Distribution Agreement

Signature:

_________________________ _______________
Li Chen Date

Statistical and informatics methods for analyzing next generation sequencing data

by

Li Chen
Doctor of Philosophy

Computer Science and Informatics

Zhaohui Steve Qin, Ph.D.
Advisor

Hao Wu, Ph.D.
Advisor

Peng Jin, Ph.D.
Committee member

Lee Cooper, Ph.D.
Committee member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical and informatics methods for analyzing next generation sequencing data

by

Li Chen
M.H.S., M.S.E. The Johns Hopkins University, 2011

Advisor: Zhaohui Steve Qin, Ph.D.
Advisor: Hao Wu, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2017

Abstract

Statistical and informatics methods for analyzing next generation sequencing data
By Li Chen

In the era of genomic big data, it is demanded to develop statistical and informatics methods for the analysis of big data. The integrative analysis of datasets generated from different sources or in different biological conditions is of particular interest. First, we develop a statistical method *ChIPComp* to perform quantitative comparison of multiple ChIP-seq datasets in different biological conditions. *ChIPComp* detects genomic regions showing differential protein binding or histone modification by considering data from control experiments, signal to noise ratios, biological variations, and multiple-factor experimental designs in a linear model framework. Simulations and real data analyses demonstrate that *ChIPComp* provides more accurate and robust results compared with existing methods. By utilizing tens of thousands of trait-associated GWAS SNPs cataloged, we present *traseR*, a computational tool that could explore the collection of trait-associated SNPs to indicate whether a given genomic interval or intervals is likely to be functionally connected with certain phenotypes or diseases. Real data results indicate that *traseR* offers a turnkey solution for enrichment analysis of trait-associated SNPs. Besides analyzing datasets from a single source (GWAS or epigenomics), we perform a joint analysis for multiple data sources by annotating GWAS SNPs using thousands of genomic and epigenomic datasets, and building *DIVAN*, a data-driven machine learning approach that aims to identify disease-specific noncoding risk variants in a genome-wide scale, which is helpful to understand the cryptic link between non-coding sequence variants and the pathophysiology of complex diseases/phenotypes. By being disease-specific, *DIVAN* demonstrates to be more powerful than competing methods in the identification of disease-specific noncoding risk variants.

Statistical and informatics methods for analyzing next generation sequencing data

by

Li Chen
M.H.S., M.S.E. The Johns Hopkins University, 2011

Advisor: Zhaohui Steve Qin, Ph.D.
Advisor: Hao Wu, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2017

Acknowledgments

First, I would like to thank my PhD advisor, Dr. Steve Qin for his support and guidance of my PhD study and research. Steve generously offered his inspiration and knowledge for helping me in seeking research topics and formalizing research problems. During the weekly meeting with him along the PhD study, I learnt lots of the scientific knowledge and how to carry out research in an independent way. I would like to thank my advisor Dr. Hao Wu, who led me to the area of next generation sequencing research by offering me the hands-on experience on data analysis and insightful suggestions on statistical modeling. Thanks to my dear advisors for working with me on paper writing and revision, which results the publication of the dissertation in high impact journals including *Genome Biology* and *Bioinformatics*. Discussion with my advisors is a rewarding experience, from which I learnt essential qualifications of being a good researcher such as persistence, curiosity and criticism. Moreover, I would like to thank the committee members Dr. Peng Jin and Dr. Lee Cooper for their helpful advice and useful comments on my dissertation.

I am also grateful to Dr. Peng Jin for offering me the opportunity to work in his lab as a part-time bioinformatician for three years. During this period, I performed bioinformatics analyses on various types of next generation sequencing data including ChIP-seq, RNA-seq and BS-seq from complex diseases, and the rewarding experience is of great help in the development of the dissertation.

One a personal note, I would like to thank my parents, for their love, constant encouragement and support all the time.

To my family

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

With the recent development of next generation sequencing (NGS), array-based technologies have being gradually replaced by sequencing-based technologies as NGS could dramatically improve the quantity and quality of high throughput genomic data compared to array-based technologies. NGS is growing more and more popular in different research areas of genomics and genetics, such as epigenetics and Genome-Wide Association Study (GWAS).

Among those NGS technologies, ChIP-seq is such as technology that combines Coupling chromatin immunoprecipitation (ChIP) and next-generation sequencing (seq), which gradually replaces ChIP-chip, a technology that combines chromatin immunoprecipitation (ChIP) with microarray (chip). Though both technologies have the advantage to investigate protein-DNA interaction *in vivo*. ChIP-seq is capable of revealing protein-DNA interaction in a whole genome-scale and could achieve high-resolution of profiling DNA-protein interaction, compared to ChIP-chip, which might introduce some bias, as an array is restricted to a fixed number of probes. Due to the advantage of ChIP-seq, it has been widely used in epigenetics research to study the mechanisms of gene expression change without the involvement of underlying DNA sequence change. Particularly, transcription factors (TFs) and histone marks play an important role in epigenetic modification on gene regulation. To be specific, it is common that TFs activate the gene expression by binding in the promoter or enhancer regions of the gene. Histone marks such as H3K4me1, H3K4me3, and H3K36me3 have the same role to activate the gene expression but differ in the modifications sites: H3K4me1 in enhancer regions, H3K4me3 in promoter regions while H3K36me3 in gene body. Other histone marks such as H3K27me3 usually repress the gene expression. By measuring the genome-wide profiling of either TF or histone modification, ChIP-seq helps reveal the status of gene regulation. Moreover, the status of gene regulation might differ in in different conditions, e.g. cell lines, tissues, cell cycle. Therefore, to compare genome-wide profiling of either TF or histone modification measured by ChIP-seq among multiple conditions will unveil the dynamic changes of gene regulation.

GWAS mainly focus on the study of associations between single-nucleotide polymorphisms (SNPs) and diseases or traits in different populations by setting up the case-control. The identified disease/trait-associated SNPs might be potential biomarkers. The high-throughput genomic technology also facilitates the GWAS, which utilizes high-density the SNP genotyping array, a type of DNA microarray, to detect the SNPs that are disease/trait-associated. With the decreasing cost of NGS, whole genome sequencing (WGS) has been frequently used in GWAS as WGS not only is presented as an alternative to, but also could overcome the limitation of genotyping array-based GWA studies.

First, WGS provides a whole genome coverage by investigating approximate 3 billion bases compared to 10 million SNPs covered by a typical genotyping array. Second, WGS provides more opportunities to discover the rare variants (MAF $< 1\%$) compared with that the genotyping array-based GWA studies focus on common variants (MAF $> 5\%$). Third, besides SNPs, WGS could detect other types of variants including structural variants and copy number variants (CNVs). Thus, due to the advantages of WGS, WGS grows more popular in the identification of genomic mutations including GWAS SNPs.

The accumulation of "omics" data from different data sources or data types provides an unique opportunity to investigate multiple datasets for novel biological discovery, which cannot be done by single dataset or datasets from single source. For GWAS, multiple data sources collect genetic variants associated with different diseases/traits in thousands of literature are publicly available such as Association Results Browser (`http://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm` and GRASP [1]. For epigenetics, the decreasing cost of NGS results in massive public available epigenetic datasets in different biological context. Large national consortiums such as ENCODE [2] and modENCODE [3] contain a comprehensive collection of genomic and epigenomic datasets including TF ChIP-seq, Histone modification ChIP-seq, Open Chromatin-seq (FAIRE-seq, DNase-seq) and DNA methylation across hundreds of cell lines. Roadmap Epigenomics ([4]) is another comprehensive consortium that collects Histone modification ChIP-seq and DNase-seq across hundreds of cell lines.

With those public datasets, it is possible to perform the study such as the change of epigenetic profiles among different cell types, the change of epigenetic profiles between risk SNPs and benign SNPs for the disease of interest. The complicated mechanisms in biological systems could only be unveiled in the era of "Big Data".

1.2 Outline of the dissertation

This thesis consists of three chapters that each addresses an independent statistical/informatics problem in high-throughput genomic data analysis.

An important problem in ChIP-seq data analysis is to detect loci that show differential binding for a transcription factor or histone modification across multiple conditions (e.g. between cancer and normal tissues). Most of the existing methods do not consider data from control experiment, signal to noise ratios, biological variations and multiple-factor experimental designs, which may lead to biased results. In chapter 2, we develop a statistical method *ChIPComp* to perform quantitative comparison of multiple ChIP-seq datasets. The key advantage of *ChIPComp* is that it considers data from control experiment, and is developed in a rigorous and coherent statistical framework. To be specific, the read counts from IP experiment at the candidate regions are assumed to follow Poisson distribution. The underlying Poisson rates are modeled as an experiment-specific function of artifacts and biological signals. Biological signals are estimated and compared through the hypothesis testing procedure in a linear model framework. Simulations and real data analyses demonstrate that the proposed method provides more accurate and robust results compared with existing ones.

Understanding the link between non-coding sequence variants especially SNPs, identified in GWAS, and the pathophysiology of complex diseases remains challenging due to a lack of annotations in non-coding regions. To overcome this, we develop a machine learning method *DIVAN* for the accurate identification of non-coding disease-specific risk variants using multi-omics profiles in chapter 3. *DIVAN* is essentially a novel feature selection and ensemble learning framework, which identifies disease-specific risk variants by leveraging a comprehensive collection of genome-wide epigenomic profiles across cell types and factors, along with other static genomic features. DIVAN accurately and robustly recognizes non-coding disease-specific risk variants under multiple testing scenarios; among all the features, histone marks, especially those mark repressed chromatin, are often more informative than others.

Tens of thousands of disease/trait-associated GWAS SNPs have already been cataloged, which we believe form a great resource for genomic research. Recent studies have demonstrated that the collection of trait-associated SNPs can be exploited to indicate whether a given genomic interval or intervals are likely to be functionally connected with certain phenotypes or diseases. Despite this importance, currently there is no ready-to-use computational tool able to connect genomic intervals to phenotypes. In chapter 4, we present *traseR*, an easy-to-use R Bioconductor package that performs enrichment analyses of trait-associated SNPs in arbitrary genomic intervals with flexible options, including testing method, type of background, and inclusion of SNPs in LD.

Chapter 2

ChIPComp : A novel statistical method for quantitative comparison of multiple ChIP-seq datasets

2.1 Introduction

Coupling chromatin immunoprecipitation (ChIP) and next-generation sequencing (seq), ChIP-seq is a powerful technology for profiling protein bindings or histone modifications in the whole genome scale. Since the introduction of the technology [5], a large number of experiments were conducted to create genome-wide profiles for many DNA-binding proteins and different types of histone modifications under various biological contexts, for example, by large national consortiums such as ENCODE [2] and modENCOD [3].

The main goal of analyzing data from a single ChIP-seq experiment is to detect protein binding or histone modification regions, often referred to as "peaks" . The raw data produced from ChIP-seq experiments are many short DNA segments called "reads". After aligning the reads to the reference genome, genomic regions with unusually large number of reads clustered are often deemed peaks. In recent years, a number of methods and software tools are developed for peak detection. Two benchmark studies have also been conducted to compare different peak calling methods [6,7]. With the continuous reduction of sequencing costs and the rapid accumulation of public data, it is now a common practice to compare data from different ChIP-seq experiments, for example, to compare the binding of certain protein under different biological conditions. Such analysis provides important information for studying the dynamics of epigenetic regulations. Results from the analysis can be further associated with other data, such as gene expressions, to better understand the gene regulation mechanisms.

The comparisons of ChIP-seq data have been widely performed. The most straightforward method is the "overlapping analysis", which is to compare the peaks called from different experiments and defines "common peaks" or "unique peaks", then represents them by Venn diagram [8]. This method, however, is highly dependent on the thresholds used for calling peaks. Genomic regions barely over the threshold in one sample but under the threshold in the other will be declared as unique peaks even if the quantitative difference is small. Moreover, it completely ignores the quantitative differences of peaks, that is, Genomic regions being peaks in both samples will be deemed common peaks, even if the quantitative difference is large. Due to these reasons, quantitative comparison is more desirable to compare ChIP-seq datasets.

The quantitative comparison of ChIP-seq can be performed by comparing the read counts among different experiments, which is similar to RNA-seq differential expression (DE) analysis. However it is a more complicated problem due to several reasons. First, the data from the IP experiments are affected by the genomic background, such as chromatin structures and DNA sequence. These backgrounds are non-uniform across the genome, and could be highly variable across different experiments. The backgrounds, measured by control experiments, need to be taken into account in quantitative comparison of multiple ChIP-seq datasets. Another complication arises from the different signal to noise ratios (SNRs) of the experiments. Many technical or biological artifacts contribute to SNRs. For example, sample with less binding sites will have taller peaks because reads are allocated into narrower genomic regions. Moreover, different SNRs may result from differences of antibody qualities, experimental protocols or lab technician skills, etc. Therefore, correctly accounting for SNRs is important in quantitative comparison of ChIP-seq. In addition, considerations for biological variance and experimental designs remain, similar to that in differential expression analysis of RNA-seq.

Quantitative comparison of ChIP-seq (often referred to as "differential binding" problem) has gained some interests recently, and several methods have been proposed for two-condition comparison. There are two methods take the approach to model the differences of normalized read counts from two IP experiments: ChIPDiff [9] applies hidden Markov model on the differences to identify differential histone modification regions, and DIME [10, 11] uses a finite exponential-normal mixture model on the differences to detect differential binding sites. However, neither the control experiment nor the biological replicates are considered in these methods. Moreover, these methods do not account for SNRs and cannot be easily extended for multiple condition comparison. MAnorm [12] and ChIPnorm [13] consider different SNRs. Both methods normalize the data before comparison: MAnorm performs normalization based on MA-plot, and ChIPnorm uses quantile normalization. But again, neither considers control data at the normalization step, and these methods cannot be easily extended to handle more complicated experimental designs.

There are two software packages provide functionalities to consider the control data: DBChIP [14] and DiffBind [15]. Both methods directly apply existing methods and software package developed for RNA-seq DE analysis. They start from a list of candidate regions which are unions of peaks called for each individual experiment. These regions are then treated like genes, and RNA-seq DE methods are directly applied for comparison. To account for the control experiment, the software provide option to subtract the normalized control counts from IP counts, then round the differences and use them as inputs for the software. There are several problems with this approach. First, the underlying assumption of the methods is that the background noise and biological signals are additive, which is not always true based on our real data observation (details in later section). Second, these methods don't consider the SNRs from different experiments. Finally, most RNA-seq DE methods are developed based on negative binomial distribution assumption of the gene counts. Subtracting control from IP counts then rounding will likely to violate that model assumption, which lead to incorrect statistical inferences.

In this work, we develop a comprehensive and rigorous statistical method, named "*ChIPComp* ", to perform quantitative comparison of multiple ChIP-seq data from experiments with narrow peaks, including data for most of the protein binding, some histone modifications if the modification regions are narrow, and the DNase-seq experiments. *ChIPComp* takes into consideration of

(1) genomic background measured by the control data; (2) SNRs in different experiments; (3) biological variances from the replicates; and (4) multiple-factor experimental designs. We demonstrate using simulations and real data analyses that *ChIPComp* provides more accurate and robust results compared with existing methods.

2.2 Methods

We use a two-step procedure for the quantitative comparison of ChIP-seq datasets. In the first step, we apply existing peak calling algorithm to each individual dataset to identify peaks. We then obtain the union of peaks called from all datasets as the candidate regions for quantitative comparison. Since the first step peak calling method is well developed, we will only present the method for quantitative comparison in this section.

2.2.1 The data model

Suppose there are D datasets and N candidate regions. For candidate region i ($i = 1, 2, \ldots, N$) in dataset j ($j = 1, 2, \ldots, D$), let Y_{ij} be the observed IP counts. We assume that Y_{ij} follow a Poisson distribution with underlying rate μ_{ij}, which is a function of the background λ_{ij} and ChIP signal S_{ij}, e.g., $\mu_{ij} = f(\lambda_{ij}, S_{ij})$. Here λ_{ij} represents the background signals caused by technical or biological artifacts. The observed read counts from the control experiment can be considered as realizations of the backgrounds, and can be used for estimating λ_{ij} (details of the estimation procedure is presented in later section). S_{ij} represents the non-control-related signals in the IP sample. Further, we assume that $S_{ij} = b_j s_{ij}$, where b_j is a constant representing the SNR from dataset j, and s_{ij} measures the relative biological signals (e.g., protein binding or histone modification strength up to a constant).

Now consider a set of general, multiple-factor experiments with design matrix $\boldsymbol{X}$. At candidate region i, the logarithm of the relative biological signals are assumed to be from a linear model:

$$\log(s_{ij}) = \boldsymbol{x}_j\boldsymbol{\beta}_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma_i^2)$$

where $\mathbf{x}_j$ is the j^{th} row of $\mathbf{X}$, and $\boldsymbol{\beta}_i$ is a vector of coefficient for the i^{th} candidate region. ϵ_{ij} is a random term accounting for the variations among biological replicates. Putting all pieces together, we have the following model for data at the candidate regions:

$$\begin{aligned} Y_{ij}|\mu_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \mu_{ij} &= f(\lambda_{ij}, b_j s_{ij}) \\ \log(s_{ij}) &= \boldsymbol{x_j}\boldsymbol{\beta_i} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma_{i^2}) \end{aligned}$$

Under this setup, the quantitative comparison for factor k at candidate region i can be achieved by testing: $H_0 : \beta_{ik} = 0$.

2.2.2 Estimate the background signal from control data

Obtaining good estimates of the background signal λ_{ij} is the crucial first step. Some existing methods (e.g., DBChIP or DiffBind) simply treat the counts from control experiment as background signals. However, the background noises are generated by artifacts such as chromatin structures and DNA sequence contexts, therefore, the noises fluctuate in genomic regions much wider than the peaks. Using the read counts at peaks regions only to estimate background is inaccurate and has large variances. The spatial correlations of the read counts from control experiment can be utilized to obtain better background estimates. Here we adopt the smoothing technique used in MACS [16] to obtain estimated background, denoted by $\hat{\lambda}_{ij}$. Once $\hat{\lambda}_{ij}$'s are obtained, we treat them as known and constant for the rest of the procedures.

2.2.3 Model the IP-background relationship

The most important component for the proposed data model is to characterize the relationship of IP and background signals, which is the f function. The approaches taken by DBChIP and DiffBind, e.g, subtracting the normalized control data, implicitly assume that the IP signal is the sum of the background and biological signals, or $\mu_{ij} = \lambda_{ij} + s_{ij}$. Another possible solution for quantitative comparison is to put the IP and background data into a 2×2 table at each candidate region, and then use χ^2 or Fisher's exact test for hypothesis testing. The underlying assumption for such approach is that the background and biological signals are multiplicative, e.g., $\mu_{ij} = \lambda_{ij} \times s_{ij}$.

In order to discover the true IP-background relationship, we obtain several public ChIP-seq datasets from ENCODE project (a description of the data is provided in the Results section) and perform exploratory analyses. For peaks in an experiment, we obtain the read counts from IP experiments and estimate backgrounds from control, then plot the IP counts versus backgrounds in the logarithm scale.

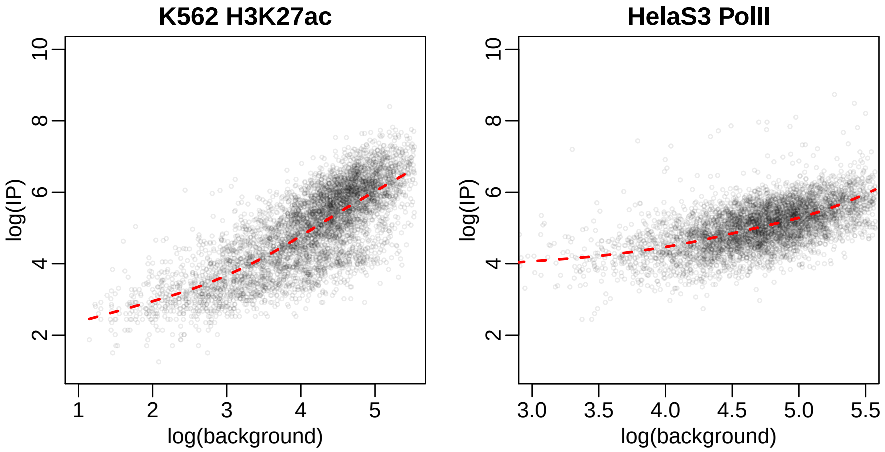


Figure 2.1: Scatterplots of IP counts versus estimated background signals from the peak regions, in logarithm scale. The red dashed line is the result from cubic smoothing spline fitting.

Figure 2.1 shows such scatterplots from two ChIP-seq dataset: H3K27 acetylation (H3K27ac) in K562 cell line and RNA polymerase II (PolII) binding in HelaS3 cell line. These figures reveal several important aspects for the IP-background relationship. First, the IP and background signals are positively correlated, as expected. Second, the IP-background relationship is neither additive nor multiplicative. The relationship is non-linear in the log scale. Finally, the IP-background relationship is different in different datasets, demonstrated by the different slopes of the scatterplots in two datasets. This emphasizes the importance of building individual background model for each dataset separately.

Based on these observations, we use a smooth function to model the IP-background relationship in logarithm scale. The IP-background response function in dataset j is described by the following model:

$$\log \mu_{ij} = g_j(\log \lambda_{ij}) + \log S_{ij} = g_j(\log \lambda_{ij}) + \log b_j + \log s_{ij}$$

Here g_j is a experimental specific smooth function. This model assumes that at a candidate region, the IP signal is the sum of background-related noise (which is a smooth function of $\log \lambda_{ij}$), SNR and biological signal in the logarithm scale.

2.2.4 The final model

Plugging in the IP-background model, the data model as described in Equation 2.1 can be written as:

$$\begin{aligned} Y_{ij}|\mu_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \log \mu_{ij} &= g_j(\log \lambda_{ij}) + \log b_j + \log s_{ij} \\ \log s_{ij} &= \boldsymbol{x}_j \boldsymbol{\beta}_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma_i^2) \end{aligned} \tag{2.1}$$

This model implies that at a candidate region in an experiment, the underlying rates for the read counts are from a lognormal distribution. The mean of the distribution depends on the genomic background, the SNR for the experiment, and the true biological signal. In this model, the observed data are Y_{ij} from IP experiment. Background λ_{ij} can be estimated from the control experiment data, and g_j can be estimated from IP-background model. $\boldsymbol{\beta}$ are parameters of interests that one wants to make inference about.

2.2.5 The procedures for quantitative comparison

In our approach for quantitative comparison of ChIP-seq datasets, the quantities to be compared across experiments are the log biological signals $\log s_{ij}$. For each dataset, we first obtain the signals by removing the estimates of background-related noises and SNRs from observed read counts from IP experiment, and then perform statistical tests. The differential protein binding or histone modifications regions are defined based on test results. The rest of this section provide detailed descriptions of the estimation and hypothesis testing procedures.

Estimating g_j and b_j

At a candidate region, the IP signals from different experiments might exhibit quantitative differences. These differences could be due to differences in biological signals, or simply because different experiments have different backgrounds, SNRs, or the IP-background responses. In order for data from different experiments to be comparable, a proper baseline is needed for normalization. A common normalization approach for ChIP-seq comparison uses the total number of reads under the peaks for adjustment. However, this approach only works for correcting technical artifacts. Biological differences such as different number of peaks cannot be corrected by this approach. For example, even if the total numbers of reads from all peaks are identical in two conditions, the peak height can still be different due to different number of peaks.

We make a crucial assumption that there exists a subset of all candidate regions, where the averages of logarithm biological signals are identical in all datasets conditional on the background signals. This is a similar assumption used by MAnorm, and by some methods for gene expression microarray data normalization where the expressions of house keeping genes are assumed constant across conditions. Denoted such set by A, $A \in \{1, 2, \dots, N\}$. By default, A is chosen as the common peaks from all datasets, or can be specified by user.

Further, we define a new function $g'_j(\log \lambda_{ij}) = g_j(\log \lambda_{ij}) + \log b_j$ to absorb the SNR into the background noise function. We take the following approach to estimate g'_j functions. For each individual dataset, we first obtain the IP counts (Y_{ij}) and estimated background signals ($\hat{\lambda}_{ij}$) for all peaks in A. Next, a cubic smoothing spline is fitted for $\log Y_{ij}$ versus $\log \hat{\lambda}_{ij}$. The fitted spline function is deemed $\hat{g}'_j$.

Hypothesis testing

The hierarchical model in equation 2.1 essentially describes the data as lognormal-Poisson compound distribution. The hypothesis testing can be performed using either likelihood ratio or Wald-based test. However either method requires numerical integration to obtain the marginal likelihood of $\boldsymbol{\beta}$, which are computationally too intensive to be practically useful given large number of candidate regions. Further, with limited number of biological replicates, it is desirable to borrow information across different candidate regions to improve the estimation of biological variances and hence statistical inference, similar to that in many other high-throughput data analysis methods [17–19]. Such information sharing is usually achieved by adding another hierarchy in the model, for example, imposing a parametric distribution on the biological variances (σ_i^2). That will further increase the complexity of the model and make the model fitting more difficult. To overcome these difficulties, we use following approximate procedures to fit the model and perform hypothesis testing at each candidate region.

We first obtain $\widehat{\log(s_{ij})}$ as

$$\widehat{\log s_{ij}} = \log(Y_{ij} + c_0) - \hat{g}_{j'}(\log \hat{\lambda}_{ij})$$

Here c_0 is a small constant (0.5) added to the IP counts to avoid having

$Y_{ij} = 0$. The estimated $\widehat{\log s_{ij}}$ can be viewed as "normalized relative log fold changes". They are quantities representing log fold changes between IP signals and background noises. They are further normalized to remove SNRs, and are values relative to the average $\log s_{ij}$'s from peaks in A and with similar background. Under our model assumptions, these quantities are directly comparable across datasets.

We then fit linear regression of $\widehat{\log s_{ij}}$ on $\boldsymbol{X}$, and obtain the estimates for coefficient $\boldsymbol{\beta}$ and residual variances σ_i^2. To overcome the small sample size problem, we apply existing variance shrinkage method developed for microarray analyses [17] to obtain the shrunk estimates of σ_i^2, denoted by $\tilde{\sigma}_i^2$. For statistical inference, an approximate estimate of the variances of $\hat{\boldsymbol{\beta}}$ with consideration of the read counts can be derived as:

$$\widehat{var(\hat{\boldsymbol{\beta}})} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\Sigma}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Here Σ is a diagonal matrix with the diagonal elements being $var(\widehat{\log s_{ij}})$. The detailed derivation of $\hat{\Sigma}$ is provided in the Appendix. In a nutshell, Σ takes into consideration of both the biological variances σ_i^2 and the uncertainty of $\log(s_{ij})$ point estimation affected by the read count Y_{ij}.

Hypothesis testing of $H_0 : \beta_{ik} = 0$ can be perform via Wald test, with the test statistics being $t = \widehat{\beta_{ik}}/\widehat{var(\widehat{\beta_{ik}})}$. The test statistics approximately follows normal distribution under null hypothesis. P-values and false discovery rate (FDR) can be obtained using canonical procedures [20].

In real data analyses, however, we found that the results from the Wald test are often influenced by the read counts, because candidate region with larger counts have greater power to detect differences. At these regions, the statistical significance is usually greater, e.g., with smaller p-values, even when the effect size is small. This is undesirable since the statistical significance doesn't necessarily imply biological significance. To overcome this problem, we provide an alternative approach in Bayesian framework. Assuming a non-informative prior on β_{ik}, e.g., $P(\beta_{ik}) \propto 1$, the following posterior probabilities are used to rank candidate regions:

$$Pr(|\beta_{ik}| > c | Y_{ij}, \hat{g}, \hat{\lambda}_{ij})) \tag{2.2}$$

Here c is a user specified threshold. In two-group comparison case, c represents the log fold change of biological signals. Under the normality assumption of $\hat{\beta}_{ik}$, the posterior probability can be obtained from normal p-values and used to rank the candidate regions. We find that this procedure often provides better results in real data analyses.

The above procedures are developed for data with biological replicates. When replicate is unavailable in the comparison, *ChIPComp* will use the difference in the estimated biological signals between two conditions, e.g., $\widehat{\log(s_{i1})} - \widehat{\log(s_{i2})}$, to rank the candidate regions.

2.3 Results

2.3.1 Data description

Both simulation and real data analysis results are based on a number of public ChIP-seq datasets. We obtain several public ChIP-seq datasets generated by ENCODE consortium [2], including three cell lines (HelaS3, GM12878, and HUVEC) for RNA polymerase II binding (PolII), and three cell lines (H1, K562, and HelaS3) for H3K27 acetylation (H3K27ac). Both the aligned sequence files (aligned to human reference genome build hg19, in BAM format) and peak calling results are obtained from ENCODE.

2.3.2 Simulation

We first perform several simulation studies, based on parameters estimated from real data, to evaluate the performance of *ChIPComp* . All simulations are for two-condition comparison, with 10,000 candidate regions, and 20% of these regions are true differential regions. Data are simulated based on two different data generative models: the proposed data model as described in equation 2.1; and an additive model where the underlying IP rates (μ_{ij}) are sums of background (λ_{ij}) and biological signals (s_{ij}). The additive model is the underlying assumption of DBChIP and DiffBind. We include the additive model in order to demonstrate that the proposed model work fine even under this assumption.

For the simulation results shown here, the simulation parameters are sampled from the real data estimates from H3K27ac and PolII. Parameters include background rate (λ_{ij}), biological signals (s_{ij}), and the IP-background relationship (g'_j) for the proposed model. For non differential candidate regions, the biological signals are made identical for two conditions. For differential regions, we randomly sample biological signals from real data for two conditions independently, so that they are different.

Since the differential analyses of ChIP-seq data is often used as a hypothesis generating tool, the goal is to have as many true positives as possible in the top-ranked candidate regions. We compare the proportions of true positives (i.e. true discovery rate, or TDR) in the top-ranked regions from different methods. The methods in comparison include the *ChIPComp* using the posterior probability in Equation 2.2, MAnorm, edgeR with and without subtracting controls, and DIME. Both DBChIP and DiffBind require aligned read files as inputs, which pose difficulties in simulations. Since both methods are based on existing RNA-seq DE detection methods, we use edgeR to approximate their performances in simulations.

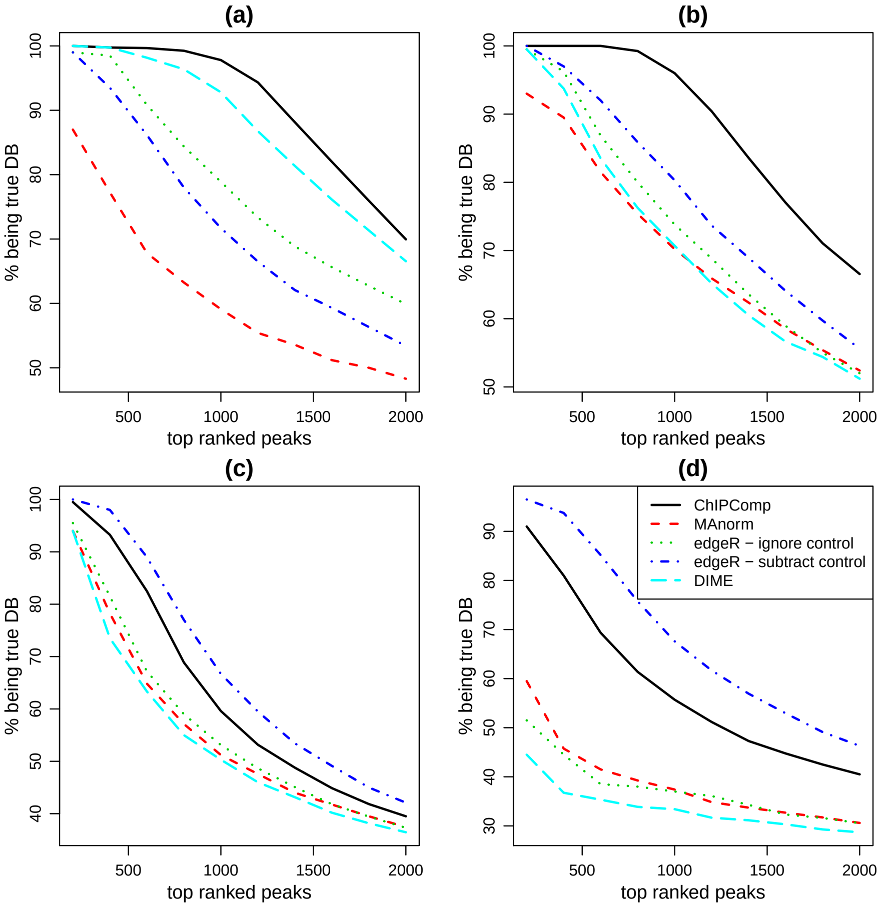


Figure 2.2: Comparison of differential peak detection accuracies from simulations. The proportions of true discovery among top-ranked candidate regions is plotted against the number of top-ranked regions. (a) and (b): data are generated based on the proposed model. (c) and (d): data are generated based on the additive model. (a) and (c) are based on H3K27ac. (b) and (d) are based on PolII.

Figure 2.2 compares the TDR curves of differential peak detection from different methods in several simulation scenarios. Figure 2.2 (a) and (b) shows the results when data are generated from the proposed model. In these cases, *ChIPComp* provides the best performance among the methods in comparison, and the gain of accuracy could be significant. It also shows that when data are generated from proposed model, subtracting control does not necessarily provide better results, that is, edgeR with and without subtracting control perform similarly. Figure 2.2 (c) and (d) shows the results when data are generated from the additive model. Not surprisingly, edgeR with subtracting control provides the best results. However, *ChIPComp* performs the second best and significantly outperforms all other method. These simulation results show good performance of *ChIPComp* . The results from the additive model further demonstrate its robustness.

In addition to detection accuracy, statistical inference is another important aspect in the differential analysis. We first investigate the empirical distribution of Wald test statistics to check whether they follow normal distribution under the null hypothesis. Figure 2.3 shows the empirical distribution of Wald test statistics by histogram and normal QQ plot. Figure 2.3 (a) and (b) are based on data simulated from the proposed model, and (c) and (d) are for data simulated from the additive model. shows that the central part of histogram of Wald test statistics approximates normal distribution even when model is mis-specified as the additive model, demonstrating the robustness of *ChIPComp* .

Another commonly used distribution assumption for sequencing read counts is negative binomial distribution, or the Gamma-Poisson compound distribution. The difference is that it assumes the underlying Poisson rate follows Gamma distribution instead of lognormal. We perform additional simulations when data are generated from negative binomial distribution, which is

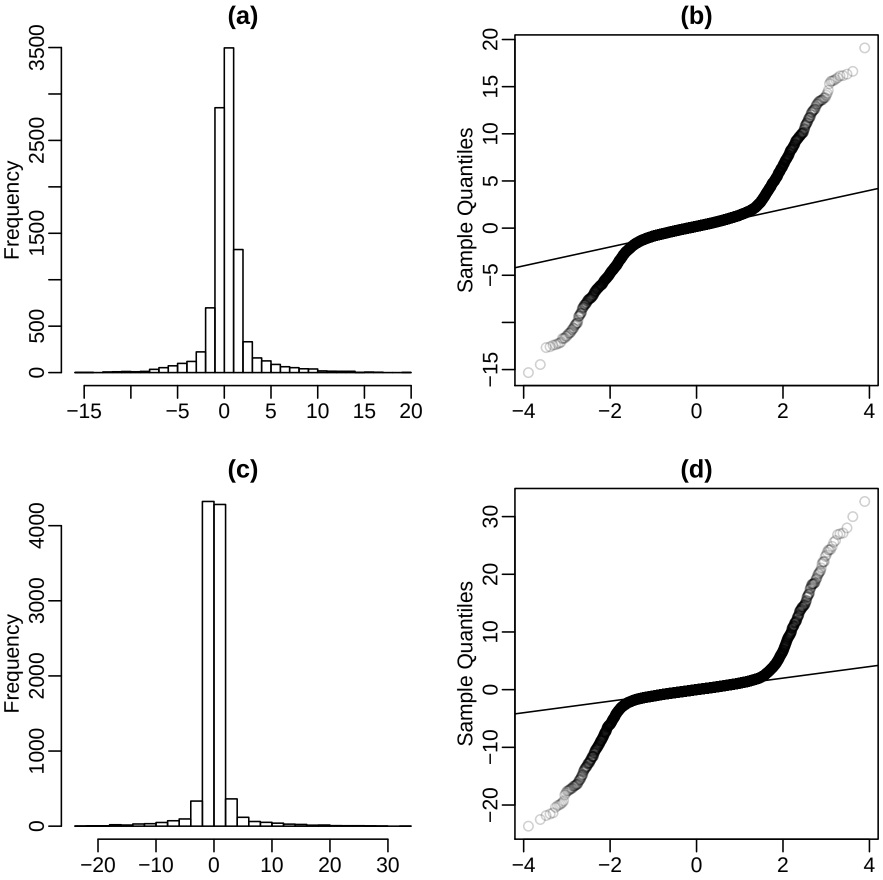


Figure 2.3: Histogram and normal QQ plot for Wald test statistics. (a) and (b) are based on data generated from the proposed model. (c) and (d) are based on data generated from the additive model.

a gamma-Poisson compound distribution. (results shown in Figure 2.4). The TDR curves show that the proposed method is robust to that distribution assumption, and *ChIPComp* still performs the best overall.

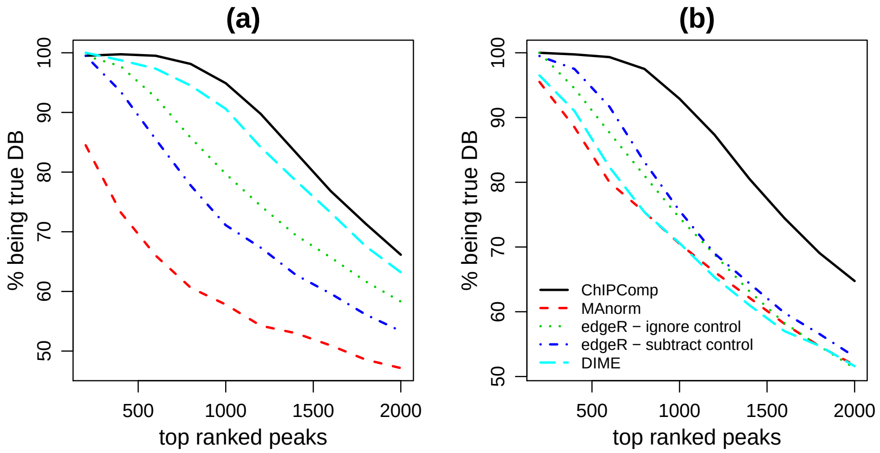


Figure 2.4: Comparison of differential peak detection accuracies from simulations when the data are generated from negative binomial distribution instead of lognormal-Poisson distribution. The proportions of true discovery among top-ranked candidate regions is plotted against the number of top-ranked regions. (a): real data-based simulation from one H3K27ac dataset. (b): real data-based simulation from one PolII dataset.

Furthermore, we perform an additional "null" simulation when there are no differential peaks. The data are generated from the proposed model using the same settings as the previous simulation. Because there are no differential peaks, the result p-values should follow uniform distribution. Figure 2.5 shows the histogram of p-values from different method, Results from edgeR ignoring control reports many false positives. P-values from MAnorm and edgeR subtracting control are heavily skewed toward 1 and tend to be over-conservative (number of false positives under different p-value threshold are shown in Table 2.1). Overall, *ChIPComp* provides the most uniform p-value distribution, which again indicates that the statistical inference will be the most accurate. Similar simulation is conducted when data are generated from additive model (results shown in Figure 2.6 and Table 2.2). Again, p-values from *ChIPComp* are more uniform compared than others when data are generated from additive model.

