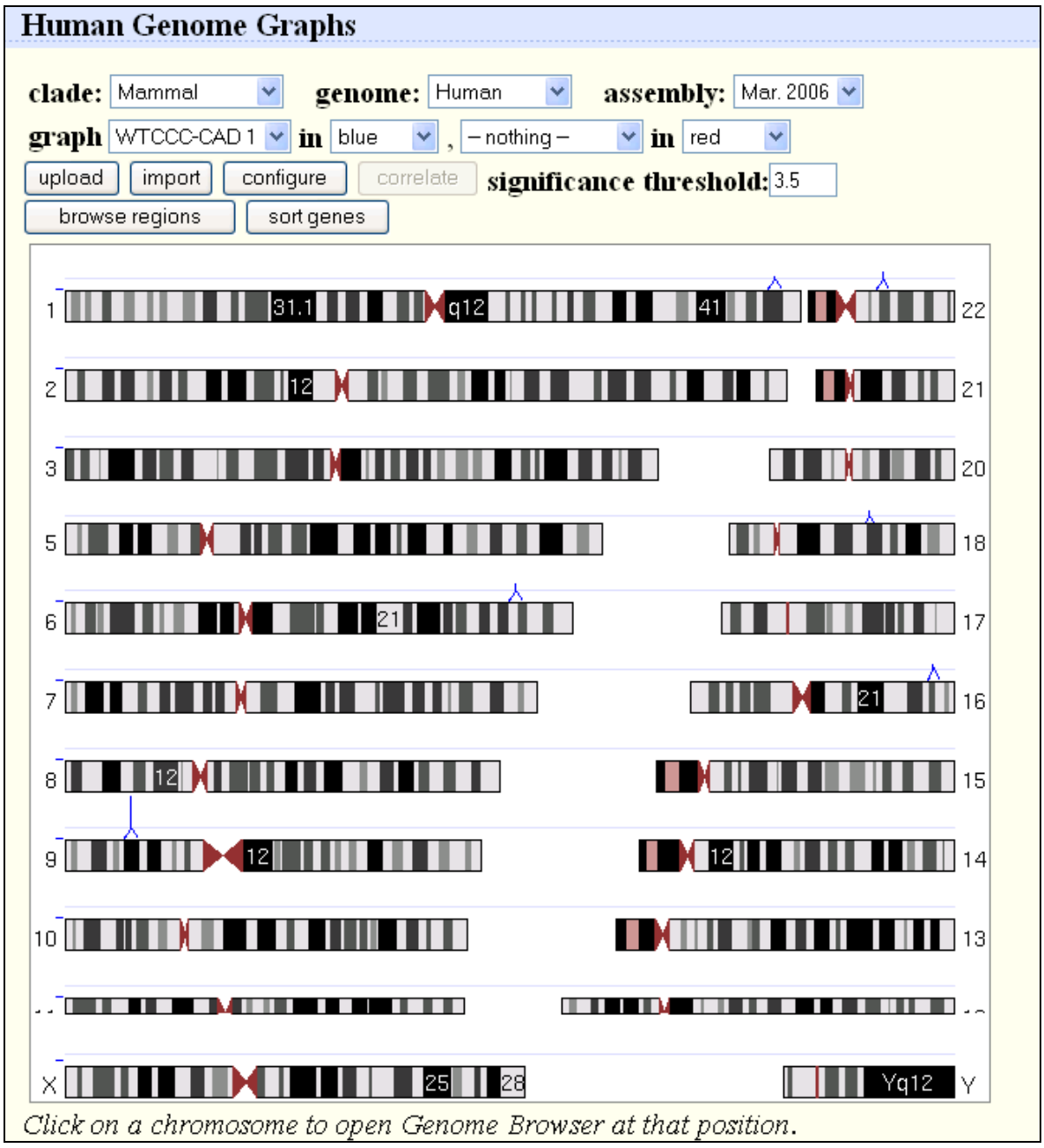
Supplemental file 2: Larger image for Figure 1: Step 1, top image.



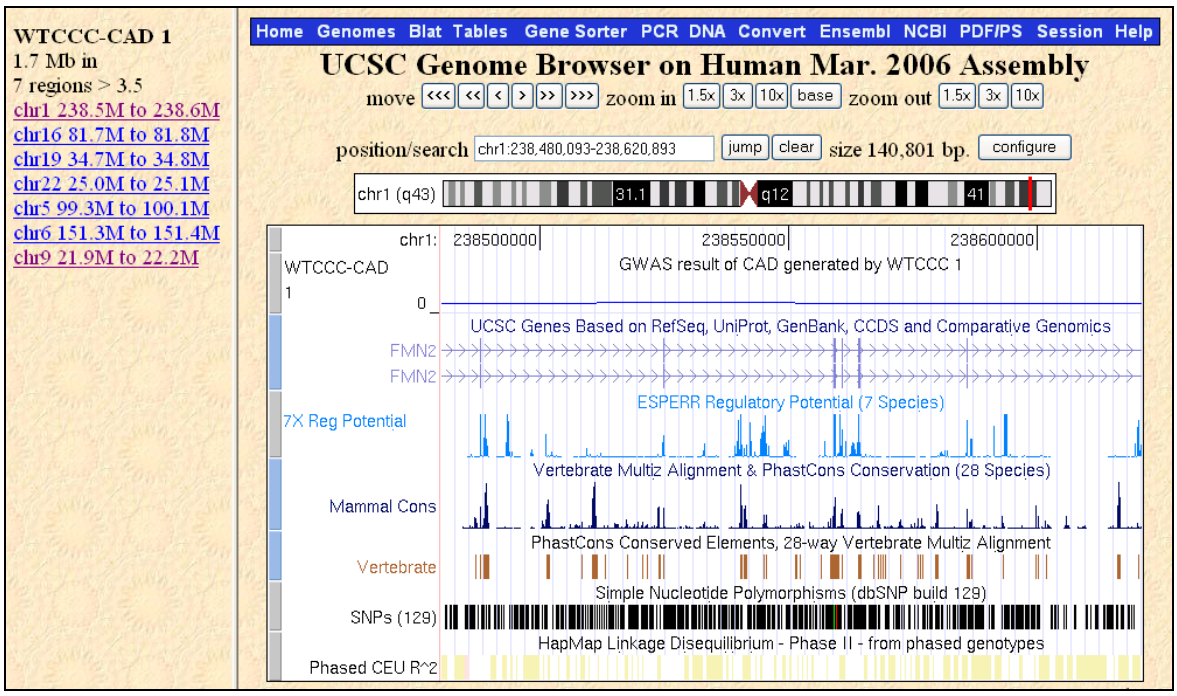
Supplemental file 3: Larger image for Figure 1: Step 1, middle image.