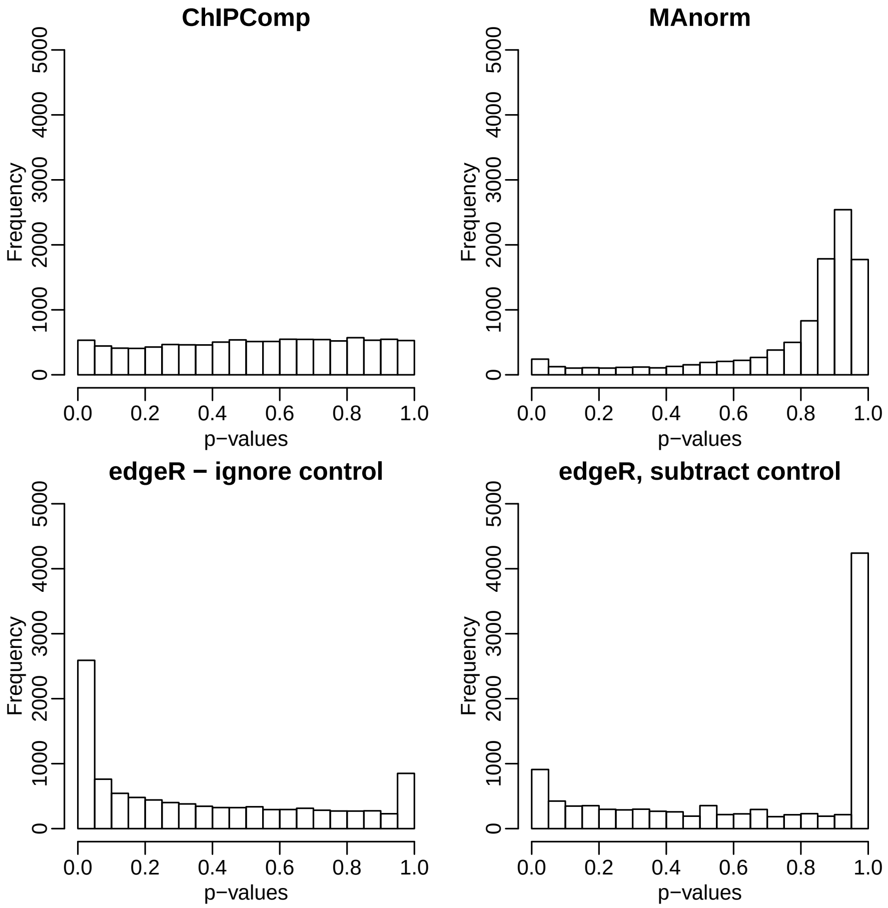


Figure 2.5: Histogram of p-values reported from different methods, based on null model that there's no differential regions. The data are generated from the proposed model.

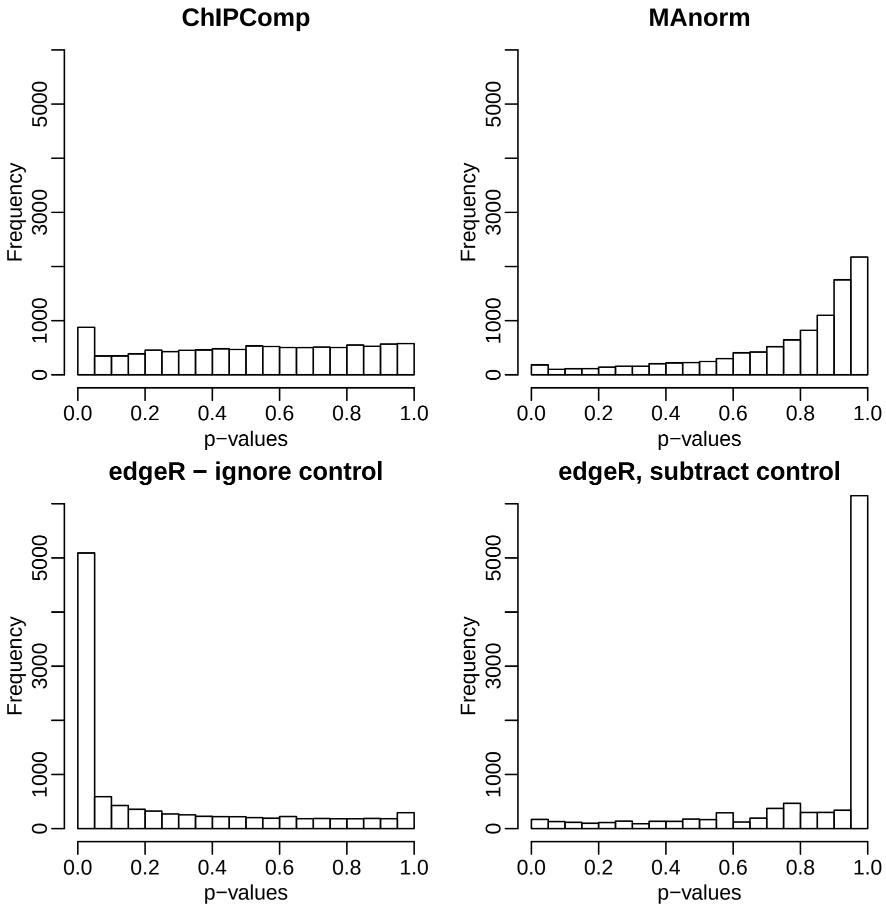


Figure 2.6: Histogram of p-values reported from different methods, based on null hypothesis that there's no differential binding sites. The data are generated from the additive model.

Specifically, Table 2.1 and Table 2.2 show that the numbers of false positives when the p-value is set at different threshold for data simulated from proposed model and additive model respectively under the null hypothesis. Since p-value is defined as the expected number of false positives, the ideal method should report number of false positives close to the expected number of false positives as much as possible. Table 2.1 shows that *ChIPComp* has the most accurate number of false positives compared to other methods. Table 2.2 shows that when the generating model is mis-specified as additive model, *ChIPComp* reports biased number of false positives, but is still on par with edgeR with

p-value	*ChIPComp*	MAnorm	edgeR - ignore control	edgeR, subtract control
0.001	36	46	838	191
0.005	91	80	1282	335
0.01	147	102	1543	456
0.05	532	241	2590	910
0.1	975	366	3351	1334

Table 2.1: Number of false positive with different p-value threshold (data generated from proposed model)

p-value	*ChIPComp*	MAnorm	edgeR - ignore control	edgeR, subtract control
0.001	351	15	3014	18
0.005	454	44	3689	36
0.01	534	62	4027	48
0.05	877	183	5091	170
0.1	1224	284	5681	300

Table 2.2: Number of false positive with different p-value threshold (data generated from additive model)

subtracting control and MAnorm.

We further investigate the FDR for different estimated models. For each candidate region, one minus the posterior probability obtained from Equation 2.2 can be viewed as local FDR. Based on the connection between the local FDR and the classical global FDR [21], the local FDR can be converted to the global FDR. Figure 2.7 shows the comparison of global FDR estimates from different methods, when data are generated from the proposed and additive model. When data are from the proposed model, *ChIPComp* provides accurate FDR estimation. DIME performs well too but all other methods have poor performances. From other methods, the estimations of the FDR in the top-ranked regions are too liberal and give overly optimistic results. When data are generated from additive models, none of the methods provide very accurate FDR estimation, but *ChIPComp* still has the best performance relatively.

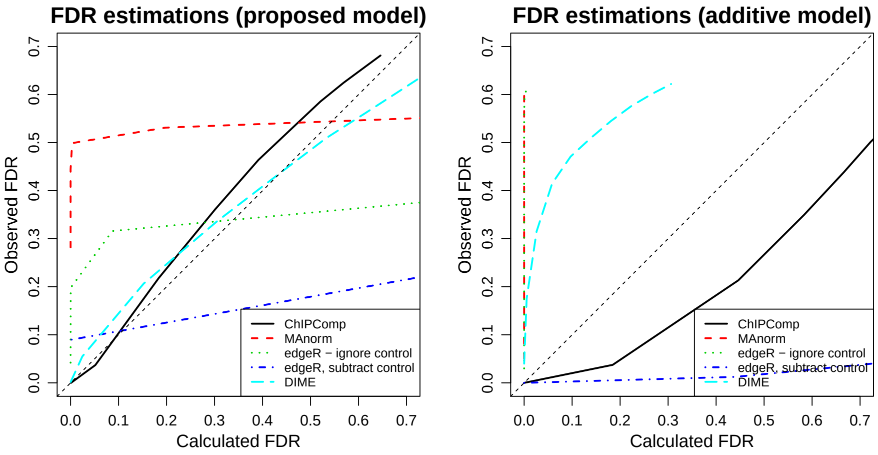


Figure 2.7: Comparison of FDR estimations from different methods, based on simulation. X-axis shows the FDR reported from different methods, and y-axis shows the observed FDR.

All simulation results demonstrate that *ChIPComp* is more accurate and robust compared to existing methods. It provides better ranking and statistical inference in detecting differential peaks. It is also fairly robust against model mis-specification, for example, when data are from additive model.

2.3.3 Implementation

The proposed method is implemented in an R Bioconductor [22] package *ChIP-Comp* , which is currently available at

http://bioconductor.org/packages/release/bioc/html/*ChIPComp* .html. The function takes detected peaks from all datasets and the aligned sequence files as inputs, and reports a list of genomic regions showing differential binding or histone modification, with estimated p-values and FDRs.

2.3.4 Real data results

We further evaluate the performance of *ChIPComp* in several real datasets. The analyses are based on two-condition comparisons. Since the gold standards for quantitative differences between ChIP-seq data are not available, we utilize other data to create "silver standard" to compare different methods. It was known that PolII binding and H3K27ac are positively correlated with gene expressions. We obtain the gene expression data from RNA-seq experiments for these samples (also from ENCODE), and then use them to create silver standard for comparison. To be specific, in a two-condition comparison, we first perform differentially expression (DE) analyses on the RNA-seq data using edgeR. Genes with FDR less than 0.01 are deemed DE, with FDR greater than 0.2 are deemed non-DE, and the rest are deemed unknown. Next, we keep candidate regions that are within 1000 base pairs of the transcriptional start sites (TSS) of a gene. Finally, a region will be deemed differential or non-differential for the protein binding or histone modification between two conditions if its corresponding gene is DE or non-DE.

Since there are three cell lines (HelaS3, GM12878, and HUVEC) for PolII and another three cell lines (H1, K562, and HelaS3) for H3K27ac, we perform following pairwise two-condition comparisons: HelaS3 versus K562, H1 versus K562, H1 versus K562 for H3K27ac; HelaS3 versus HUVEC, GM12878 versus HelaS3, GM12878 versus HUVEC for PolII. The performance of the proposed method is compared with MAnorm, DBChIP ignoring or subtracting control, and DIME. DiffBind essentially uses the same algorithm as DBChIP (apply existing RNA-seq DE methods), so they are not included in the comparison. We use $c = 1$ in *ChIPComp* to generate and rank the differential binding regions for the results presented below.

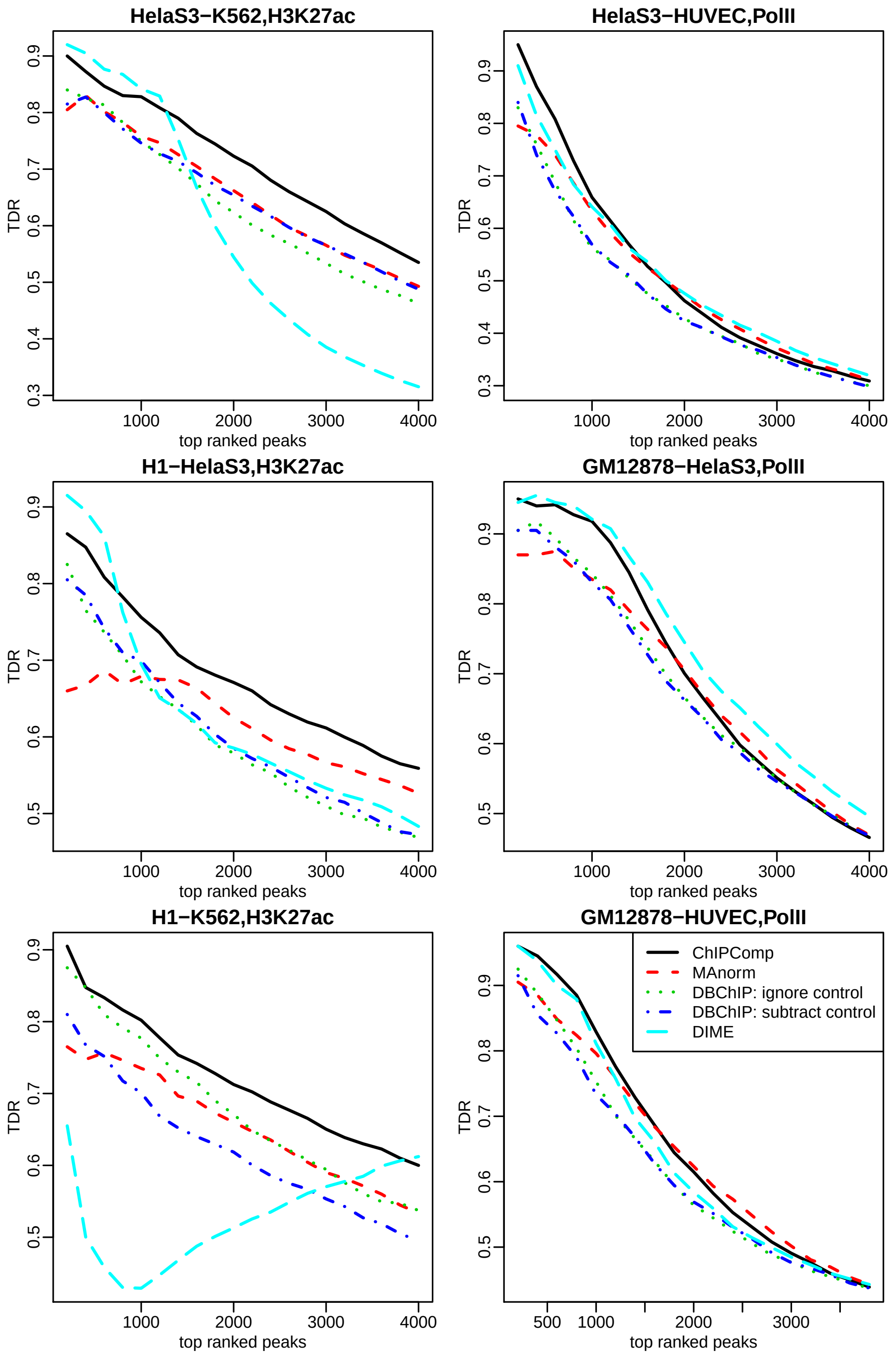


Figure 2.8: Comparison of differential peak accuracies from real datasets. All results are for two-condition comparisons on different histone modification or protein binding, as marked in figure titles.

Figure 2.8 shows the detection accuracies from all the comparisons. For H3K27ac comparisons, *ChIPComp* performs best except that DIME slightly outperforms at a small number of top peaks in HelaS3 versus K562 and H1 versus HelaS3. However, DIME fails badly in the H1 versus K562 comparison. In practice, we found that DIME is sometimes unstable, perhaps due to the convergence problem in EM algorithm. Compared with other methods, gains of detection accuracies from *ChIPComp* is usually over 10%. For comparisons in PolII binding data, *ChIPComp* and DIME are the best performers across three cell lines. Overall, these real data analyses demonstrate that *ChIPComp* provides the most accurate and robust results compared with other methods. In addition, we notice that subtracting control, based on the assumption of additive model, does not necessarily provide better performance than ignoring control from the results of DBChIP. In H1 versus K562 comparison for H3K27ac data, ignoring control actually provides much better performance than subtracting control. This is consistent with the results from simulation studies, and further demonstrates that simply subtracting control from IP is not an optimal way to use the data from control experiment in quantitative comparison of ChIP-seq data.

Although the posterior probability threshold c could have some impacts on the performance of *ChIPComp* , the overall performance of *ChIPComp* remains stable with reasonable choice of c value. Since the default c value is 1, we try using different c values (0.5 and 2), and obtain similar TDR curves as using default c value (Figure 2.9 and Figure 2.10). We also plot the TDR curves by using p-values from hypothesis test instead of the posterior probabilities to rank peaks, and find similar results (Figure 2.11).

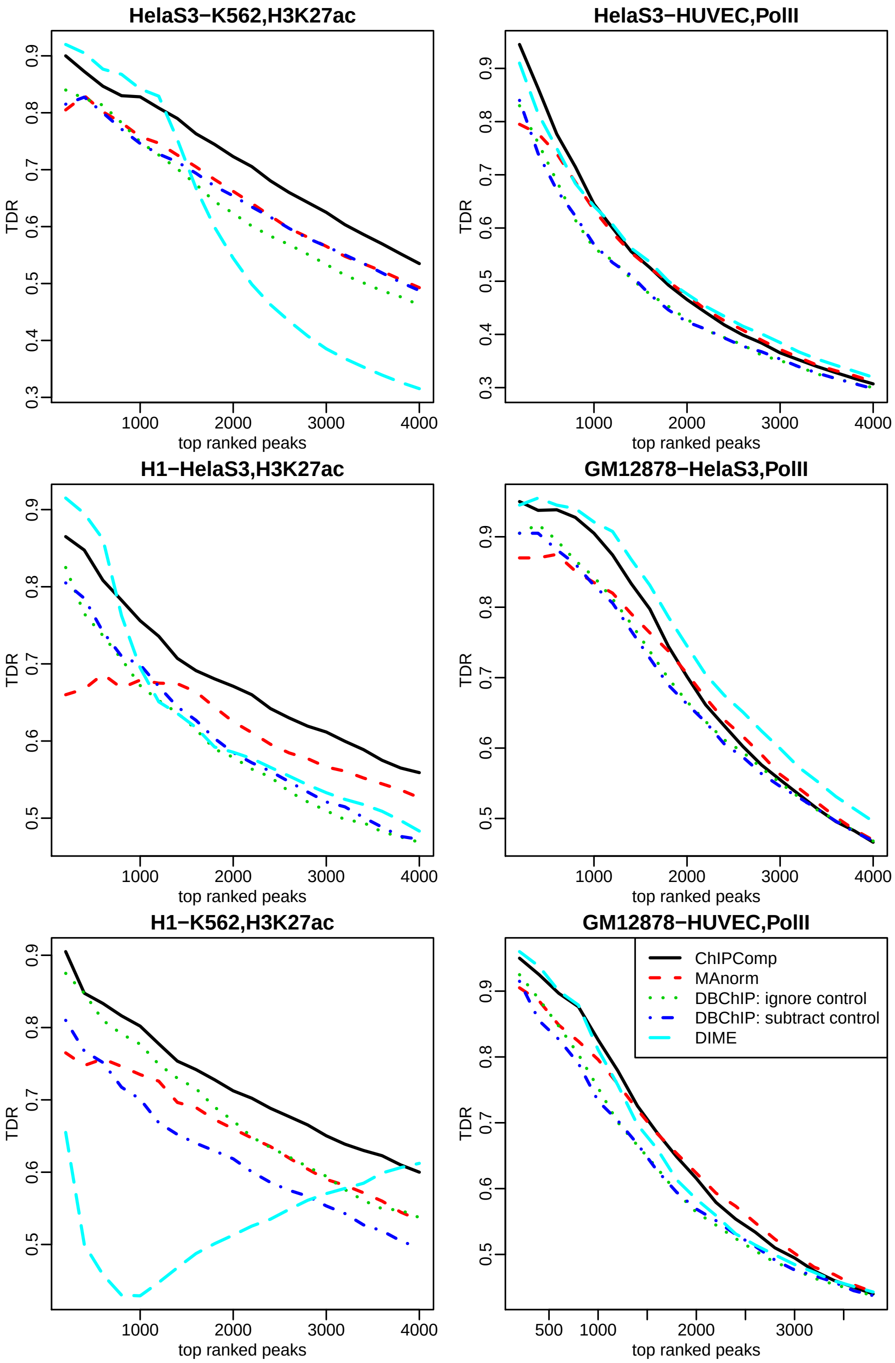


Figure 2.9: Comparison of differential peak accuracies from real datasets ($c = 0.5$)

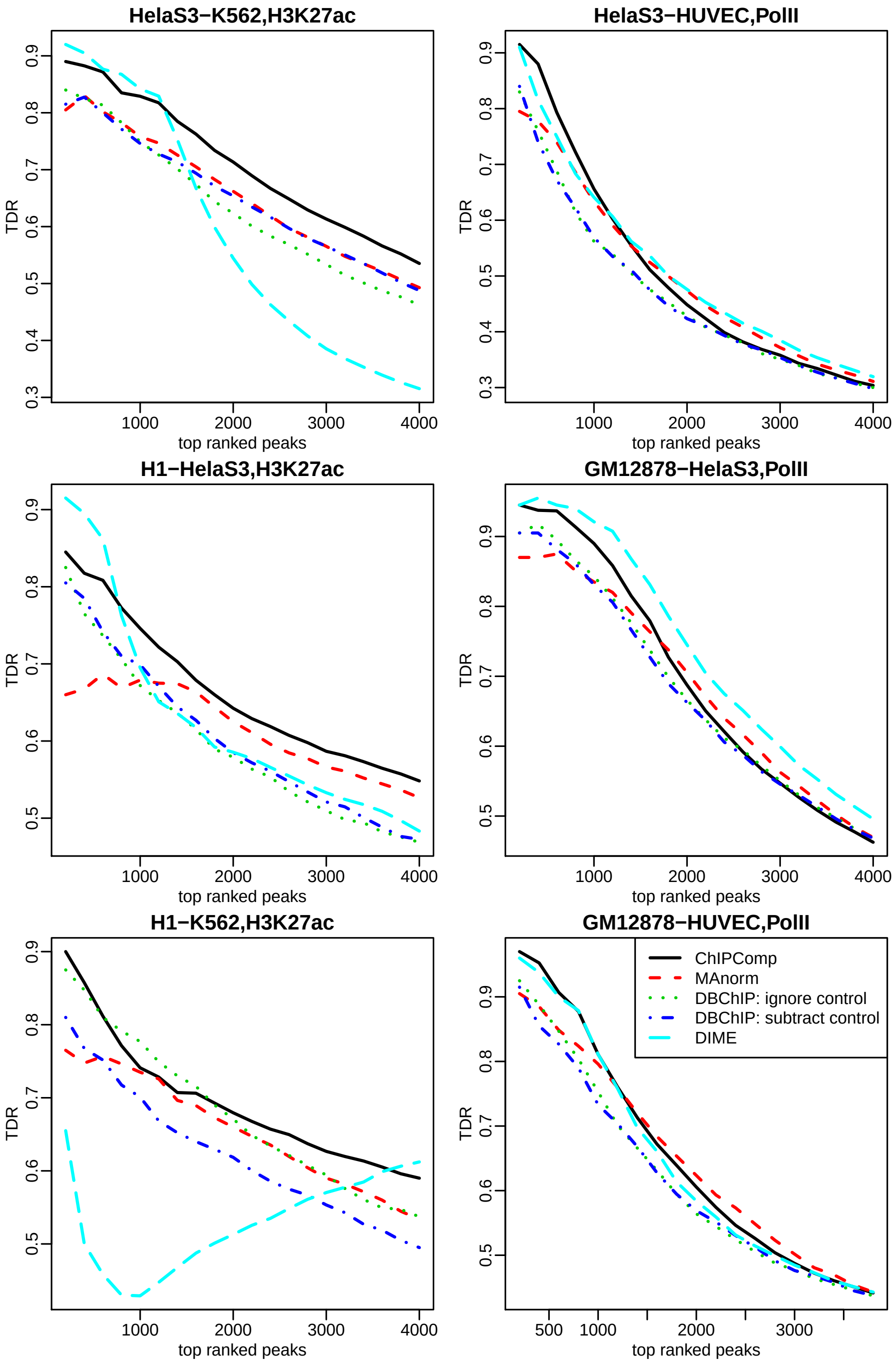


Figure 2.10: Comparison of differential peak accuracies from real datasets ($c = 2$)

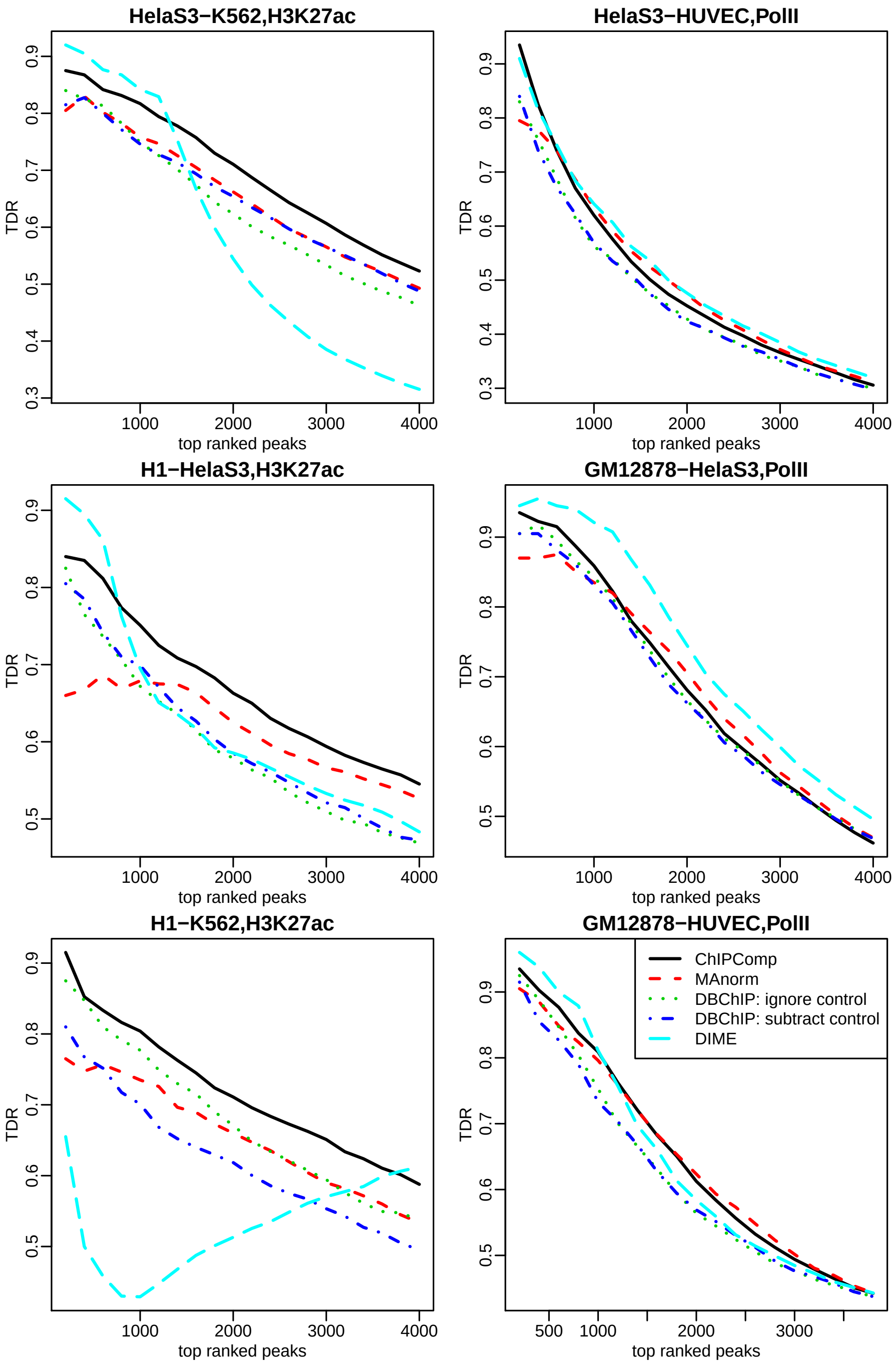


Figure 2.11: Comparison of differential peak accuracies from real datasets using *ChIPComp* p-value instead of posterior probability

In addition, we exam the FDR estimation accuracies from all methods in real datasets, using gene expression as gold standard. We plot the observed vs.

reported FDR from different methods (Figure 2.12). In general, none of the methods provide very accurate FDR estimation, but *ChIPComp* still has the best performance overall.

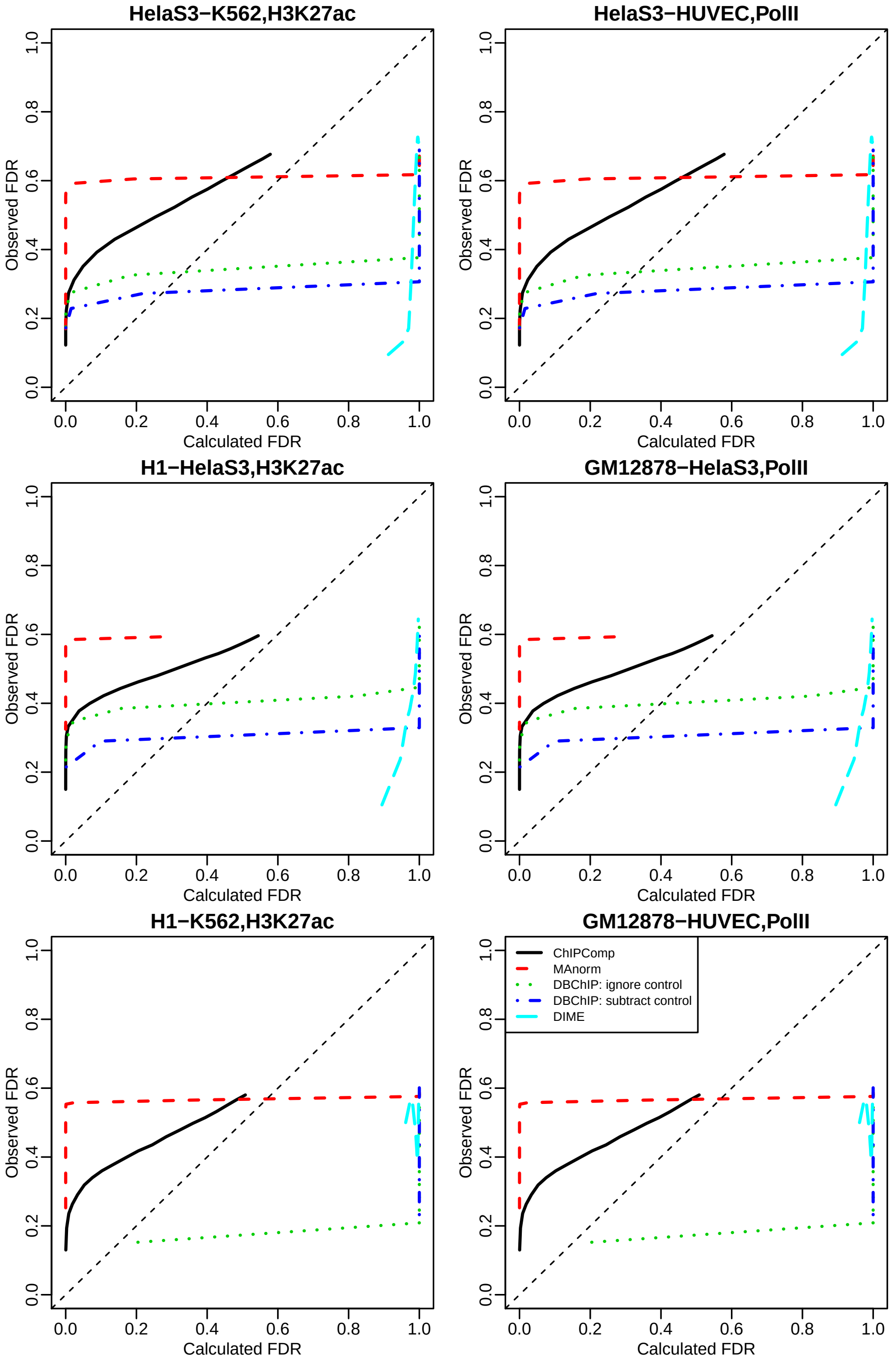


Figure 2.12: Comparison FDR estimation for real datasets

Furthermore, we generate the ROC curves and use AUC (Area Under the Curve) as another criteria to compare the performance of different methods (Figure 2.13). Overall *ChIPComp* has the highest AUC value.

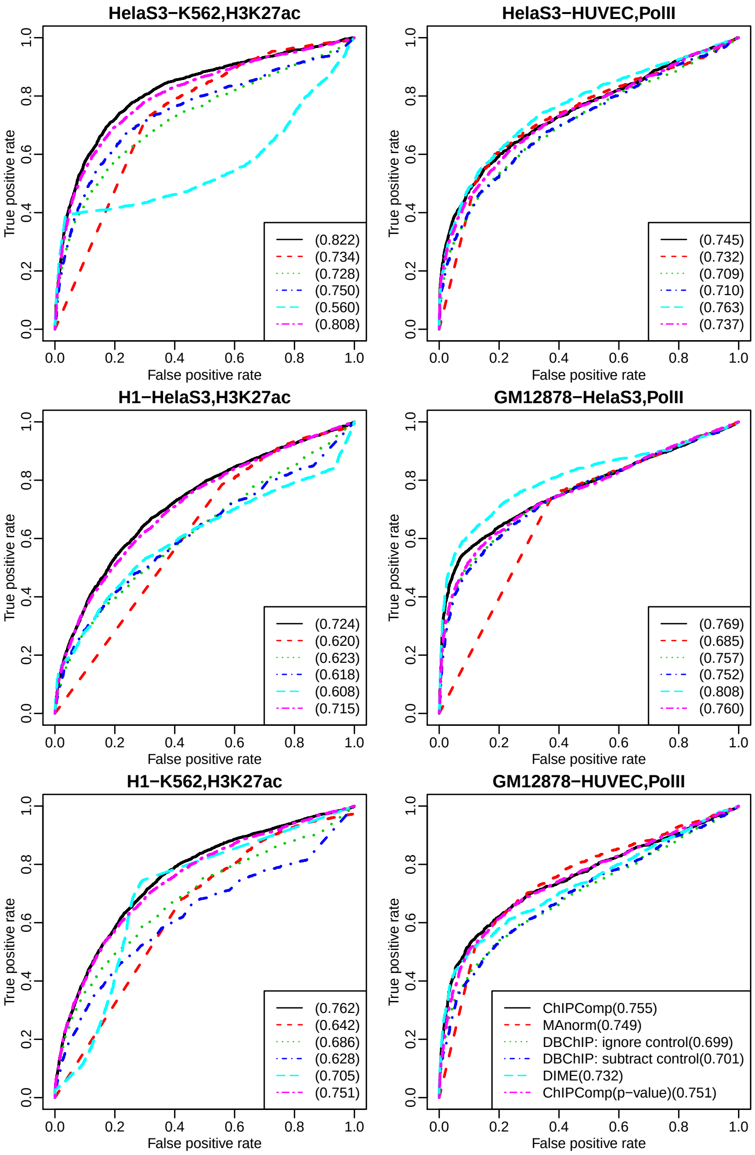


Figure 2.13: Comparison of differential peak accuracies from real datasets using ROC curve

2.4 Discussion

In this work, we develop a novel statistical method to perform quantitative comparisons of multiple ChIP-seq datasets and detect differential protein binding or histone modification regions. Statistical methods of differential analysis for other sequencing data such as RNA-seq have been well developed. The comparison of ChIP-seq data, however, is more complicated because of different background noises and signal to noise ratios in distinct experiments. Existing methods either ignore the data from control experiments (such as MAnorm or DIME), or directly apply RNA-seq methods without proper normalization (such as DBChIP or DiffBind). The proposed method describes the data by a rigorous statistical model with the considerations of control data, signal to noise ratios, biological variations, and general experimental designs. Statistical test procedures are developed for detecting differential regions. Simulation and real data analyses results demonstrate that *ChIPComp* provides more accurate and robust results compared with existing methods.

The essence of the method is to extract biological signals from different experiments and then compare. The process involves estimating and removing biological and technical artifacts, and normalization of the biological signals. In order to ensure that the estimated biological signals are comparable across different experiments, proper references are needed for normalization and put the biological signals in a common baseline. In that regard, the proposed method relies on two important assumptions. First, the ChIP-seq datasets in comparison need to have a non-trivial number of common peaks. In fact, when there are very few common peaks among datasets, a simple overlapping analysis of the peak will be adequate. Second, it is assumed that there's no global difference in the true biological signals for the common peaks across all datasets, which is the same assumption used by MAnorm. This assumption provides a common baseline for different datasets for comparison. Similar assumption has been used in differential expression analysis for many years: a majority of the genes show no differential expression.

The hypothesis test is performed base on the log biological signals, which is derived based on log counts. When the counts at candidate regions are very small, this procedure could bring some biases and high variance. To overcome that, we added a small constant in the counts to “squeeze” the lower end of the log count distribution, and carefully derived the variances for estimated parameters to take the raw counts into consideration. A similar approach has been proposed in a recently developed RNA-seq DE method, voom [23], and proved to have good performance.

The proposed method describes the count data from replicated samples through a lognormal-Poisson model. More often, these data are described by negative binomial, which is a Gamma-Poisson compound distribution. In our model, the underlying Poisson rate is assumed to follow a lognormal, instead of Gamma distribution. This is mainly motivated by methodological convenience. However, when the shape parameter in Gamma distribution is reasonably large, the Gamma and lognormal distributions become very similar. Simulation results show that the results from *ChIPComp* is robust and still provides good results when the data are from negative binomial. So we believe that our method will perform well in real data settings.

The method is specifically designed for comparison of ChIP-seq with short peaks, including most of the protein binding data, some histone modification data and DNase-seq. For histones modification data with long peaks/blocks such as H3K9me3, the method is not directly applicable. However the problems presented in those data are similar: consideration of backgrounds, different signal to noise ratios, biological variances, etc. To design method working for the quantitative comparison of data with long peaks is our research plan in the near future.

2.5 Appendix

As described in the manuscript, the data model is:

$$\begin{aligned}
Y_{ij}|\mu_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log \mu_{ij} &= \log s_{ij} + g_j^{'}(\lambda_{ij}) \\
\log s_{ij} &= \boldsymbol{x}_j \boldsymbol{\beta}_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma_i^2)
\end{aligned}$$

We denote Y_{ij} as the observed read counts in candidate region i in dataset j, and μ_{ij} as the corresponding underlying poisson rate. The approach for differential binding analysis is to estimate $\boldsymbol{\beta}_i$ and $var(\hat{\boldsymbol{\beta}}_i)$ and then perform Wald test for each candidate region i. First of all, $\log s_{ij}$, the "normalized relative log fold changes", is estimated as $\widehat{\log s_{ij}} = \log(Y_{ij}) - \hat{g}_j'(\log \hat{\lambda}_{ij})$. Here $\hat{g}_j'(\log \hat{\lambda}_{ij})$ is estimated previously and deemed constant in below derivations. Then, we obtain the estimates of $\boldsymbol{\beta}$ through a linear model:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \widehat{\log(\boldsymbol{S})}$$

The variances of $\hat{\boldsymbol{\beta}}$, according to linear model theory, is:

$$Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Var(\widehat{\log \boldsymbol{S}}) \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \equiv \boldsymbol{\Sigma}$$

To perform Wald test, we need to obtain estimates of the variances. Notice that

$$Var(\widehat{\log \boldsymbol{S}}) = Var(\log \boldsymbol{Y} - \hat{g}(\hat{\boldsymbol{\lambda}})) = Var(\log \boldsymbol{Y}),$$

where $Var(\log \boldsymbol{Y})$ is a matrix with $Var(\log Y_{ij})$ in the diagonal. The derivation of $Var(\log Y_{ij})$ is provided below.

We expand $\log Y_{ij}$ around $\log \mu_{ij}$ by Taylor expansion to the first order term:

$$\log(Y_{ij}) \approx \log \mu_{ij} + \frac{1}{\mu_{ij}}(Y_{ij} - \mu_{ij})$$

Then we have:

$$\begin{aligned} Var(\log Y_{ij}) &\approx Var\left[\log \mu_{ij} + \frac{1}{\mu_{ij}}(Y_{ij} - \mu_{ij}\right] \\ &= E\left\{Var\left[\log \mu_{ij} + \frac{1}{\mu_{ij}}(Y_{ij} - \mu_{ij})\middle|\mu_{ij}\right]\right\} + Var\left\{E\left[\log \mu_{ij} + \frac{1}{\mu_{ij}}(Y_{ij} - \mu_{ij})\middle|\mu_{ij}\right]\right\} \\ &= E\left\{Var\left[\frac{Y_{ij}}{\mu_{ij}}\middle|\mu_{ij}\right]\right\} + Var(\log \mu_{ij}) \\ &= E\left[\frac{1}{\mu_{ij}}\right] + Var(\log \mu_{ij}) \end{aligned}$$

The above derivations use the fact that $E(Y_{ij}) = Var(Y_{ij}) = \mu_{ij}$, since Y_{ij} is from Poisson distribution. The second term $Var(\log \mu_{ij}) = \sigma_i^2$ because of the linear model assumption. The first term can be directly obtained from log-normal distribution, since it is assumed that $\log \mu_{ij} \sim N(\mu_0, \sigma_i^2)$. Here μ_0 is defined to be $\mu_0 = g_j(\lambda_{ij}) + \boldsymbol{x_j \beta_i}$. We have:

$$E\left[\frac{1}{\mu_{ij}}\right] = e^{-\mu_0 + \sigma_i^2/2}$$

Put all pieces together, we have

$$Var(\log Y_{ij}) \approx e^{-g_j(\lambda_{ij}) - \boldsymbol{x_j}\boldsymbol{\beta_i} + \sigma_i^2/2} + \sigma_i^2$$

Plugging in the estimates of g_j, β_i, and σ_i^2, we obtain the estimated variance as:

$$\widehat{Var(\log Y_{ij})} \approx e^{-\hat{g}_j(\hat{\lambda}_{ij}) - \boldsymbol{x_j}\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} + \hat{\sigma}_i^2/2} + \hat{\sigma}_i^2.$$

The variance/covariance matrix of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\Sigma}$, is then a diagonal matrix with diagonal elements being $\widehat{Var(\log Y_{ij})}$.

Chapter 3

DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles

3.1 Introduction

With the development of high-density genotyping arrays, over the past ten years, investigators have conducted thousands of GWAS studies, which have identified tens of thousands of loci associated with a host of human traits and diseases. There are now resources established to catalog a comprehensive collection of trait-associated SNPs. One example, the Association Results Browser (ARB) (`http://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm`, accessed May 28, 2016) currently contains 44,124 SNP trait association results, which correspond to 30,553 (autosomes plus chromosome X) unique trait-associated SNPs linked to 573 phenotypes. Overall, 90 percent of those SNPs are located in non-coding regions (introns and intergenic regions), which is consistent with the observation that over 70% of the risk-association loci in the National Human Genome Research Institute (NHGRI) GWAS catalog lack variants that map to exons within their haplotype block [24].

Unlike coding variants, whose functional impact can be gauged by checking whether the DNA sequence variant affects the translated protein sequence [25], there is little we can say for non-coding variants, except about evolutionary conservation at the loci. Therefore, one needs information beyond the DNA sequence level to identify variants that functionally link to a disease or phenotype. Since non-coding SNPs are suspected of disrupting normal regulatory control mechanisms of target genes, and we know that epigenetic information, such as DNase hypersensitivity and histone modifications, is closely related to regulatory function [26–28] and has been linked to the enrichment of GWAS SNPs [29], epigenetics data have thus been recognized as an important source of functional annotation for non-coding variants [26].

Taking advantage of the powerful high-throughput technologies, such as next-generation sequencing (NGS), experimental assays have been developed to comprehensively survey the entire genome for such regulatory events. Major experiments in this category include ChIP-seq (coupling chromatin immunoprecipitation and next-generation sequencing) [30–32] to identify in vivo binding of transcription factors (TFs) and histone marks; DNase-seq(DNase I hypersensitive sites sequencing) [33, 34] and FAIRE-seq(formaldehyde-assisted isolation of regulatory elements sequencing) [35], both for identifying open chromatin regions. Given the importance of such regulatory information, large international consortia, like the Encyclopedia of DNA Elements (ENCODE) [36] and the

Roadmap Epigenomics Mapping Consortium (REMC) [4] have been formed to systematically conduct these experiments to identify functional elements with regulatory activities across hundreds of cell lines/tissues. These datasets offer a great opportunity to link sequence variants to regulatory elements, including TF binding, histone modification, and open chromatin.

Taking advantage of these resources, researchers have developed multiple computational approaches to identify non-coding risk variations. Ritchie et al. developed a supervised approach called Genome-Wide Annotation of Variants (GWAVA) [37], which is a modified random forest classifier [38], to distinguish disease-implicated variants from benign variants using various static genomic and epigenomic annotations, such as genic context, phylogenetic conservation scores, TF binding sites, and histone modifications. Kircher et al. developed a supervised learning approach named CADD [39], which is a support vector machine classifier that integrates 63 annotations, including phylogenetic conservation scores, genic context, and scaled p-values derived from ENCODE, as features of the classifier. Lu et al. developed an EM-based algorithm called GenoCanyon [40] that models the non-coding variant using a two-component mixture model (risk or benign). Recently, Ionita-Laza et al. developed Eigen [41], another unsupervised approach adopting a more sophisticated two-component mixture model by imposing a predefined block-wise structure among features in the model-fitting process.

A common feature of all the above methods is that they are disease/phenotype neutral; that is, variants associated with all diseases/phenotypes are included in the training set. As an example, GWAVA uses all regulatory mutations from the public release of the Human Gene Mutation Database (HGMD) [42]. Eigen and CADD use GWAS index SNPs found in the US National Human

Genome Research Institute's GWAS catalog. GenoCanyon uses all the annotated variants from ClinVar [43]. However, it is likely that the biological functions underlying a risk variant for type 2 diabetes, a metabolic disorder, is different from that for Alzheimer's disease, a neurodegenerative disorder. Furthermore, the regulatory activities of TFs and histone marks are different in different cell lines/tissues, sometimes dramatically, so it is not clear which combination of cell line/tissue and TFs/histone modifications could better distinguish risk variants of a particular disease/phenotype from benign variants. Therefore, we believe it is desirable and appropriate to develop a method that can identify disease-specific risk variants. This is particularly important for interpreting variants identified via personal genome sequencing (PGS), since most of the variants identified by PGS are rare variants (minor allele frequency less than 1%), making their association with disease difficult to measure using GWAS.

Here we present *DIVAN* (DIsease-specific Variant ANnotation), a novel method to identify disease-specific risk variants. *DIVAN* adopts an ensemble learning framework with a feature selection step to annotate and prioritize non-coding variants using a large collection of genomic and epigenomic annotations. To evaluate *DIVAN*'s performance, we conduct comprehensive analyses using data from two different databases. One study involves 45 different diseases/phenotypes across 12 disease/phenotype classes, and the other includes 36 diseases/phenotypes.

In this work, we treat the trait-associated index SNPs identified by GWAS and reported in the ARB as surrogates for the functional SNPs. This is because validated or annotated bona fide functional SNPs are too rare for most diseases/phenotypes to form a meaningful training set. Furthermore, the belief is that real functional variants are enriched among GWAS index SNPs than random background SNPs.

3.2 Methods

3.2.1 Software and data package availability

To maximize *DIVAN*'s utility, we pre-computed *DIVAN* score for every base of the human genome (hg19), and for each of the 45 diseases, using either the TSS-matched criterion or the region-matched criterion. *DIVAN* offers two options to query and retrieve these scores: by variant identifier (for known variants) or by genomic regions. For known variants, *DIVAN* allows the user to retrieve scores for all known variants found in the Ensembl variation database (release 70, including 49,999,357 variants), COSMIC database [44] (v78, including 3,153,949 variants by excluding variants on Mitochondrial DNA and variants without genomic position) and 1000 Genome variants (Phase I, including 17,076,840 variants). For genomic regions, users can select either to retrieve scores from all known variants within the genomic regions or obtain the average base-level scores for each genomic region. Correspondingly, *DIVAN* provides R scripts for both options. The *DIVAN* software toolkit and the pre-computed scores are freely available at `https://sites.google.com/site/emoryDIVAN` under the GNU General Public License v3.

3.2.2 Data sources

Construction of disease-specific risk variants and benign variants

The risk variants chosen from ARB include 28,713 unique non-coding SNPs (12,159 intronic SNPs and 16,803 intergenic SNPs) spanning 555 diseases/phenotypes across 33 disease/phenotype classes. In the present study, to maintain enough risk variants in the training set, we chose 45 diseases/phenotypes spanning 12 disease/phenotype classes, with at least 50 disease-SNP associations. The 45 diseases/phenotypes with the numbers of risk variants are summarized in Table 3.2.

To construct a set of benign variants for each disease/phenotype, we apply the same strategy used in GWAVA by sampling variants not reported to be disease-implicated and by requiring the distances between these benign variants and their nearest transcription start sites (TSSs) to have the same empirical distribution as those risk variants. All benign variants are sampled from the 1000 Genomes Project Phase I catalog (with minor allele frequency higher than 5%), excluding all variants found in the ARB. Similar to GWAVA, ten times more benign variants than risk variants are selected for each disease/phenotype.

Merge replicates

Most of the experiments in ENCODE and RMEC contain biological replicates. To simplify the analysis, we merge reads produced from replicated ChIP-seq experiments if both the factor (TF/Histone) and cell line are the same, and reads from open chromatin experiments conducted on the same cell line are also merged. Since all ENCODE/REMC ChIP-seq experiments are performed with ChIP and matched input samples, we calculate the normalized read count by subtracting the number of input reads from the ChIP reads after adjusting the sequencing depth. For open chromatin experiments, DNase-seq and FAIRE-seq, we use the ChIP reads directly, as there is no matching input sample. For preprocessed peak files of the same factor and the same cell line, overlapped peaks are merged by taking the union.

Annotation sources

Open chromatin

ENCODE conducts two types of sequencing experiments to profile genome-wide open chromatin regions: DNase-seq and FAIRE-seq. We include both in the feature collection for *DIVAN*. To be specific, for mapped read files, we collect 230 DNase-seq datasets (merged into 80 features) and 78 FAIRE-seq datasets (merged into 31 features) from ENCODE, and 350 DNase-seq datasets (merged into 73 features) from REMC; for corresponding preprocessed peak files, we collect 100 DNase-peak files and 38 FAIRE-peak files (merged into 31 features) from ENCODE, and 39 DNase-peak files from REMC.

Transcription factor binding sites (TFBS)

For mapped read files, we obtain 650 TF ChIP-seq datasets (merged into 292 features) from ENCODE/HAIB and 681 TF ChIP-seq datasets (merged into 279 features) from ENCODE/SYDH; for corresponding preprocessed peak files, we collect 638 TF-peak files (merged into 295 features) from ENCODE/HAIB and 321 TF-peak files (merged into 288 features) from ENCODE/SYDH.

RNA polymerase binding

For mapped read files, we collect 156 RNA polymerase binding ChIP-seq datasets (merged into 49 features); for corresponding preprocessed peak files, we collect 92 peak files (merged into 53 features) from ENCODE.

Histone modification

We include histone ChIP-seq datasets from both ENCODE and REMC. For mapped read files, we collect 549 histone ChIP-seq datasets (merged into 267 features) from ENCODE and 1,407 histone ChIP-seq datasets (merged into 735 features) from REMC; for corresponding preprocessed peak files, we collect 280 histone-peak files (merged into 270 features) from ENCODE and 979 histone ChIP-peak files from REMC.

Genomic features

Two types of static genomic features are included in *DIVAN*: repeated elements and conservation scores (genomic evolutionary rate profiling (GERP) element [45] and phastCon scores [46]). We consider all repeated elements collected in the UCSC Genome Browser, including LINE, low complexity, satellite, simple repeat, SINE, LTR, etc. Conservation annotations include GERP elements and phastCon score, which are known to influence the functional consequences of genetic variants, such as phylogenetic conservation and level of selective constraint. GREP elements are downloaded from the Sidow Lab (`http://mendel.stanford.edu/SidowLab/downloads/gerp/`) and further treated as a binary annotation for each variant investigated. The phastCon scores are calculated for variants of interest using Bioconductor package phastCons100way.UCSC.hg19.

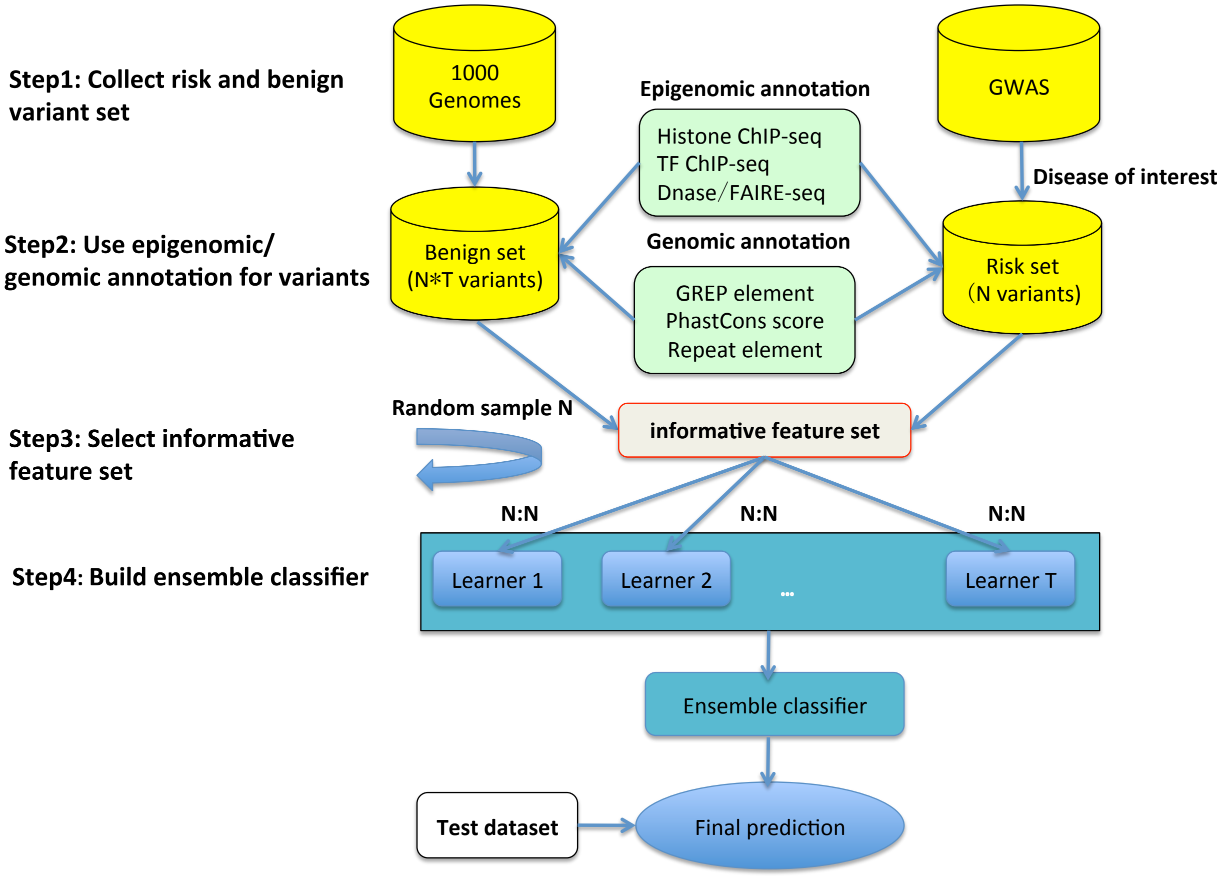
Annotation segmentation

To simplify the computation, we first cut the whole genome into 200-bp bins and calculate the feature value, i.e., normalized mapped read count or the peak presence for each bin. Therefore, the result is a genome-wide annotation matrix with rows as 200-bp bins across the whole genome, and columns as genomic and epigenomic features. With the pre-built genome-wide annotation matrix, we could easily retrieve feature values for each variant by simply determining which bin the variant falls into.

3.2.3 Feature selection-based ensemble-learning framework

The workflow of *DIVAN* is illustrated in Figure 3.1, which consists of four steps. The first step is to build the risk variant set and the benign variant set. All risk variants from the selected 45 diseases/phenotypes are retrieved from ASB. The benign variants are obtained from the 1000 Genomes Project. In the second step, variants in both sets are annotated by genomic and epigenomic sources, including GERP elements, phastCon scores, repeat elements, and genome-wide epigenomics profiling data collected from ENCODE and RMEC. The third step is selecting informative features. In the last step, an ensemble module, which is a collection of ensemble base learners, is developed to adjust the class imbalance between risk variant set and benign variant set. The base learner could be an arbitrary binary classifier. The default option is the decision tree. With the test variants annotated by the same source in the second step, the trained model would output the probability of being disease-implicated for each test variant.

Figure 1


Step1: Collect risk and benign variant set
1000 Genomes
Epigenomic annotation
GWAS
Histone ChIP-seq
TF ChIP-seq
Dnase/FAIRE-seq
Disease of interest
Step2: Use epigenomic/ genomic annotation for variants
Benign set (N*T variants)
Genomic annotation
Risk set (N variants)
GREP element
PhastCons score
Repeat element
Random sample N
Step3: Select informative feature set
informative feature set
N:N
N:N
N:N
Step4: Build ensemble classifier
Learner 1
Learner 2
...
Learner T
Ensemble classifier
Test dataset
Final prediction

Feature selection

We perform feature selection to avoid over-fitting since the number of features is far greater than the number of variants, which is a typical large p, small n problem.

As the confidence of a feature is measured by p-values, we use different tests for different types of annotations to obtain the p-values. For continuous features, e.g. number of reads, we use a two-sided t-test; for binary features, e.g., peak presence, we use Fisher's exact test by constructing a two-by-two contingency table. Figure 3.17A shows the distribution of t-statistics for all epigenomic features, with the heavy tail corresponding to the informative features. The distribution of corresponding p-values is shown in Figure 3.17B, while the p-values obtained from Fisher's exact test can be found in Figure 3.17C. By comparing the distribution of p-values for the two tests, we find that p-values from Fisher's exact test are right-skewed compared to the left-skewed t-test p-values. This observation indicates fewer informative features would be selected if peak is used as the feature.

After obtaining the p-values for all features, we use cross-validation to define the p-value threshold in the feature selection step, and features with a p-value below the threshold are considered as informative features. To be specific, we set a sequence of possible p-value thresholds. For each threshold, the mean of the predicted AUC values is calculated using five-fold cross-validation on the training set, and the p-value threshold is chosen as the one with the largest predicted AUC value. Actually, the selected p-value threshold could be considered as a tuning parameter.

Choosing the appropriate base learner

Three classifier engines have been evaluated as a base learner in the ensemble module of *DIVAN*: decision tree, support vector machine (SVM), and Lasso. For SVM, we use nonlinear classifiers with radial kernel. For Lasso, we perform five-fold cross-validation to choose the best tuning parameter for penalty. Figure 3.18 shows that even if decision tree, Lasso, and SVM have comparable AUC values, decision tree shows a better precision-recall curve. Thus, decision tree is chosen as the default base learner for the ensemble module.

Ensemble method for class imbalance adjustment

The number of benign variants far exceeds the number of disease-associated variants, which makes the task of discriminating disease-specific risk variants from benign ones an inherent imbalanced two-class classification problem. A single binary classifier usually has poor predictive performance without adjusting the class imbalance. To build a balanced classifier without downsizing or duplicating the training set, we adopt an ensemble learning approach, which not only keeps all variants in the training set but also overcomes the class imbalance issue. We formularize the ensemble method as below.

We denote the benign set as N , the risk variant set as P, and the number of base learners as C. Specifically, we create two balanced classes by sampling the same number of variants N_i with replacement from the benign set as the number of variants C in the risk variant set to form one training set $N_i \cup P$ for base learner C_i . The choice of number of base learner would be large enough to ensure the unions of all N_i $(N_1 \cup N_2, ..., N_C)$ could cover most of N. The default C is set to be twice the number of benign variants in N over risk variants in P. We further denote the annotation matrix for variants in $N_i \cup P$ as X^i_{train} and the labeled $N_i \cup P$ as Y^i_{train}, the trained ensemble module is formulated as a function of training sets, which is $\boldsymbol{f}(\boldsymbol{X},\boldsymbol{Y}) = c(f(X^1_{train}, Y^1_{train}), f(X^2_{train}, Y^2_{train}), ..., f(X^C_{train}, Y^C_{train}))$. With a given variant with annotation matrix X_{test}, the probability of the given variant being disease-implicated is the average of all predictive probabilities of base learners,

$$E(Y_{test} = 1|X_{test}, \boldsymbol{f}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{C}\sum_{i=1}^{C} E(Y_{test} = 1|X_{test}, f(X^i_{train}, Y^I_{train}))$$

3.2.4 Competing methods

We compare *DIVAN* with four existing risk variant annotation and prioritization methods: GWAVA, CADD, Eigen, and GenoCanyon. For each of the below methods, we download and retrieve the pre-computed scores for the risk and benign variants. The scores are designed such that the higher the score, the better chance the variant is disease-associated. For GWAVA, we only report the set of scores with the best performance. For Eigen, we include both Eigen and EigenPC scores in the method comparison.

Supervised methods: GWAVA and CADD

CADD is a SVM-based supervised learning method. It maintains a database of pre-computed C-scores for 1000 Genomes variants and base levels for the whole human genome. GWAVA is a random forest-based supervised learning method. It maintains a database containing three sets of pre-computed scores for 1000 Genomes variants (minor allele frequency $> 1\%$) based on different choices of benign variants (TSS, unmatched, and region).

Unsupervised methods: GenoCanyon and Eigen

GenoCanyon is an unsupervised learning method, which is a two-component mixture model. It maintains a database of base-level pre-computed scores across the whole human genome. Eigen, another unsupervised learning method, is also a two-component mixture model; however, it considers feature correlation. Eigen maintains a database containing two sets of pre-computed scores for 1000 Genomes variants. One is Eigen score and another is a variation of Eigen score-EigenPC score.

3.3 Results

3.3.1 Overview of the *DIVAN* approach

The main challenge to disease-specific variant annotation is that the size of the training set is often small as the disease-specific risk variants identified by GWAS with high confidence (stringent p-values) is often very limited as the median of trait-SNP associations is only 8 for the 573 traits in the ARB. On the other hand, to improve predictive performance, we attempt to include as many genome-wide genomic and epigenomic features as possible, often thousands of them (made possible given the abundant TFs/histone modifications across many cell lines/tissues), resulting in a typical large p, small n problem [47]. Thus, simply fitting the predictive model with all features would easily cause over-fitting. To accommodate as many features as possible while avoiding over-fitting, we employ two important machine learning strategies in *DIVAN*: feature selection and ensemble learning [48]. Feature selection is used to select the informative set of features that contribute most to the predictive performance, and ensemble learning enables better predictive performance by creating a balanced risk/benign variant set in each base learner. The entire procedure of *DIVAN* is illustrated in Figure 3.1.

Diseases studied

We conduct extensive real data analyses to evaluate the performance of *DIVAN* in detecting disease-specific risk variants. Out of the total of 573 diseases/traits found in ARB, 45 of them, spanning 12 disease classes, contain at least 50 reported disease-SNP associations. These diseases are included in our study. A complete list of the diseases/phenotypes along with the number of associated risk variants are summarized in Table 3.2.

Features considered

As shown in Table 3.1, we use 1,806 epigenomic features in this study, including features related to histone modification (1002), TF binding (571), open chromatin (184), and RNA Pol II/III binding (49), spanning 261 cell lines. Features are represented by read counts in the neighborhoods of each variant, and reads from biological replicates (same factor and same cell line) are further merged. More detailed descriptions of these features can be found in the Methods section.

3.3.2 Characteristics of epigenomic profiles around risk variants

Open chromatin regions marked by selected histone marks or DNA hypersensitivity are known to harbor GWAS risk variants [4]. For demonstration

Table 3.1: Summary of feature categories in *DIVAN*

Data Source	Cell Lines	Factors	Features
REMC DNase	73	-	73
REMC Histone	109	31	735
ENCODE DNase	80	-	80
ENCODE FAIRE	31	-	31
ENCODE TF(HAIB)	19	76	292
ENCODE TF(SYDH)	31	100	279
ENCODE Histone	18	42	267
ENCODE RNA Polymerase	31	2	49
Total	261*	217*	1806

* The same cell lines or factors may appear in multiple sources

purposes, we present the sequencing read abundance pattern of selected epigenomic marks in the neighborhoods of a type 1 diabetes-associated risk variant (rs3024505) and a benign variant (rs114490664) on chromosome 1 (Figure 3.2A). One can see that the neighborhoods of risk variant rs3024505 are enriched in the active chromatin marks, H3K27ac and H3K4me1, as well as an open chromatin regions defined by DNase-seq and FAIRE-seq in the CD14 or K562 cell line. In contrast, repressive chromatin marks, such as H3K9me3 and H3K27me3 in the CD14 cell line, are depleted around risk variant rs3024505 versus benign variant rs114490664.

We further investigate whether some epigenomic features differ in terms of the distribution of neighborhood read counts between risk variants and benign ones. Those epigenomic features showing a significant distribution difference are considered informative features. As an example, FAIRE-seq in the K562 cell line shows significant read enrichment (t-test statistics 6.03, p-value $< 10^{-8}$) around risk variants associated with type 1 diabetes compared to benign ones, while H3K9me3 in the CD14 cell line shows significant read depletion around risk variants (t-test statistics -6.65, p-value $< 10^{-10}$) (Figure 3.2B).

For illustration purposes, Figure 3.2C shows 200 epigenomic profiles represented by read counts in the neighborhoods of 147 risk variants associated with type I diabetes and 147 randomly selected benign variants. The top 100 features that are mostly enriched in risk variants compared to benign ones and the bottom 100 features that are mostly depleted in risk variants. Clearly, there exist different enrichment patterns for the two sets of variants in these selected features.

For the informative features with p-values of t-test below 0.09 (0.09 is the selected p-value threshold for type 1 diabetes using the method described in the Methods section), we find that more features associated with open chromatin or TF binding show enrichment around risk variants, while more features associated with histone modifications show depletion around risk variants (Figure 3.2D). As type 1 diabetes is an immune-related disease, it is interesting to observe that all eight features associated with open chromatin in the cluster of differentiation (CD) cell line show enrichment in risk variants, while 14 features associated with H3K9me3 in the CD cell line show depletion.

Figure 2

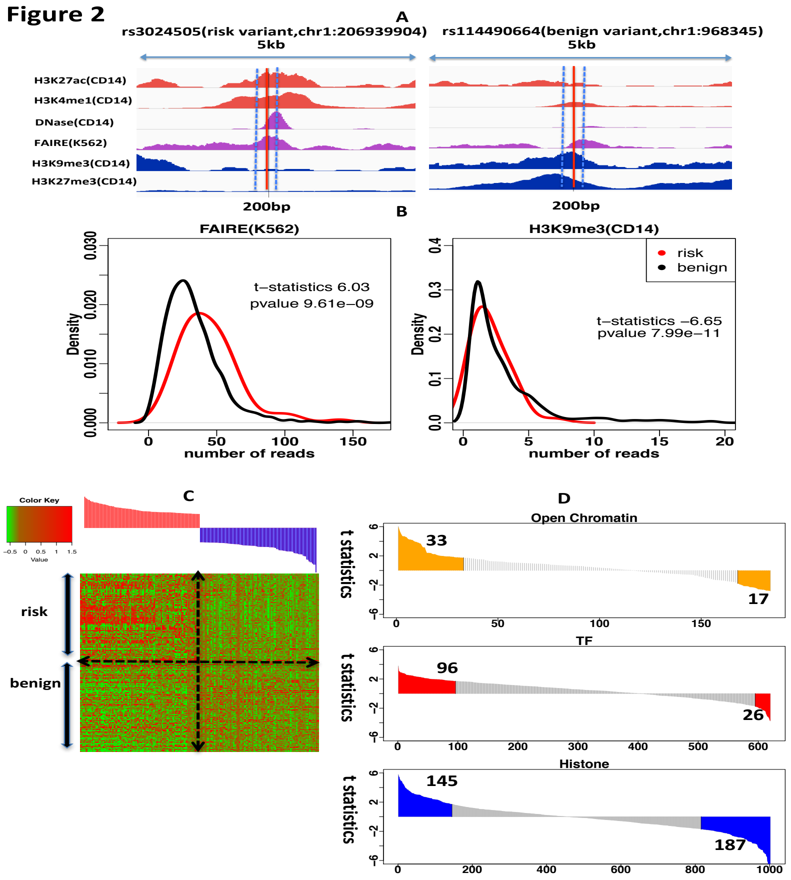


Figure 3.2: Epigenomic profiles of risk variants and benign variants. (A) Epigenomic profiles of active chromatin marks, H3K27ac and H3K4me1, repressive chromatin marks, H3K9me3 and H3K27me3, open chromatin regions in the CD14 and K562 cell lines in the neighborhoods of a risk variant, rs3024505 (chr1:206939904), associated with type 1 diabetes, and a benign variant, rs114490664 (chr1:968345). (B) Distribution of read counts for FAIRE-seq in the K562 cell line across 147 risk variants associated with type 1 diabetes and corresponding benign variants; distribution of read counts of H3K9me3 ChIP-seq in the CD14 cell line across 147 risk variants associated with type 1 diabetes and corresponding benign variants. (C) Heatmap of standardized read counts of top 100 epigenomic features and bottom 100 epigenomic features across 147 risk variants associated with type 1 diabetes and 147 corresponding benign risk variants. Epigenomic features are ranked by the t-statistics from the most enriched to the most depleted in risk variants compared to benign variants. Read counts are standardized by subtracting the average of read counts of each feature and divided by the standard deviation of read counts of each feature. (D) Distribution of t-statistics for three types of epigenomic features: TF binding, histone modification, and open chromatin. Within the informative features, 33 informative open chromatin-associated features are enriched while 17 informative open chromatin-associated features are depleted; 96 TF-associated informative features are enriched while 26 TF-associated informative features are depleted; 145 informative histone-associated features are enriched while 187 informative histone-associated features are depleted.

3.3.3 Disease-specific variant prioritization evaluation using cross-validation

Five-fold cross-validation is used to evaluate the predictive performance of different methods, and results are presented in the form of receiver operator characteristics (ROC) curves with corresponding area under the curve (AUC) values. For demonstration purposes, we present here results from four diseases: carotid artery disease (cardiovascular disease), macular degeneration (eye disease), ulcerative colitis (digestive system disease/immune disease), and multiple sclerosis (immune disease) in Figure 3.3A.Figure 3.8 shows the corresponding precision recall curves for the four diseases. The remaining 41 ROC curves are presented in Figure 3.9. Overall, *DIVAN* achieves the best predictive performance among all methods, with AUC values ranging from 0.65 to 0.88 (median 0.74), followed by GWAVA and GenoCanyon. For a comprehensive comparison, we present the AUC values of all methods compared across 45 diseases in a heatmap (Figure 3.4A). The AUC values are included in Table 3.3, and the average Matthews correlation coefficient (MCC) values of different methods across 45 diseases are shown in Table 3.4. Moreover, we find *DIVAN* performs the best among immune-related diseases, followed by multiple eye diseases and urogenital disorders. On the other hand, identifying risk variants associated with mental disorders and cardiovascular diseases seems more challenging for *DIVAN* (Figure 3.4B).

Figure 3

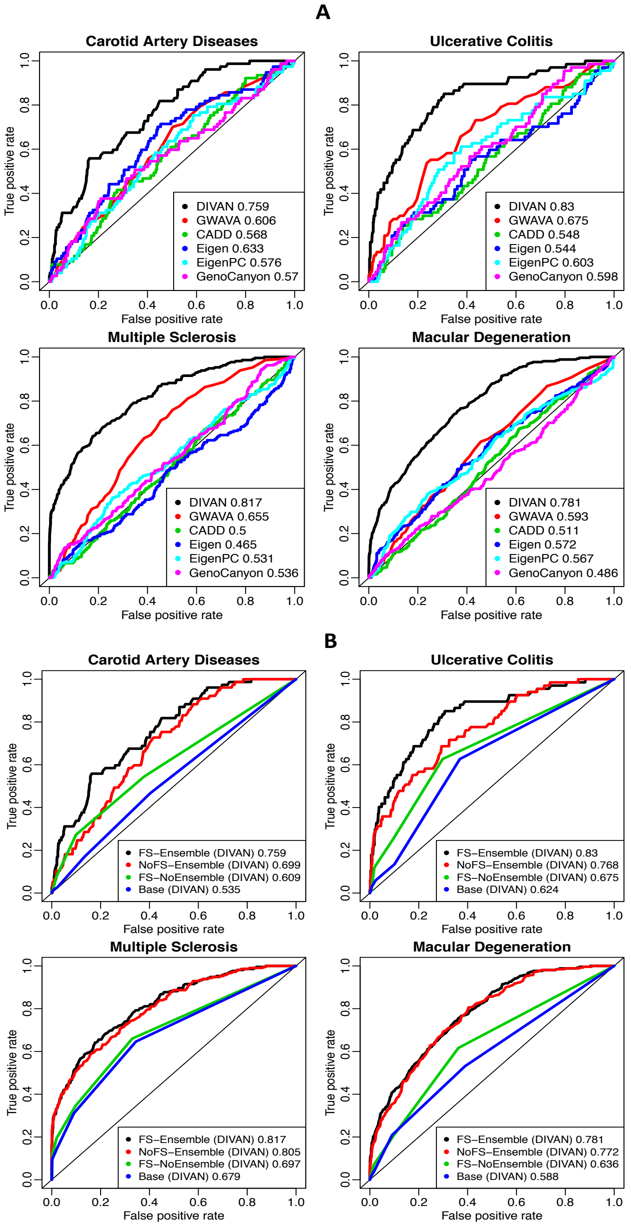


Figure 3.3: Predictive performance of five-fold cross-validation on four diseases: carotid artery disease, macular degeneration, ulcerative colitis, and multiple sclerosis. (A) ROC curves comparing the predictive performance among *DIVAN* and CADD, GWAVA, Eigen, Eigen-PC, and GenoCanyon for the four diseases. (B) ROC curves showing the effectiveness of feature selection and ensemble method by comparing feature selection and ensemble combined, feature selection only, ensemble only, and the baseline case: neither feature selection nor ensemble.

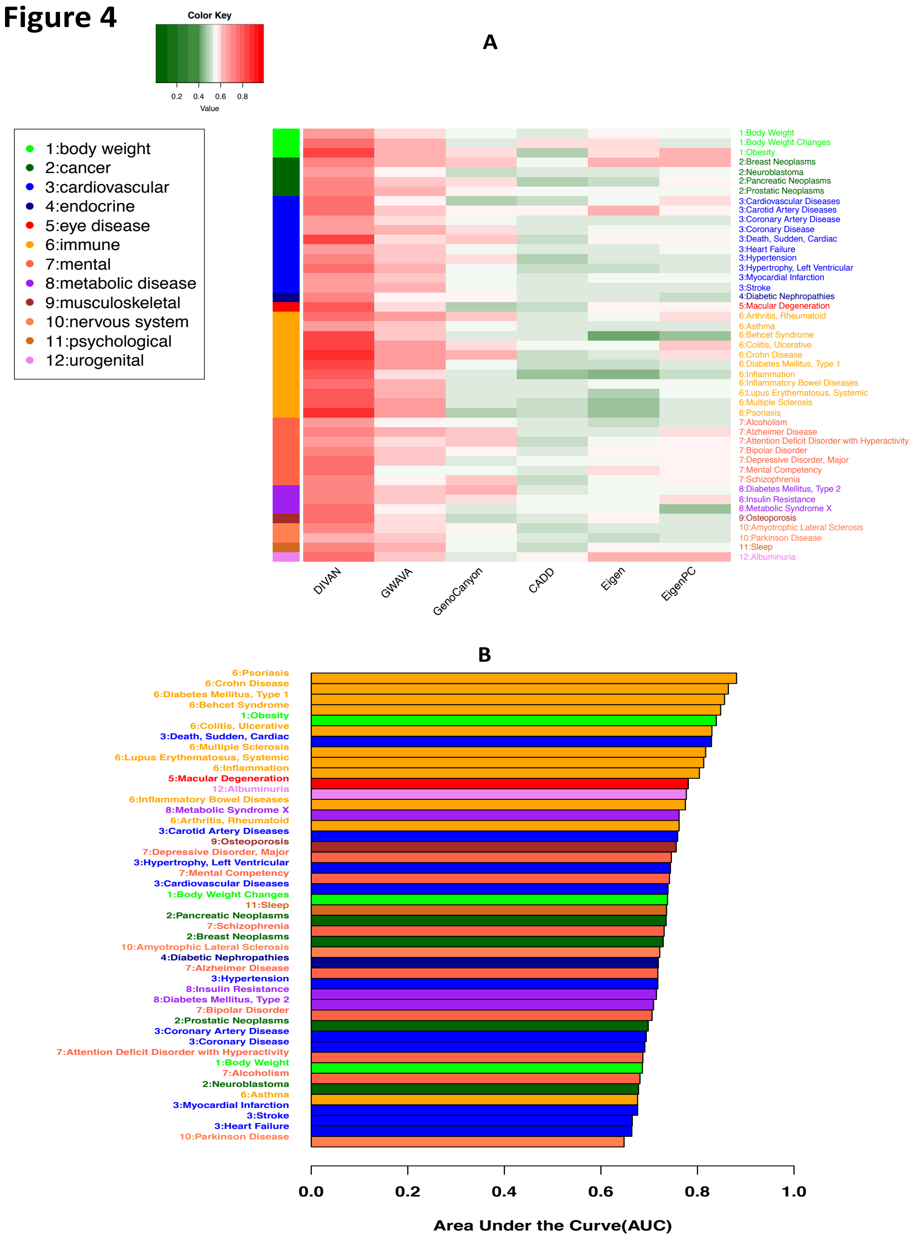


Figure 3.4: Predictive performance of five-fold cross-validation across 45 diseases in 12 disease classes. (A) Heatmap of five-fold cross-validation AUC values for predictive performance comparison among *DIVAN* and CADD, GWAVA, Eigen, Eigen-PC, and GenoCanyon across 45 diseases in 12 disease classes. (B) Bar charts of five-fold cross-validation AUC values of *DIVAN* across 45 diseases in 12 disease classes ranked in decreasing order. Disease classes are color-coded.

Disease specificity of variant annotation

A key feature of *DIVAN* lies on its disease-specificity, which means the predictive model is trained disease by disease using annotated disease-specific variants. To justify the necessity of the disease-specific assumption, we conduct an experiment in which a model trained using variants from one disease is subsequently applied to classify variants annotated for a different disease. In the experiment, we use four diseases from distinct disease classes: carotid artery disease (cardiovascular disease), macular degeneration (eye disease), Alzheimer's disease (mental disease), and multiple sclerosis (immune disease). For the same disease training and testing, we report the AUC values of five-fold cross-validation. As expected, we find decreased AUC values when a model trained in one disease is applied to a different disease (Figure 3.10), which confirms the advantage of using the disease-specific model adopted by *DIVAN*.