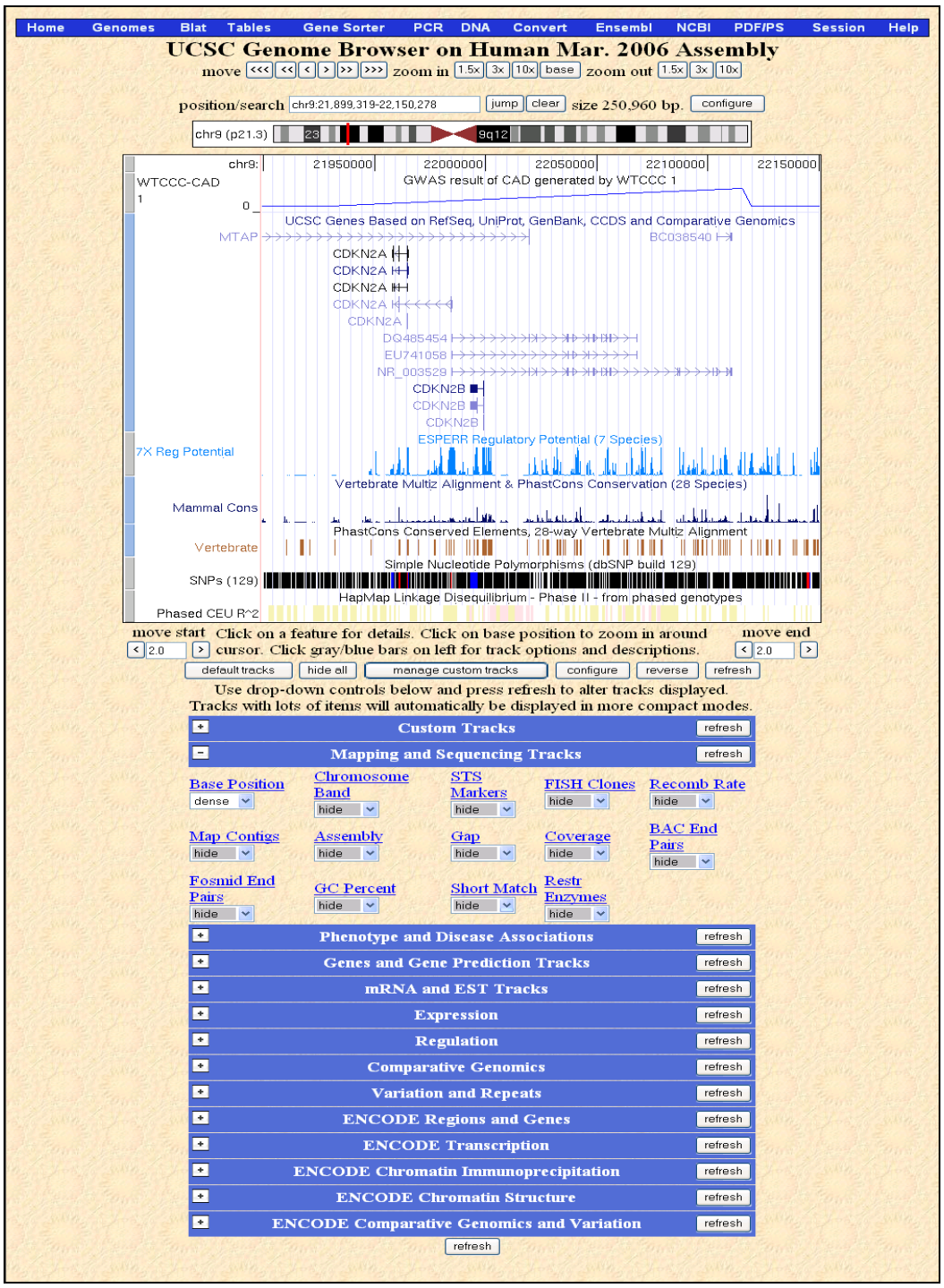
Supplemental file 4: Larger image for Figure 1: Step 1, bottom image.



Supplemental file 5: Larger image for Figure 1: Step 2, top image.



Supplemental file 6: Larger image for Figure 1: Step 2, bottom image.



Supplemental file 7: Larger image for Figure 1: Step 3.

Human Gene CDKN2A (uc003zpk.1) Description and Page Index

Description: cyclin-dependent kinase inhibitor 2A isoform 1
RefSeq Summary (NM_000077)**:** This gene generates several transcript variants which differ in their first exons. At least three alternatively spliced variants encoding distinct proteins have been reported, two of which encode structurally related isoforms known to function as inhibitors of CDK4 kinase. The remaining transcript includes an alternate first exon located 20 Kb upstream of the remainder of the gene; this transcript contains an alternate open reading frame (ARF) that specifies a protein which is structurally unrelated to the products of the other variants. This ARF product functions as a stabilizer of the tumor suppressor protein p53 as it can interact with, and sequester, MDM1, a protein responsible for the degradation of p53. In spite of the structural and functional differences, the CDK inhibitor isoforms and the ARF product encoded by this gene, through the regulatory roles of CDK4 and p53 in cell cycle G1 progression, share a common functionality in cell cycle G1 control. This gene is frequently mutated or deleted in a wide variety of tumors, and is known to be an important tumor suppressor gene. [provided by RefSeq].
Strand: - **Genomic Size:** 7288 **Exon Count:** 3 **Coding Exon Count:** 3

Page Index	Sequence and Links	UniProt Comments	Genetic Associations	CTD	Microarray
RNA Structure	Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Pathways
Other Names	Model Information	Methods			