Effectiveness of feature selection and ensemble learning

To demonstrate the effectiveness of adopting the feature selection and ensemble learning strategies, we conduct a performance comparison using four different settings: baseline (no feature selection, no ensemble learning), feature selection only, ensemble learning only, and feature selection combined with ensemble learning. Again, we use the four aforementioned diseases as representatives and five-fold cross-validation to evaluate the predictive performance, and results are presented in the form of ROC curves with corresponding AUC values (Figure 3.3B), as well as precision recall curves (Figure 3.11). The results confirm that feature selection combined with ensemble learning achieves the best performance. Moreover, either feature selection or ensemble learning alone improves the predictive performance compared to the baseline.

Contribution of different feature groups

Since most epigenomic features used in *DIVAN* come from three groups: TF binding, histone modifications, as well as open chromatin (DNase-seq and FAIRE-seq), it would be interesting to investigate which feature group contributes relatively more to risk variant identification. In addition, existing methods use called peaks from sequencing-based assays to represent epigenomics features, which is a binary indicator of whether a variant overlaps with any peak (referred to as peak hereafter). Instead, by default *DIVAN* uses read counts in the neighborhood of the variant as the feature representation (referred to as read hereafter) for the robustness of predictive performance when limited features are available.

To compare performance with different feature groups and different feature representations, we apply *DIVAN* to the aforementioned four diseases in different settings. We find that no matter whether peak or read is used, using all feature groups achieves the best performance, as expected; and using features related to histone modifications alone could achieve better predictive performance than any other feature group. However, the contribution of each feature group when using peak and read differs slightly (Figure 3.12). Specifically, using features related to histone modifications alone achieves comparable predictive performance no matter whether peak or read is used, whereas using read shows much better performance than using peak as the feature representation for TF binding and open chromatin. A possible explanation is that the continuous read counts are more sensitive than peak overlap in detecting subtle differences between risk and benign variants, especially when genome-wide coverage of the feature is relatively sparse, such as TF binding or open chromatin.

3.3.4 Disease-class variant prioritization

Diseases/phenotypes in the same disease/phenotype class are believed to be likely more phenotypically related to each other, and we want to investigate the predictive performance when including risk variants from diseases/phenotypes that belong to the same class into the training set. This strategy is called disease-class specificity, which is an extension of the disease-specificity strategy adopted so far. Because only a handful of risk variants have been identified by GWAS for most of the diseases/phenotypes, this strategy is rather attractive since it allows the critically needed boost to the training set when only a few variants have been identified.

To demonstrate the utility of this assumption, we perform a "leave-one-disease-out" testing approach; that is, we build the model using known risk variants of all but one disease within the disease class, and apply the model to identify risk variants for the omitted disease. To illustrate the performance of this strategy, we take five immune diseases reported in ARB, including rheumatoid arthritis, asthma, type 1 diabetes mellitus, systemic lupus erythematosus, and multiple sclerosis, as examples. We observe promising predictive performance since all AUC values are above 0.8, except for asthma (Figure 3.13).

3.3.5 Applying *DIVAN* to disease-specific variants in the GRASP database

To further evaluate the performance of *DIVAN*, we take on a different testing set using risk variants in the GRASP database [1], which includes around 8.87 million SNPs identified from 2,082 GWAS studies (accessed Mar 30th, 2016). The large size of the database is mainly due to the fact that a less stringent p-value threshold (0.05) is used for risk SNP inclusion. For the testing set, we are able to match 36 out of 45 ARB diseases in GRASP, and for each disease, we only keep risk variants in non-coding regions with a p-value less than 10^{-4}, and further exclude risk variants collected in ARB for the same disease/phenotype, we further remove duplicated variants (the same SNP being reported multiples times from different platforms or different studies) in GRASP. The corresponding benign variants are selected by randomly sampling ten times the number of risk variants of each disease from the catalog of the 1000 Genomes Project, excluding all GRASP variants.

For each of the 36 diseases, we use the same set of risk variants in ARB as the training set, and the risk variants in GRASP but not in ARB as the testing set. The number of training and testing variants for the 36 diseases are summarized in Table 3.2. To avoid possible bias due to sampling variability, for each disease, we repeat the whole procedure ten times with a different set of benign variants (by random sampling) each time and calculate the average AUC values. For illustration purposes, we compare the AUC values of different methods for the four representative diseases (Figure 3.5A). *DIVAN* shows the highest AUC values once again. For an overview, we present the average AUC values of different methods across all 36 diseases in a heatmap (Figure 3.5B) and in Table 3.5, and the average MCC values of different methods across 36 diseases are shown in Table 3.6. Overall, *DIVAN* shows the best performance as it achieves the highest AUC values in 27 out of 36 diseases, and is close to the best in the remaining nine diseases. GWAVA has the second-best predictive performance for obtaining the highest AUC values in four diseases, followed by GenoCanyon, with the highest AUC values in three diseases. For AUC values achieved by *DIVAN*, we find it performs the best for immune-related diseases, which is consistent with the findings from the 45 ARB diseases using five-fold cross-validation.

Figure 5

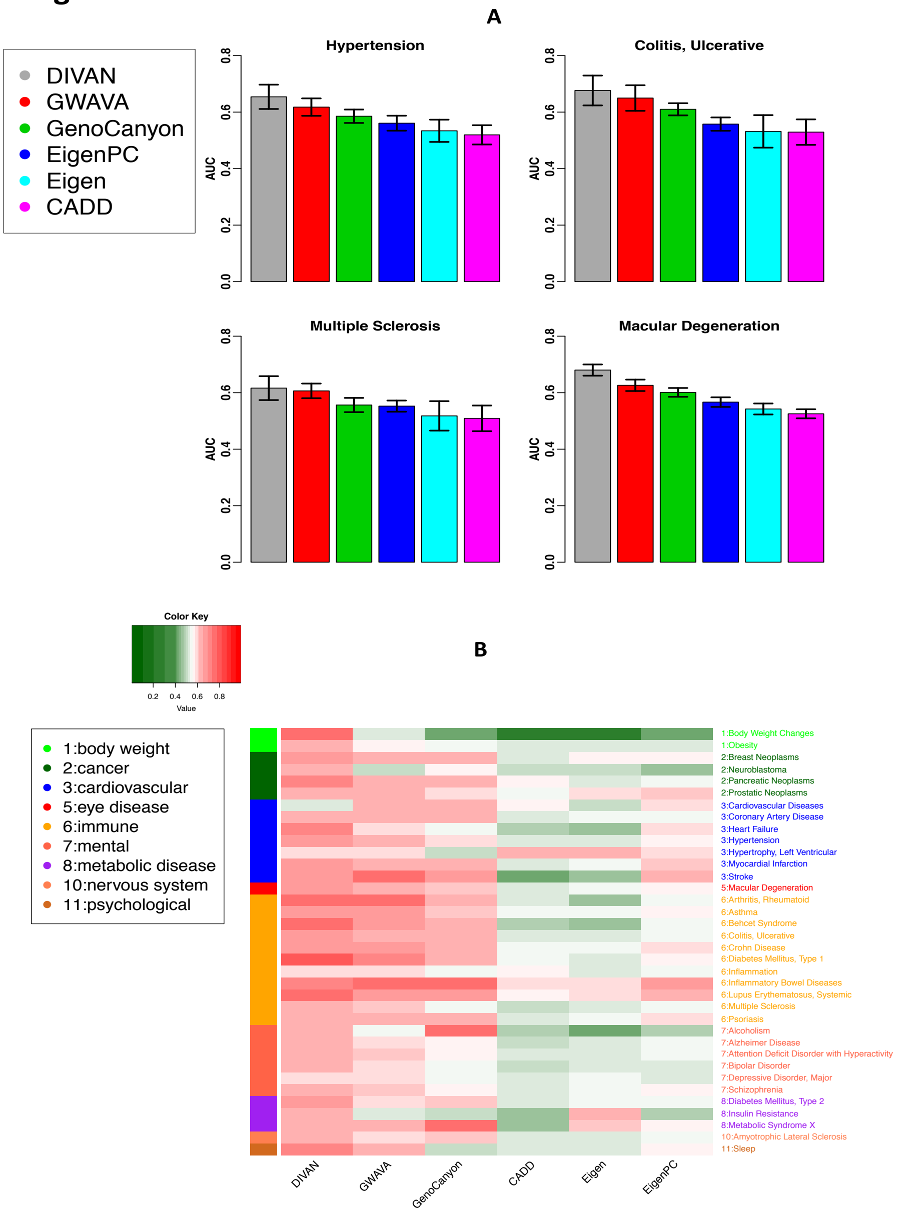


Figure 3.5: Predictive performance on 36 diseases in GRASP database (A) Bar charts of AUC values among *DIVAN* and CADD, GWAVA, Eigen, Eigen-PC, and GenoCanyon for four diseases: hypertension, macular degeneration, ulcerative colitis, and multiple sclerosis. The bar charts are sorted by the mean AUC values, and the error bar describes the standard deviation. The training set is risk variants collected from the ARB, and the testing set is the risk variants collected from GRASP. (B) Heatmap of mean AUC values for predictive performance comparison among *DIVAN* and CADD, GWAVA, Eigen, Eigen-PC, and GenoCanyon across 36 diseases investigated.

3.3.6 Applying *DIVAN* to regulatory variants in the HGMD database

So far the risk variants we use are collected from ARB and GRASP databases where the variants are disease-implicated by GWAS. It is also of great interest to test variants from other sources. There are well-known databases available that contain curated variants, which are often carefully selected by experts. For example, variants with pathogenic or non-pathogenic effects in ClinVar are collected from literature evaluation, clinical testing and research, and reviewed by different expert groups. Mutations in HGMD are collected from the literature. Unfortunately, among the collected 194 non-coding ClinVar variants used by GWAVA, none of them are associated with any of the 45 diseases in ARB used in the training set. This might be attributed to the fact that the majority of the variants in ClinVar are either coding variants or associated with Mendelian diseases. Because *DIVAN* is disease-specific, and requires training and testing set from the same disease, we choose not to test *DIVAN* on ClinVar variants.

For HGMD, we collect 1614 disease-associated regulatory variants used by GWAVA. In order to find out which disease is associated with each variant, we manually query each of the 45 diseases on HGMD website to retrieve all regulatory variants in HGMD that are associated with any of the 45 diseases. By looking for the overlap between the two sets of variants, we identify 117 unique autosomal variants (excluding sex chromosomes and mitochondria) associated with at least one of the 45 diseases.

Among these 117 variants, very few of them (less than 15) map to any one of the 45 diseases individually, which is not enough to get meaningful comparison results for a disease-specific study. Fortunately, we find that there are 34 variants associated with at least one disease in the immune disease class including Asthma, Behcet syndrome, Ulcerative Colitis, Crohn's disease, Inflammatory bowel diseases and Systemic lupus erythematosus. Hence we group the 34 variants associated with diseases in the immune disease class as an independent testing set, conduct a disease class-specific analysis using *DIVAN* and compare the predictive performance with other methods. The corresponding benign variants of the 34 immune disease-associated variants in the testing set are chosen in the same way as for GRASP testing set. For this experiment, we do not include GWAVA since it uses the 1614 HGMD variants as its training set. For *DIVAN*, we train a disease class-specific model by pooling all the variants in ARB that are associated with any of the aforementioned six immune-related diseases together in the training set. For other methods that are not disease-specific, we use their pre-computed scores. The AUC values are summarized in Table 3.7. There we see that *DIVAN* virtually tied with GenoCanyon, and is better than CADD, Eigen and EigenPC. The results demonstrate *DIVAN*'s robust performance on different independent testing sets.

3.3.7 Applying *DIVAN* on synonymous mutations

Though *DIVAN* is designed for the identification of non-coding variants, it is interesting to see how *DIVAN* performs on coding variants especially synonymous mutations.

We collect synonymous mutations from the online database dbDSM [49], which is a manuallycurated database that collects 1936 synonymous mutations-disease association entries, In total, we have 1109 autosomal synonymous mutations (excluding sex chromosomes and mitochondria). We find seven diseases associated with more than 20 synonymous mutations in dbDSM are also among

45 diseases in ARB; hence we use the seven diseases for performance comparison. The corresponding benign variants for each disease in the testing set are chosen in the same way as for GRASP testing set. The AUC values are reported in Table 3.8.

The results show that overall GWAVA performs the best while *DIVAN* is on par with the other methods, suggesting *DIVAN* is not as good in predicting coding variants as it predict non-coding variants. This is not surprising since all the features and the training procedure used by *DIVAN* are optimized for prioritizing non-coding variants. On the other hand, GWAVA uses HGMD regulatory mutations as the training set in which 75% of them lies within a 2kb window around TSS, indicating majority of HGMD mutations is close to the coding regions. That might explain the better performance of GWAVA. In the future, we plan to extend *DIVAN*'s functionality to identify disease-specific coding variants, by perhaps adding coding-region specific features.

3.3.8 Exploration and interpretation of features

Variability of factors across cell types

A key merit of *DIVAN* is its ability to consider a large number of cell type-specific epigenomic profiles as features to accommodate the cell type-specific nature of the epigenome, which aims to include as many features as possible, without any screening up front, and let the algorithm select informative features automatically. For some existing methods, such as GenoCanyon and Eigen, epigenomic profiles of the same factor across different cell types are collapsed to simplify the model or speed up computation. That way, the plastic epigenomic profiles across cell types are ignored.

To show the variability of epigenomic factors across cell types and the dynamic profiles of epigenomic factors across diseases, we obtain the p-values from t-tests conducted between the risk and benign variants across 1,806 epigenomic features for the four aforementioned diseases. We sort the factors profiled in more than ten cell types by the number of features remaining in the informative feature set and plot the log-transformed p-values (Figure 3.6A). One can see that there is considerable variability of the p-values for the same factor across different cell types, which confirms the necessity of considering the combination of factors and cell types as the epigenomic features. Moreover, the rank of factors varies from disease to disease, further reflecting the variable nature of these factors.

Overall, we see that the top-ranked factors for the four diseases are two repressive chromatin marks, H3K9me3 and H3K27me3, followed by open chromatin, and two active chromatin marks, H3K4me1 and H3K36me3. The top factor is H3K9me3 for carotid artery disease and macular degeneration, and H3K27me3 for ulcerative colitis and multiple sclerosis. Both factors are repressive chromatin marks. JunD, Pol2, and p300 also frequently rank high. On the other hand, active chromatin marks, e.g., H3K4me3 and H3K27ac, do not always appear among the top factors. Moreover, it is interesting to see that EZH2 and H3K27me3 both top rank in multiple sclerosis and ulcerative colitis as EZH2 represses gene transcription by mediating H3K27me3 methylation [50].

Informative features across different diseases

As the informative feature set helps improve the predictive performance, we further investigate the number of informative features selected within three feature groups: histone modification, TF binding, and open chromatin (Figure 3.6B). Overall, the total numbers of informative features selected vary from disease to disease, ranging from 664 (body weight) to 34 (inflammation) if read is used as the feature, while the overall numbers of informative features decrease, ranging from 549 (type 2 diabetes) to 41 (obesity) if peak is used as the feature (Figure 3.14)

We also observe that the histone modifications feature group contributes more to informative features than the TF binding or open chromatin feature group. Moreover, more TF-associated and fewer histone-associated features show up in the informative feature set when read rather than peak is used as the feature (Figure 3.6B and Figure 3.14).

Figure 6

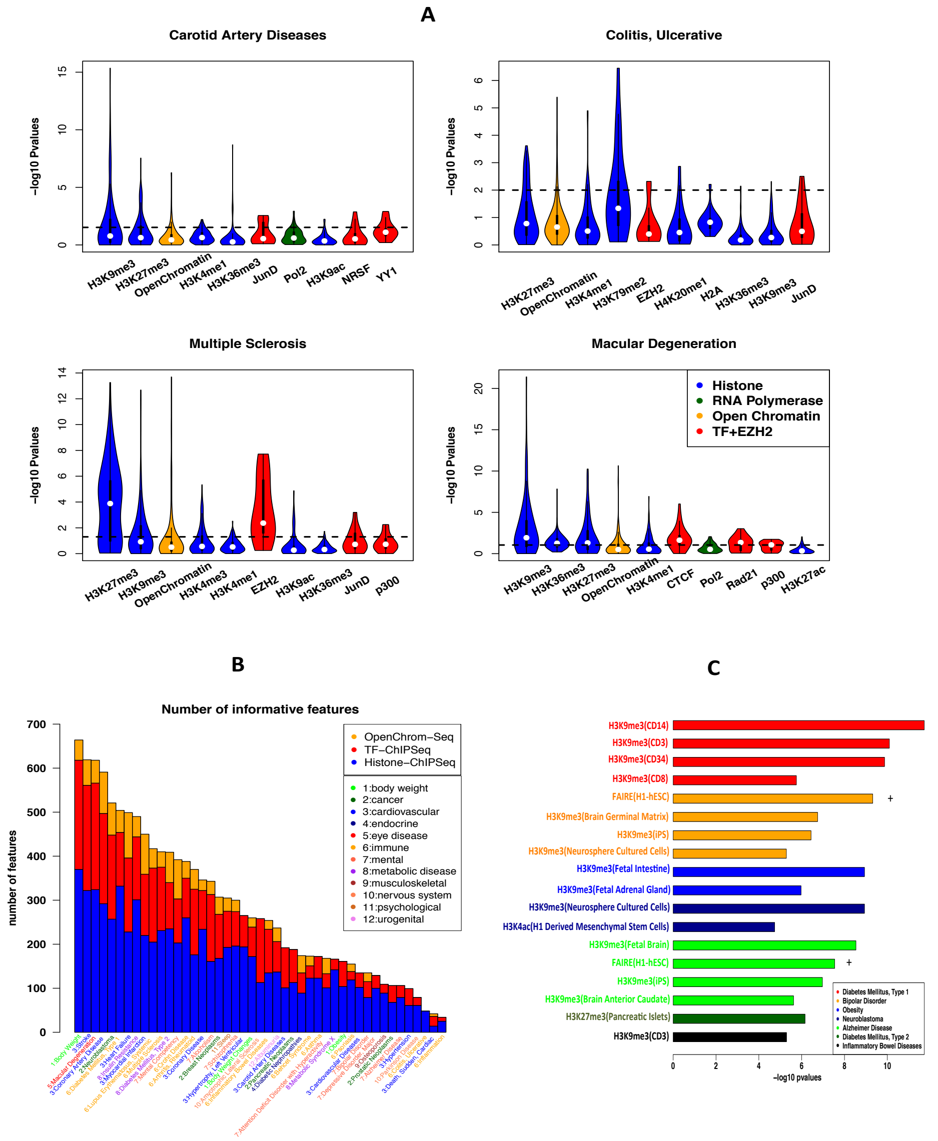


Figure 3.6: Exploration and interpretation of epigenomic features. (A) Violin plot for the distribution of -log10 p-values of the top 10 factors (TF binding/histone modification/open chromatin/RNA polymerase) associated with more than 10 epigenomic features for four diseases: carotid artery disease, macular degeneration, ulcerative colitis, and multiple sclerosis. P-values are calculated by t-test on the read counts in the neighborhoods of the risk variants and benign variants. (B) Number of informative features for three feature categories (TF binding/histone modification/open chromatin) for 45 diseases across 12 disease classes using read as the feature value. (C) Bar chart of -log10 p-values for top-ranked features for selected diseases: type 1/type 2 diabetes, bipolar disorder, obesity, neuroblastoma, Alzheimer's disease, and inflammatory bowel disease.

Interpretation of top features

Although the main goal of *DIVAN* is to distinguish disease-specific risk variants from the vast pool of benign ones, we demonstrate that the feature selection step could also help identify top features that are biologically meaningful.

To illustrate, we present some of the top features identified from selected diseases, and the observed enrichment/depletion patterns are readily interpretable (Figure 3.6C). For example, we find that H3K9me3 in CD cells, known to be on the cell lineage that leads to immune-related disease, is depleted around the risk variants associated with type 1 diabetes. Interestingly, H3K9me3 in CD cells is also depleted around risk variants associated with another immune-related disease: inflammatory bowel disease. H3K27me3, another repressive chromatin mark, in pancreatic islet cells is found to be depleted around risk variants associated with type 2 diabetes, a disease caused by pancreatic islet dysfunction. For bipolar disorder, we find open chromatin regions in H1 cells measured by FAIRE-seq are enriched, while H3K9me3 in the brain germinal matrix, iPS, and neurosphere cultured cells is depleted in the neighborhoods of their risk variants. Risk variants associated with another mental disorder, Alzheimer's disease, are also depleted of H3K9me3 in fetal brain, iPS, and brain anterior caudate cells, but enriched of open chromatin regions in H1 cells measured by FAIRE-seq. Risk variants associated with obesity are depleted of H3K9me3 in fetal intestine and fetal adrenal gland cells. H3K9me3 in neurosphere cultured cells and H3K4ac in H1-derived mesenchymal stem cells are depleted around risk variants associated with neoplasms. For the above diseases investigated, we find that H3K9me3 consistently shows depletion, while open chromatin consistently shows enrichment around risk variants.

H3K9me3 is the most informative factor for risk variant identification across diseases/phenotypes

In addition to identify informative epigenomic factors for differentiating risk variants from benign variants in each individual disease, we also want to identify the "frequent fliers," i.e., the epigenomic factors that contribute to a wide spectrum of diseases. To find out, for each disease, we check which factors are over-represented in the list of identified informative features using a binomial test. Let n_i represent the number of informative features in disease i; N the total number of features in this study (1806); m_{ij} represent the number of features associated with factor j in disease i; k_{ij} represent the number of informative features associated with factor j in disease i. The p-value for factor j over-represented in disease i could be calculated as,

$$p(x > k_{ij}|n_i, p_{ij}) = \sum_{x=k_{ij}+1}^{n_i} \binom{n_i}{x} \cdot \hat{p_{ij}}^x (1 - \hat{p_{ij}})^{n_i - x}$$

$$\hat{p_{ij}} = \frac{n_i}{N}$$

Any factor with p-value less than the Bonferroni corrected threshold (0.05/45) is said to be over-represented in the disease i. At the end, for each factor, we tally the number of times it is over-represented across all 45 diseases (Figure 3.7A). We find that H3K9me3 and open chromatin are the top informative factors; H3K9me3 is over-represented in 34 out of 45 diseases, while open chromatin is over-represented in 25 out of 45 diseases.

Consistent with previous finding that histone marks are the most frequent features to be ranked at the top among the three types of epigenomic features (Figure 3.6B), Figure 3.7A shows that histone marks are associated with more diseases than TFs overall; however, to our surprise, among the histone marks that are most significant, most of them are associated with repressive chromatin, such as H3K9me3 and H3K27me3, and H3K9me3 in particular. We also confirm the well-documented fact that open chromatin marked by DNase-seq and FAIRE-seq is enriched around risk variants [51].

To further illustrate the dominance of H3K9me3 compared to other histone marks among top features, we plot the enrichment of different histone marks sorted by p-values for type 1 diabetes (Figure 3.7B). H3K9me3 is the most over-represented factor among the informative features, associated with 40% of the top 100 features, followed by H3K27me3 (29%), and H3K4me1 (4%). Other marks associated with active chromatin, H3K4me3, H3K27ac, and H3K9ac, are not significantly enriched among the top features.

It has been shown that genomic regions marked by active chromatin, such as H3K4me1, are enriched near GWAS-identified risk variants [4, 29], so we are interested to see whether regions marked by repressive chromatin, such as H3K9me3, are depleted by risk variants. To do this, we collect the called peaks of H3K9me3 and H3K4me1 in the CD14 cell line, known to be from the cell lineage that leads to immune-related diseases, and calculate the enrichment of risk variants associated with each of the 45 diseases in those peaks using traseR [52], an R package that is capable of searching and ranking diseases/phenotypes for a given set of genomic regions based on the enrichment level of trait-associated SNPs. We plot the p-values on the logarithm scale of the enrichment test across 11 immune diseases (Figure 3.7C). We find that none of the immune-related diseases are statistically significantly enriched in H3K9me3, while all but asthma and inflammation are statistically significantly enriched in H3K4me1.

Figure 7

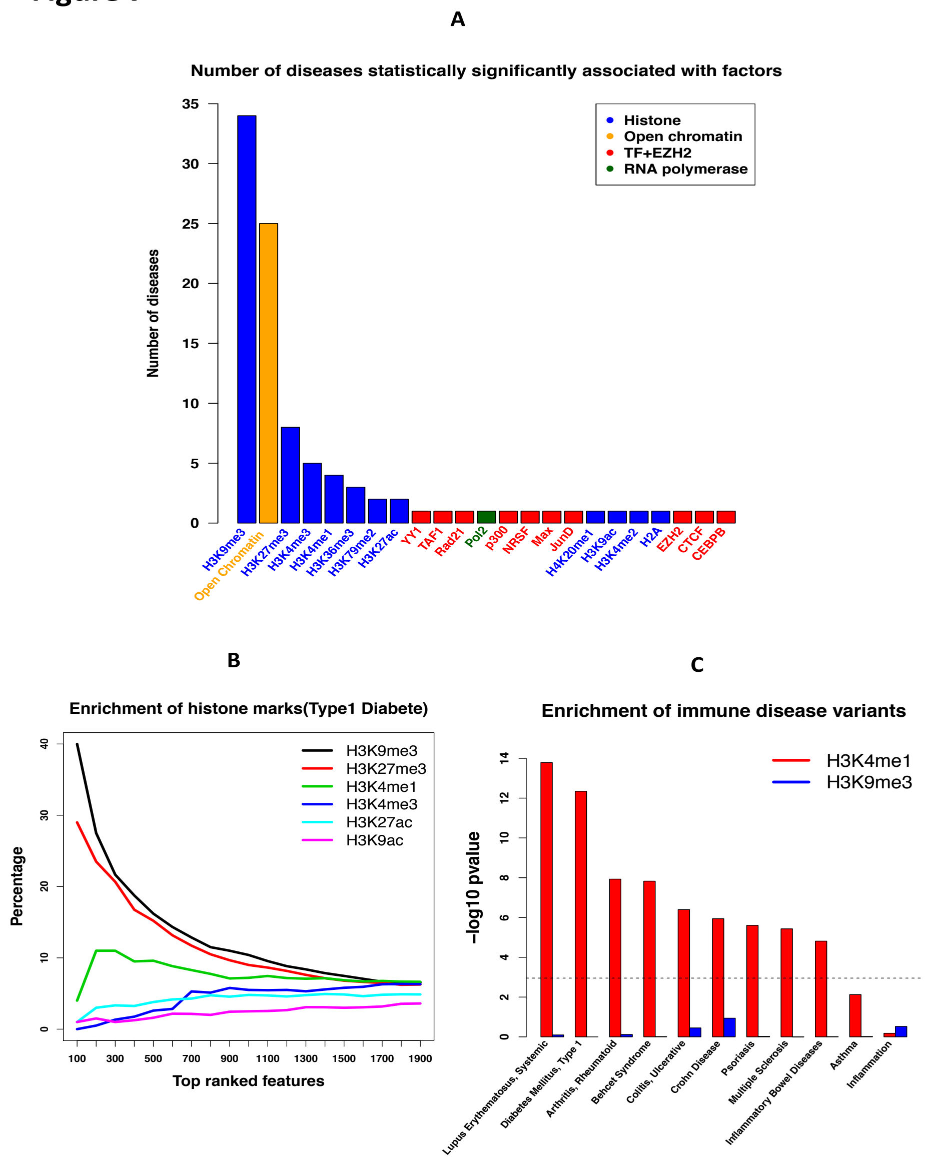


Figure 3.7: Association between factors and diseases. (A) Number of diseases statistically significantly associated with different factors (TF binding/histone modification/open chromatin/RNA polymerase). (B) Enrichment of different histone marks among top features for type1 diabetes. (C) Enrichment of risk variants associated with immune disease in peaks of active chromatin mark H3K4me1 in the CD14 cell line and peaks of repressive chromatin mark H3K9me3 in the CD14 cell line.

3.3.9 Additional tests on more settings of *DIVAN*

For a complex machine learning problem like what we are tackling, different settings in training and testing might cause overestimate or underestimate of the actual performance. Here we carry out additional tests under different experimental settings to investigate the robustness of *DIVAN*'s performance.

Different sources of benign variants in the training set

Currently, the set of benign variants are chosen from the 1000 Genomes (phase I). Since the risk variants are mostly GWAS SNPs, to avoid picking up features that might be a by-product of SNP design and selection, we instead choose benign variants from GWAS SNPs as well, found on one of the latest GWAS genotyping array-Affymetrix Genome-Wide Human SNP array 6.0. To be specific, we collect 900,611 non-coding GWAS SNPs out of 934,968 GWAS SNPs from the SNP annotation file (`http://www.affymetrix.com/Auth/analysis/downloads/na35/genotyping/GenomeWideSNP_6.na35.annot.csv.zip`) to construct the set of benign variants for each disease. Using the new set of benign variants, we retrain the disease-specific model for the 45 diseases in ARB, obtain the CV-AUC values (Table 3.9) for the five-fold cross-validation and the predicted AUC values for the 36 diseases in GRASP in the independent test (Table 3.10 and 3.11).

For the 45 diseases found in ARB, the Pearson correlation coefficient between the two sets of AUC values is 0.979, (p-value $< 2.2\text{e-}16$). The average CV-AUC values for the 45 diseases changes from 0.745 (sd: 0.060) to 0.742 (sd: 0.061). For the 36 diseases found in GRASP, the Pearson correlation coefficient between the two sets of AUC values is 0.950, (p-value $< 2.2\text{e-}16$). The average predicted AUC values for the 36 diseases changes from 0.661 (sd: 0.055) to

0.658 (sd: 0.061). The results show that the AUC values are similar either using SNPs from the GWAS genotyping array or using SNPs 1000 Genomes to form the set of benign variants in the construction of disease-specific model.

Different criteria of choosing benign variants in the training set

By default, *DIVAN* uses distance to the nearest TSS as the criterion to choose a set of benign variants such that distances to the nearest TSS matched (have a similar empirical distribution) with those of the risk variants. The distance to TSS-matched criterion keeps the two sets (risk and benign) on leveled grounds in their chromatin profiles because non-coding disease-associated variants in ARB tend to locate close to TSS (mostly within 200kb, Figure 3.15) and chromatin landscape is quite different between promoter regions and intergenic regions. The same criterion has also been adopted by GWAVA.

We also adopt an alternative and perhaps more stringent criterion to choose the set of benign variants in the training set in which we require that all benign variants have to be located within 10kb of a risk variant. Here we use a slightly wider region than the 1kb region used by GWAVA but narrower than the 100kb region used by Eigen. This is because the histone mark profiles, which *DIVAN* used predominantly, typically extend to a few kbs.

We conduct another test using the new region-matched benign set (denoted as region) and compare the results with the results obtained earlier using the distance to TSS matched benign set (denoted as TSS). We find average CV-AUC values for the 45 diseases in ARB changes from 0.745 (sd: 0.060) to 0.680 (sd: 0.037); and the average AUC values for the 36 diseases in GRASP changes from 0.661 (sd: 0.055) to 0.637 (sd: 0.043). The CV-AUC values are shown in Table 3.12. The decrease of predictive performance using the region-matched benign set is consistent with what is observed in GWAVA. Despite the slight drop in performance when using the region-matched criterion, *DIVAN* still maintains its lead over all the competitors tested. In the independent test, among the 36 diseases in GRASP, *DIVAN* is the best performer in 23 diseases, followed by GWAVA (7 diseases), GenoCanyon (4 diseases) and Eigen (2 diseases). The predicted AUC values are shown in Table 3.13 and 3.14.

Impact of nearby variants on cross-validation

In the cross validating study described earlier, although there is no overlap of variants between the training and the testing sets, it is possible that a risk variant in the testing set is located near a risk variant in the training set which may potentially inflate the CV performance. In order to eliminate such influence, before preforming CV, we further remove risk variants that are too close to each other and do the same thing for benign variants as well. To be specific, we first sort all risk variants (or benign variants) based on their genomic locations, and only keep one variant if multiple variants happen to be less than 10kb away. That way, we make sure that neither training folds nor the testing fold contains risk variants (or benign variants) at the same or nearby location (10kb). The updated numbers of risk variants for the 45 diseases in ARB are shown in Table 3.2. The numbers of risk variants of the 45 diseases in ARB decrease around 16% on average.

To evaluate the impact of this change, we conduct an experiment using the new rule and compare the results with those obtained before. We retrain all the disease-specific models and calculate the CV-AUC values for the 45 diseases in ARB (Table 3.15). We find that using the new rule, the average AUC values for the 45 diseases changes from 0.745 (sd: 0.060) to 0.736 (sd: 0.056) and the

Pearson correlation coefficient between the two sets of CV-AUC values is 0.917 (p-value $<$ 2.2e-16). In conclusion, we see little difference the new rule has on the outcome of CV-AUC values. *DIVAN* still outperforms all the competitors by a comfort margin.

Impact of nearby variants on independent test

For the independent test described earlier, although we have excluded all ARB variants from the GRASP testing set, it is possible that some variants in the GRASP testing set are located near ARB variants used in the training set, which may affect the independent test performance. Therefore, to eliminate such influence, for each disease, we further exclude risk variants in the GRASP testing set that are close to risk variants found in the ARB training set. The updated numbers of disease-associated SNPs for the 36 diseases in GRASP can be found in Table 3.2. The numbers of risk variants of 36 diseases in GRASP decrease around 7% on average.

To be specific, for each disease, hypertension for example, we exclude any hypertension-associated variants in the GRASP testing set that fall within 10kb of any hypertension-associated variants found in the ARB training set. We then repeat the performance comparison experiment using the newly reduced testing set. The results are summarized in Table 3.16. The predicted AUC values are shown in Table 3.17. The average MCC values are shown in Table 3.18.

From Table 3.16, we see that despite slightly dampened performance, removing variants in the testing set that are close to variants in the training set does not change the fact that *DIVAN* significantly outperforms all the other competing methods that have been tested.

Different size of benign set

Because there are much more benign variants than risk variants, it is an interesting question that how many benign variants should be included in training set. In the CV described earlier, we choose the size of the benign variants to be ten times that of the risk variants. Here we investigate whether increasing the size of the benign set to 100 times of the risk set has any effect on the predictive performance. We calculate Pearson correlation coefficient between the two sets of CV-AUC values obtained from the two settings. We also summarize the mean and standard deviation of the CV-AUC values for each setting in Table 3.19. The CV-AUC values are shown in Table 3.20. The predicted AUC values are shown in Table 3.21 and 3.22. Our result suggests that overall, increasing the size of the benign variant set when set up the training model does not change much in terms of the predictive performance in CV.

For the independent test, we also experiment with increasing the number of benign variants from 10 times that of the risk variants to 100 times for each disease and check whether the different level of imbalance in the testing set has any effect on the prediction performance. The new predicted AUC values are shown in Table 3.23 and 3.24, where we could see that the AUC values remain stable on the 36 diseases in GRASP. The Pearson correlation coefficient between the two sets of 36 predicted AUC values is 0.999 (p-value $< 2.2\text{e-}16$) when the number of benign variants is 10 and 100 times of risk variants respectively. Thus, we see that increasing the size of the benign variants has little effect on the predictive performance for the independent test, which suggests that the performance of *DIVAN* is not significantly affected with different level of risk/benign imbalance.

3.4 Discussion

In this work we describe *DIVAN*, a feature selection-based ensemble learning framework for identifying disease-specific, non-coding risk variants. *DIVAN* performs favorably when compared to existing state-of-the-art methods, both supervised (CADD, GWAVA) and unsupervised (GenoCanyon, Eigen), for detecting disease-specific non-coding risk variants. From a clinical perspective, it is of great practical and conceptual value to evaluate the impact of a variant on individual disease/phenotype. Because the number of disease-implicated variants is far fewer than the number of static genomic and epigenomic annotations for most diseases, to avoid potential over-fitting in the high-dimensional setting, we employ model selection to remove non-informative features. Besides feature selection, the ensemble method is adopted to improve the predictive performance due to the nature of the imbalance between risk variants and benign ones. This combination of feature selection and ensemble method makes *DIVAN* more powerful and robust.

Another major finding of the study is that the depletion of H3K9me3, a histone mark associated with repressed chromatin, is the most prominent hallmark around risk variants. Overall, histone marks contribute more informative features in risk variant identification than transcription factors and open chromatin in *DIVAN*. We believe the above findings have profound implications for understanding the mechanism behind the way non-coding variants make their impact on diseases/phenotypes via epigenetic modifications.

A key emphasis of *DIVAN* lies on disease specificity. We believe this can be achieved by using variants that are specific to that disease in the training set as opposed to including all variants that have shown associations with some diseases. Despite a small training set, we show that advanced statistical learning techniques can help us overcome this challenge and achieve better performance in identifying variants specific to that disease. Unlike existing approaches, *DIVAN* uses thousands of annotations from various public resources, including DNase-seq, FAIRE-seq, and TF/Histone ChIP-seq, across different cell types. The more annotations collected, the better the chance informative annotations will be discovered, resulting in a better chance of discriminating risk variants from benign ones. There is still room to improve *DIVAN* further. Other types of genomic and epigenomic features, including eQTL, DNA methylation, and pre-computed scores from GWAVA, CADD, and GenoCanyon, will also be added into *DIVAN*. Another important regulatory mechanism through which non-coding variants influence diseases is the disruption of splice junction and splicing enhancer [53]. The mutations effect on splice sites is similar to nonsense or missense mutations. A myriad of cases about splice site variants have been reported in the literature [54–58]. Because of this, we have decided to add a splicing-related feature, which is the distance to the splice sites (586,795 such sites can be found in Ensemble [59] release 70), into the next release of *DIVAN*. The same feature has been used in GWAVA.

Currently, to represent epigenomic features, existing methods use binary indicators showing whether a ChIP-seq peak overlaps with the variant. In *DIVAN*, we apply an alternative method in which continuous ChIP-seq read count in the vicinity of the variant are used to represent epigenomic features. The advantage of using read count rather than peak presence as the feature lies on the former's better sensitivity and ability to distinguish risk variants from benign ones with a limited number of epigenomic features and to detect significant differences in both enrichment and depletion (Figure 3.16). Moreover, our analyses also show that using read count as the feature results in more informative features being included in the model, especially for features associated with TF binding.

One of the key findings from this study is that histone marks associated with repressive chromatin, in particular, H3K9me3, turns out to be an important feature for risk variant identification. For most of the diseases, we find that this particular repressive mark is often among the top-ranked features, showing significant depletion around the risk variants compared to benign ones. Such a finding is consistent with what has been reported in the literature. In a recent study, Pickrell found that repressed chromatin is significantly depleted around SNPs associated with multiple phenotypes [27]. Chen et al. found that the binding regions of another repressive histone mark, H3K27me3, are significantly less likely to overlap with risk SNP blocks of prostate cancer [60]. Despite these findings, repressive chromatin marks do not play an important role in existing methods for risk variant annotation. For histone marks, almost all attention has been focused on the enrichment of active chromatin marks. For example, the three histone marks used in CADD and Eigen are H3K27ac, H3K4me1, and H3K4me3. A primary reason why only active chromatin marks are used is that it is easier to detect enrichment of a factor, but not depletion when peak is used as the feature. In contrast, using read around the variants as the feature, we are able to detect enrichment as well as depletion.

It is worth clarifying that the risk variants considered in this study are not necessarily "causal" variants since in most cases, no evidence beyond significant association p-values derived from GWAS studies separates them from the millions of variants found throughout the genome. It would be interesting to test *DIVAN* using functionally validated variants as the training set. However, the number of such variants is very limited and insufficient for study on individual diseases today.

A potential application of *DIVAN* is personal genome sequencing interpretation. In the genome of an individual patient, it is expected that many novel, rare, and non-coding variants will be detected. Due to the sample size limitation, little information can be learned from GWASs for these rare variants. Alternatively, by looking at the surrounding regions of such variants and comparing to the genomic and epigenomic profiles of GWAS-associated risk variants represented by *DIVAN*, we can potentially gauge their impact on a particular disease. We have pre-computed *DIVAN* scores for every base in the human genome, which we believe will be a great resource for annotating rare and non-coding variants that would be identified in personal genome sequencing studies.

3.5 Appendix

3.5.1 Availability of data and material

The variants used in the training set are available at Association Results Browser (`https://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm`). The variants used in the GRASP testing set are available at `https://s3.amazonaws.com/NHLBI_Public/GRASP/GraspFullDataset2.zip`. The noncoding ClinVar variants, noncoding HGMD variants and variants in the Ensembl variation database (release 70) are available at `ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/`. Vcf files in 1000 Genomes Project (Phase I) could be downloaded from `ftp://share.sph.umich.edu/1000genomes/fullProject/2012.03.14/`. The variants are detected in the same way as traseR [52]. Variants in COSMIC (v78) could be downloaded from `http://cancer.sanger.`

ac.uk/cosmic/download.

For the pre-computed scores used in the study, scores for variants in CADD (1000 Genomes Phase III) are available at http://krishna.gs.washington.edu/download/CADD/v1.3/1000G_phase3.tsv.gz. Scores for variants in GWAVA are available ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/VEP_plugin/gwava_scores.bed.gz. Scores for variants in Eigen/EigenPC could be downloaded from https://xioniti01.u.hpc.mssm.edu/v1.1/. Scores for GenoCanyon (every base in the human genome (hg19)) could be downloaded from http://genocanyon.med.yale.edu/GenoCanyon_Downloads.html.

For the annotations used in the study, GREP elements are downloaded from http://mendel.stanford.edu/SidowLab/downloads/gerp/. The repeated elements are retrieved from UCSC genome browser (https://genome.ucsc.edu/). The mapped read bam files (hg19) in ENCODE are downloaded from http://genome.ucsc.edu/ENCODE/downloads.html for several collections including Broad Histone, SYDH Histone, UNC FAIRE, Duke DNaseI HS, HAIB TFBS and SYDH TFBS. The mapped read bam files (hg19) in Roadmap Epigenomes Project are downloaded from https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/.

The pre-computed *DIVAN* scores and source codes of *DIVAN* toolkit are freely available under the GNU Public License v3 at https://sites.google.com/site/emoryDIVAN/. The source codes of *DIVAN* toolkit are additionally deposited at GitHub (https://github.com/lichenbiostat86/DIVAN/releases) and have been assigned an MIT open source license with the DOI 10.5281/zenodo.165849.

3.5.2 Supplementary figures

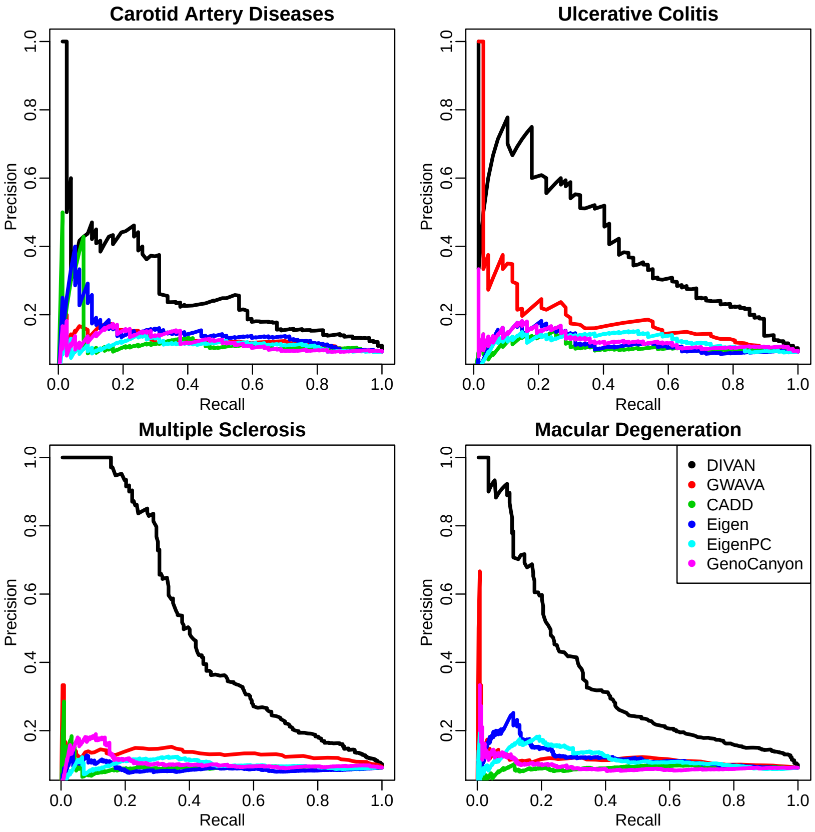
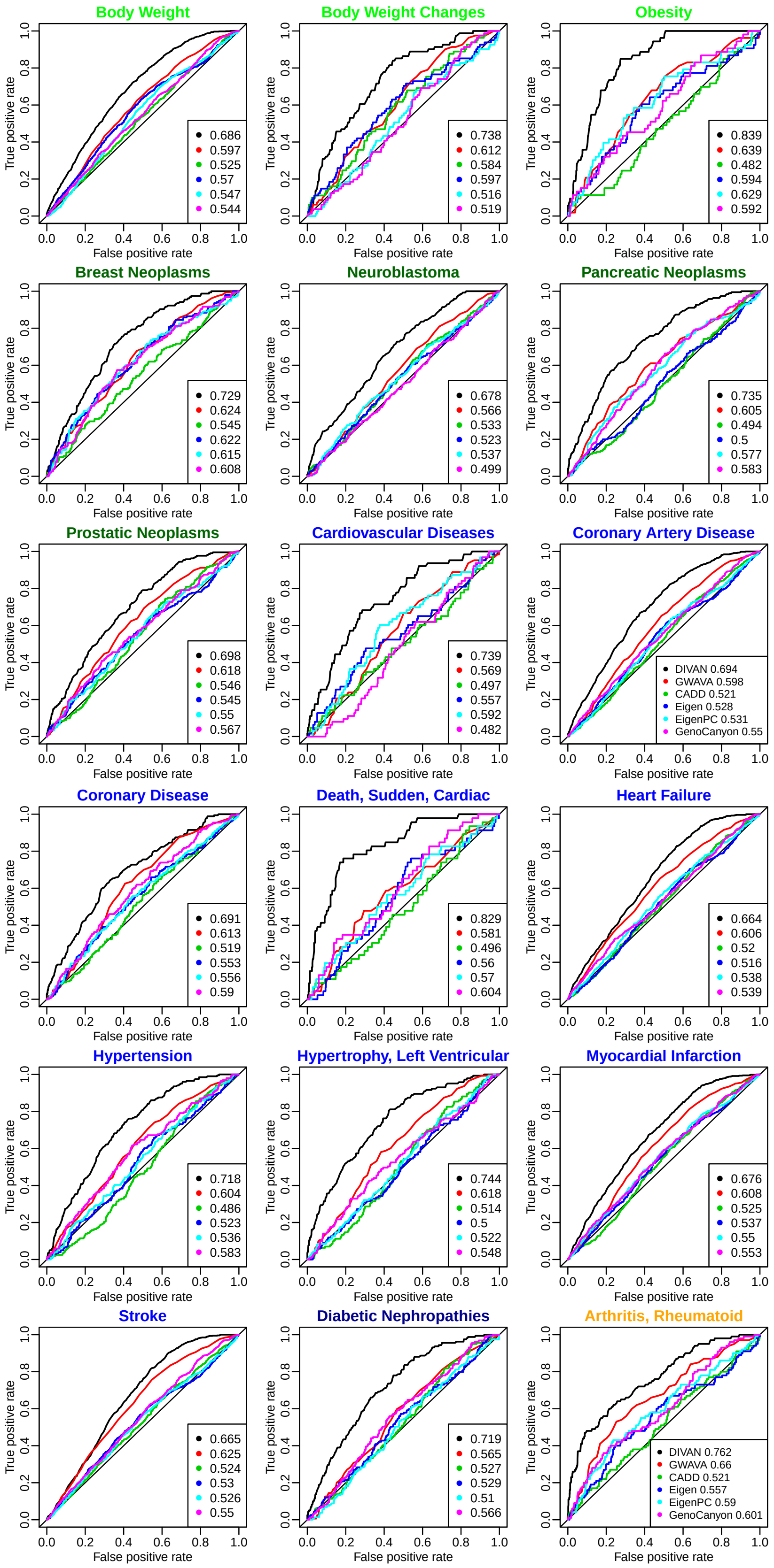
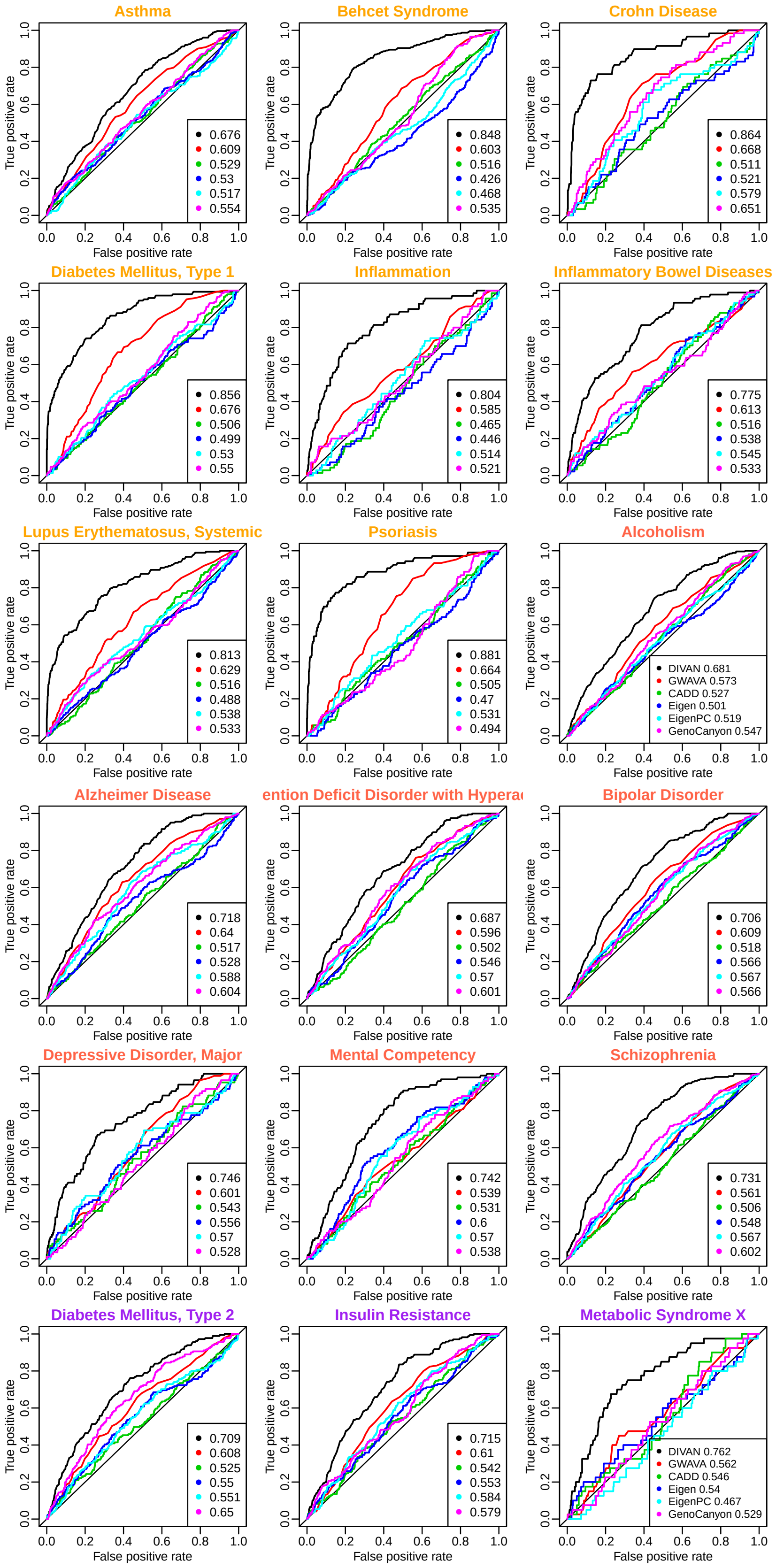
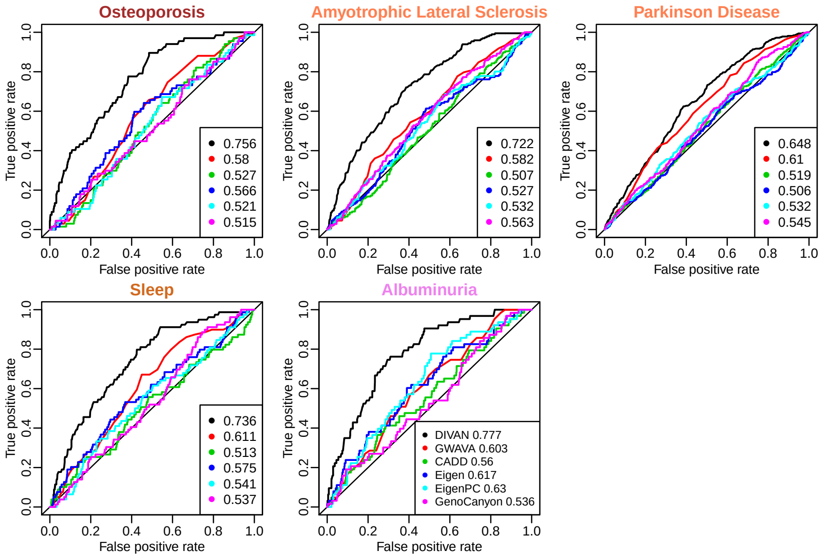


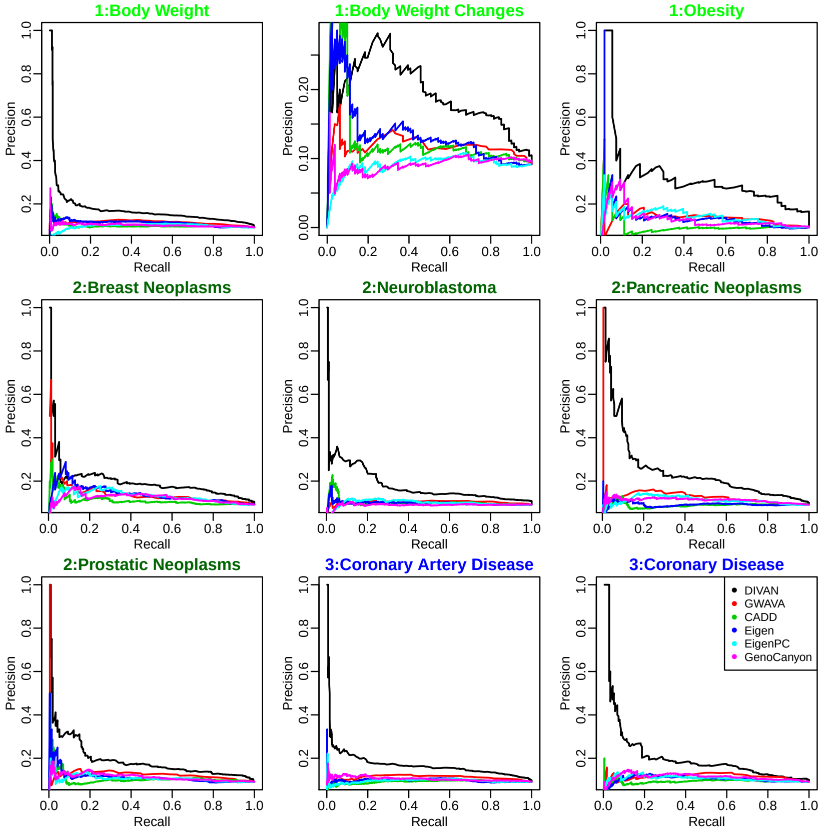
Figure 3.8: Precision-recall curves of five-fold cross-validation on four diseases: carotid artery disease, macular degeneration, ulcerative colitis, and multiple sclerosis.

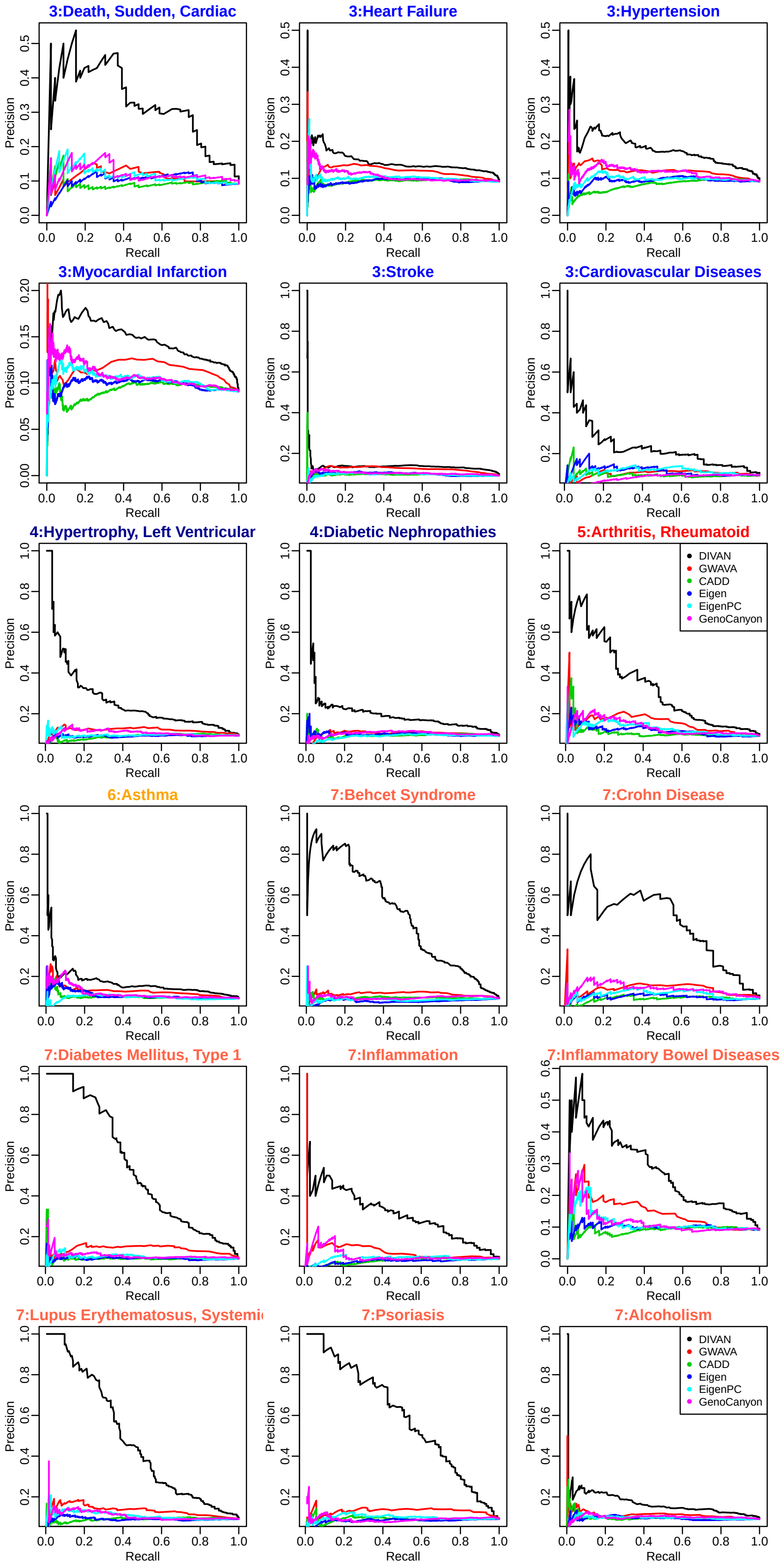
Figure 3.9: ROC curves of five-fold cross-validation for 41 diseases




Asthma
0.676
0.609
0.529
0.53
0.517
0.554
Behcet Syndrome
0.848
0.603
0.516
0.426
0.468
0.535
Crohn Disease
0.864
0.668
0.511
0.521
0.579
0.651
Diabetes Mellitus, Type 1
0.856
0.676
0.506
0.499
0.53
0.55
Inflammation
0.804
0.585
0.465
0.446
0.514
0.521
Inflammatory Bowel Diseases
0.775
0.613
0.516
0.538
0.545
0.533
Lupus Erythematosus, Systemic
0.813
0.629
0.516
0.488
0.538
0.533
Psoriasis
0.881
0.664
0.505
0.47
0.531
0.494
Alcoholism
DIVAN 0.681
GWAVA 0.573
CADD 0.527
Eigen 0.501
EigenPC 0.519
GenoCanyon 0.547
Alzheimer Disease
0.718
0.64
0.517
0.528
0.588
0.604
ention Deficit Disorder with Hypera
0.687
0.596
0.502
0.546
0.57
0.601
Bipolar Disorder
0.706
0.609
0.518
0.566
0.567
0.566
Depressive Disorder, Major
0.746
0.601
0.543
0.556
0.57
0.528
Mental Competency
0.742
0.539
0.531
0.6
0.57
0.538
Schizophrenia
0.731
0.561
0.506
0.548
0.567
0.602
Diabetes Mellitus, Type 2
0.709
0.608
0.525
0.55
0.551
0.65
Insulin Resistance
0.715
0.61
0.542
0.553
0.584
0.579
Metabolic Syndrome X
DIVAN 0.762
GWAVA 0.562
CADD 0.546
Eigen 0.54
EigenPC 0.467
GenoCanyon 0.529
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0


Osteoporosis
True positive rate
False positive rate
● 0.756
● 0.58
● 0.527
● 0.566
● 0.521
● 0.515
Amyotrophic Lateral Sclerosis
● 0.722
● 0.582
● 0.507
● 0.527
● 0.532
● 0.563
Parkinson Disease
● 0.648
● 0.61
● 0.519
● 0.506
● 0.532
● 0.545
Sleep
● 0.736
● 0.611
● 0.513
● 0.575
● 0.541
● 0.537
Albuminuria
● DIVAN 0.777
● GWAVA 0.603
● CADD 0.56
● Eigen 0.617
● EigenPC 0.63
● GenoCanyon 0.536
0.0
0.2
0.4
0.6
0.8
1.0


1:Body Weight
1:Body Weight Changes
1:Obesity
2:Breast Neoplasms
2:Neuroblastoma
2:Pancreatic Neoplasms
2:Prostatic Neoplasms
3:Coronary Artery Disease
3:Coronary Disease
Precision
Recall
● DIVAN
● GWAVA
● CADD
● Eigen
● EigenPC
● GenoCanyon
0.00
0.10
0.20
0.0
0.2
0.4
0.6
0.8
1.0


3:Death, Sudden, Cardiac
3:Heart Failure
3:Hypertension
3:Myocardial Infarction
3:Stroke
3:Cardiovascular Diseases
4:Hypertrophy, Left Ventricular
4:Diabetic Nephropathies
5:Arthritis, Rheumatoid
6:Asthma
7:Behcet Syndrome
7:Crohn Disease
7:Diabetes Mellitus, Type 1
7:Inflammation
7:Inflammatory Bowel Diseases
7:Lupus Erythematosus, Systemi
7:Psoriasis
7:Alcoholism
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.05
0.10
0.15
0.20
● DIVAN
● GWAVA
● CADD
● Eigen
● EigenPC
● GenoCanyon

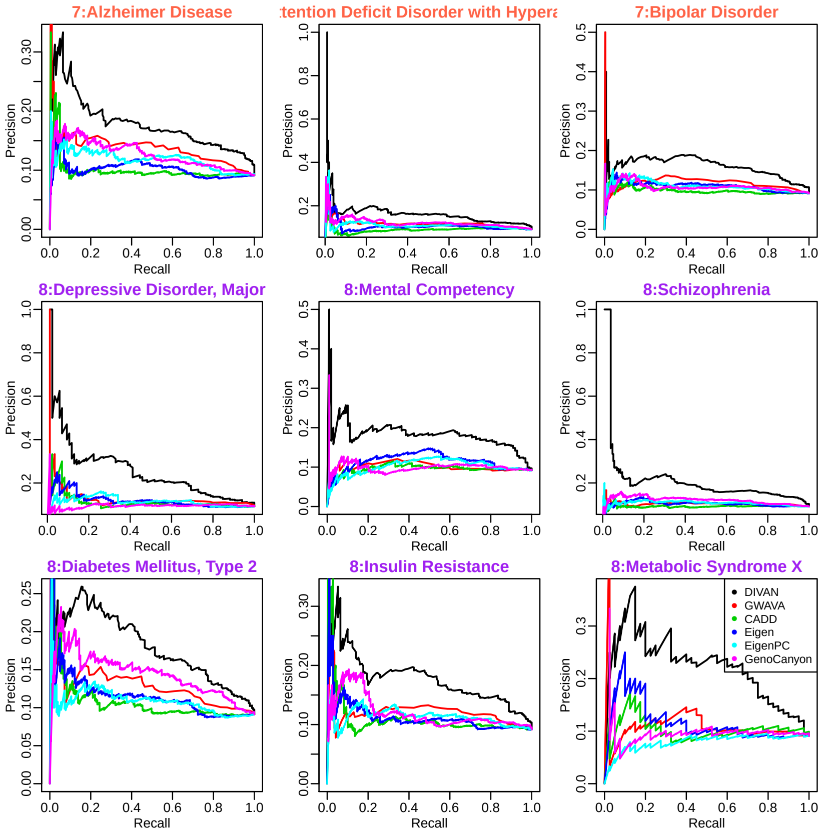
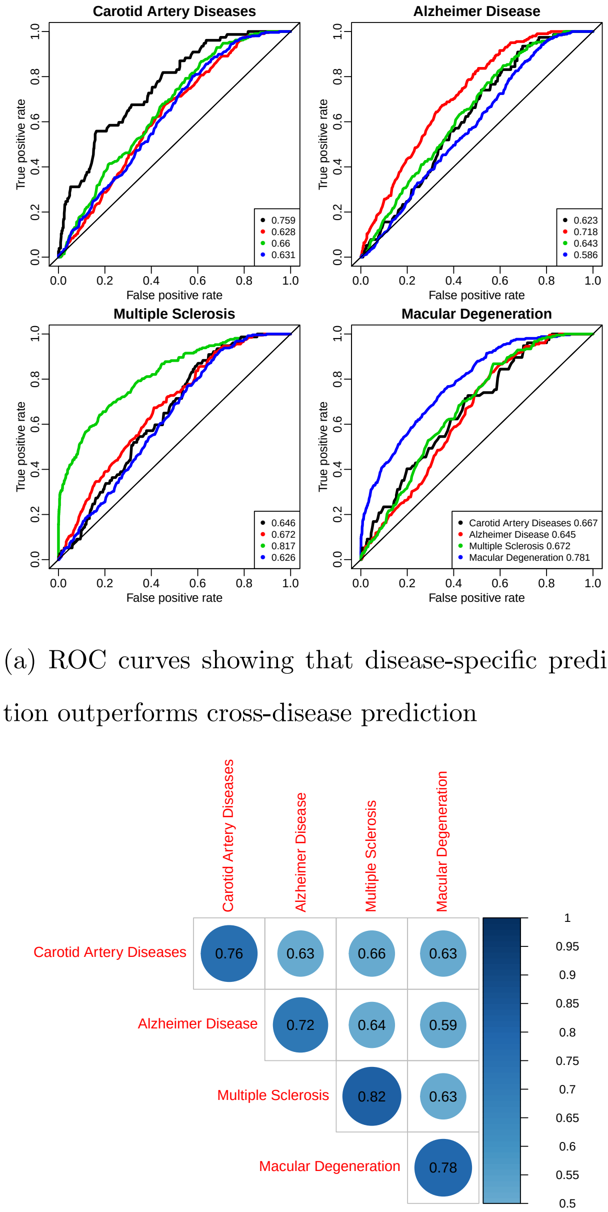

7:Alzheimer Disease
:tention Deficit Disorder with Hypera
7:Bipolar Disorder
8:Depressive Disorder, Major
8:Mental Competency
8:Schizophrenia
8:Diabetes Mellitus, Type 2
8:Insulin Resistance
8:Metabolic Syndrome X
Precision
Recall
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
1.0
● DIVAN
● GWAVA
● CADD
● Eigen
● EigenPC
● GenoCanyon

Figure 3.10: Cross-disease prediction shows the necessity for disease specificity of variant prioritization by using four diseases from four different disease classes: carotid artery disease (cardiovascular disease), Alzheimer's disease (mental disease), multiple sclerosis (immune disease), and macular degeneration (eye disease).



(a) ROC curves showing that disease-specific prediction outperforms cross-disease prediction

(b) Heatmap showing that disease-specific prediction outperforms cross-disease prediction

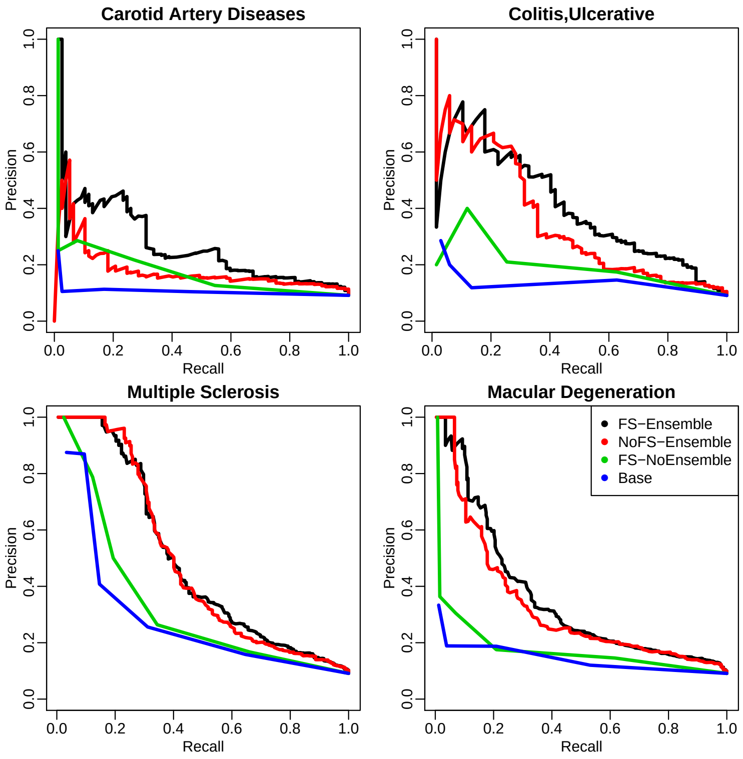
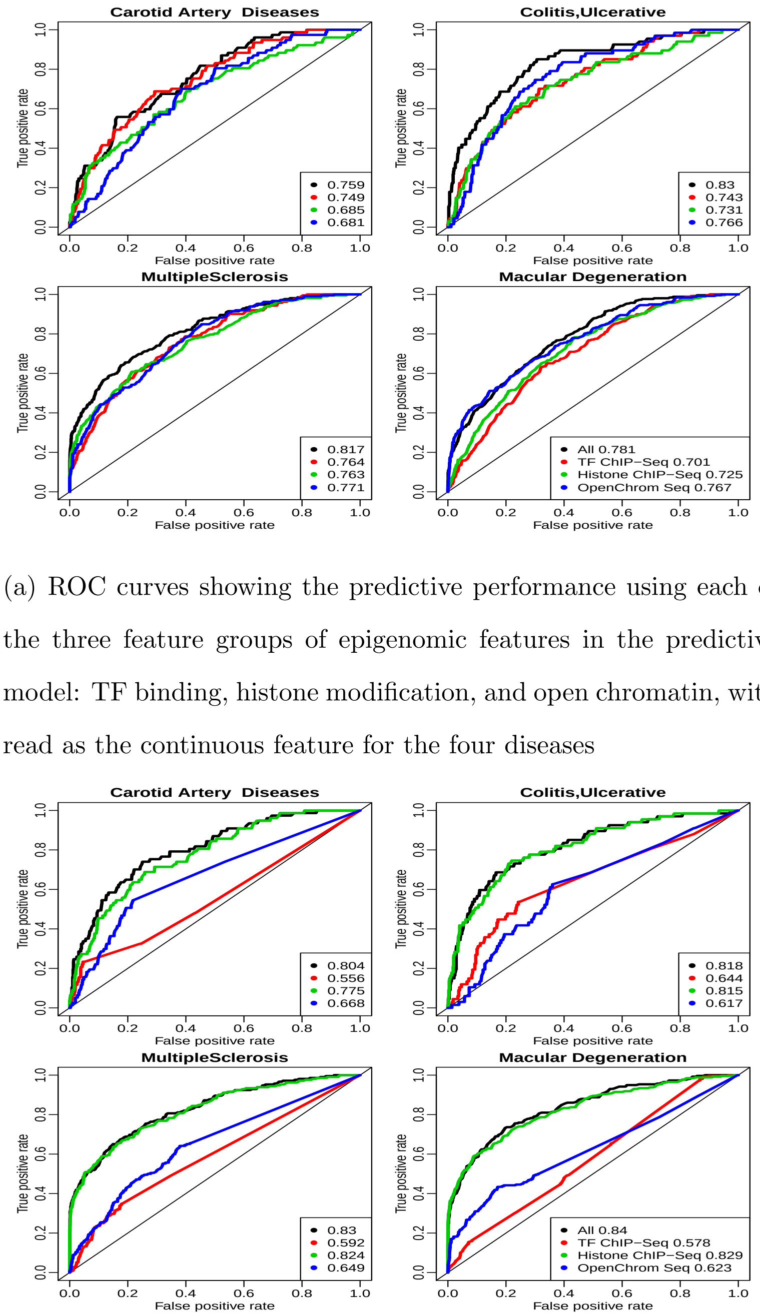


Figure 3.11: Precision-recall curves showing the effectiveness of the feature selection and ensemble method by comparing feature selection and ensemble combined, feature selection only, ensemble only, and the baseline case: neither feature selection nor ensemble.

Figure 3.12: Contribution of three different feature groups, TF binding/histone modification/open chromatin, on prediction in four diseases: carotid artery disease, macular degeneration, ulcerative colitis, and multiple sclerosis



(a) ROC curves showing the predictive performance using each of the three feature groups of epigenomic features in the predictive model: TF binding, histone modification, and open chromatin, with read as the continuous feature for the four diseases

(b) ROC curves showing the predictive performance using each of the three feature groups of epigenomic features in the predictive model, TF binding, histone modification, and open chromatin, with peak as the binary feature for the four diseases

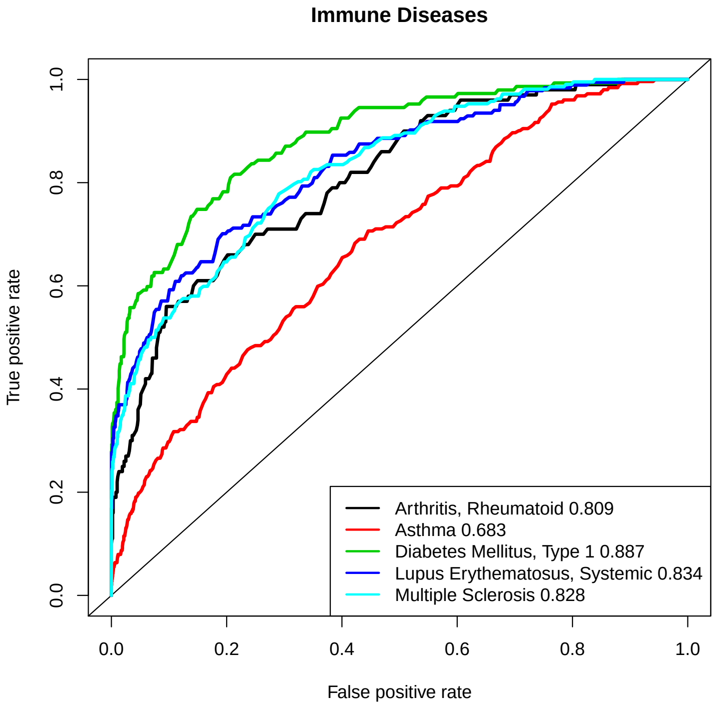


Figure 3.13: ROC curves showing predictive performance of disease-class variant prioritization for immune diseases, including rheumatoid arthritis, asthma, type 1 diabetes mellitus, systemic lupus erythematosus, and multiple sclerosis. The ROC curves are generated by using a "leave-one-disease-out" approach; that is, the predictive model is built using variants of all other diseases within the disease class, tested on the variants of the disease being left out.