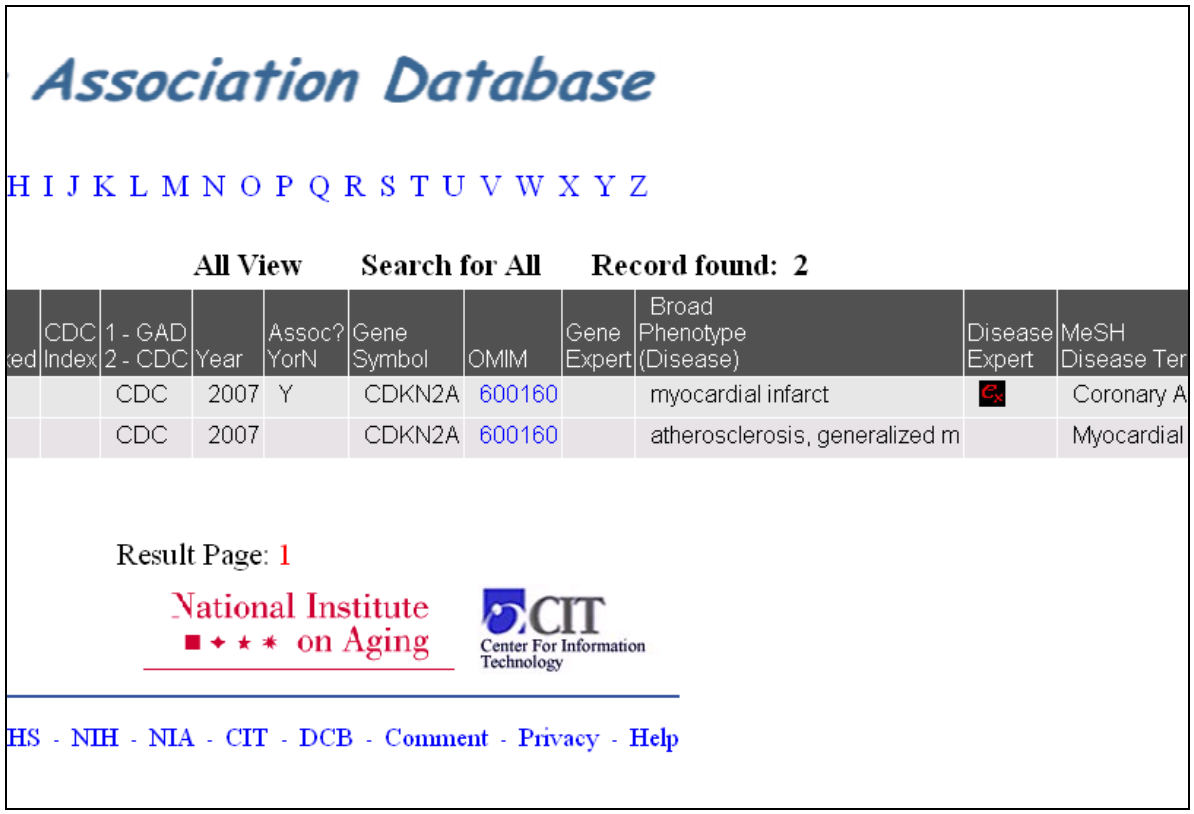
Supplemental file 8: Larger image for Figure 1: Step 2, top image.

Genetic Association Studies of Complex Diseases and Disorders

Genetic Association Database: CDKN2A
CDC HuGE Published Literature: CDKN2A
Positive Disease Associations: adult T-cell leukemia , bladder cancer , breast cancer melanoma , diabetes, type 2 , diabetes, type 2 hypertension lipoprotein , esophageal squamous cell carcinoma , familial melanoma , lung carcinoma , melanoma , myocardial infarct , neurofibromatosis 1 , oligodendrogliomas , ovarian cancer , pancreatic cancer , physical function
Related Studies:

1. **adult T-cell leukemia**
   Fujiwara H et al. 1999, Alteration of p16 (CDKN2) gene is associated with interleukin-2-induced tumor cell growth in adult T-cell leukemia., Experimental hematology. 1999 Jun;27(6):1004-9. [PubMed 10378889]
2. **bladder cancer**
   Sakano, S. et al. 2003, Clinical course of bladder neoplasms and single nucleotide polymorphisms in the CDKN2A gene., International journal of cancer. Journal international du cancer. 2003 Mar;104(1):98-103. [PubMed 12532425]
   Our results corroborate the earlier findings that single base mutation is not the prime mode of inactivation of the CDKN2A gene in bladder cancer. Further, the results indicate, a role for the 3' UTR polymorphisms in the CDKN2A gene in tumor invasiveness.
3. **breast cancer melanoma**
   Debniak, T. et al. 2006, MC1R common variants, CDKN2A and their association with melanoma and breast cancer risk, Int J Cancer 2006. [PubMed 16988943]

Supplemental file 9: Larger image for Figure 1: Step 1, middle image.



Supplemental file 10: Larger image for Figure 1: Step 4, bottom image.