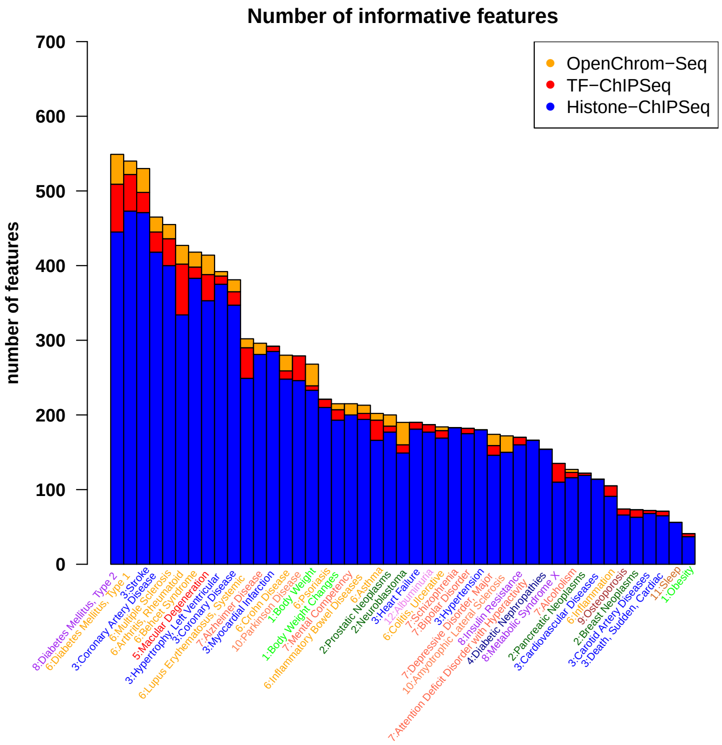


Figure 3.14: Bar chart showing the number of informative features for three feature categories (TF binding/histone modification/open chromatin) for 45 diseases when using peak as the feature

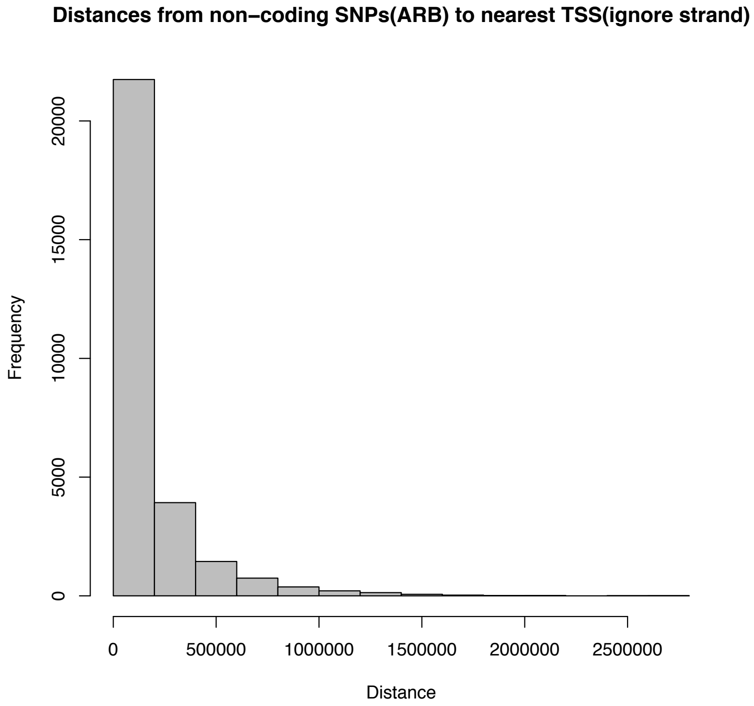


Figure 3.15: Distribution of distances between non-coding SNPs associated with the 45 diseases in ARB to their nearest TSS (ignoring stand)

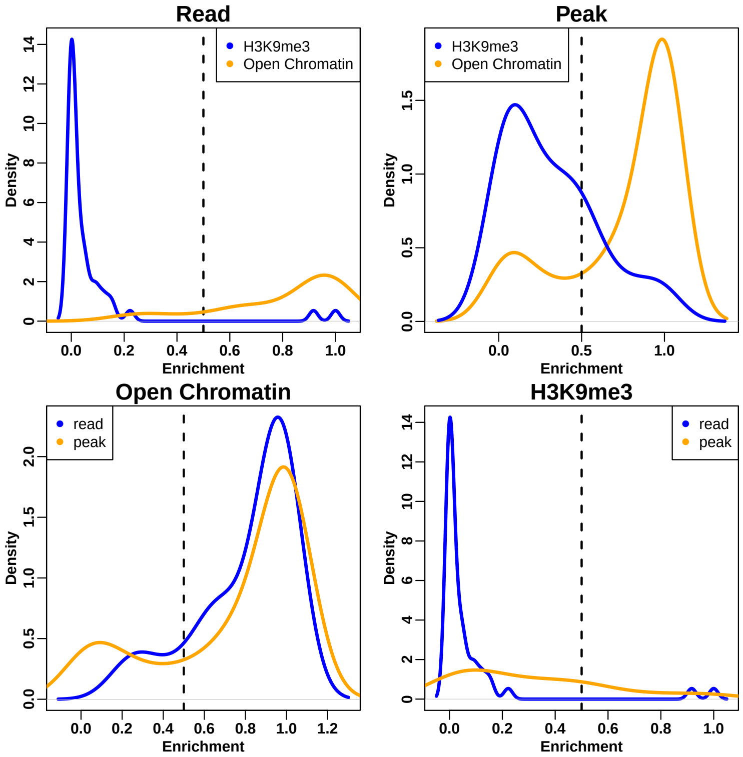
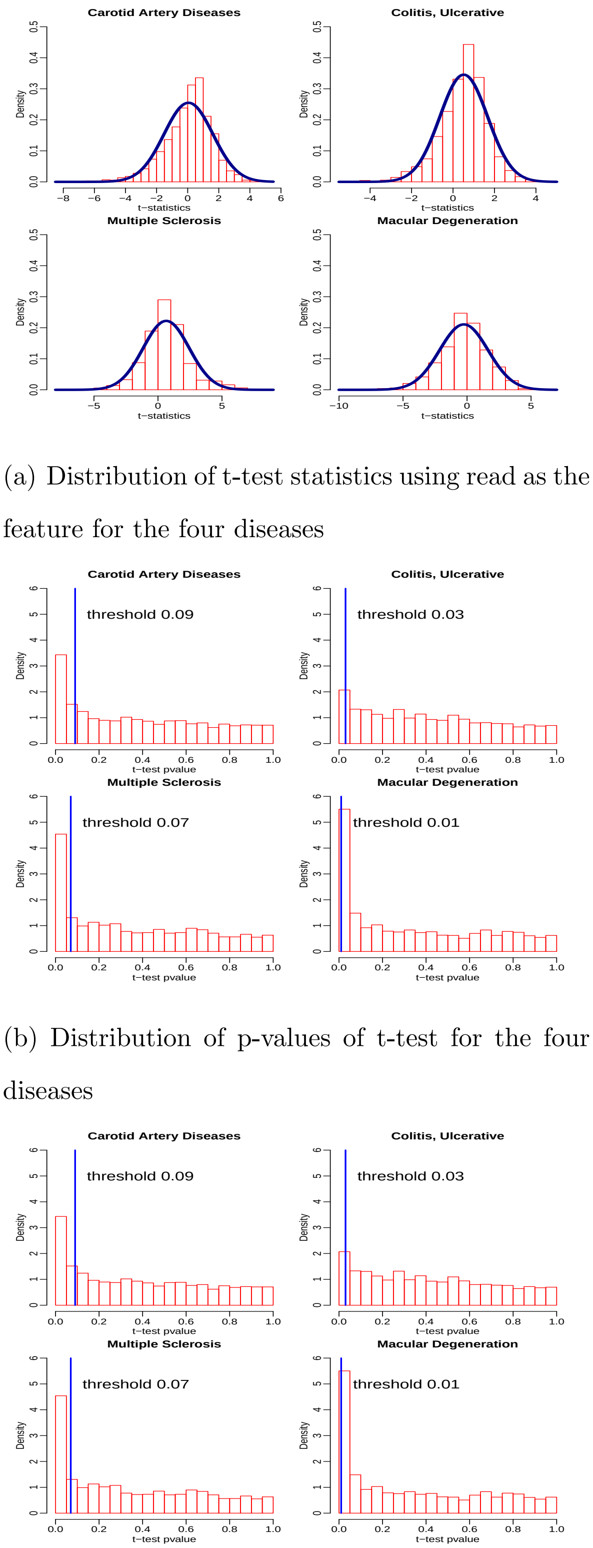


Figure 3.16: Distribution of enrichment proportions of H3K9me3 features and open chromatin features when using peak and read as the feature. The enrichment proportion is defined as the ratio between average read counts of each H3K9me3/open chromatin feature on risk variants and benign variants respectively.

Figure 3.17: Distributions of test statistics and p-values calculated between risk variants and benign variants for four diseases: carotid artery disease, macular degeneration, ulcerative colitis, and multiple sclerosis

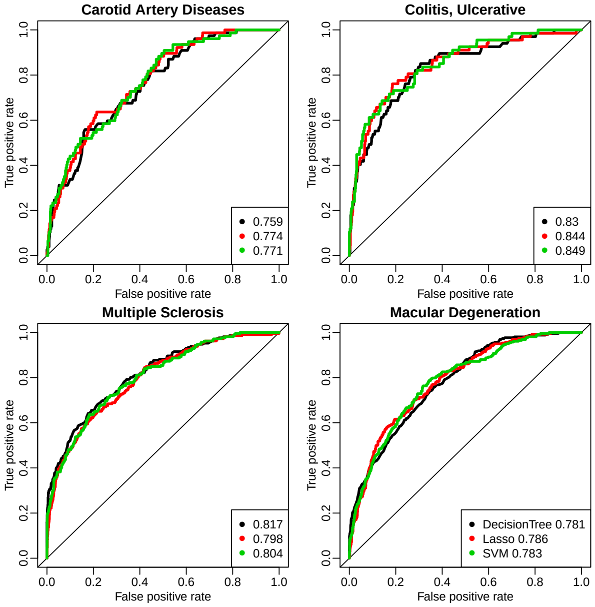


(a) Distribution of t-test statistics using read as the feature for the four diseases

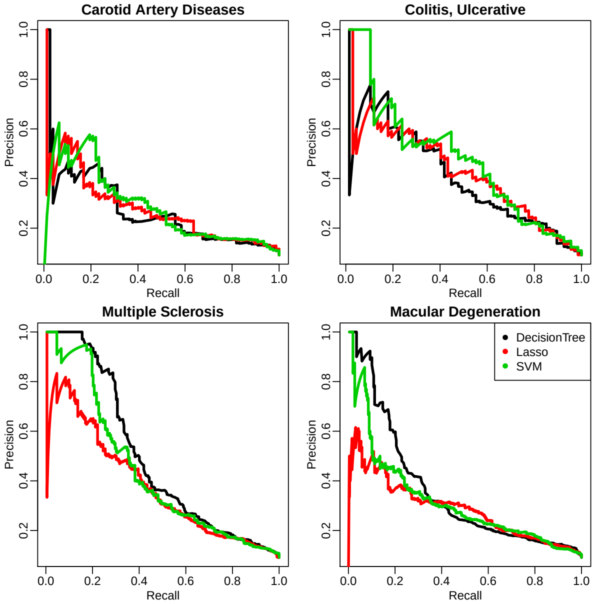
(b) Distribution of p-values of t-test for the four diseases

(c) Distribution of p-values of Fisher's exact test using peak as the feature for the four diseases

Figure 3.18: ROC curves and precision-recall curves of five-fold cross-validation using three base learners, decision tree, SVM, and Lasso, for four diseases: carotid artery disease, macular degeneration, ulcerative colitis, and multiple sclerosis



(a) ROC curves of five-fold cross-validation for three base learners: decision tree, SVM, and Lasso on the four diseases



(b) Precision-recall curves of five-fold cross-validation for three base learners, decision tree, SVM, and Lasso, on the four diseases

3.5.3 Supplementary tables

Table 3.2: Summary of the number of risk variants, the number of reduced risk variants in ARB such that none is within 10kb of another; the number of risk variants in the GRASP database but not in ARB, and the number of risk variants in the GRASP database 10kb away from the corresponding disease-specific risk variants in ARB

disease/trait	disease class	#SNP(ASB)	#SNP(ASB,one in 10kb window)	#SNP(GRASP,non-overlapping from ASB)	#SNP(GRASP, 10kb away from ASB)
Body Weight	body weight	857	744	X	X
Body Weight Changes	body weight	81	74	15	15
Obesity	body weight	53	47	600	577
Breast Neoplasms	cancer	155	135	381	339
Neuroblastoma	cancer	268	231	53	43
Pancreatic Neoplasms	cancer	190	150	66	37
Prostatic Neoplasms	cancer	215	198	463	423
Carotid Artery Diseases	cardiovascular	80	70	X	X
Coronary Artery Disease	cardiovascular	584	467	3185	3184
Coronary Disease	cardiovascular	176	152	X	X
Death, Sudden, Cardiac	cardiovascular	46	46	X	X
Heart Failure	cardiovascular	530	491	22	22
Hypertension	cardiovascular	201	165	491	480
Myocardial Infarction	cardiovascular	584	529	396	392
Stroke	cardiovascular	725	680	21	21
Cardiovascular Diseases	cariovascular	63	63	33	31
Hypertrophy, Left Ventricular	cariovascular	143	112	13	13
Diabetic Nephropathies	endocrine	159	107	X	X
Macular Degeneration	eye disease	258	170	2989	2283
Arthritis, Rheumatoid	immune	100	94	6638	6017
Asthma	immune	252	232	654	578
Behcet Syndrome	immune	229	101	212	144
Colitis, Ulcerative	immune	67	58	385	350
Crohn Disease	immune	59	53	1049	960
Diabetes Mellitus, Type 1	immune	147	100	764	585
Inflammation	immune	70	55	76	76
Inflammatory Bowel Diseases	immune	91	76	104	103
Lupus Erythematosus, Systemic	immune	184	133	194	175
Multiple Sclerosis	immune	212	141	444	433
Psoriasis	immune	106	60	374	360
Alcoholism	mental	261	206	128	128
Alzheimer Disease	mental	202	188	888	852
Attention Deficit Disorder with Hyperactivity	mental	197	190	294	289
Bipolar Disorder	mental	268	230	1937	1873
Depressive Disorder, Major	mental	85	77	954	929
Mental Competency	mental	99	89	X	X
Schizophrenia	mental	233	187	1270	1239
Diabetes Mellitus, Type 2	metabolic disease	181	159	2224	2197
Insulin Resistance	metabolic disease	170	146	95	95
Metabolic Syndrome X	metabolic disease	40	30	20	20
Osteoporosis	musculoskeletal	67	65	X	X
Amyotrophic Lateral Sclerosis	nervous system	197	160	364	354
Parkinson Disease	nervous system	325	282	X	X
Sleep	psychological	79	78	99	99
Albuminuria	urogenital	63	60	X	X

Table 3.3: Summary of CV-AUC values of five-fold cross-validation for the 45 diseases in ARB for all risk variants when benign variants are 10 times of the risk variants in the training set

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC
Body Weight	body weight	0.686	0.597	0.544	0.525	0.57	0.547
Body Weight Changes	body weight	0.738	0.612	0.519	0.584	0.597	0.516
Obesity	body weight	0.839	0.639	0.592	0.482	0.594	0.629
Breast Neoplasms	cancer	0.729	0.624	0.608	0.545	0.622	0.615
Neuroblastoma	cancer	0.678	0.566	0.499	0.533	0.523	0.537
Pancreatic Neoplasms	cancer	0.735	0.605	0.583	0.494	0.5	0.577
Prostatic Neoplasms	cancer	0.698	0.618	0.567	0.546	0.545	0.55
Carotid Artery Diseases	cardiovascular	0.759	0.606	0.57	0.568	0.633	0.576
Coronary Artery Disease	cardiovascular	0.694	0.598	0.55	0.521	0.528	0.531
Coronary Disease	cardiovascular	0.691	0.613	0.59	0.519	0.553	0.556
Death, Sudden, Cardiac	cardiovascular	0.829	0.581	0.604	0.496	0.56	0.57
Heart Failure	cardiovascular	0.664	0.606	0.539	0.52	0.516	0.538
Hypertension	cardiovascular	0.718	0.604	0.583	0.486	0.523	0.536
Myocardial Infarction	cardiovascular	0.676	0.608	0.553	0.525	0.537	0.55
Stroke	cardiovascular	0.665	0.625	0.55	0.524	0.53	0.526
Cardiovascular Diseases	cardiovascular	0.739	0.569	0.482	0.497	0.557	0.592
Hypertrophy, Left Ventricular	cardiovascular	0.744	0.618	0.548	0.514	0.5	0.522
Diabetic Nephropathies	endocrine	0.719	0.565	0.566	0.527	0.529	0.51
Macular Degeneration	eye disease	0.781	0.593	0.486	0.511	0.572	0.567
Arthritis, Rheumatoid	immune	0.762	0.66	0.601	0.521	0.557	0.59
Asthma	immune	0.676	0.609	0.554	0.529	0.53	0.517
Behcet Syndrome	immune	0.848	0.603	0.535	0.516	0.426	0.468
Colitis, Ulcerative	immune	0.83	0.675	0.598	0.548	0.544	0.603
Crohn Disease	immune	0.864	0.668	0.651	0.511	0.521	0.579
Diabetes Mellitus, Type 1	immune	0.856	0.676	0.55	0.506	0.499	0.53
Inflammation	immune	0.804	0.585	0.521	0.465	0.446	0.514
Inflammatory Bowel Diseases	immune	0.775	0.613	0.533	0.516	0.538	0.545
Lupus Erythematosus, Systemic	immune	0.813	0.629	0.533	0.516	0.488	0.538
Multiple Sclerosis	immune	0.817	0.655	0.536	0.5	0.465	0.531
Psoriasis	immune	0.881	0.664	0.494	0.505	0.47	0.531
Alcoholism	mental	0.681	0.573	0.547	0.527	0.501	0.519
Alzheimer Disease	mental	0.718	0.64	0.604	0.517	0.528	0.588
Attention Deficit Disorder with Hyperactivity	mental	0.687	0.596	0.601	0.502	0.546	0.57
Bipolar Disorder	mental	0.706	0.609	0.566	0.518	0.566	0.567
Depressive Disorder, Major	mental	0.746	0.601	0.528	0.543	0.556	0.57
Mental Competency	mental	0.742	0.539	0.538	0.531	0.6	0.57
Schizophrenia	mental	0.731	0.561	0.602	0.506	0.548	0.567
Diabetes Mellitus, Type 2	metabolic disease	0.709	0.608	0.65	0.525	0.55	0.551
Insulin Resistance	metabolic disease	0.715	0.61	0.579	0.542	0.553	0.584
Metabolic Syndrome X	metabolic disease	0.762	0.562	0.529	0.546	0.54	0.467
Osteoporosis	musculoskeletal	0.756	0.58	0.515	0.527	0.566	0.521
Amyotrophic Lateral Sclerosis	nervous system	0.722	0.582	0.563	0.507	0.527	0.532
Parkinson Disease	nervous system	0.648	0.61	0.545	0.519	0.506	0.532
Sleep	psychological	0.736	0.611	0.537	0.513	0.575	0.541
Albuminuria	urogenital	0.777	0.603	0.536	0.56	0.617	0.63

Table 3.4: Summary of MCC values of five-fold cross-validation for 45 diseases in ARB for all risk variants when benign variants are 10 times of the risk variants in the training set

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC
Body Weight	body weight	0.124	0.0164	0.0295	0.0101	0.0361	-0.0246
Body Weight Changes	body weight	0.224	0.0293	-0.0221	0.0509	0.0705	-0.0295
Obesity	body weight	0.297	-0.0171	0.0689	0.0347	0.04	0.0578
Breast Neoplasms	cancer	0.167	0.066	0.114	0.0202	0.088	0.0649
Neuroblastoma	cancer	0.119	-0.01	0.000205	0.00428	0.00921	0.03
Pancreatic Neoplasms	cancer	0.199	0.00965	0.059	0.0152	0.0147	0.0171
Prostatic Neoplasms	cancer	0.146	0.0284	0.0588	-0.00455	0.0267	0.0141
Carotid Artery Diseases	cardiovascular	0.244	-0.0068	0.0821	-0.00432	0.0785	0.00379
Coronary Artery Disease	cardiovascular	0.13	0.00183	0.0406	-0.00566	0.00567	-0.0057
Coronary Disease	cardiovascular	0.144	0.0187	0.0651	0.00497	-0.00246	0.0126
Death, Sudden, Cardiac	cardiovascular	0.347	0.00235	0.0427	0.016	-0.0513	0.0363
Heart Failure	cardiovascular	0.0884	-0.00381	0.0254	-0.00689	-0.0156	0.00243
Hypertension	cardiovascular	0.161	0.0302	0.0772	-0.05	-0.0304	-0.015
Myocardial Infarction	cardiovascular	0.113	0.0125	0.0388	-0.00667	-0.0106	0.0256
Stroke	cardiovascular	0.0672	0.0133	0.0321	0.00296	0.0113	0.000875
Cardiovascular Diseases	cardiovascular	0.217	-0.035	-0.0274	0.0142	0.0382	-0.0232
Hypertrophy, Left Ventricular	cardiovascular	0.205	0.00783	0.0534	-0.0262	-0.000493	0.023
Diabetic Nephropathies	endocrine	0.168	-0.0512	0.0599	-0.00283	0.00409	-0.0365
Macular Degeneration	eye disease	0.254	0.0236	-0.0275	0.00245	0.0791	0.0949
Arthritis, Rheumatoid	immune	0.31	0.0193	0.0379	0.0427	0.0583	0.108
Asthma	immune	0.122	0.0547	0.0452	0.0124	0.0415	0.00242
Behcet Syndrome	immune	0.382	0.0196	-0.0121	-0.00781	-0.0309	-0.000477
Colitis, Ulcerative	immune	0.339	0.145	0.0526	0.0251	0.0482	0.0561
Crohn Disease	immune	0.455	0.00435	0.133	-0.043	0.0017	-0.00543
Diabetes Mellitus, Type 1	immune	0.403	0.0794	0.0245	-0.000191	0.00058	0.0117
Inflammation	immune	0.317	0.0238	0.00999	-0.0683	-0.0517	-0.0318
Inflammatory Bowel Diseases	immune	0.271	0.103	0.0342	-0.0221	0.0317	0.0615
Lupus Erythematosus, Systemic	immune	0.352	0.0844	0.00953	-0.0265	0.0167	0.0475
Multiple Sclerosis	immune	0.355	0.0682	0.0227	-0.0247	0.0269	-0.00478
Psoriasis	immune	0.489	0.032	-0.0556	-0.02	-0.0434	0.00221
Alcoholism	mental	0.134	0.0463	0.0404	0.00817	0.00271	0.0124
Alzheimer Disease	mental	0.161	0.0321	0.0774	0.0131	0.0226	0.0416
Attention Deficit Disorder with Hyperactivity	mental	0.134	0.0476	0.0895	-0.0241	0.0199	0.0371
Bipolar Disorder	mental	0.17	0.00225	0.0328	0.0249	0.0251	0.046
Depressive Disorder, Major	mental	0.21	0.0523	0.00309	0.0416	0.08	0.0321
Mental Competency	mental	0.172	-0.0127	-0.0229	-0.00893	-0.0242	-0.0209
Schizophrenia	mental	0.174	0.0182	0.0814	-0.00471	-0.0025	0.0137
Diabetes Mellitus, Type 2	metabolic disease	0.184	0.0438	0.134	0.0364	0.0497	0.0232
Insulin Resistance	metabolic disease	0.17	0.00539	0.0498	0.017	0.0535	0.00281
Metabolic Syndrome X	metabolic disease	0.227	0.0263	0.0423	0.0618	0.0884	-0.0564
Osteoporosis	musculoskeletal	0.199	-0.0394	0.00672	-0.0607	-0.0389	-0.00955
Amyotrophic Lateral Sclerosis	nervous system	0.186	-0.0167	0.0409	-0.0244	0.0223	0.000415
Parkinson Disease	nervous system	0.0901	0.0112	0.0257	0.0141	0.0187	0.0234
Sleep	psychological	0.202	0.0197	0.0101	0.0235	0.0448	-0.0362
Albuminuria	urogenital	0.24	0.0466	0.0161	0.036	0.0731	0.0449

Table 3.5: Summary of predicted AUC values for 36 diseases in GRASP using the predictive models built from risk variants of corresponding diseases in ARB when benign variants are 10 times of the risk variants in the testing set. For each disease, the risk variants in GRASP are non-overlapping with the risk variants in ARB

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.738	0.517	0.43	0.252	0.268	0.421	*DIVAN*
Obesity	body weight	0.632	0.573	0.556	0.531	0.53	0.534	*DIVAN*
Breast Neoplasms	cancer	0.672	0.617	0.634	0.534	0.563	0.577	*DIVAN*
Neuroblastoma	cancer	0.642	0.495	0.572	0.504	0.5	0.464	*DIVAN*
Pancreatic Neoplasms	cancer	0.697	0.616	0.65	0.562	0.526	0.548	*DIVAN*
Prostatic Neoplasms	cancer	0.629	0.627	0.592	0.555	0.591	0.607	*DIVAN*
Cardiovascular Diseases	cardiovascular	0.529	0.622	0.636	0.579	0.504	0.594	GenoCanyon
Coronary Artery Disease	cardiovascular	0.647	0.627	0.614	0.535	0.537	0.558	*DIVAN*
Heart Failure	cardiovascular	0.708	0.59	0.555	0.482	0.467	0.581	*DIVAN*
Hypertension	cardiovascular	0.654	0.618	0.585	0.52	0.534	0.561	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.589	0.598	0.495	0.616	0.631	0.584	Eigen
Myocardial Infarction	cardiovascular	0.657	0.633	0.641	0.529	0.557	0.602	*DIVAN*
Stroke	cardiovascular	0.662	0.748	0.677	0.412	0.472	0.627	GWAVA
Macular Degeneration	eye disease	0.68	0.626	0.601	0.525	0.542	0.567	*DIVAN*
Arthritis, Rheumatoid	immune	0.766	0.759	0.64	0.53	0.469	0.547	*DIVAN*
Asthma	immune	0.67	0.665	0.603	0.542	0.552	0.569	*DIVAN*
Behcet Syndrome	immune	0.74	0.679	0.619	0.492	0.461	0.553	*DIVAN*
Colitis, Ulcerative	immune	0.677	0.65	0.61	0.529	0.532	0.558	*DIVAN*
Crohn Disease	immune	0.674	0.654	0.641	0.544	0.546	0.58	*DIVAN*
Diabetes Mellitus, Type 1	immune	0.802	0.725	0.648	0.552	0.516	0.577	*DIVAN*
Inflammation	immune	0.59	0.596	0.556	0.567	0.525	0.542	GWAVA
Inflammatory Bowel Diseases	immune	0.695	0.778	0.77	0.582	0.592	0.659	GWAVA
Lupus Erythematosus, Systemic	immune	0.748	0.682	0.682	0.573	0.581	0.634	*DIVAN*
Multiple Sclerosis	immune	0.616	0.606	0.556	0.509	0.518	0.552	*DIVAN*
Psoriasis	immune	0.636	0.65	0.627	0.532	0.544	0.58	GWAVA
Alcoholism	mental	0.637	0.553	0.769	0.492	0.415	0.488	GenoCanyon
Alzheimer Disease	mental	0.626	0.591	0.568	0.513	0.524	0.54	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.65	0.601	0.56	0.533	0.535	0.541	*DIVAN*
Bipolar Disorder	mental	0.612	0.589	0.544	0.503	0.52	0.523	*DIVAN*
Depressive Disorder, Major	mental	0.599	0.579	0.544	0.522	0.54	0.533	*DIVAN*
Schizophrenia	mental	0.647	0.607	0.579	0.526	0.541	0.564	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.666	0.589	0.604	0.531	0.54	0.547	*DIVAN*
Insulin Resistance	metabolic disease	0.622	0.524	0.5	0.456	0.637	0.478	Eigen
Metabolic Syndrome X	metabolic disease	0.619	0.628	0.743	0.467	0.61	0.562	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.639	0.592	0.605	0.52	0.517	0.544	*DIVAN*
Sleep	psychological	0.736	0.641	0.514	0.535	0.536	0.574	*DIVAN*

Table 3.6: Summary of predicted MCC values for 36 diseases in GRASP using the predictive models built from risk variants of corresponding diseases in ARB when benign variants are 10 times of the risk variants in the testing set. For each disease, the risk variants in GRASP are non-overlapping with the risk variants in ARB

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethod
Body Weight Changes	body weight	0.118	-0.0514	-0.0594	-0.0996	-0.0755	-0.0188	*DIVAN*
Obesity	body weight	0.0685	0.0141	0.0457	0.00234	0.0104	-0.006	*DIVAN*
Breast Neoplasms	cancer	0.103	0.0339	0.118	0.00852	0.073	0.0434	GenoCanyon
Neuroblastoma	cancer	0.0577	0.0527	-0.0689	0.0642	0.00142	-0.0545	CADD
Pancreatic Neoplasms	cancer	0.094	0.036	0.119	0.0217	0.0275	0.0478	GenoCanyon
Prostatic Neoplasms	cancer	0.0395	0.0292	0.0624	0.00487	0.0283	0.0279	GenoCanyon
Cardiovascular Diseases	cardiovascular	0.0189	0.0685	0.0918	0.0375	-0.0308	0.0671	GenoCanyon
Coronary Artery Disease	cardiovascular	0.063	0.0496	0.0887	-0.00137	0.0118	0.0309	GenoCanyon
Heart Failure	cardiovascular	0.12	0.00265	-0.0215	0.00102	-0.0567	0.0215	*DIVAN*
Hypertension	cardiovascular	0.0765	0.028	0.0612	0.00436	0.00274	0.0173	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	-0.014	0.208	-0.0315	-0.0168	0.129	0.0695	GWAVA
Myocardial Infarction	cardiovascular	0.0815	0.0871	0.118	0.00237	0.0402	0.0715	GenoCanyon
Stroke	cardiovascular	0.0571	0.171	0.0203	-0.0431	-0.014	0.0677	GWAVA
Macular Degeneration	eye disease	0.146	0.0549	0.0726	-0.00997	0.0703	0.0778	*DIVAN*
Arthritis, Rheumatoid	immune	0.228	0.166	0.0905	0.00771	-0.0011	0.0514	*DIVAN*
Asthma	immune	0.112	0.0893	0.0681	0.0276	0.0398	0.0304	*DIVAN*
Behcet Syndrome	immune	0.257	0.0489	0.0603	-0.00121	-0.0269	0.0103	*DIVAN*
Colitis, Ulcerative	immune	0.111	0.0349	0.0991	0.0232	0.0304	0.0635	*DIVAN*
Crohn Disease	immune	0.116	0.0785	0.119	0.00616	0.0129	0.0413	GenoCanyon
Diabetes Mellitus, Type 1	immune	0.365	0.164	0.112	-0.00672	0.00944	0.0567	*DIVAN*
Inflammation	immune	0.0486	-0.00533	0.0856	-0.0345	0.0451	0.0293	GenoCanyon
Inflammatory Bowel Diseases	immune	0.133	0.233	0.243	0.0939	0.118	0.155	GenoCanyon
Lupus Erythematosus, Systemic	immune	0.161	0.0961	0.143	-0.0358	0.0159	0.108	*DIVAN*
Multiple Sclerosis	immune	0.0426	0.0361	0.0382	-0.00714	0.0226	0.000787	*DIVAN*
Psoriasis	immune	0.0744	0.0362	0.11	-0.00301	0.0183	0.0528	GenoCanyon
Alcoholism	mental	0.0535	-0.0402	0.294	-0.0577	-0.0507	-0.0686	GenoCanyon
Alzheimer Disease	mental	0.078	0.0288	0.0533	-0.0112	-0.00145	-0.00242	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.0599	0.0186	0.0411	-0.00419	0.0315	-0.0124	*DIVAN*
Bipolar Disorder	mental	0.0461	0.0255	0.0306	-0.00556	0.00686	0.00115	*DIVAN*
Depressive Disorder, Major	mental	0.0436	0.00457	0.0139	4.3e-05	0.0136	-0.00601	*DIVAN*
Schizophrenia	mental	0.0839	0.037	0.0479	-0.000961	0.0207	0.0334	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.0978	0.0218	0.0902	-0.00154	0.0192	0.0107	*DIVAN*
Insulin Resistance	metabolic disease	0.0509	-0.0309	-0.0346	-0.0559	0.189	-0.0745	Eigen
Metabolic Syndrome X	metabolic disease	-0.0756	-0.0154	0.203	-0.0428	0.152	0.0237	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.0354	0.032	0.0785	-0.00426	0.00918	0.0275	GenoCanyon
Sleep	psychological	0.171	0.0789	-0.0318	-0.014	0.0086	0.078	*DIVAN*

Table 3.7: Summary of predicted AUCs for immune-related HGMD regulatory variants. Summary of predicted AUCs of different methods for 34 variants related to diseases in immune disease class including Asthma, Behcet syndrome, Ulcerative Colitis, Crohn's disease, Inflammatory bowel diseases and Systemic lupus erythematosus. *DIVAN* uses the immune disease class specific model by pooling all variants of the aforementioned six diseases from ARB together.

Method	AUC
GenoCanyon	0.79
DIVAN	0.788
EigenPC	0.736
Eigen	0.629
CADD	0.583

Table 3.8: Summary of predicted AUCs for synonymous mutations. Summary of predicted AUCs of different methods for synonymous mutations in seven diseases including Macular degeneration, Alzheimer disease, Asthma, Metabolic syndrome X, obesity, Parkinson's disease and Rheumatoid arthritis.

Disease	#SM	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC
Macular Degeneration	63	0.747	0.775	0.496	0.705	0.757	0.755
Alzheimer Disease	44	0.537	0.579	0.433	0.512	0.553	0.527
Asthma	22	0.525	0.753	0.602	0.577	0.623	0.631
Metabolic Syndrome X	91	0.564	0.586	0.519	0.544	0.597	0.656
Obesity	52	0.534	0.667	0.583	0.596	0.586	0.612
Parkinson Disease	43	0.449	0.505	0.707	0.444	0.577	0.542
Arthritis, Rheumatoid	121	0.713	0.802	0.641	0.686	0.632	0.63

[*] #SM denotes somatic mutation

Table 3.9: Summary of CV-AUC values of five-fold cross-validation for 45 diseases in ARB when the benign variants are selected from Affymetrix Genome-Wide Human SNP Array 6.0

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC
Body Weight	body weight	0.683	0.519	0.495	0.513	0.521	0.505
Body Weight Changes	body weight	0.742	0.528	0.481	0.493	0.516	0.45
Obesity	body weight	0.822	0.529	0.515	0.562	0.622	0.61
Breast Neoplasms	cancer	0.723	0.574	0.546	0.491	0.573	0.526
Neuroblastoma	cancer	0.686	0.535	0.484	0.497	0.553	0.554
Pancreatic Neoplasms	cancer	0.733	0.52	0.522	0.465	0.508	0.515
Prostatic Neoplasms	cancer	0.686	0.564	0.528	0.512	0.544	0.537
Carotid Artery Diseases	cardiovascular	0.751	0.518	0.541	0.467	0.512	0.507
Coronary Artery Disease	cardiovascular	0.689	0.521	0.508	0.503	0.515	0.516
Coronary Disease	cardiovascular	0.712	0.546	0.525	0.497	0.545	0.542
Death, Sudden, Cardiac	cardiovascular	0.831	0.538	0.567	0.445	0.533	0.529
Heart Failure	cardiovascular	0.654	0.531	0.504	0.497	0.525	0.538
Hypertension	cardiovascular	0.684	0.586	0.562	0.504	0.547	0.544
Myocardial Infarction	cardiovascular	0.667	0.537	0.506	0.501	0.529	0.537
Stroke	cardiovascular	0.669	0.54	0.507	0.519	0.507	0.509
Cardiovascular Diseases	cardiovascular	0.759	0.484	0.452	0.472	0.56	0.552
Hypertrophy, Left Ventricular	cardiovascular	0.743	0.564	0.518	0.528	0.527	0.512
Diabetic Nephropathies	endocrine	0.708	0.55	0.554	0.494	0.532	0.515
Macular Degeneration	eye disease	0.795	0.511	0.434	0.516	0.476	0.508
Arthritis, Rheumatoid	immune	0.748	0.615	0.568	0.552	0.547	0.574
Asthma	immune	0.687	0.529	0.516	0.485	0.511	0.501
Behcet Syndrome	immune	0.845	0.574	0.479	0.515	0.467	0.479
Colitis, Ulcerative	immune	0.829	0.607	0.593	0.605	0.533	0.557
Crohn Disease	immune	0.863	0.576	0.611	0.483	0.479	0.507
Diabetes Mellitus, Type 1	immune	0.846	0.628	0.523	0.514	0.466	0.479
Inflammation	immune	0.774	0.516	0.449	0.532	0.533	0.5
Inflammatory Bowel Diseases	immune	0.783	0.559	0.47	0.479	0.547	0.544
Lupus Erythematosus, Systemic	immune	0.816	0.6	0.474	0.527	0.432	0.469
Multiple Sclerosis	immune	0.808	0.605	0.506	0.522	0.466	0.482
Psoriasis	immune	0.888	0.615	0.44	0.425	0.402	0.444
Alcoholism	mental	0.677	0.538	0.519	0.491	0.509	0.522
Alzheimer Disease	mental	0.695	0.563	0.582	0.505	0.512	0.547
Attention Deficit Disorder with Hyperactivity	mental	0.682	0.551	0.565	0.524	0.559	0.562
Bipolar Disorder	mental	0.704	0.56	0.54	0.537	0.527	0.52
Depressive Disorder, Major	mental	0.75	0.509	0.479	0.477	0.52	0.548
Mental Competency	mental	0.732	0.506	0.5	0.539	0.538	0.552
Schizophrenia	mental	0.709	0.557	0.564	0.524	0.581	0.581
Diabetes Mellitus, Type 2	metabolic disease	0.694	0.55	0.618	0.491	0.608	0.567
Insulin Resistance	metabolic disease	0.72	0.511	0.523	0.514	0.514	0.536
Metabolic Syndrome X	metabolic disease	0.784	0.545	0.581	0.521	0.499	0.468
Osteoporosis	musculoskeletal	0.742	0.536	0.47	0.562	0.511	0.496
Amyotrophic Lateral Sclerosis	nervous system	0.705	0.504	0.49	0.494	0.511	0.493
Parkinson Disease	nervous system	0.657	0.533	0.505	0.531	0.517	0.517
Sleep	psychological	0.741	0.513	0.483	0.537	0.571	0.514
Albuminuria	urogenital	0.769	0.545	0.512	0.498	0.516	0.537

Table 3.10: Summary of predicted AUC values for 36 diseases in GRASP when the benign variants are selected from Affymetrix Genome-Wide Human SNP Array 6.0. For each disease, the risk variants in GRASP are non-overlapping with the risk variants in ARB

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.73	0.56	0.364	0.396	0.437	0.453	*DIVAN*
Obesity	body weight	0.612	0.536	0.515	0.542	0.538	0.539	*DIVAN*
Breast Neoplasms	cancer	0.654	0.572	0.579	0.516	0.561	0.564	*DIVAN*
Neuroblastoma	cancer	0.649	0.539	0.514	0.635	0.492	0.468	*DIVAN*
Pancreatic Neoplasms	cancer	0.678	0.459	0.57	0.454	0.421	0.423	*DIVAN*
Prostatic Neoplasms	cancer	0.639	0.548	0.55	0.481	0.518	0.543	*DIVAN*
Cardiovascular Diseases	cardiovascular	0.547	0.636	0.615	0.551	0.558	0.588	GWAVA
Coronary Artery Disease	cardiovascular	0.664	0.582	0.569	0.513	0.521	0.53	*DIVAN*
Heart Failure	cardiovascular	0.739	0.517	0.442	0.508	0.388	0.395	*DIVAN*
Hypertension	cardiovascular	0.649	0.555	0.549	0.511	0.509	0.525	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.593	0.498	0.421	0.509	0.526	0.536	*DIVAN*
Myocardial Infarction	cardiovascular	0.653	0.587	0.603	0.488	0.523	0.552	*DIVAN*
Stroke	cardiovascular	0.636	0.637	0.578	0.435	0.33	0.519	GWAVA
Macular Degeneration	eye disease	0.68	0.607	0.559	0.498	0.527	0.549	*DIVAN*
Arthritis, Rheumatoid	immune	0.81	0.738	0.592	0.508	0.457	0.518	*DIVAN*
Asthma	immune	0.667	0.634	0.57	0.508	0.536	0.551	*DIVAN*
Behcet Syndrome	immune	0.752	0.671	0.565	0.549	0.478	0.507	*DIVAN*
Colitis, Ulcerative	immune	0.687	0.601	0.567	0.538	0.528	0.55	*DIVAN*
Crohn Disease	immune	0.661	0.624	0.602	0.52	0.526	0.549	*DIVAN*
Diabetes Mellitus, Type 1	immune	0.794	0.703	0.607	0.514	0.505	0.538	*DIVAN*
Inflammation	immune	0.599	0.533	0.526	0.53	0.471	0.537	*DIVAN*
Inflammatory Bowel Diseases	immune	0.673	0.728	0.745	0.538	0.637	0.683	GenoCanyon
Lupus Erythematosus, Systemic	immune	0.767	0.635	0.638	0.505	0.53	0.564	*DIVAN*
Multiple Sclerosis	immune	0.604	0.548	0.528	0.501	0.512	0.54	*DIVAN*
Psoriasis	immune	0.644	0.582	0.563	0.496	0.516	0.53	*DIVAN*
Alcoholism	mental	0.623	0.631	0.769	0.386	0.537	0.605	GenoCanyon
Alzheimer Disease	mental	0.621	0.532	0.527	0.49	0.509	0.517	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.649	0.577	0.529	0.467	0.534	0.517	*DIVAN*
Bipolar Disorder	mental	0.618	0.538	0.502	0.484	0.508	0.509	*DIVAN*
Depressive Disorder, Major	mental	0.587	0.525	0.495	0.527	0.56	0.536	*DIVAN*
Schizophrenia	mental	0.633	0.555	0.536	0.51	0.507	0.528	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.666	0.553	0.562	0.524	0.553	0.535	*DIVAN*
Insulin Resistance	metabolic disease	0.553	0.521	0.488	0.49	0.609	0.451	Eigen
Metabolic Syndrome X	metabolic disease	0.615	0.647	0.74	0.636	0.627	0.639	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.619	0.541	0.55	0.528	0.506	0.515	*DIVAN*
Sleep	psychological	0.712	0.57	0.505	0.533	0.476	0.563	*DIVAN*

Table 3.11: Summary of predicted AUC values for 36 diseases in GRASP when the benign variants are selected from Affymetrix Genome-Wide Human SNP Array 6.0. For each disease, the risk variants in GRASP are 10kb away from the risk variants in ARB.

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.688	0.498	0.332	0.47	0.486	0.503	*DIVAN*
Obesity	body weight	0.602	0.51	0.488	0.536	0.54	0.544	*DIVAN*
Breast Neoplasms	cancer	0.607	0.555	0.57	0.51	0.55	0.545	*DIVAN*
Neuroblastoma	cancer	0.577	0.489	0.54	0.55	0.254	0.304	*DIVAN*
Pancreatic Neoplasms	cancer	0.602	0.591	0.546	0.559	0.532	0.521	*DIVAN*
Prostatic Neoplasms	cancer	0.607	0.527	0.546	0.492	0.511	0.519	*DIVAN*
Cardiovascular Diseases	cardiovascular	0.526	0.679	0.645	0.454	0.518	0.541	GWAVA
Coronary Artery Disease	cardiovascular	0.648	0.571	0.569	0.523	0.508	0.518	*DIVAN*
Heart Failure	cardiovascular	0.725	0.472	0.457	0.458	0.425	0.441	*DIVAN*
Hypertension	cardiovascular	0.619	0.527	0.532	0.493	0.491	0.511	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.594	0.544	0.407	0.511	0.587	0.561	*DIVAN*
Myocardial Infarction	cardiovascular	0.642	0.604	0.61	0.501	0.542	0.565	*DIVAN*
Stroke	cardiovascular	0.596	0.574	0.552	0.456	0.343	0.49	*DIVAN*
Macular Degeneration	eye disease	0.617	0.581	0.558	0.498	0.506	0.528	*DIVAN*
Arthritis, Rheumatoid	immune	0.696	0.687	0.578	0.508	0.437	0.489	*DIVAN*
Asthma	immune	0.612	0.593	0.558	0.517	0.527	0.534	*DIVAN*
Behcet Syndrome	immune	0.644	0.587	0.524	0.548	0.453	0.474	*DIVAN*
Colitis, Ulcerative	immune	0.63	0.567	0.543	0.531	0.515	0.522	*DIVAN*
Crohn Disease	immune	0.641	0.607	0.598	0.485	0.501	0.532	*DIVAN*
Diabetes Mellitus, Type 1	immune	0.703	0.629	0.597	0.512	0.487	0.531	*DIVAN*
Inflammation	immune	0.566	0.487	0.514	0.438	0.384	0.43	*DIVAN*
Inflammatory Bowel Diseases	immune	0.634	0.702	0.727	0.509	0.547	0.624	GenoCanyon
Lupus Erythematosus, Systemic	immune	0.672	0.628	0.613	0.478	0.532	0.548	*DIVAN*
Multiple Sclerosis	immune	0.595	0.557	0.53	0.52	0.505	0.541	*DIVAN*
Psoriasis	immune	0.634	0.578	0.574	0.508	0.51	0.54	*DIVAN*
Alcoholism	mental	0.503	0.754	0.824	0.529	0.6	0.645	GenoCanyon
Alzheimer Disease	mental	0.606	0.528	0.513	0.505	0.502	0.508	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.621	0.55	0.516	0.49	0.517	0.533	*DIVAN*
Bipolar Disorder	mental	0.616	0.529	0.496	0.489	0.489	0.494	*DIVAN*
Depressive Disorder, Major	mental	0.598	0.535	0.489	0.52	0.531	0.515	*DIVAN*
Schizophrenia	mental	0.622	0.546	0.521	0.513	0.503	0.518	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.668	0.542	0.556	0.532	0.544	0.526	*DIVAN*
Insulin Resistance	metabolic disease	0.613	0.514	0.438	0.496	0.613	0.478	Eigen
Metabolic Syndrome X	metabolic disease	0.59	0.659	0.728	0.563	0.583	0.615	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.607	0.555	0.554	0.514	0.507	0.525	*DIVAN*
Sleep	psychological	0.644	0.577	0.498	0.53	0.402	0.578	*DIVAN*

Table 3.12: Summary of CV-AUC values of five-fold cross-validation for 45 diseases in ARB when benign variants are selected within 10kb of risk variants in the training set for each disease

disease/trait	class	*DIVAN*.region10kb	GWAVA.region10kb	GenoCanyon.region10kb	CADD.region10kb	Eigen.region10kb	EigenPC.region10kb
Body Weight	body weight	0.643	0.518	0.498	0.532	0.529	0.518
Body Weight Changes	body weight	0.687	0.532	0.495	0.547	0.547	0.515
Obesity	body weight	0.795	0.522	0.496	0.508	0.539	0.536
Breast Neoplasms	cancer	0.68	0.511	0.496	0.512	0.512	0.478
Neuroblastoma	cancer	0.641	0.543	0.51	0.514	0.524	0.512
Pancreatic Neoplasms	cancer	0.709	0.526	0.536	0.546	0.552	0.54
Prostatic Neoplasms	cancer	0.643	0.548	0.481	0.513	0.518	0.49
Carotid Artery Diseases	cardiovascular	0.707	0.544	0.517	0.507	0.501	0.501
Coronary Artery Disease	cardiovascular	0.649	0.515	0.497	0.506	0.513	0.51
Coronary Disease	cardiovascular	0.665	0.532	0.529	0.545	0.541	0.561
Death, Sudden, Cardiac	cardiovascular	0.76	0.497	0.503	0.525	0.487	0.494
Heart Failure	cardiovascular	0.63	0.519	0.519	0.491	0.515	0.512
Hypertension	cardiovascular	0.656	0.531	0.516	0.514	0.52	0.508
Myocardial Infarction	cardiovascular	0.644	0.525	0.49	0.535	0.537	0.508
Stroke	cardiovascular	0.627	0.527	0.513	0.509	0.51	0.516
Cardiovascular Diseases	cardiovascular	0.7	0.546	0.526	0.537	0.534	0.532
Hypertrophy, Left Ventricular	cardiovascular	0.666	0.507	0.505	0.489	0.499	0.485
Diabetic Nephropathies	endocrine	0.681	0.566	0.515	0.509	0.528	0.502
Macular Degeneration	eye disease	0.662	0.549	0.502	0.522	0.524	0.521
Arthritis, Rheumatoid	immune	0.713	0.528	0.505	0.509	0.502	0.48
Asthma	immune	0.634	0.529	0.508	0.53	0.518	0.515
Behcet Syndrome	immune	0.667	0.545	0.552	0.501	0.521	0.52
Colitis, Ulcerative	immune	0.735	0.546	0.514	0.49	0.526	0.541
Crohn Disease	immune	0.693	0.557	0.479	0.459	0.492	0.493
Diabetes Mellitus, Type 1	immune	0.689	0.518	0.501	0.528	0.554	0.522
Inflammation	immune	0.719	0.533	0.427	0.488	0.454	0.453
Inflammatory Bowel Diseases	immune	0.676	0.534	0.517	0.475	0.517	0.513
Lupus Erythematosus, Systemic	immune	0.675	0.538	0.476	0.495	0.51	0.482
Multiple Sclerosis	immune	0.691	0.52	0.493	0.438	0.501	0.497
Psoriasis	immune	0.714	0.5	0.49	0.495	0.514	0.546
Alcoholism	mental	0.648	0.518	0.514	0.52	0.533	0.534
Alzheimer Disease	mental	0.665	0.505	0.482	0.531	0.496	0.495
Attention Deficit Disorder with Hyperactivity	mental	0.638	0.54	0.519	0.523	0.518	0.524
Bipolar Disorder	mental	0.634	0.498	0.513	0.523	0.513	0.513
Depressive Disorder, Major	mental	0.678	0.557	0.516	0.545	0.539	0.516
Mental Competency	mental	0.674	0.507	0.499	0.482	0.473	0.473
Schizophrenia	mental	0.658	0.511	0.484	0.512	0.491	0.481
Diabetes Mellitus, Type 2	metabolic disease	0.662	0.536	0.523	0.521	0.501	0.526
Insulin Resistance	metabolic disease	0.712	0.486	0.486	0.509	0.5	0.486
Metabolic Syndrome X	metabolic disease	0.675	0.531	0.543	0.556	0.549	0.544
Osteoporosis	musculoskeletal	0.723	0.515	0.497	0.501	0.538	0.544
Amyotrophic Lateral Sclerosis	nervous system	0.679	0.526	0.511	0.51	0.53	0.519
Parkinson Disease	nervous system	0.642	0.516	0.502	0.521	0.519	0.51
Sleep	psychological	0.682	0.583	0.53	0.464	0.536	0.529
Albuminuria	urogenital	0.758	0.55	0.529	0.55	0.563	0.571

Table 3.13: Summary of predicted AUC values for 36 diseases in GRASP when benign variants are selected within 10kb of risk variants in the training set for each disease. For each disease, the risk variants in GRASP are non-overlapping with the risk variants in ARB

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.653	0.517	0.43	0.252	0.268	0.421	*DIVAN*
Obesity	body weight	0.616	0.573	0.556	0.531	0.53	0.534	*DIVAN*
Breast Neoplasms	cancer	0.652	0.617	0.634	0.534	0.563	0.577	*DIVAN*
Neuroblastoma	cancer	0.674	0.495	0.572	0.504	0.5	0.464	*DIVAN*
Pancreatic Neoplasms	cancer	0.677	0.616	0.65	0.562	0.526	0.548	*DIVAN*
Prostatic Neoplasms	cancer	0.6	0.627	0.592	0.555	0.591	0.607	GWAVA
Cardiovascular Diseases	cardiovascular	0.585	0.622	0.636	0.579	0.504	0.594	GenoCanyon
Coronary Artery Disease	cardiovascular	0.653	0.627	0.614	0.535	0.537	0.558	*DIVAN*
Heart Failure	cardiovascular	0.623	0.59	0.555	0.482	0.467	0.581	*DIVAN*
Hypertension	cardiovascular	0.656	0.618	0.585	0.52	0.534	0.561	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.573	0.598	0.495	0.616	0.631	0.584	Eigen
Myocardial Infarction	cardiovascular	0.627	0.633	0.641	0.529	0.557	0.602	GenoCanyon
Stroke	cardiovascular	0.622	0.748	0.677	0.412	0.472	0.627	GWAVA
Macular Degeneration	eye disease	0.675	0.626	0.601	0.525	0.542	0.567	*DIVAN*
Arthritis, Rheumatoid	immune	0.652	0.759	0.64	0.53	0.469	0.547	GWAVA
Asthma	immune	0.667	0.665	0.603	0.542	0.552	0.569	*DIVAN*
Behcet Syndrome	immune	0.676	0.679	0.619	0.492	0.461	0.553	GWAVA
Colitis, Ulcerative	immune	0.667	0.65	0.61	0.529	0.532	0.558	*DIVAN*
Crohn Disease	immune	0.627	0.654	0.641	0.544	0.546	0.58	GWAVA
Diabetes Mellitus, Type 1	immune	0.739	0.725	0.648	0.552	0.516	0.577	*DIVAN*
Inflammation	immune	0.623	0.596	0.556	0.567	0.525	0.542	*DIVAN*
Inflammatory Bowel Diseases	immune	0.668	0.778	0.77	0.582	0.592	0.659	GWAVA
Lupus Erythematosus, Systemic	immune	0.704	0.682	0.682	0.573	0.581	0.634	*DIVAN*
Multiple Sclerosis	immune	0.601	0.606	0.556	0.509	0.518	0.552	GWAVA
Psoriasis	immune	0.662	0.65	0.627	0.532	0.544	0.58	*DIVAN*
Alcoholism	mental	0.597	0.553	0.769	0.492	0.415	0.488	GenoCanyon
Alzheimer Disease	mental	0.623	0.591	0.568	0.513	0.524	0.54	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.623	0.601	0.56	0.533	0.535	0.541	*DIVAN*
Bipolar Disorder	mental	0.626	0.589	0.544	0.503	0.52	0.523	*DIVAN*
Depressive Disorder, Major	mental	0.598	0.579	0.544	0.522	0.54	0.533	*DIVAN*
Schizophrenia	mental	0.636	0.607	0.579	0.526	0.541	0.564	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.658	0.589	0.604	0.531	0.54	0.547	*DIVAN*
Insulin Resistance	metabolic disease	0.602	0.524	0.5	0.456	0.637	0.478	Eigen
Metabolic Syndrome X	metabolic disease	0.493	0.628	0.743	0.467	0.61	0.562	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.625	0.592	0.605	0.52	0.517	0.544	*DIVAN*
Sleep	psychological	0.675	0.641	0.514	0.535	0.536	0.574	*DIVAN*

Table 3.14: Summary of predicted AUC values for 36 diseases in GRASP when benign variants are selected within 10kb of risk variants in the training set for each disease. For each disease, the risk variants in GRASP are 10kb away from the risk variants in ARB

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.641	0.446	0.37	0.255	0.278	0.372	*DIVAN*
Obesity	body weight	0.607	0.542	0.533	0.522	0.506	0.511	*DIVAN*
Breast Neoplasms	cancer	0.61	0.606	0.643	0.528	0.555	0.604	GenoCanyon
Neuroblastoma	cancer	0.562	0.458	0.602	0.549	0.532	0.494	GenoCanyon
Pancreatic Neoplasms	cancer	0.64	0.567	0.655	0.552	0.527	0.534	GenoCanyon
Prostatic Neoplasms	cancer	0.57	0.609	0.583	0.527	0.564	0.563	GWAVA
Cardiovascular Diseases	cardiovascular	0.574	0.699	0.674	0.6	0.59	0.612	GWAVA
Coronary Artery Disease	cardiovascular	0.64	0.614	0.604	0.52	0.519	0.541	*DIVAN*
Heart Failure	cardiovascular	0.616	0.628	0.576	0.49	0.407	0.58	GWAVA
Hypertension	cardiovascular	0.645	0.592	0.562	0.511	0.507	0.519	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.623	0.542	0.512	0.418	0.355	0.374	*DIVAN*
Myocardial Infarction	cardiovascular	0.629	0.63	0.64	0.517	0.538	0.581	GenoCanyon
Stroke	cardiovascular	0.571	0.681	0.695	0.485	0.494	0.545	GenoCanyon
Macular Degeneration	eye disease	0.623	0.628	0.606	0.527	0.509	0.538	GWAVA
Arthritis, Rheumatoid	immune	0.598	0.739	0.629	0.545	0.458	0.519	GWAVA
Asthma	immune	0.64	0.639	0.599	0.545	0.547	0.54	*DIVAN*
Behcet Syndrome	immune	0.66	0.613	0.599	0.535	0.499	0.571	*DIVAN*
Colitis, Ulcerative	immune	0.635	0.621	0.582	0.525	0.517	0.534	*DIVAN*
Crohn Disease	immune	0.605	0.639	0.63	0.545	0.527	0.549	GWAVA
Diabetes Mellitus, Type 1	immune	0.671	0.673	0.648	0.573	0.515	0.57	GWAVA
Inflammation	immune	0.642	0.602	0.556	0.563	0.527	0.565	*DIVAN*
Inflammatory Bowel Diseases	immune	0.646	0.737	0.747	0.568	0.551	0.636	GenoCanyon
Lupus Erythematosus, Systemic	immune	0.63	0.657	0.672	0.593	0.594	0.619	GenoCanyon
Multiple Sclerosis	immune	0.585	0.595	0.554	0.505	0.52	0.547	GWAVA
Psoriasis	immune	0.654	0.622	0.621	0.52	0.527	0.551	*DIVAN*
Alcoholism	mental	0.589	0.665	0.812	0.477	0.436	0.495	GenoCanyon
Alzheimer Disease	mental	0.601	0.585	0.558	0.535	0.532	0.534	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.619	0.607	0.557	0.521	0.551	0.555	*DIVAN*
Bipolar Disorder	mental	0.614	0.581	0.536	0.491	0.509	0.513	*DIVAN*
Depressive Disorder, Major	mental	0.604	0.591	0.548	0.509	0.523	0.511	*DIVAN*
Schizophrenia	mental	0.625	0.597	0.571	0.526	0.529	0.55	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.664	0.573	0.592	0.53	0.534	0.532	*DIVAN*
Insulin Resistance	metabolic disease	0.664	0.581	0.501	0.429	0.605	0.451	*DIVAN*
Metabolic Syndrome X	metabolic disease	0.539	0.618	0.755	0.519	0.562	0.643	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.603	0.593	0.595	0.54	0.531	0.546	*DIVAN*
Sleep	psychological	0.654	0.638	0.519	0.444	0.496	0.528	*DIVAN*

Table 3.15: Summary of CV-AUC values of five-fold cross-validation for the 45 diseases in ARB for the reduced set of variants such that none is within 10kb of another

disease/trait	class	*DIVAN*.10kb	GWAVA.10kb	GenoCanyon.10kb	CADD.10kb	Eigen.10kb	EigenPC.10kb
Body Weight	body weight	0.691	0.579	0.542	0.536	0.589	0.551
Body Weight Changes	body weight	0.735	0.588	0.504	0.543	0.604	0.51
Obesity	body weight	0.871	0.616	0.548	0.553	0.637	0.631
Breast Neoplasms	cancer	0.733	0.6	0.558	0.539	0.626	0.596
Neuroblastoma	cancer	0.669	0.537	0.486	0.559	0.578	0.571
Pancreatic Neoplasms	cancer	0.742	0.578	0.554	0.486	0.531	0.576
Prostatic Neoplasms	cancer	0.693	0.591	0.578	0.523	0.584	0.565
Carotid Artery Diseases	cardiovascular	0.766	0.581	0.587	0.51	0.546	0.531
Coronary Artery Disease	cardiovascular	0.686	0.586	0.527	0.529	0.579	0.56
Coronary Disease	cardiovascular	0.721	0.576	0.585	0.505	0.567	0.57
Death, Sudden, Cardiac	cardiovascular	0.829	0.566	0.598	0.511	0.551	0.567
Heart Failure	cardiovascular	0.662	0.58	0.532	0.523	0.572	0.56
Hypertension	cardiovascular	0.703	0.604	0.592	0.5	0.55	0.584
Myocardial Infarction	cardiovascular	0.674	0.582	0.543	0.518	0.56	0.556
Stroke	cardiovascular	0.659	0.603	0.542	0.515	0.573	0.548
Cardiovascular Diseases	cardiovascular	0.739	0.587	0.481	0.575	0.647	0.63
Hypertrophy, Left Ventricular	cardiovascular	0.737	0.579	0.545	0.539	0.602	0.567
Diabetic Nephropathies	endocrine	0.694	0.522	0.499	0.487	0.519	0.471
Macular Degeneration	eye disease	0.738	0.585	0.499	0.494	0.575	0.577
Arthritis, Rheumatoid	immune	0.744	0.625	0.604	0.523	0.575	0.613
Asthma	immune	0.679	0.59	0.548	0.514	0.57	0.545
Behcet Syndrome	immune	0.778	0.606	0.544	0.559	0.535	0.556
Colitis, Ulcerative	immune	0.809	0.659	0.598	0.586	0.599	0.645
Crohn Disease	immune	0.83	0.649	0.634	0.486	0.537	0.616
Diabetes Mellitus, Type 1	immune	0.816	0.66	0.551	0.519	0.548	0.55
Inflammation	immune	0.834	0.567	0.517	0.503	0.556	0.533
Inflammatory Bowel Diseases	immune	0.762	0.624	0.543	0.518	0.591	0.597
Lupus Erythematosus, Systemic	immune	0.758	0.608	0.569	0.526	0.532	0.581
Multiple Sclerosis	immune	0.775	0.633	0.594	0.549	0.58	0.597
Psoriasis	immune	0.846	0.63	0.55	0.513	0.541	0.576
Alcoholism	mental	0.674	0.554	0.533	0.51	0.545	0.547
Alzheimer Disease	mental	0.691	0.632	0.597	0.528	0.565	0.6
Attention Deficit Disorder with Hyperactivity	mental	0.685	0.596	0.595	0.556	0.601	0.61
Bipolar Disorder	mental	0.683	0.577	0.558	0.499	0.558	0.568
Depressive Disorder, Major	mental	0.689	0.579	0.517	0.518	0.57	0.574
Mental Competency	mental	0.76	0.544	0.543	0.547	0.607	0.563
Schizophrenia	mental	0.685	0.545	0.591	0.486	0.556	0.57
Diabetes Mellitus, Type 2	metabolic disease	0.71	0.578	0.601	0.569	0.6	0.589
Insulin Resistance	metabolic disease	0.723	0.589	0.566	0.53	0.577	0.611
Metabolic Syndrome X	metabolic disease	0.792	0.55	0.505	0.602	0.688	0.607
Osteoporosis	musculoskeletal	0.742	0.582	0.502	0.512	0.613	0.574
Amyotrophic Lateral Sclerosis	nervous system	0.701	0.563	0.553	0.525	0.55	0.571
Parkinson Disease	nervous system	0.658	0.6	0.544	0.524	0.555	0.556
Sleep	psychological	0.75	0.546	0.504	0.543	0.644	0.591
Albuminuria	urogenital	0.787	0.559	0.498	0.555	0.605	0.609

Table 3.16: Number of diseases for which the method has the best predictive performance. The two settings are: SNPs in training and testing sets are non-overlapping; for each disease, any SNP in the training set (from ARB) is at least 10kb away from all SNPs in the testing set (from GRASP).

Top method	Non-overlapping	10kb away
DIVAN	27	20
GWAVA	4	8
GenoCanyon	3	8
CADD	0	0
Eigen	2	0
EigenPC	0	0

Table 3.17: Summary of predicted AUC values for 36 diseases in GRASP using the predictive models built from risk variants of corresponding diseases in ARB. For each disease, the risk variants in GRASP are 10kb away with the risk variants in ARB. The benign variants are 10 times of the risk variants in the testing set

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.755	0.446	0.37	0.255	0.278	0.372	*DIVAN*
Obesity	body weight	0.643	0.542	0.533	0.522	0.506	0.511	*DIVAN*
Breast Neoplasms	cancer	0.618	0.606	0.643	0.528	0.555	0.604	GenoCanyon
Neuroblastoma	cancer	0.598	0.458	0.602	0.549	0.532	0.494	GenoCanyon
Pancreatic Neoplasms	cancer	0.553	0.567	0.655	0.552	0.527	0.534	GenoCanyon
Prostatic Neoplasms	cancer	0.592	0.609	0.583	0.527	0.564	0.563	GWAVA
Cardiovascular Diseases	cardiovascular	0.535	0.699	0.674	0.6	0.59	0.612	GWAVA
Coronary Artery Disease	cardiovascular	0.634	0.614	0.604	0.52	0.519	0.541	*DIVAN*
Heart Failure	cardiovascular	0.579	0.628	0.576	0.49	0.407	0.58	GWAVA
Hypertension	cardiovascular	0.616	0.592	0.562	0.511	0.507	0.519	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.573	0.542	0.512	0.418	0.355	0.374	*DIVAN*
Myocardial Infarction	cardiovascular	0.642	0.63	0.64	0.517	0.538	0.581	*DIVAN*
Stroke	cardiovascular	0.647	0.681	0.695	0.485	0.494	0.545	GenoCanyon
Macular Degeneration	eye disease	0.602	0.628	0.606	0.527	0.509	0.538	GWAVA
Arthritis, Rheumatoid	immune	0.682	0.739	0.629	0.545	0.458	0.519	GWAVA
Asthma	immune	0.608	0.639	0.599	0.545	0.547	0.54	GWAVA
Behcet Syndrome	immune	0.649	0.613	0.599	0.535	0.499	0.571	*DIVAN*
Colitis, Ulcerative	immune	0.635	0.621	0.582	0.525	0.517	0.534	*DIVAN*
Crohn Disease	immune	0.646	0.639	0.63	0.545	0.527	0.549	*DIVAN*
Diabetes Mellitus, Type 1	immune	0.701	0.673	0.648	0.573	0.515	0.57	*DIVAN*
Inflammation	immune	0.593	0.602	0.556	0.563	0.527	0.565	GWAVA
Inflammatory Bowel Diseases	immune	0.672	0.737	0.747	0.568	0.551	0.636	GenoCanyon
Lupus Erythematosus, Systemic	immune	0.655	0.657	0.672	0.593	0.594	0.619	GenoCanyon
Multiple Sclerosis	immune	0.589	0.595	0.554	0.505	0.52	0.547	GWAVA
Psoriasis	immune	0.627	0.622	0.621	0.52	0.527	0.551	*DIVAN*
Alcoholism	mental	0.506	0.665	0.812	0.477	0.436	0.495	GenoCanyon
Alzheimer Disease	mental	0.61	0.585	0.558	0.535	0.532	0.534	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.628	0.607	0.557	0.521	0.551	0.555	*DIVAN*
Bipolar Disorder	mental	0.614	0.581	0.536	0.491	0.509	0.513	*DIVAN*
Depressive Disorder, Major	mental	0.605	0.591	0.548	0.509	0.523	0.511	*DIVAN*
Schizophrenia	mental	0.64	0.597	0.571	0.526	0.529	0.55	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.669	0.573	0.592	0.53	0.534	0.532	*DIVAN*
Insulin Resistance	metabolic disease	0.659	0.581	0.501	0.429	0.605	0.451	*DIVAN*
Metabolic Syndrome X	metabolic disease	0.633	0.618	0.755	0.519	0.562	0.643	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.618	0.593	0.595	0.54	0.531	0.546	*DIVAN*
Sleep	psychological	0.703	0.638	0.519	0.444	0.496	0.528	*DIVAN*

Table 3.18: Summary of predicted MCC values for 36 diseases in GRASP using the predictive models built from risk variants of corresponding diseases in ARB. For each disease, the risk variants in GRASP are 10kb away with the risk variants in ARB. The benign variants are 10 times of the risk variants in the testing set

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethod
Body Weight Changes	body weight	0.118	-0.0514	-0.101	-0.0996	-0.0755	-0.0188	*DIVAN*
Obesity	body weight	0.0615	0.00519	0.0445	0.00598	-0.00314	-0.0223	*DIVAN*
Breast Neoplasms	cancer	0.0427	0.062	0.123	-0.00969	0.0449	0.0387	GenoCanyon
Neuroblastoma	cancer	0.111	-0.0225	0.0337	-0.0411	-0.0727	-0.0858	*DIVAN*
Pancreatic Neoplasms	cancer	-0.0303	-0.0217	0.156	-0.00567	0.0322	0.0355	GenoCanyon
Prostatic Neoplasms	cancer	0.0116	0.0254	0.0492	-0.00843	0.0106	0.0202	GenoCanyon
Cardiovascular Diseases	cardiovascular	-0.076	0.132	0.187	-0.0191	0.0885	0.178	GenoCanyon
Coronary Artery Disease	cardiovascular	0.0537	0.0696	0.0767	-3.69e-05	0.00672	0.0193	GenoCanyon
Heart Failure	cardiovascular	0.0071	-0.0523	0.00576	0.00102	-0.0567	-0.0281	*DIVAN*
Hypertension	cardiovascular	0.0387	0.0374	0.0379	-0.0306	-0.0285	0.00202	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.0617	-0.061	-0.0315	-0.0968	-0.0722	-0.0924	*DIVAN*
Myocardial Infarction	cardiovascular	0.056	0.0723	0.117	-0.00736	0.0444	0.0575	GenoCanyon
Stroke	cardiovascular	0.0571	-0.052	0.13	-0.0944	-0.0755	-0.0879	GenoCanyon
Macular Degeneration	eye disease	0.0641	0.101	0.067	0.0151	0.0345	0.0489	GWAVA
Arthritis, Rheumatoid	immune	0.143	0.186	0.0824	0.0139	0.00658	0.0363	GWAVA
Asthma	immune	0.0632	0.0794	0.0648	0.032	0.0432	0.0232	GWAVA
Behcet Syndrome	immune	0.177	0.0712	0.0675	0.0109	-0.00752	0.0294	*DIVAN*
Colitis, Ulcerative	immune	0.0661	0.0355	0.0606	-0.0155	0.016	0.032	*DIVAN*
Crohn Disease	immune	0.0554	0.0778	0.114	0.0111	0.0271	0.0259	GenoCanyon
Diabetes Mellitus, Type 1	immune	0.157	0.128	0.108	0.0289	0.0315	0.0573	*DIVAN*
Inflammation	immune	0.00224	-0.00533	0.0933	-0.00692	0.0131	0.00144	GenoCanyon
Inflammatory Bowel Diseases	immune	0.118	0.14	0.218	0.0566	0.0648	0.0544	GenoCanyon
Lupus Erythematosus, Systemic	immune	0.0966	0.04	0.134	-0.0117	-0.00309	0.0883	GenoCanyon
Multiple Sclerosis	immune	0.0125	0.00432	0.0387	-0.000781	0.0221	0.00559	GenoCanyon
Psoriasis	immune	0.0602	0.0563	0.112	0.0161	0.0285	0.0257	GenoCanyon
Alcoholism	mental	0.0995	-0.0542	0.338	-0.0662	-0.0403	-0.0686	GenoCanyon
Alzheimer Disease	mental	0.0592	0.0168	0.0438	-0.000725	0.00274	-0.00175	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.0456	0.02	0.0424	-0.00954	0.0384	0.00809	*DIVAN*
Bipolar Disorder	mental	0.058	0.0061	0.0166	-0.0111	-0.00656	-0.00835	*DIVAN*
Depressive Disorder, Major	mental	0.0442	0.00775	0.0194	0.00909	0.00919	-0.0157	*DIVAN*
Schizophrenia	mental	0.0846	0.0602	0.0443	0.0192	0.0187	0.0251	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.12	0.0314	0.0871	0.0107	0.0249	0.000926	*DIVAN*
Insulin Resistance	metabolic disease	0.0854	-0.0469	-0.0666	-0.0331	0.2	-0.0745	Eigen
Metabolic Syndrome X	metabolic disease	-0.00949	0.0776	0.246	0.00784	0.0572	0.0237	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.0558	0.038	0.0765	0.0152	0.0321	0.00324	GenoCanyon
Sleep	psychological	0.121	0.0935	-0.0196	-0.0246	-0.0211	0.081	*DIVAN*

Table 3.19: Summary statistics of CV-AUC values for the 45 diseases in ARB and predicted AUC values of 36 diseases in GRASP when number of benign variants varies from 10 times to 100 times of the risk variants in the training set for *DIVAN*

	CV (ARB)	Non-overlapping (GRASP)	10kb away (GRASP)
10 times	0.745	0.661	0.626
AUC (mean/sd)	(sd: 0.060)	(sd: 0.055)	(sd: 0.047)
100 times	0.74	0.666	0.633
AUC (mean/sd)	(sd: 0.057)	(sd: 0.056)	(sd: 0.046)

Table 3.20: Summary of CV-AUC values of five-fold cross-validation for 45 diseases in ARB when benign variants are 100 times of risk variants in the training set for each disease

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC
Body Weight	body weight	0.691	0.575	0.538	0.53	0.582	0.546
Body Weight Changes	body weight	0.734	0.603	0.52	0.549	0.612	0.523
Obesity	body weight	0.822	0.626	0.593	0.558	0.626	0.637
Breast Neoplasms	cancer	0.718	0.591	0.592	0.526	0.627	0.601
Neuroblastoma	cancer	0.671	0.545	0.494	0.538	0.561	0.571
Pancreatic Neoplasms	cancer	0.745	0.596	0.592	0.51	0.555	0.593
Prostatic Neoplasms	cancer	0.688	0.589	0.575	0.519	0.571	0.557
Carotid Artery Diseases	cardiovascular	0.748	0.596	0.594	0.503	0.553	0.534
Coronary Artery Disease	cardiovascular	0.692	0.586	0.537	0.532	0.584	0.563
Coronary Disease	cardiovascular	0.703	0.588	0.589	0.497	0.555	0.563
Death, Sudden, Cardiac	cardiovascular	0.811	0.553	0.576	0.531	0.57	0.586
Heart Failure	cardiovascular	0.67	0.578	0.533	0.534	0.577	0.564
Hypertension	cardiovascular	0.696	0.613	0.589	0.525	0.583	0.585
Myocardial Infarction	cardiovascular	0.676	0.586	0.552	0.513	0.566	0.559
Stroke	cardiovascular	0.669	0.609	0.545	0.508	0.569	0.546
Cardiovascular Diseases	cardiovascular	0.725	0.593	0.495	0.591	0.655	0.629
Hypertrophy, Left Ventricular	cardiovascular	0.755	0.591	0.564	0.523	0.592	0.573
Diabetic Nephropathies	endocrine	0.718	0.535	0.571	0.501	0.536	0.506
Macular Degeneration	eye disease	0.781	0.575	0.485	0.503	0.572	0.583
Arthritis, Rheumatoid	immune	0.775	0.63	0.603	0.498	0.584	0.62
Asthma	immune	0.681	0.587	0.552	0.508	0.571	0.535
Behcet Syndrome	immune	0.847	0.59	0.521	0.532	0.504	0.511
Colitis, Ulcerative	immune	0.814	0.637	0.608	0.584	0.614	0.642
Crohn Disease	immune	0.839	0.653	0.671	0.49	0.553	0.608
Diabetes Mellitus, Type 1	immune	0.849	0.657	0.539	0.515	0.546	0.558
Inflammation	immune	0.791	0.56	0.493	0.481	0.535	0.534
Inflammatory Bowel Diseases	immune	0.763	0.615	0.53	0.511	0.587	0.585
Lupus Erythematosus, Systemic	immune	0.813	0.614	0.522	0.524	0.526	0.564
Multiple Sclerosis	immune	0.811	0.625	0.516	0.526	0.536	0.552
Psoriasis	immune	0.879	0.661	0.511	0.52	0.521	0.544
Alcoholism	mental	0.675	0.547	0.537	0.512	0.54	0.541
Alzheimer Disease	mental	0.712	0.62	0.602	0.521	0.551	0.585
Attention Deficit Disorder with Hyperactivity	mental	0.675	0.595	0.603	0.544	0.594	0.597
Bipolar Disorder	mental	0.697	0.58	0.586	0.501	0.559	0.563
Depressive Disorder, Major	mental	0.703	0.577	0.529	0.493	0.552	0.573
Mental Competency	mental	0.736	0.543	0.522	0.559	0.618	0.553
Schizophrenia	mental	0.724	0.569	0.609	0.501	0.565	0.574
Diabetes Mellitus, Type 2	metabolic disease	0.719	0.6	0.642	0.563	0.615	0.595
Insulin Resistance	metabolic disease	0.729	0.592	0.583	0.527	0.592	0.614
Metabolic Syndrome X	metabolic disease	0.759	0.548	0.559	0.56	0.62	0.553
Osteoporosis	musculoskeletal	0.728	0.577	0.508	0.501	0.61	0.591
Amyotrophic Lateral Sclerosis	nervous system	0.708	0.571	0.548	0.521	0.549	0.557
Parkinson Disease	nervous system	0.666	0.592	0.54	0.52	0.555	0.553
Sleep	psychological	0.723	0.532	0.511	0.535	0.635	0.578
Albuminuria	urogenital	0.769	0.567	0.538	0.523	0.595	0.583

Table 3.21: Summary of predicted AUC values for 36 diseases in GRASP when benign variants are 100 times of risk variants in the training set for each disease. For each disease, the risk variants in GRASP are non-overlapping with the risk variants in ARB

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.715	0.517	0.43	0.252	0.268	0.421	*DIVAN*
Obesity	body weight	0.621	0.573	0.556	0.531	0.53	0.534	*DIVAN*
Breast Neoplasms	cancer	0.664	0.617	0.634	0.534	0.563	0.577	*DIVAN*
Neuroblastoma	cancer	0.67	0.495	0.572	0.504	0.5	0.464	*DIVAN*
Pancreatic Neoplasms	cancer	0.689	0.616	0.65	0.562	0.526	0.548	*DIVAN*
Prostatic Neoplasms	cancer	0.65	0.627	0.592	0.555	0.591	0.607	*DIVAN*
Cardiovascular Diseases	cardiovascular	0.583	0.622	0.636	0.579	0.504	0.594	GenoCanyon
Coronary Artery Disease	cardiovascular	0.66	0.627	0.614	0.535	0.537	0.558	*DIVAN*
Heart Failure	cardiovascular	0.729	0.59	0.555	0.482	0.467	0.581	*DIVAN*
Hypertension	cardiovascular	0.669	0.618	0.585	0.52	0.534	0.561	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.6	0.598	0.495	0.616	0.631	0.584	Eigen
Myocardial Infarction	cardiovascular	0.654	0.633	0.641	0.529	0.557	0.602	*DIVAN*
Stroke	cardiovascular	0.656	0.748	0.677	0.412	0.472	0.627	GWAVA
Macular Degeneration	eye disease	0.677	0.626	0.601	0.525	0.542	0.567	*DIVAN*
Arthritis, Rheumatoid	immune	0.802	0.759	0.64	0.53	0.469	0.547	*DIVAN*
Asthma	immune	0.688	0.665	0.603	0.542	0.552	0.569	*DIVAN*
Behcet Syndrome	immune	0.758	0.679	0.619	0.492	0.461	0.553	*DIVAN*
Colitis, Ulcerative	immune	0.692	0.65	0.61	0.529	0.532	0.558	*DIVAN*
Crohn Disease	immune	0.675	0.654	0.641	0.544	0.546	0.58	*DIVAN*
Diabetes Mellitus, Type 1	immune	0.798	0.725	0.648	0.552	0.516	0.577	*DIVAN*
Inflammation	immune	0.621	0.596	0.556	0.567	0.525	0.542	*DIVAN*
Inflammatory Bowel Diseases	immune	0.684	0.778	0.77	0.582	0.592	0.659	GWAVA
Lupus Erythematosus, Systemic	immune	0.758	0.682	0.682	0.573	0.581	0.634	*DIVAN*
Multiple Sclerosis	immune	0.619	0.606	0.556	0.509	0.518	0.552	*DIVAN*
Psoriasis	immune	0.651	0.65	0.627	0.532	0.544	0.58	*DIVAN*
Alcoholism	mental	0.598	0.553	0.769	0.492	0.415	0.488	GenoCanyon
Alzheimer Disease	mental	0.63	0.591	0.568	0.513	0.524	0.54	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.661	0.601	0.56	0.533	0.535	0.541	*DIVAN*
Bipolar Disorder	mental	0.628	0.589	0.544	0.503	0.52	0.523	*DIVAN*
Depressive Disorder, Major	mental	0.606	0.579	0.544	0.522	0.54	0.533	*DIVAN*
Schizophrenia	mental	0.65	0.607	0.579	0.526	0.541	0.564	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.674	0.589	0.604	0.531	0.54	0.547	*DIVAN*
Insulin Resistance	metabolic disease	0.593	0.524	0.5	0.456	0.637	0.478	Eigen
Metabolic Syndrome X	metabolic disease	0.586	0.628	0.743	0.467	0.61	0.562	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.637	0.592	0.605	0.52	0.517	0.544	*DIVAN*
Sleep	psychological	0.737	0.641	0.514	0.535	0.536	0.574	*DIVAN*

Table 3.22: Summary of predicted AUC values for 36 diseases in GRASP when benign variants are 100 times of risk variants in the training set for each disease. For each disease, the risk variants in GRASP are 10kb away from the risk variants in ARB

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.729	0.446	0.37	0.255	0.278	0.372	*DIVAN*
Obesity	body weight	0.623	0.542	0.533	0.522	0.506	0.511	*DIVAN*
Breast Neoplasms	cancer	0.611	0.606	0.643	0.528	0.555	0.604	GenoCanyon
Neuroblastoma	cancer	0.623	0.458	0.602	0.549	0.532	0.494	*DIVAN*
Pancreatic Neoplasms	cancer	0.643	0.567	0.655	0.552	0.527	0.534	GenoCanyon
Prostatic Neoplasms	cancer	0.609	0.609	0.583	0.527	0.564	0.563	GWAVA
Cardiovascular Diseases	cardiovascular	0.514	0.699	0.674	0.6	0.59	0.612	GWAVA
Coronary Artery Disease	cardiovascular	0.638	0.614	0.604	0.52	0.519	0.541	*DIVAN*
Heart Failure	cardiovascular	0.677	0.628	0.576	0.49	0.407	0.58	*DIVAN*
Hypertension	cardiovascular	0.642	0.592	0.562	0.511	0.507	0.519	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.607	0.542	0.512	0.418	0.355	0.374	*DIVAN*
Myocardial Infarction	cardiovascular	0.639	0.63	0.64	0.517	0.538	0.581	GenoCanyon
Stroke	cardiovascular	0.678	0.681	0.695	0.485	0.494	0.545	GenoCanyon
Macular Degeneration	eye disease	0.615	0.628	0.606	0.527	0.509	0.538	GWAVA
Arthritis, Rheumatoid	immune	0.68	0.739	0.629	0.545	0.458	0.519	GWAVA
Asthma	immune	0.623	0.639	0.599	0.545	0.547	0.54	GWAVA
Behcet Syndrome	immune	0.651	0.613	0.599	0.535	0.499	0.571	*DIVAN*
Colitis, Ulcerative	immune	0.637	0.621	0.582	0.525	0.517	0.534	*DIVAN*
Crohn Disease	immune	0.653	0.639	0.63	0.545	0.527	0.549	*DIVAN*
Diabetes Mellitus, Type 1	immune	0.706	0.673	0.648	0.573	0.515	0.57	*DIVAN*
Inflammation	immune	0.587	0.602	0.556	0.563	0.527	0.565	GWAVA
Inflammatory Bowel Diseases	immune	0.653	0.737	0.747	0.568	0.551	0.636	GenoCanyon
Lupus Erythematosus, Systemic	immune	0.656	0.657	0.672	0.593	0.594	0.619	GenoCanyon
Multiple Sclerosis	immune	0.604	0.595	0.554	0.505	0.52	0.547	*DIVAN*
Psoriasis	immune	0.648	0.622	0.621	0.52	0.527	0.551	*DIVAN*
Alcoholism	mental	0.487	0.665	0.812	0.477	0.436	0.495	GenoCanyon
Alzheimer Disease	mental	0.615	0.585	0.558	0.535	0.532	0.534	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.628	0.607	0.557	0.521	0.551	0.555	*DIVAN*
Bipolar Disorder	mental	0.622	0.581	0.536	0.491	0.509	0.513	*DIVAN*
Depressive Disorder, Major	mental	0.604	0.591	0.548	0.509	0.523	0.511	*DIVAN*
Schizophrenia	mental	0.639	0.597	0.571	0.526	0.529	0.55	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.675	0.573	0.592	0.53	0.534	0.532	*DIVAN*
Insulin Resistance	metabolic disease	0.651	0.581	0.501	0.429	0.605	0.451	*DIVAN*
Metabolic Syndrome X	metabolic disease	0.598	0.618	0.755	0.519	0.562	0.643	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.613	0.593	0.595	0.54	0.531	0.546	*DIVAN*
Sleep	psychological	0.694	0.638	0.519	0.444	0.496	0.528	*DIVAN*

Table 3.23: Summary of predicted AUC values for 36 diseases in GRASP using the predictive models built from risk variants of corresponding diseases in ARB. The benign variants are 100 times of the risk variants in the testing set

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.729	0.525	0.434	0.256	0.272	0.422	*DIVAN*
Obesity	body weight	0.633	0.572	0.556	0.529	0.528	0.533	*DIVAN*
Breast Neoplasms	cancer	0.672	0.616	0.633	0.531	0.562	0.577	*DIVAN*
Neuroblastoma	cancer	0.647	0.498	0.574	0.504	0.502	0.465	*DIVAN*
Pancreatic Neoplasms	cancer	0.695	0.617	0.651	0.558	0.524	0.546	*DIVAN*
Prostatic Neoplasms	cancer	0.626	0.625	0.591	0.552	0.589	0.607	*DIVAN*
Cardiovascular Diseases	cardiovascular	0.529	0.63	0.64	0.577	0.507	0.596	GenoCanyon
Coronary Artery Disease	cardiovascular	0.647	0.627	0.614	0.534	0.537	0.557	*DIVAN*
Heart Failure	cardiovascular	0.717	0.595	0.561	0.484	0.474	0.584	*DIVAN*
Hypertension	cardiovascular	0.653	0.616	0.585	0.516	0.532	0.56	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.59	0.596	0.505	0.625	0.644	0.585	Eigen
Myocardial Infarction	cardiovascular	0.656	0.633	0.64	0.526	0.555	0.602	*DIVAN*
Stroke	cardiovascular	0.668	0.753	0.68	0.414	0.478	0.626	GWAVA
Macular Degeneration	eye disease	0.68	0.626	0.601	0.525	0.542	0.566	*DIVAN*
Arthritis, Rheumatoid	immune	0.766	0.758	0.64	0.53	0.469	0.546	*DIVAN*
Asthma	immune	0.67	0.665	0.603	0.539	0.55	0.568	*DIVAN*
Behcet Syndrome	immune	0.74	0.678	0.619	0.488	0.459	0.553	*DIVAN*
Colitis, Ulcerative	immune	0.677	0.649	0.609	0.526	0.53	0.557	*DIVAN*
Crohn Disease	immune	0.672	0.653	0.64	0.542	0.544	0.578	*DIVAN*
Diabetes Mellitus, Type 1	immune	0.801	0.724	0.648	0.55	0.514	0.576	*DIVAN*
Inflammation	immune	0.59	0.598	0.556	0.565	0.523	0.543	GWAVA
Inflammatory Bowel Diseases	immune	0.693	0.78	0.77	0.58	0.591	0.661	GWAVA
Lupus Erythematosus, Systemic	immune	0.752	0.683	0.683	0.57	0.58	0.636	*DIVAN*
Multiple Sclerosis	immune	0.616	0.605	0.556	0.506	0.516	0.552	*DIVAN*
Psoriasis	immune	0.635	0.65	0.627	0.528	0.542	0.58	GWAVA
Alcoholism	mental	0.635	0.556	0.77	0.489	0.414	0.49	GenoCanyon
Alzheimer Disease	mental	0.624	0.59	0.567	0.511	0.522	0.539	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.647	0.598	0.559	0.529	0.533	0.54	*DIVAN*
Bipolar Disorder	mental	0.611	0.589	0.543	0.503	0.519	0.522	*DIVAN*
Depressive Disorder, Major	mental	0.599	0.578	0.543	0.52	0.538	0.532	*DIVAN*
Schizophrenia	mental	0.645	0.606	0.578	0.526	0.54	0.563	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.665	0.588	0.603	0.53	0.539	0.546	*DIVAN*
Insulin Resistance	metabolic disease	0.622	0.525	0.498	0.454	0.636	0.48	Eigen
Metabolic Syndrome X	metabolic disease	0.614	0.629	0.745	0.472	0.617	0.561	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.639	0.59	0.604	0.516	0.515	0.544	*DIVAN*
Sleep	psychological	0.737	0.644	0.514	0.534	0.535	0.576	*DIVAN*

Table 3.24: Summary of predicted AUC values for 36 diseases in GRASP using the predictive models built from risk variants of corresponding diseases in ARB. For each disease, the risk variants in GRASP are 10kb away with the risk variants in ARB. The benign variants are 100 times of the risk variants in the testing set

disease/trait	class	*DIVAN*	GWAVA	GenoCanyon	CADD	Eigen	EigenPC	topmethods
Body Weight Changes	body weight	0.747	0.455	0.372	0.26	0.285	0.374	*DIVAN*
Obesity	body weight	0.645	0.541	0.533	0.519	0.504	0.51	*DIVAN*
Breast Neoplasms	cancer	0.616	0.604	0.642	0.524	0.552	0.603	GenoCanyon
Neuroblastoma	cancer	0.605	0.459	0.608	0.548	0.535	0.493	GenoCanyon
Pancreatic Neoplasms	cancer	0.56	0.568	0.657	0.554	0.53	0.534	GenoCanyon
Prostatic Neoplasms	cancer	0.591	0.607	0.583	0.524	0.562	0.563	GWAVA
Cardiovascular Diseases	cardiovascular	0.532	0.706	0.678	0.597	0.592	0.612	GWAVA
Coronary Artery Disease	cardiovascular	0.634	0.614	0.604	0.519	0.519	0.54	*DIVAN*
Heart Failure	cardiovascular	0.59	0.636	0.58	0.493	0.412	0.582	GWAVA
Hypertension	cardiovascular	0.615	0.592	0.562	0.507	0.505	0.518	*DIVAN*
Hypertrophy, Left Ventricular	cardiovascular	0.576	0.545	0.522	0.426	0.366	0.372	*DIVAN*
Myocardial Infarction	cardiovascular	0.641	0.63	0.639	0.514	0.537	0.581	*DIVAN*
Stroke	cardiovascular	0.651	0.688	0.697	0.485	0.5	0.545	GenoCanyon
Macular Degeneration	eye disease	0.602	0.627	0.606	0.526	0.509	0.537	GWAVA
Arthritis, Rheumatoid	immune	0.682	0.739	0.629	0.545	0.458	0.519	GWAVA
Asthma	immune	0.609	0.639	0.599	0.542	0.545	0.539	GWAVA
Behcet Syndrome	immune	0.65	0.614	0.6	0.533	0.498	0.573	*DIVAN*
Colitis, Ulcerative	immune	0.634	0.62	0.581	0.522	0.514	0.534	*DIVAN*
Crohn Disease	immune	0.644	0.638	0.629	0.543	0.526	0.548	*DIVAN*
Diabetes Mellitus, Type 1	immune	0.701	0.673	0.648	0.571	0.513	0.569	*DIVAN*
Inflammation	immune	0.592	0.604	0.556	0.561	0.525	0.566	GWAVA
Inflammatory Bowel Diseases	immune	0.67	0.739	0.748	0.565	0.55	0.637	GenoCanyon
Lupus Erythematosus, Systemic	immune	0.658	0.658	0.672	0.59	0.593	0.62	GenoCanyon
Multiple Sclerosis	immune	0.59	0.593	0.555	0.502	0.518	0.547	GWAVA
Psoriasis	immune	0.626	0.621	0.62	0.517	0.525	0.55	*DIVAN*
Alcoholism	mental	0.505	0.668	0.813	0.473	0.434	0.497	GenoCanyon
Alzheimer Disease	mental	0.608	0.584	0.557	0.534	0.531	0.533	*DIVAN*
Attention Deficit Disorder with Hyperactivity	mental	0.626	0.604	0.555	0.517	0.548	0.554	*DIVAN*
Bipolar Disorder	mental	0.613	0.581	0.535	0.49	0.508	0.513	*DIVAN*
Depressive Disorder, Major	mental	0.605	0.59	0.547	0.507	0.521	0.51	*DIVAN*
Schizophrenia	mental	0.638	0.596	0.57	0.526	0.529	0.549	*DIVAN*
Diabetes Mellitus, Type 2	metabolic disease	0.669	0.572	0.591	0.529	0.534	0.531	*DIVAN*
Insulin Resistance	metabolic disease	0.659	0.583	0.5	0.426	0.604	0.453	*DIVAN*
Metabolic Syndrome X	metabolic disease	0.629	0.621	0.758	0.523	0.571	0.642	GenoCanyon
Amyotrophic Lateral Sclerosis	nervous system	0.618	0.593	0.594	0.537	0.529	0.546	*DIVAN*
Sleep	psychological	0.704	0.64	0.519	0.443	0.496	0.53	*DIVAN*

Chapter 4

traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals

4.1 Introduction

Genome-wide association studies (GWAS) have been conducted en masse in the past decade and have been tremendously successful in identifying sequence variants that are significantly associated with common diseases and traits [61]. To this day, thousands of GWAS have been conducted and reported, across diverse spectrums of diseases as well as qualitative and quantitative phenotypes. Resources, such as association result browser (`http://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm`) and NHGRI GWAS catalog [62] have been established to catalog all the trait-associated variants.

Currently, the association result browser contains 44,124 association results (checked on October 10, 2015), which corresponds to 30,553 (autosomes plus chromosome X) unique trait-associated single nucleotide polymorphisms (taSNPs), linking to 573 diseases or phenotypes. We believe such a catalog of taSNPs offers scientists a unique perspective to explore and annotate the functional potential of any given genomic intervals.

Maurano et al. showed that regulatory DNA marked by deoxyribonuclease I (DNase I) hypersensitive sites (DHSs) was enriched with noncoding GWAS SNPs [29]. Recent studies from ENCODE and Roadmap Epigenome consortia systematically examined enrichment of taSNPs in ChIP-seq peaks of transcription factors and histone marks, and unveiled biologically interesting connections [63].

These results indicate the utilities of conducting taSNP enrichment analyses in genomic intervals of interest. We believe that it will be a powerful tool for researchers to be able to query any given set of genomic intervals to see whether taSNPs are enriched in these particular neighborhoods of the genome and more importantly, which specific traits show significant enrichment. Typical genomic intervals include ChIP-seq peaks, differentially methylated regions and putative enhancers. In this way, we can build hypotheses linking these intervals to phenotypes. This is similar to the Gene Ontology (GO) [64] term enrichment analysis or the Gene Set Enrichment Analysis (GSEA) [65] except that GO terms or functional categories are replaced by traits, and a set of genes is replaced by a set of genomic intervals. We believe taSNPs can bring important functional insights to genomic regions, especially intergenic regions. Despite the great utilities, currently there is no off-the-shelf computational tool available to carry out this non-trivial test. To cater for this demand, we developed an

R Bioconductor package named *traseR*, TRait-Associated SNP EnRichment analysis, which offers a turnkey solution for enrichment analysis of taSNPs.

4.2 Method

traseR provides multiple options, including testing method, type of background and inclusion of SNPs in linkage disequilibrium (LD), to conduct statistical tests of taSNP enrichment for a given set of query genomic intervals. We here provide a brief description.

4.2.1 Background SNPs

All SNPs from the CEU panel of the phase I 1000 Genomes with minor allele frequency (MAF) greater than 0.05 are used as background SNPs (6,571,512 SNPs genome-wide excluding those from the Y chromosome). These SNPs have a comparable MAF distribution to the taSNP collection.

Obtain taSNPs

Association Results Browser (`http://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm`) combines identified trait-associated single nucleotide polymorphisms (taSNPs) from dbGaP and NHGRI genome-wide association study (GWAS) catalog [66], which together provide 44,124 SNP-trait associations. This list contains 30,553 unique taSNPs associated with 573 different traits. We build this resource into a GRanges object named taSNP in *traseR*, which could be loaded into R console by typing data(taSNP).

Obtain linkage disequilibrium taSNPs from 1000 Genomes

We first download all vcf files from the 1000 Genomes consortium (`ftp://share.sph.umich.edu/1000genomes/fullProject/2012.03.14/`). Next, we use PLINK [67] to find nearby SNPs (within 100kb) that are in tight linkage disequilibrium (LD) (> 0.8) with at least one taSNP. At the end, we collect 78,247 unique SNPs that every SNP is either a taSNP itself or is in LD with a taSNP. We build these LD taSNPs into another GRanges object named taSNPLD in *traseR*, which could be loaded into R console by typing data(taSNPLD).

Obtain background SNPs from 1000 Genomes

All SNPs from the CEU panel of the phase I 1000 Genomes with minor allele frequency (MAF) greater than 0.05 are used as background SNPs (6,571,512 SNPs genome-wide excluding those from the Y chromosome). We build these SNPs in another GRanges object named CEU in *traseR*, which could be loaded into R console by typing data(CEU).

4.2.2 Enrichment tests

All SNPs can be categorized in a contingency table (Table 4.1) as the following,

Table 4.1: The contingency table for enrichment test

	#taSNP	#non-taSNP	
Inside	n_{11}	n_{12}	$n_{1.}$
Outside	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

The null hypothesis is,

$$H_0 : p_1 = p_2$$

If the enrichment is of interest, the alternative hypothesis is,

$$H_0 : p_1 > p_2$$

$p_1$is the probability of observing a SNP being a taSNP inside the query genomic intervals, p_2 is the probability of observing a SNP being a taSNP outside the query genomic intervals.

Contingency table-based tests

The null hypothesis assumes that the proportion of taSNPs among all SNPs is the same both within and outside of the query genomic intervals. We classify all SNPs into either within/outside (query genomic intervals), or taSNPs/non-taSNPs, then construct a two-by-two contingency table and run a test or Fisher's exact test to assess the enrichment level of the taSNPs. Alternatively, we classify every single base in the genome (except for chromosome Y) into either within/outside (query genomic intervals), or taSNPs/non-taSNPs, and conduct the test accordingly.

Chi-squared Test

Agoodness of fittest establishes whether the observed taSNPsfrequency distributioninside the query genomic intervals differs from the taSNPs frequency distribution outside the genomic intervals.

$$\hat{p}_{i.} = \frac{n_{i.}}{n}, i = 1, 2$$

$$\hat{p}_{.j} = \frac{n_{.j}}{n}, j = 1, 2$$

$$E_{ij} = \hat{p}_{i.}\hat{p}_{.j}$$

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Fisher's exact Test

The number of taSNPs inside the query genomic intervals is assumed to follow a hypergeometric distribution, and the probability of observing n_{11} taSNPs is,

$$p(X = n_{11})\frac{\binom{n_{.1}}{n_{11}}\binom{n_{.2}}{n_{12}}}{\binom{n}{n_{1.}}}$$

The p-value calculated using the formula below indicates whether the enrichment of taSNPs inside the genomic intervals is statistical significant.

$$p(X \geqslant n_{11}) = \sum_{n_i=n_{11}}^{n_{.1}} p(X = n_i)$$

Binomial test

The null hypothesis states that the chance of observing a SNP being a taSNP is the same in query genomic intervals as in the whole genome (excluding chromosome Y). Therefore, under the null hypothesis, the number of observed taSNPs out of all SNPs in the query genomic intervals follows a binomial distribution with probability equal to the genome-wide proportion of all taSNPs among all SNPs. Alternatively, we can use all bases in the genome as the background and conduct the test accordingly.

The proportion of taSNPs across the whole genome is used to estimate the probability of observing a taSNP inside the genomic intervals. Under the null hypothesis $p_1 = p$, where is the probability of observing a SNP being a taSNP across the whole genome, we have,

$$\hat{p}_1 = \hat{p} = \frac{n_{.1}}{n}$$

The p-values calculated using the formula below indicates whether the enrichment of taSNPs inside the query genomic intervals is statistical significant.

$$p(X \geqslant n_{11}) = \sum_{n_i = n_{11}}^{n_{1.}} \hat{p}_1^X (1 - \hat{p}_1)^{n_{1.} - X}$$

Nonparametric statistical testing procedure

Because typical query genomic intervals only span a small fraction of the whole genome, and the genome-wide distribution of SNPs is not uniform, it is desirable to conduct a nonparametric test in which a set of randomly selected control intervals is compared to the query genomic intervals for taSNP enrichment, rather than imposing distribution assumptions. For each permutation, *traseR* generates a matching control interval of the same size and on the same chromosome as each query genomic interval. The process is repeated 10,000 times (or a number specified by the user) to obtain an empirical p-value.

In this test, for each query genomic interval, *traseR* generates a matching control interval on the same chromosome, with the same size. Then, how many times the number of taSNPs inside the query genomic intervals is larger/smaller than the number of taSNPs inside the set of control genomic intervals are counted. This procedure is repeated a large number of times (e.g. 10,000) to obtain the corresponding empirical p-value. Suppose sets of matching control intervals are generated. For each set i, n_i taSNPs are observed inside the matching control intervals, then the empirical p-value calculated using the formula below indicates whether the enrichment of taSNPs inside the query genomic intervals is statistical significant.

$$\frac{\sum_{i=1}^{N} I(n_i \geqslant n_{11})}{N}$$

4.2.3 Linkage disequilibrium

As an option, *traseR* allows users to expand the taSNP set to include all the SNPs that are in tight linkage disequilibrium (LD) (>0.8) with any of the taSNPs. The extended taSNP set contains 78,247 unique SNPs. Inclusion of SNPs in LD with the taSNPs is preferable if there is a limited number of taSNPs associated with the traits of interest.

4.3 Results

4.3.1 SNP collection

We collect a compendium of up-to-date taSNPs from dbGaP and NHGRI. There are 30,553 unique taSNPs spanning 573 phenotypes, all of which have been preloaded into the *traseR* package. *traseR* takes in a bed format input file that contains the query genomic intervals, then performs a user-specified enrichment test on all taSNPs combined, as well as taSNPs associated with each of the 573 traits. In the output, *traseR* reports the overall enrichment level of all taSNPs in the query genomic intervals, followed by a ranked list of traits that show statistically significant enrichment. Accordingly, p-values, FDR q-values and odds ratios are also reported for each trait.

Table 4.2: Top-ranked traits for T cell H3K4me1 peaks

Trait	p-value	OR	#taSNP hits	#taSNP
All	2.70E-48	1.5	1,846	30,553
Behcet Syndrome	4.40E-23	6.3	59	274
Diabetes Mellitus, Type 1	1.70E-11	5	33	185
Lupus Erythematosus	6.20E-09	3.9	32	223
Arthritis, Rheumatoid	1.40E-07	5.1	20	112
Multiple Sclerosis	1.60E-05	2.9	26	236
Autoimmune Diseases	5.20E-05	15.9	6	15

4.3.2 Real data analyses

H3K4me1 peaks in T cell

We demonstrate *traseR*'s utilities by displaying a sample result (H3K4me1 peak regions in peripheral T cell) [4]. The 198,162 H3K4me1 ChIP-seq peaks in human peripheral blood T cell are downloaded from Roadmap Epigenome website. The peak regions span 128 MB and account for around 4% of the human genome. We run *traseR* using binomial test, with 30,553 taSNPs, using whole genome as the background. The top-ranked traits are all immune-related (Table 4.2). Moreover, peaks are significantly enriched with overall taSNPs. Here we use the whole genome as background and binomial test as the testing method.

H3K4me1 peaks in liver cell

The 233,386 H3K4me1 ChIP-seq peaks in human liver cell are downloaded from Roadmap Epigenome website. The peak regions span 144 MB and account for around 4.8% of the human genome. We ran *traseR* using binomial test, with 30,553 taSNPs, using whole genome as the background. The results in Table 4.3 show that six traits are significantly enriched with taSNPs in these peak regions after Bonferroni correction.

Table 4.3: Top-ranked traits for liver cell H3K4me1 peaks

Trait	p-value	OR	#taSNP	#taSNP
All	1.20E-132	1.8	2480	30,553
Macular Degeneration	1.10E-08	3.1	43	331
gamma-Glutamyltransferase	3.50E-08	11.7	13	36
Lipoproteins, HDL	5.50E-07	4.3	22	129
Cholesterol	1.10E-06	2.1	61	649
Coronary Artery Disease	2.00E-05	2	56	639
Lipoproteins, LDL	4.70E-05	3.7	17	113

Ultra Conserved Elements

The 481 Ultra Conversed Elements (UCEs) are downloaded from Ultra Conserved Elements website `http://ultraconserved.org/`. The regions span 0.13MB and account for around 0.004% of the human genome. We ran *traseR* using binomial test, with the 78,247 taSNP set (including SNPs in LD), using whole genome as the background. The results in Table 4.4 show that UCEs lack taSNPs, which confirms the functional importance of the UCEs.

Table 4.4: Top-ranked traits for Ultra Conserved Elements

Trait	p-value	OR	#taSNP	#taSNP
All	0.83	0.93	2	78,247

Differential Methylated Regions

These Differentially Methylated Regions (DMRs) are formed by extending 3,601 differentially methylated CpG sites [68] upstream and downstream 1kb in the analysis of subcutaneous adipose before and after weight loss. The regions span 7.2 MB and account for around 0.24% of whole human genome. We run *traseR* using binomial test, with 30,553 taSNPs, using whole genome as the background. The results in Table refc4:table5 show that only one trait, Type 1

Diabetes, survives Bonferroni correction. The results are interesting since adipose dysfunction has been linked to insulin resistance and Type 2 Diabetes [69], and there are also studies supporting a connection between obesity and Type 1 Diabetes [70].

Table 4.5: Top-ranked traits for adipose DMRs

All	5.7e-07	1.6	118	30,553
Type1 Diabetes	6.8e-06	16.4	6	185

4.3.3 Computational time

traseR runs very fast even on a personal PC or laptop. Taking the H3K4me1 peaks in T cell as an example, *traseR* only takes about 1.5 seconds on a MacBook laptop with a 1.7 GHz processor and 8 GB memory for all testing methods except for the nonparametric testing option. For the nonparametric testing option, *traseR* costs around 5 minutes for 100 permutations due to the large number of H3K4me1 peaks.

Chapter 5

Conclusion and future work

This dissertation presents the utilization of statistical and informatics methods for the analysis of high-throughput genomic data especially next generation sequencing data and GWAS data. In the era of “Big Data”, the accumulation of massive datasets provides an opportunity to integrate the datasets from different sources for novel discovery in biological systems. However, different modeling approaches should be used in different scenarios. In this dissertation, we present a combination of statistical modeling and informatics approaches for different types of datasets.

Chapter 2 describes a hierarchical linear model *ChIPComp* to quantitatively compare ChIP-seq datasets in multiple conditions. The novelty of the method is that it considers the unique characteristic of ChIP-seq data by modeling the control data in a rigorous. In contrast, other competing methods do not consider control data or fail to model the control data in a correct way. The true biological signals (TF binding/Histone modification) are derived from both IP and control data, and the hypothesis testing is performed on the true biological signals. Moreover, since there is only a small number of biological replicates in each condition, an empirical bayesian shrunken variance estimate is utilized in the variance estimation for the test statistics. *ChIPComp* could be easily modified and extended to other type of sequencing data with matched control data that measures the background noise. For example, the ribosome profiling technique (Ribo-seq) data targets only mRNA sequences protected by the ribosome during in protein translation. The protected mRNA regions in decoding are usually open reading frames (ORFs), and therefore, Ribo-seq could identify ORFs differentially translated between experimental conditions. The non-active ORFs in normal condition could be considered as the control data since the non-active ORFs measure the RNA structure and RNA sequence. Therefore, it is feasible to modify *ChIPComp* for multiple Ribo-seq comparison in multiple conditions.

Chapter 3 describes a machine learning method *DIVAN* to identify and prioritize disease-specific non-coding variants by building a predictive model that integrates GWAS SNPs and thousands of epigenomic profiles as the training set. The motivation for the development of *DIVAN* is that the biological significances of GWAS SNPs are not only captured by p-value in the case-control study. Especially for SNPs in highly LD, the SNPs with most stringent p-values might not be the causal ones. Thus, additional information should be combined along with the GWAS summary statistics to re-prioritize the SNPs, and the top ranked SNPs have better chance to be causal. Besides re-prioritizing variants, *DIVAN* is also useful in the identification of novel disease-associated variants, especially rare variants with the assumption that the epigenomic and genomic profiles among causal variants are similar. And the epigenomic and genomic profiles for either novel variants or known variants could be simply obtained from the pre-built genome-wide annotation matrix. We could pre-computed the *DIVAN* score for each base in entire human genome for 45 diseases/traits studied, and the *DIVAN* could be considered as the probability of a variant being disease-associated. Therefore, bases with high D-score could be potential the novel variants that are disease-associated.

Particularly, *DIVAN* could be potentially widely used in cancer research. The variants reported in GWAS are germline variants in majority. However, somatic mutations are of a particular interest in cancer research since they may be the drivers and could be potential drug biomarkers. *DIVAN* could be applied to re-prioritize somatic mutations and identify novel somatic mutations by using recurrent somatic mutations as the training set.

At the moment, only GWAS SNPs with p-values meet stringent significant level (e.g. 10^{-4}) are considered as casual in the training set. However, the quantitative difference of SNPs, which meet the significant level, are ignored. To incorporate the p-value information, we plan to extend *DIVAN* in a supervised-unsupervised machine learning framework to improve the accuracy of risk variant identification and prioritization. We directly use disease-specific D-scores assigned by *DIVAN* in the supervised learning step. Bayesian method will be used as the unsupervised method to calculate the posterior probability of each single nucleotide being disease-associated conditioned on GWAS summary statistics (p-values). Besides D-scores, other pre-computed scores of regulatory variants from multiple existing variant annotation methods such as CADD, Eigen, and GWAVA could be used as the prior as well. The rationale is that the variant with high probability of being regulatory or disease-associated in GWAS is more likely to be disease casual in reality. This strategy could also be extended to tissue-specific functional element prioritization such as eQTLs. In this extension, eQTLs with cell-type specific epigenomic annotations could be used in the supervised model, while the scores of regulatory variants could be served as the prior in the unsupervised model.

Bibliography

[1] J. D. Eicher, C. Landowski, B. Stackhouse, A. Sloan, W. Chen, N. Jensen, J. P. Lien, and R.and others Leslie. Grasp v2.0: an update on the genome-wide repository of associations between snps and phenotypes. *Nucleic Acids Res*, 43(Database issue):D799–804, 2015.

[2] D.J. Thomas, K.R. Rosenbloom, H. Clawson, A.S. Hinrichs, H. Trumbower, B.J. Raney, D. Karolchik, G.P. Barber, R.A. Harte, J. Hillman-Jackson, et al. The encode project at uc santa cruz. *Nucleic acids research*, 35(suppl 1):D663, 2006.

[3] Susan E Celniker, Laura AL Dillon, Mark B Gerstein, Kristin C Gunsalus, Steven Henikoff, Gary H Karpen, Manolis Kellis, Eric C Lai, Jason D Lieb, David M MacAlpine, et al. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.

[4] Consortium Roadmap Epigenomics, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, and P.and others Kheradpour. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[5] D.S. Johnson, A. Mortazavi, R.M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497,

2007.

[6] Teemu D Laajala, Sunil Raghav, Soile Tuomela, Riitta Lahesmaa, Tero Aittokallio, and Laura L Elo. A practical comparison of methods for detecting transcription factor binding sites in chip-seq experiments. *BMC genomics*, 10(1):618, 2009.

[7] Elizabeth G Wilbanks and Marc T Facciotti. Evaluation of algorithm performance in chip-seq peak detection. *PLoS One*, 5(7):e11471, 2010.

[8] Xi Chen, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B Vega, Eleanor Wong, Yuriy L Orlov, Weiwei Zhang, Jianming Jiang, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008.

[9] Han Xu, Chia-Lin Wei, Feng Lin, and Wing-Kin Sung. An hmm approach to genome-wide identification of differential histone modification sites from chip-seq data. *Bioinformatics*, 24(20):2344–2349, 2008.

[10] Cenny Taslim, Jiejun Wu, Pearlly Yan, Greg Singer, Jeffrey Parvin, Tim Huang, Shili Lin, and Kun Huang. Comparative study on chip-seq data: normalization and binding pattern characterization. *Bioinformatics*, 25(18):2334–2340, 2009.

[11] Cenny Taslim, Tim Huang, and Shili Lin. Dime: R-package for identifying differential chip-seq based on an ensemble of mixture models. *Bioinformatics*, 27(11):1569–1570, 2011.

[12] Zhen Shao, Yijing Zhang, Guo-Cheng Yuan, Stuart H Orkin, and David J Waxman. Manorm: a robust model for quantitative comparison of chip-seq data sets. *Genome biology*, 13(3):R16, 2012.

[13] Nishanth Ulhas Nair, Avinash Das Sahu, Philipp Bucher, and Bernard ME Moret. Chipnorm: a statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries. *PloS one*, 7(8):e39573, 2012.

[14] Kun Liang and Sündüz Keleş. Detecting differential binding of transcription factors with chip-seq. *Bioinformatics*, 28(1):121–122, 2012.

[15] Rory Stark and Gordon Brown. Diffbind: differential binding analysis of chip-seq peak data, 2013.

[16] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nussbaum, R.M. Myers, M. Brown, W. Li, et al. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137, 2008.

[17] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[18] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, 14(2):232–243, 2013.

[19] Hao Feng, Karen N Conneely, and Hao Wu. A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69–e69, 2014.

[20] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[21] Bradley Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465), 2004.

[22] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[23] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Preprint 2013*, 2013.

[24] Consortium Genomes Project, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, and M. E.and others Hurles. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[25] S. E. Flanagan, A. M. Patch, and S. Ellard. Using sift and polyphen to predict loss-of-function and gain-of-function mutations. *Genet Test Mol Biomarkers*, 14(4):533–537, 2010.

[26] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*, 16(2):85–97, 2015.

[27] Y. G. Tak and P. J. Farnham. Making sense of gwas: using epigenomics and genome engineering to understand the functional relevance of snps in non-coding regions of the human genome. *Epigenetics Chromatin*, 8:57, 2015.

[28] F. Zhang and J. R. Lupski. Non-coding genetic variants in human disease. *Hum Mol Genet*, 24(R1):R102–110, 2015.

[29] Matthew T. Maurano, Richard Humbert, and Eric Rynes. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337:1190–1195, 2012.

[30] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, and I.and others Chepelev. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

[31] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, and B.and others Bernier. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–657, 2007.

[32] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[33] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, 2008.

[34] G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, and Y.and others Chen. Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). *Genome Res*, 16(1):123–131, 2006.

[35] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb. Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res*, 17(6):877–885, 2007.

[36] Encode Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[37] G. R. Ritchie, I. Dunham, E. Zeggini, and P. Flicek. Functional annotation of noncoding sequence variants. *Nat Methods*, 11(3):294–296, 2014.

[38] Breiman L. Random forests. *Machine Learning*, 45:5–32, 2001.

[39] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3):310–315, 2014.

[40] Q. Lu, Y. Hu, J. Sun, Y. Cheng, K. H. Cheung, and H. Zhao. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep*, 5:10576, 2015.

[41] I. Ionita-Laza, K. McCallum, B. Xu, and J. D. Buxbaum. A spectral approach integrating functional genomic annotations for coding and non-coding variants. *Nat Genet*, 48(2):214–220, 2016.

[42] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas, and D. N. Cooper. The human gene mutation database: 2008 update. *Genome Med*, 1(1):13, 2009.

[43] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott. Clinvar: public archive of relationships among

sequence variation and human phenotype. *Nucleic Acids Res*, 42(Database issue):D980–985, 2014.

[44] S. A. Forbes, D. Beare, N. Bindal, S. Bamford, S. Ward, C. G. Cole, M. Jia, and C.and others Kok. Cosmic: High-resolution cancer genetics using the catalogue of somatic mutations in cancer. *Curr Protoc Hum Genet*, 91:10 11 11–10 11 37, 2016.

[45] G. M. Cooper, E. A. Stone, G. Asimenos, Nisc Comparative Sequencing Program, E. D. Green, S. Batzoglou, and A. Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*, 15(7):901–913, 2005.

[46] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, and J.and others Spieth. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, 2005.

[47] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010.

[48] Zhi-Hua Zhou. *Ensemble methods : foundations and algorithms*. Chapman & Hall/CRC machine learning & pattern recognition series. Taylor & Francis, 2012.

[49] P. Wen, P. Xiao, and J. Xia. dbdsm: a manually curated database for deleterious synonymous mutations. *Bioinformatics*, 32(12):1914–1916, 2016.

[50] K. Lund, P. D. Adams, and M. Copland. Ezh2 in normal and malignant hematopoiesis. *Leukemia*, 28(1):44–49, 2014.

[51] J. K. Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American journal of human genetics*, 94(4):559–573, 2014.

[52] L. Chen and Z. S. Qin. traser: an r package for performing trait-associated snp enrichment analysis in genomic intervals. *Bioinformatics*, 32(8):1214–1216, 2016.

[53] L. D. Ward and M. Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol*, 30(11):1095–1106, 2012.

[54] N. A. Faustino and T. A. Cooper. Pre-mrna splicing and human disease. *Genes Dev*, 17(4):419–437, 2003.

[55] J. F. Caceres and A. R. Kornblihtt. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet*, 18(4):186–193, 2002.

[56] N. Lopez-Bigas, B. Audit, C. Ouzounis, G. Parra, and R. Guigo. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett*, 579(9):1900–1903, 2005.

[57] S. Barbaux, P. Niaudet, M. C. Gubler, J. P. Grunfeld, F. Jaubert, F. Kuttenn, C. N. Fekete, and N.and others Souleyreau-Therville. Donor splice-site mutations in wt1 are responsible for frasier syndrome. *Nat Genet*, 17(4):467–470, 1997.

[58] C. L. Lorson, E. Hahnen, E. J. Androphy, and B. Wirth. A single nucleotide in the smn gene regulates splicing and is responsible for spinal muscular atrophy. *Proc Natl Acad Sci U S A*, 96(11):6307–6311, 1999.

[59] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, and J.and others Cuff. The ensembl genome database project. *Nucleic Acids Res*, 30(1):38–41, 2002.

[60] H. Chen, H. Yu, J. Wang, Z. Zhang, Z. Gao, Z. Chen, Y. Lu, and W.and others Liu. Systematic enrichment analysis of potentially functional regions for 103 prostate cancer risk-associated loci. *Prostate*, 75(12):1264–1276, 2015.

[61] B. E. Stranger, E. A. Stahl, and T. Raj. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383, 2011.

[62] Welter D. The nhgri gwas catalog. *Nucleic Acids Research*, 42(Database):1001–1006, 2014.

[63] M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder. Linking disease associations with regulatory information in the human genome. *Genome research*, 22(9):1748–1759, 2012.

[64] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, and K.and others Dolinski. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, 2000.

[65] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, and S. L.and others Pomeroy. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[66] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–9367, 2009.

[67] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, and P.and others Sklar. Plink: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–575, 2007.

[68] M. C. Benton, A. Johnstone, D. Eccles, B. Harmon, M. T. Hayes, R. A. Lea, L. Griffiths, and E. P.and others Hoffman. An analysis of dna methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. *Genome Biol*, 16:8, 2015.

[69] A. Guilherme, J. V. Virbasius, V. Puri, and M. P. Czech. Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes. *Nat Rev Mol Cell Biol*, 9(5):367–377, 2008.

[70] K. C. Verbeeten, C. E. Elks, D. Daneman, and K. K. Ong. Association between childhood obesity and subsequent type 1 diabetes: a systematic review and meta-analysis. *Diabet Med*, 28(1):10–18, 2011.